Introduction to Multilevel Analysis: Theory, Measurement, and Two-Level Nested Models

Dr. James LeBreton

Recommended Reading List

Module 1 – Multilevel Theory: Constructs, Inferences, and Composition Models

Gully, S. M., & Phillips, J. M. (2019). On finding your level. In S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook of multilevel theory, measurement, and analysis* (pp. 11-38). Washington, D.C.: American Psychological Association.

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.), Multi-level Theory, Research and Methods in Organizations. San Francisco: Jossey-Bass.

Module 2 – Multilevel Measurement: Aggregation, Aggregation Bias, and Cross-Level Inference

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

Module 3 – Multilevel Measurement: Estimating Interrater Agreement & Reliability

LeBreton, J. M., & Senter, J. L. (2008). Answers to twenty questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815-852.

LeBreton, J. M., Moeller, A. N., & Wittmer, J. L. S. (2023). Data aggregation in multilevel research: Best practice recommendations and tools for moving forward. *Journal of Business and Psychology, 38*, 239-258.

Module 4 – Multilevel Analysis: Multilevel Regression (2-Level Nested)

Bliese, P. (2022). *Multilevel modeling in R (2.7): A brief introduction to R, the multilevel package and the nlme package*. URL: https://cran.r-project.org/doc/contrib/Bliese_Multilevel.pdf

Raudenbush, S. W., & Bryk, A. S. (2002). Chapter 2: The Logic of HLM. In *Hierarchical linear models: Application and data analysis methods* (2nd ed., pp.16-37). Newbury Park: Sage.

Shiverdecker, L. K., & LeBreton, J. M. (2019). A primer on multilevel (random coefficient) regression modeling. In S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook of multilevel theory, measurement, and analysis* (p. 389-422). Washington, D.C.: American Psychological Association.

PART I

MULTILEVEL THEORY

Chapter 1

ON FINDING YOUR LEVEL

Stanley M. Gully and Jean M. Phillips

We encourage scholars to think more deeply about the concept of the appropriate level of analysis in research. By *appropriate level*, we are not referring to the stated level of interest, the level of data collection, or the appropriate level of statistical analysis. For the purpose of this chapter, *identifying the appropriate level* means finding the level that has the most explanatory power for the outcome of interest. When we wish to identify the level that explains what is happening in a dynamic and complex system, how do we know where the "action" is in a multilevel framework? To answer this question, we focus on three key issues. First, we discuss why identifying the appropriate level is more challenging than it might initially appear. Some scholars have suggested that contiguity in levels creates the strong interactions necessary to create bonds, close coupling, and embeddedness. We agree in general but propose that focusing on tightly coupled levels can overlook key determinants of process dynamics that may skip levels. Second, we examine the notions of causality and variance and discuss the implications of these concepts in a multilevel framework. Third, we offer ideas and examples of how we can expand our thinking to consider what is meant when we refer to the appropriate level of analysis. By examining variance creation and variance restriction, we highlight a number of key processes that lead to the generation and dissolution of higher and lower level effects. We provide several examples and propose some potential solutions to the thorny challenge of thinking about levels in the social sciences.

The concept *level of analysis* is pertinent to all social sciences, including research on family dynamics (Snijders & Kenny, 1999), education (Raudenbush & Bryk, 1988), health (Blakely & Woodward, 2000), crime (Groff, Weisburd, & Yang, 2010), emotions (Keltner & Haidt, 1999), international relations (Singer, 1961), community psychology (Shinn & Rapkin, 2000), applied psychology (Chan, 1998), and social and personality psychology (Nezlek, 2001). Similarly, we see levels issues in the micro–macro divide in the disciplines of sociology and economics (Gerstein, 1987; Hodgson, 1998; Jepperson & Meyer, 2011). The following discussion pertains to levels issues in all of these fields and more. We use a variety of examples drawn from different fields but applicable to all of the above social science disciplines, as well as others.

We begin with the following simple question: What is the appropriate level of analysis? To answer this question, we must first define what is meant by *level*. We define a *level* as a focal plane in social, psychological, or physical space that exists within a hierarchical structure among things or constructs (Gully & Phillips, 2005; K. J. Klein, Dansereau, & Hall, 1994; Rousseau, 1985). *Levels* refer to distinct hierarchical structures within a system, with some entities existing within or as a part of others. We can conceive of the hierarchical structure as a series

http://dx.doi.org/10.1037/0000115-002
The Handbook of Multilevel Theory, Measurement, and Analysis, S. E. Humphrey and J. M. LeBreton (Editors-in-Chief)

of nested relationships with, for example, repeated observations nested within individuals, individuals nested within families, and families within communities. Communities, in turn, can be nested within regions, and regions can be nested within countries.

Typically, what is defined as a higher or lower level depends on the phenomenon of interest. In a business context, teams are a higher level of analysis when compared to individuals, but they are a lower level of analysis when compared to organizations, unless the "teams" are policy makers in governments that affect organizational regulation, in which case the teams may be at a higher level than organizations. To be clear, *higher* and *lower levels* do not refer to power relationships or echelons (Rousseau, 1985), although they may be related. Also, we do not mean to imply that all levels are equally present in all organizations or all contexts. Some organizations may not have teams within their hierarchical structure (i.e., flatter organizations), whereas other organizations may or may not have distinct business units within their overall organizational structure. We are merely trying to highlight the possibilities in levels as we pursue our discussion.

Figure 1.1 shows the levels most often considered by organizational scientists. Given our backgrounds in industrial and organizational psychology and management, we draw from examples that tend to focus largely on the individual (e.g., individual attitudes, individual behavior, and individual performance) or team levels (e.g., team cohesion, team potency, team performance, team conflict), but the same principles we describe also apply to other levels and other contexts of social science. For example, we can envision students within classrooms within schools within school districts within states or regions within nations. Also, as we seek to answer the question "what is the appropriate level," we find that our focal level may not be the individual level, even if we wish to understand individual outcomes. In any case, much of our attention as social scientists has been targeted toward the lower regions (as indicated by the shading), ranging from social collectives (e.g., organizations) to individuals. Some scholars might suggest that we look at levels below that of individuals by trying to understand what happens to individuals over time (e.g., mood). This is a lower level because it is within the person

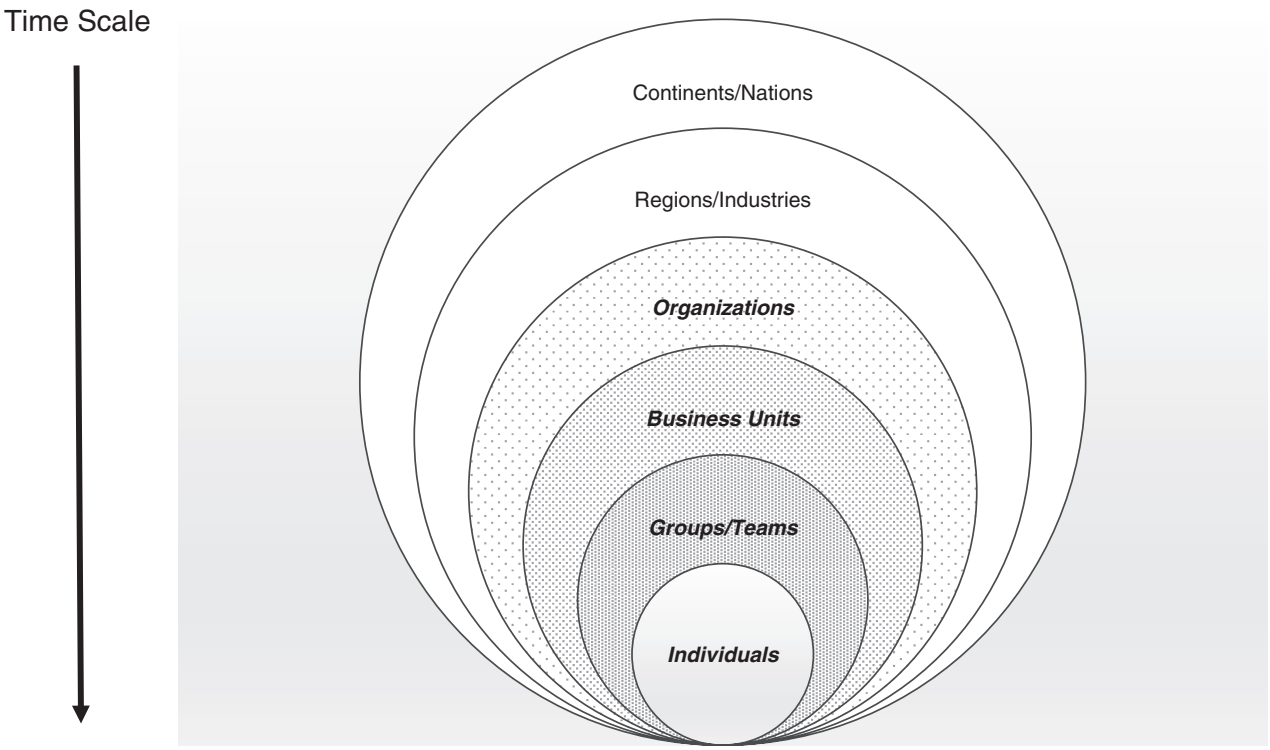


FIGURE 1.1. Representation of levels.

over time. This argument is correct, but we address time as a distinct topic in subsequent discourse for reasons that will soon become clear. For now, we focus on physical, psychological, or social structures that exist above and below the level of groups–teams.

One issue that should be apparent is that levels have semipermeable boundaries. For example, some individuals are members of multiple teams within organizations and some countries are somewhat homogenous internally, whereas others may have varying characteristics across regions (as in the United States). We see the same thing in high schools with students attending several classes within the same school or teachers working in more than one school within the same district.

Additionally, it is not always clear where some levels reside. Some levels cut across other levels. Consider for example, the notions of "job" or "vocation." Does a job or vocation exist within (i.e., hierarchically nested under) an organization? Often, we place jobs at a lower level than the work unit (different jobs within the unit) and the individual who fills a given job at the level below the job level (one type of job can be filled by different people, so job is the higher level). The organization would be at a higher level than the individual, job, or work unit, because the work unit is nested within the organization. Yet individuals have allegiances to their jobs that transcend work units and even organizations. Moreover, many jobs share common characteristics that transcend organizations or even geographic regions (O*Net, 2017). Similarly, many of us see our membership within professional societies as part of our job, but our membership both defines and is defined by our roles outside the university or college to which we belong.

Consider the job of professor. A person in such a position is nested in multiple sets of structures, including formal hierarchical job structures (e.g., rank, department, college, university) and professional semi- or nonhierarchical structures (leadership in professional societies, editorial boards, multiple research teams). Thus, the person is nested within the job (e.g., associate professor) nested within the department (e.g., Sociology) within the university (e.g., Penn State), but that same person may also be nested within several research teams that may be further nested within distinct geographic or cultural regions. Professional structures and research teams often exist independently of the department or university within which the person is employed, yet the particular individual and associated job of professor is relevant to all these contexts. This is what we mean when we say that some levels cut across levels and that levels are semi-permeable. One challenge is trying to pick the correct level of analysis for understanding an outcome when the level itself may be semi-permeable (and as we shall see in the section Dual Processes in Multilevel Systems, impermanent). As scientists, we need to be clear about our basis for grouping entities and articulate reasons for ignoring other potentially relevant grouping systems (Mathieu & Chen, 2011).

PROBLEMS WITH IDENTIFYING THE LEVEL

It may seem that a chapter exploring the identification of the appropriate level isn't needed because if we want to know the level, we need only think about the outcome we are trying to understand. If we want to understand individual behavior (e.g., student achievement), then the level is the individual (i.e., the student). There may be factors residing both above the individual level (e.g., teacher experience, availability of classroom technology) and below the individual level (e.g., day-to-day patterns of sleep, nutrition, or affect) that may impact or influence individual behavior, but the level is clearly the individual. Likewise, if we want to understand outcomes residing at the level of a social collective (e.g., team performance), then the level is the team. Similarly, a focus on trust in married couples implies that the level is the couple; a focus on organizational profitability implies that the "appropriate" level is the organization.

However, what if it isn't so simple? When we ask, "What is the appropriate level?" we don't mean "What is the level of your dependent variable?" or "What is the level at which I should analyze my data?" Rather, we are asking, "What is the level that will enable deep understanding of a particular phenomenon?" Or, in plain language, "Where's the

action?" This question is more difficult to answer for a variety of reasons. First, our theory and disciplinary backgrounds affect what we attend to. Perhaps if we want to understand individual student health, then we might surmise that the level is the individual student. We could incorporate a level up, so we could examine social structures (e.g., peer social support, school lunch programs) to understand individual student stress and wellness. This is an obvious approach to the levels question. Think about the outcome of interest, and that tells you the level of action.

Unless, of course, you work at the Centers for Disease Control and Prevention (CDC) and are trying to understand student health; then the level might be within-person, at the biological or cellular level. Or suppose you work at the CDC and you are trying to understand a pandemic in schools, in which case the level might be international or global travel patterns. We may not be able to understand individual health without understanding global infection patterns (global level), and we can't understand global patterns if we don't understand the biological pathways through which an infection occurs (viral or bacterial level). Additionally, we may need to understand individual travel patterns (within or across national levels) as well as within-person level behaviors and genetics (Are particular people predisposed? Does the individual wash his or her hands and when? What did he or she eat? With whom did she or he interact?). Understanding student health at the CDC to stop a global pandemic requires an understanding of all these levels.

Our choice of theory affects everything. It determines what level of analysis we use for sampling, who or what we sample, how we measure, what we measure, when we measure, and how we analyze the data. If we believe that student health is an individual-level phenomenon, then we might go to a single school; gather data on 1,000 students, including classrooms, teachers, classmates, and so forth; and then analyze the data. Given that we believe it is an individual-level phenomenon, we will ask how the *individual* feels, thinks, and acts. But what if it is a group phenomenon? Perhaps instead of asking whether the individual worries about wellness, we need to ask whether other students in the classroom worry about wellness because that sets the social context for taking care of oneself. We know the wording matters. Asking whether I feel self-efficacious is not the same as asking whether classmates feel efficacious (Chan, 1998), even if we aggregate individual perceptions to the level of the classroom. How we word the questions affects the variability we observe and the intrinsic meaning of the construct (Baltes, Zhdanova, & Parker, 2009; Chen, Mathieu, & Bliese, 2004; K. J. Klein, Conn, Smith, & Sorra, 2001). What if student health is influenced by school-level or districtwide phenomena, as it almost certainly is (e.g., McNeely, Nonnemaker, & Blum, 2002), but we focus on the individual?

For another context, consider employees working in the technology industry. Gathering data (even a large sample; $N = 5{,}000$) from individuals nested within a single company or individuals nested within two companies (e.g., IBM and Google; $N = 5{,}000$ per organization) necessarily excludes key factors that determine employee health such as industrial safety records. Our initial focus on the individual may overlook the large-scale patterns of health and wellness across industries, organizations, and occupations (Bureau of Labor Statistics, 2015a, 2015b, 2016; Johnson et al., 2005), and we miss the big picture. Our theory determines what we sample and what we measure. If we believe that occupations (or grades), organizations (or schools), or industries (or school districts) matter, then we gather data on units at that level. If we believe individuals matter, then that's where we seek our samples.

If our theory determines how we sample, measure, and analyze, and if the way in which we approach these decisions then constricts (or enhances) our ability to see effects at a particular level, then how do we know when we've got the right level or levels? There have been multilevel analytic systems that purport to tell us the level at which effects exist (Dansereau, Alutto, & Yammarino, 1984). Although such approaches can be useful because they can tell us the level at which we observe variance, the findings are unlikely to generalize to other, more diverse samples or data sets. These approaches cannot tell you where the variance lives in the system, out there in the wild, because the variance

of the data is itself determined by what and how we measure and sample. We won't see variance, even if it's important, if we don't sample and measure in ways that enable us to see it.

In a way, James, Demaree, and Wolf (1984) recognized this conundrum when they introduced their measure of agreement: r_{WG}. Simply because there is insufficient variance across units to see differences doesn't mean that agreement isn't present or that a higher level construct doesn't exist. Most existing measures for agreement at the time examined it from a reliability perspective or from a group effects perspective [e.g., F tests, intraclass correlation coefficient (ICC)(1), ICC(2)] and thus relied on the proportion of variance across units as compared to the proportion of variance within units to tell us whether members within units responded reliably or consistently (James, 1982). Simply because all units are relatively similar in terms of their mean levels on a construct (and thus, there is very little between-unit variance) doesn't mean that there isn't a higher level construct worth assessing and understanding.

Our choice of theory can and will lead us to focus on a given level, but we bring with us a certain myopia for looking at a given research problem. A public health researcher might look at employee health at the governmental and community levels (e.g., health policy), and a sociologist might look at employee health as a social problem, at the level of structures within our society (e.g., poverty, access to health care). A management researcher studying teams might look at employee wellness as a function of social and team dynamics or leadership (e.g., social support, abusive supervision, access to wellness programs). A personality theorist might look at employees' mental health as a function of individual characteristics (e.g., self-identity, self-esteem, sensation seeking, depression). Which level is "correct"? The point is that selecting the focal level of interest can excise meaningful other variables from the equation.

How then do we find the correct level? For now, it's enough to assert that no statistical system can identify whether we have the correct level because our statistical results are determined by what we do. Furthermore, theory drives what we do, which then tells us what we see and have done. It is difficult to determine the correct level theoretically because the theory gives us tunnel vision. Unfortunately, there's no recipe to follow to know that we've found the right level. Instead, we have to operate as detectives, picking up a clue here or there, fencing off areas for careful examination to reduce contamination of evidence, making observations over time. It is only then that we will uncover where (i.e., at what level or levels) the (majority of) variance truly resides. In most circumstances, we are not looking for a single level. Instead, we have to think about identifying a multilevel nomological network (Cronbach & Meehl, 1955). The following sections provide thoughts on how we can proceed more effectively as detectives and provide some fodder for thinking beyond our respective disciplinary foci.

ASSUMPTIONS OF TIGHT COUPLING

Previous scholars (Hackman, 2003; Kozlowski & Klein, 2000; Simon, 1973) have suggested that contiguity in levels creates the strong interactions necessary to create bonds, close coupling, and embeddedness. In general, we agree with this sentiment, but we caution that by focusing on tightly coupled levels we may overlook key determinants of process dynamics that may skip levels. Scholars (ourselves included) frequently assume that stronger relationships may be detected among variables that exist at the same level of analysis. So, for example, it is thought that the best predictors of individual performance will be individual factors (e.g., intelligence, conscientiousness, work ethic; Schmidt & Hunter, 1998). Likewise, many scholars would argue that we need to examine team-level input and process factors to understand team-level outcomes, even when those input or process factors represent aggregations of individual-level variables (e.g., team ability or team personality; Barrick, Stewart, Neubert, & Mount, 1998; Baysinger, Scherer, & LeBreton, 2014). It's such a compelling argument that it nearly seems tautological, and there is a long history of this type of thinking.

Indik (1968) was an early systems scholar who presented a figure with panels representing variables for structure, process, and function at

the organizational, group, and individual levels. He stated,

> Specifically, we expect larger relationships to occur between variables in panels closer together in the schema. For example, organization size, a Panel One variable, should be more clearly related to variables of organization function or process such as communication, control, or coordination than to Panel Five [individual] variables such as attitude toward the organization, attitude toward the work group, and achievement motivation. (p. 22)

Here he is making the clear argument that we will find stronger connections for variables connected at a given level, or across contiguous levels, than jumping across levels.

Simon (1973) made a similar argument when describing the nature of hierarchical systems. He described a hierarchy that "leads from human societies to organizations, to small groups, to individual human beings, to cognitive programs in the central nervous system, to elementary information processes" (p. 9). He went on to state the following:

> Most interactions that occur in nature, between systems of all kinds, decrease in strength with distance. Hence, any given "particle" has most of its strong interactions with nearby particles. As a result, a system is likely to behave either as made up of a collection of localized subsystems or to form a more or less uniform "tissue" of equally strong interactions. (p. 9)

This perspective suggests that everything is connected but that some things are more connected than others, and those factors closest to others at the same level are most important for understanding a particular phenomenon. We once believed this as well, and perhaps it is often true. But there is evidence to suggest otherwise, and we think the "skip level effect" may be more common than we typically consider.

Hackman (2003) questioned the utility of seeking explanations solely at the same or lower levels of analysis. Data were collected on 300 flight crews in nine different types of aircraft at seven different airlines. The conceptual model tested was Hackman's own (1987), which posited that the design of the flying task and the design of the aircrew would determine whether aircrews developed into self-correcting high-performing units. As he put it,

> We knew we were in trouble when we performed a simple one-way analysis of variance on our measures of crew structure and behavior across the seven diverse airlines. There was almost no variation across airlines on precisely those crew-level variables that we had expected to be most consequential for performance. (p. 910)

Where was the team level (i.e., crew-level) variance? When they turned to the organizational and institutional contexts within which aircrews operated, they found variation in perceived contextual factors, including adequacy of material resources, clarity of performance objectives, recognition and reinforcement for excellent crew performance, availability of educational and technical assistance, and availability of informational resources. Between-airline differences accounted for 23% of the variation in the composite measure of context supportiveness and context related to pilot satisfaction. However, none of these factors predicted pilot behavior and aircrew effectiveness. Why?

According to Hackman (2003), there were three dominant influences on aircrew performance, none of which were under the control of the flight crews or even the airlines for which they worked. First, there was aircraft and cockpit technology. Standard cockpit technology was generated by designers and engineers at Airbus and Boeing, and there are finite sets of configurations and technologies, all of which were designed for efficiency and safety. They found no aircrew differences associated with aircraft differences, because the consistency of technology determined how aircrew members interacted with one another and with others external to the flight crew. Even though there was limited

observable variance, there was a higher level aircraft manufacturer effect that determined how aircrews behaved.

Second, there is a clearly defined set of regulatory procedures and standards developed over the years by the U.S. Federal Aviation Administration (FAA), in cooperation with airline manufacturers and flight operations departments of U.S. airlines. These FAA procedures and standards have been adopted globally, often with little modification, by airlines and regulatory agencies around the world. This creates consistency in how aircrews operate in widely varying contexts. Again, although we see limited variance, there is a higher level effect of institutionalized regulatory procedures and standards on aircrew behavior.

Hackman (2003) described the pervasive "culture of flying" that is rooted in a shared individualized experience that affects how pilots perceive and behave. Every pilot has a shared, nearly institutionalized experience as a pilot. All pilots worry about medical checks, and each pilot recognizes the importance of proficiency checks. This creates a consistency in the values and mindset of pilots regardless of background or location. There is limited variance across individuals for this "pilot" cultural effect, but the consistency across individual experiences affects what happens in flight crews. There is empirical evidence in other domains that when decision makers engage in consistent behaviors and interventions across time and situations, restriction of variance in performance can result. In other words, effective organizational or policy interventions attenuate or restrict variance in behavior or performance (Lebreton, Burgess, Kaiser, Atchley, & James, 2003).

Hackman (2003) detailed other examples of how factors at higher or lower levels of analysis than the team impinge on team processes and outcomes across a diverse array of examples including orchestral performance (influenced by national level cultures regarding gender and strength of ties with the orchestral community) and hospital patient care teams (influenced by nurse managers' reporting of errors).

Upon reflection, these findings make intuitive sense. Higher and lower level factors can shape the variance and outcomes observed at a particular level, sometimes many levels apart. Organizational differences can enhance or impede variation in team processes, so it may be the case that team effectiveness is not an outgrowth of team-level process or input factors but a result of contextual factors that determine how the team is structured and behaves. Likewise, family dynamics within a given society may be equally or more influenced by cultural factors than by individual factors. This is actually an old idea, but one that we believe theorists should re-embrace.

We concur with Hackman (2003) that we should consider "bracketing" higher and lower levels when investigating processes at a given level. Bracketing can enhance one's understanding of targeted phenomena; help discover factors that drive those phenomena, even when variance doesn't initially appear to exist; uncover boundary conditions and interactions that shape an outcome of interest; and inform choice of constructs within a nomological network (Cronbach & Meehl, 1955) representing a multilevel theoretic framework.

We hope that we've convinced you that (a) phenomena often cannot be easily understood with a focus on a single level of analysis; (b) we can't simply use theory to know the correct levels to consider because theory can be blinding; (c) we can't simply use statistics to know the correct levels to consider because our observations are driven by what we do (i.e., what data we consider and how we analyze it); and (d) simply assuming that factors at the same or contiguous levels are the most important for understanding outcomes at a given level may be misleading. Next, we turn to some thorny issues associated with finding the correct levels to consider and offer some possible solutions to these challenges.

EXPLANATIONS OF CAUSE AND EFFECT

From an epistemological perspective, why do scholars and researchers examine multilevel issues, or for that matter, any research domain? Presumably we do so because we wish to understand particular phenomena of interest. Inherent in this statement, despite limitations of our research methodologies

and data analytic techniques, is the notion that we wish to understand *causality or causal systems*, or to answer the "why" question. If we aren't trying to understand what causes what, and why, then what is the point of theory and research? We could merely describe what we see.

Implicit in the work we do is the idea that we want to understand cause and effect relationships. If we adopt John Stuart Mill's approach to determining causality (Cook & Campbell, 1979), then three conditions must hold: (a) the cause must precede the effect in time, (b) the cause and effect have to be related (covary), and (c) other explanations of cause and effect have to be eliminated. These three conditions pose unique challenges and opportunities in a multilevel context. We begin with condition (c), that other explanations of cause and effect have to be eliminated.

Almost by definition, our world is made up of complex multilevel systems, including geographical tracts, nations, religions, schools, families, and neighborhoods. Organizations and the people who work in them are a sliver of that complex multilevel context. This is important because by acknowledging complexity across levels, we are acknowledging that most levels (actually, we would argue all levels) operate as part of an open rather than a closed system. As Katz and Kahn (1978) noted, a closed system walls off external influences on the relationships between inputs, processes, and outcomes, so that it becomes easier to see when movement in one part of the system leads to predictable patterns of movement or outcomes in other parts of the system. In contrast, in an open system the inputs, processes, and outcomes of a given part of the system may be influenced by inputs and disruptions from factors external to the system.

For example, in a closed system we could raise an individual in a box and control all aspects of the environment so that we could see how inputs (e.g., rewards) translate into outcomes (e.g., productivity) in a relatively deterministic pattern. This would be Skinner's box applied to an individual's life. Reality, however, is quite different. Individuals are buffeted by a variety of factors ranging from genetic and family effects to work, school, and national effects. We see external influences on individuals resulting from major wars (e.g., World War II, Vietnam, Iraq), unusual opportunities (e.g., the Internet), and disasters (e.g., tsunamis), as well as other geophysical events that shape food and resource accessibility (e.g., volcanic eruptions, water scarcity). Our planet is the ultimate open system, with the constant energization on earth provided by the sun allowing life to fight the constant downhill ride toward entropy and enabling it to build structure where otherwise there would be disorder. In open systems, higher levels can be affected by lower level inputs as well. For example, individuals can become infected by parasites or bombarded by high-energy particles that can alter DNA, eventually leading to adaptations or cancer.

Within such dynamic, open systems, how are we to resolve condition (c), eliminating other explanations of cause and effect? It seems impossible because of the nearly infinite array of multilevel influences that can serve as an alternative explanation of cause and effect. To address this issue, we adopt a different but related point of view. Rather than completely eliminating other explanations of cause and effect, we suggest that we should incorporate them, theorize about them, and analyze and test them. We must consider bracketing levels (Hackman, 2003) by imagining a larger box (a theoretically closed system that doesn't exist in reality) that incorporates input and disruptive factors at levels below and above the focal level of interest. In this system, we may need to bracket levels much higher or lower than the outcome of interest. This is the foundation for the multilevel nomological network we mentioned in the section Assumptions of Tight Coupling.

For example, we have studies examining how organizational factors influence happiness and job satisfaction (Grant, Christianson, & Price, 2007; Takeuchi, Chen, & Lepak, 2009), how cross-national differences are related to happiness (Schyns, 1998), how teams affect happiness (Cheshin, Rafaeli, & Bos, 2011), how marital and family factors relate to happiness over time (Tsang, Harvey, Duncan, & Sommer, 2003), and how individual attributes are related to happiness 10 years later (Costa & McCrae, 1980). Our point is that such examinations have been piecemeal, taking one or two points of a

larger system. This is not enough because happiness is likely to be influenced by all of these factors plus other factors such as macroeconomic patterns (Di Tella, MacCulloch, & Oswald, 2003).

It's fine enough if we, as scholars, tackle different components of a problem, but we tend to stay on our disciplinary tracks, rarely peering up, down, or across to take in the perspectives of scientists in other disciplines working on the same or related problems. At a recent conference on multilevel issues, we saw some of this in action. Some scholars focused on factors proximal to the individual, whereas others focused on organizational contextual factors. This led to an interesting dialogue about what matters more: organizational context or proximal team characteristics. The reality is that both perspectives are valid and we need to bridge them (Hitt, Beamish, Jackson, & Mathieu, 2007; House, Rousseau, & Thomas-Hunt, 1995). The effort to create precision in our theorizing may have the unintended side effect of creating ambiguity in what is the appropriate level of focus, because it is rarely a single level.

It is often said that the *level of theory* refers to the focal level to which generalizations are meant to apply (Mathieu & Chen, 2011; Rousseau, 1985). We can and should attempt to specify levels of theoretical interest, but we fear the process creates blinders that inhibit our ability to see the system as a system, particularly one with open inputs and disruptions. As Kozlowski and Klein (2000) stated: "The system is sliced into organization, group, and individual levels, each level the province of different disciplines, theories, and approaches. The organization may be an integrated system, but organizational science is not" (p. 3). We wholeheartedly concur and would argue that this criticism is not unique to organizational science, but more generally to social science (inclusively defined). Furthermore, at the higher level the system is also sliced into clusters (e.g., industries, economies, social and political structures), and at the lower level individuals are sliced into within-person constructs (e.g., moods, intraindividual variation in performance, neural patterns, even genetics). Scientists (including ourselves) are typically trained to slice and dice; we are trained to isolate and test. However, even theories that purport to be "integrative" rarely offer a unified treatment of phenomena considering possible antecedents, correlates, and consequences of the phenomena across four, five, or even more interacting "levels."

Rousseau (1985) pointed out an idea on which House et al. (1995) later elaborated: Processes across several different levels are often connected, and the appropriate units of analysis may span from individuals to teams to organizations to clusters of organizations. Phenomena such as marital satisfaction, learning, group decision making, altruism, and technological systems are not single-level systems. To understand them and their associated causal determinants and effects, we must adopt a more holistic and integrative mode of thinking. The "appropriate" level is the system. It is all of the levels. Truly understanding phenomena requires a comprehensive understanding of many parts of a system or a clearer definition of what you are interested in studying (recognizing that you may be excising important determining factors). There is no single focal level because the phenomenon transcends the distinct slices. The mind is more than a cluster of neurons, the individual is more than a pile of organs, the family is more than an aggregation of individuals, the organization is more than a group of people with similar attitudes or goals, and an organizational strategy is more than the CEO's vision.

Consider another example to further elucidate the importance of this type of thinking. It is well accepted that the laboratory experiment is the epitome of research designs to assess causality because of its high internal validity and ability to eliminate or control extraneous influences (Shadish, Cook, & Campbell, 2002). However, experiments are part of an open system and exist within higher levels (Hanges & Wang, 2012). Consider the psychology undergraduate who expected to participate in laboratory studies as part of the requirements for course credit. The broader context is the cause for participation in the study, and the student carries these external influences into the closed box of the laboratory environment. The student will almost certainly behave differently if he or she feels like a lab rat forced to participate as compared to a student who feels that his or her participation is making a

fundamental contribution to science. Additionally, it's possible that there is a social influence process operating, perhaps at the department or university level, so that in some universities students are happy to participate whereas at others they begrudgingly give their time. Laboratories are not as pure as we might like to believe. How might the outcome of an experiment vary as a function of the open system input factors?

How do we build a science of understanding causal connections to individual behavior when such effects can cascade across systems and levels? Our answer is to build theory that examines, tests, and incorporates the possible effects of such multilevel input factors into our models of causality. Failure to do so may lead us terribly astray because we cannot know what to conclude from even an excellently designed experiment without understanding the context.

How does this line of thinking help us with the causal problem at hand? By integrating causal factors across levels, we eliminate or reduce the likelihood that other unspecified causal factors are at play. "Other unspecified causal factors" is analogous to the omitted variables problem (see James, 1980; James, Mulaik, & Brett, 1982). Failure to include key causal variables in our models creates bias and misspecification of effects. By integrating other causal mechanisms across levels, we reduce the likelihood of omitted variables and overlooking other causal factors. As Katz and Kahn (1978) suggested,

> The closed system view implies that irregularities in the functioning of a system due to environmental influences are error variances and should be treated accordingly. According to this conception, they should be controlled out of studies of organizations. From the organization's own operations they should be excluded as irrelevant and should be guarded against. . . . Open system theory, on the other hand, would maintain that environmental influences are not sources of error variance but are integral to the functioning of a social system, and that we cannot understand a system without a constant study of the forces that impinge upon it. (p. 32)

Additionally, if we are able to understand the operation of complex causal open systems across levels, then it becomes inordinately difficult for an unspecified and unmeasured causal factor to explain the interwoven chains of events unfolding across levels. Similarly, if there is an important yet overlooked causal variable in the system, then we should be able to note its presence by perturbations within some particular level or across levels.

Pragmatically, no scholar can know or study everything. For this integration we need partnerships across disciplines, and we must learn to speak the language of other scholars investigating the same or similar phenomena. We may learn that when economists speak of shirking (Kim, Han, Blasi, & Kruse, 2015), it appears similar to social loafing (Lount & Wilk, 2014). And when economists speak of moral hazard and the 1/*n* problem (Thompson, McWilliams, & Shanley, 2014), it reminds us of social dilemmas in social psychology (Bridoux & Stoelhorst, 2016; Van Lange, Joireman, Parks, & Van Dijk, 2013). In the process, we may begin to recognize how economic forces at the national and global level such as the labor market affect the feelings, decision making, and actions of individuals within families or employees in particular organizations. Working with engineers, we may discover that they think about the effect of individual "forgetting" over time on team productivity (Nembhard, 2014; Shafer, Nembhard, & Uzumeri, 2001) and that failure to consider forgetting as a process leads to an incomplete picture about what makes some teams more effective than others. We propose that social scientists should make a concerted effort to work with scholars across diverse disciplines to explore other levels of analysis and to think more broadly about the causal systems at play. This may improve our collective understanding of why things happen the way they do.

> **Principle 1**: Adopting a holistic, open systems view of social systems and building multilevel theory to account

for phenomena will reduce the likelihood that important causal variables will be omitted.

CAUSE AND EFFECT COVARIANCE

The next issue we contemplate as a vexing challenge is condition (b): that the cause and effect have to be related (covary). It is nice to believe that when *X* happens, *Y* follows consistently and deterministically. This certainly would allow us to build a strong causal theory for *X* to *Y* effects. However, two key issues limit our ability to see such relationships: probabilistic outcomes and equifinality. First, when dealing with social phenomena, most outcomes we observe are probabilistic. It's probably nearly axiomatic that if you are hungry, you eat. There's a clear causal connection: hunger → eating. If you are hungry and it's lunchtime, do you always eat? Do you always stop your work to go get lunch? Do all people eat at lunchtime when they are hungry? Do people only eat when they are hungry? If we required hunger to precede eating and hunger to be associated with eating each and every time that hunger is present, and if we further required that in the absence of hunger, eating would not occur, how well could we establish the causal link between hunger and eating? The simple answer is that we cannot do it if we require such deterministic and absolute relationships.

In a multilevel system, we must take into account multiple factors across levels and assume that they relate to the outcome in a probabilistic, not deterministic, fashion. In addition to hunger, eating behavior at lunchtime may be influenced by contextual factors such as workload (is there a chapter to be finished?), social influence (are others going to lunch and inviting you?), and cost (is lunch expensive at the diner?). It is influenced by individual factors such as memory (did I bring my lunch today?), self-image (is it okay to eat?), and so forth. When we combine the various factors, we can begin to see the causal patterns at play, but it's important to recognize that outcomes are caused by combinations of input factors, often across levels.

One approach to determining causal patterns is to conduct the necessary condition analysis (Dul, 2016). In multicausal situations in which many factors contribute to outcomes, identifying those factors that are necessary but not sufficient to generate the outcome is helpful. Within a multilevel system, this can allow us to begin to parse the causal system into its dynamic components. Necessary conditions include those that are essential, critical, and not easily replaced by other factors but that are not sufficient by themselves to generate the outcome.

The second issue is equifinality. Multilevel and open systems are often characterized by the principle of equifinality, which means that a system can reach the same final state from differing initial conditions and by a variety of paths (Katz & Kahn, 1978). Katz and Kahn (1978) stated, "The equifinality principle simply asserts that there are more ways than one of producing a given outcome" (p. 30). If there are different paths to the outcome, then it also means that a given cause will inconsistently covary with the phenomenon of interest. In closed systems, the same initial operating parameters and inputs yield consistent outcomes. In open systems, there are many paths to any given outcome, and thus, many ways to achieve any given outcome. For example, employees might become more committed by treating them fairly and justly (Colquitt & Zipay, 2015), involving them in important decisions (Cox, Zagelmeyer, & Marchington, 2006), or empowering them to work in a more autonomous manner (Avolio, Zhu, Koh, & Bhatia, 2004). In short, there may be no single *X* that results in *Y*. There are many Xs and permutations of *X*, as well as other factors such as *Z*, *W*, and *Q*. To make the point, K. H. Roberts, Hulin, and Rousseau (1978) quoted Piaget (1971, p. 37): "Behavior is at the mercy of every possible disequilibrating factor, since it is always dependent on an environment which has no fixed limits and is constantly fluctuating" (p. 57).

How, then, do we establish causality when we don't know when or whether *X* will covary with *Y* (or whether *Z*, *W*, or *Q* will also covary with *Y*)? Again, the solution is to adopt a more holistic and integrated view of the phenomenon of interest. If we incorporate factors across levels and see that *X* probabilistically covaries with *Y* and also that *A*, *B*, and *C* covary with *X* and *Y*, then we can begin to build models of the conditions under which *X* will or

will not covary with the *Y* outcome. We suggest that even though inputs from multiple levels are complex and dynamic, there are predictable dynamics that can be quantified and understood if we examine the system as a system. We can also begin to identify the necessary conditions for phenomena (Dul, 2016).

It's not easy, but we must find and explore the multiple paths to the outcomes so that we can better understand the conditions (and boundary conditions) in which the cause and effect are related. This is more important than scholars might readily acknowledge. For example, within applied psychology, there is a belief that hiring top performers (*X*) will yield high performance (*Y*). This tends to be true . . . except when it isn't (Groysberg, Nanda, & Nohria, 2004). What might affect the ability of stars to perform in a new job? Factors include the technology of the system, the social support of the new work group, firm resources, and developmental culture (Groysberg, 2010). We cannot assume simple $X \rightarrow Y$ causal patterns when equifinality and probabilistic outcomes exist. Yet, just because it's difficult doesn't mean we shouldn't try. As K. H. Roberts et al. (1978) stated,

> We can and should try to observe, quantify, and explain regularities in responses of individuals and groups in organizational contexts. By regularities we do not mean there must be a one-to-one correspondence between a stimulus and a response or between two responses. (p. 6)

These efforts will result in the development of multilevel nomological networks. Nomological networks are interlocking systems of constructs and relationships that constitute the fundamental components of our theories (cf. Binning & Barrett, 1989; Cronbach & Meehl, 1955; Messick, 1995).

> **Principle 2:** Adopting a holistic, open systems view of social systems and building multilevel theory to account for phenomena will improve our understanding and increase the likelihood that causal relationships will be properly specified.

CAUSE AND TEMPORAL PRECEDENCE

We now turn our attention to the final remaining condition for establishing causality—condition (a), that the cause must temporally precede the effect. This requirement seems both obvious and reasonable. If *X* is to cause *Y*, then surely *X* must precede *Y* in time. In some ways, however, this condition of determining causality may well be the most difficult of all to establish in multilevel open systems for three reasons: scaling of time, lagged outcomes, and fragile homeostasis.

Scaling of Time

Scaling of time refers to the notion that the rate at which processes unfold often varies across the level of hierarchy within a system. To discuss time, we must first recognize that time can cut across all focal levels yet can also exist at a lower level than the focal level. All units, whatever the level, change or evolve in some way. This means that repeated observations of a given entity (e.g., person, group, organization) reside at a lower level of analysis than the level of the entity itself (e.g., within-person, within-group, within-organization). For example, if data on student reading proficiency were collected four times throughout the academic year, then the repeated observations over time would be considered a within-person (student) factor.

However, this is not to imply that all observations over time reside at the lowest level. To make this statement clear, we provide an example. Assume we are interested in employee performance and believe it is affected by the employees' levels of task-specific self-efficacy and their individual levels of conscientiousness. In addition, we hypothesize that individual-level performance is also influenced by organizational culture. To test these hypotheses, we measure employee performance over four quarters with self-efficacy assessed each time (within-persons). We measure employee conscientiousness once, at the beginning of the study (person-level). We obtain a measure of organizational culture based on a survey completed the previous year by employees and subsequently aggregated up to the organizational level (organizational level; i.e., all individuals residing in a given organization are assigned the

score for culture). In this model, time exists at the bottommost level, because observations take place within individuals (time or quarters), individual conscientiousness is at Level 2 (individual), and organizational success (i.e., previous performance) at Level 3 (organizational):

$$\text{Level 1} \quad Y_{tij} = \pi_{0ij} + \pi_{1ij}(\text{Self-Efficacy}_{tij}) + e_{tij}$$

$$\text{Level 2} \quad \pi_{0ij} = \beta_{00j} + \beta_{01k}(\text{Conscientiousness}_{ij}) + r_{0ij}$$

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

$$\text{Level 3} \quad \beta_{00j} = \gamma_{000} + \gamma_{001}(\text{Organizational Culture}_{k}) + u_{00j}$$

$$\beta_{01j} = \gamma_{010} + u_{01j}$$

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

In most hierarchical linear or random coefficient models, time is treated at the lowest level. Now imagine the same study but instead of a single measure of organizational culture based on the previous year's annual culture survey, we instead obtained ratings of culture each quarter. This situation creates substantial complexity not present in the previous example because we cannot simply stick several observations of organizational culture under the level of individuals. Time is a lower level than each entity, but multiple observations over time do not necessarily exist at the bottom rung of levels effects. In this latter example, we might be interested in the magnitude and trajectory of organizational culture. For example, if the organization is increasing in its culture for innovation, then perhaps that stimulates individuals to try harder and perform better, whereas decreasing culture for innovation might have the opposite effect. Conversely, from a social loafing perspective, perhaps individuals try less hard when the organization is doing increasingly well, but they work harder if the trend is diminishing because they perceive a threat to the organization (and therefore, their jobs).

Our typical multilevel model is not generally well suited to handling analyses of this type because we are now looking at intercepts and slope coefficients of the higher level construct (i.e., culture) over time as inputs to the lower level effects. Normally the intercepts and slope coefficients of the lower level are used as outcomes to be predicted by higher level effects. Here we have intercepts and slope coefficients of both lower level and higher level entities. It can be done, but it's not business as usual, and we can't simply tuck repeated observations under individuals as a lower level factor.

We introduced this example because entities evolve over time, and temporal effects may reside at higher or lower levels depending on what's being measured (and when it is being measured). However, temporal assessments always reside at a lower level than the entity being measured because they are nested within the entity or unit. We also used this example because it brings up a second factor: the scaling of time. It is generally understood that processes at lower levels take place more rapidly than higher levels. Chemical transition states have incredibly short lifetimes of a few femtoseconds (10^{-15} seconds), the time required for electron redistribution (Schramm, 2011), and neurons fire in milliseconds (Diba, Amarasingham, Mizuseki, & Buzsáki, 2014); people may make decisions in split-seconds (G. A. Klein, 1998), whereas other behaviors or decisions might take minutes, hours, days, or even longer to unfold (e.g., Bragger, Hantula, Bragger, Kirnan, & Kutcher, 2003). If we wish to say *X* must precede *Y* to establish causality, what does this mean when all parts of a multilevel system are in motion and some parts move or evolve more rapidly than others? Furthermore, various relationships may have rhythms or patterns over time in addition to having different scaling (Mitchell & James, 2001; Zaheer, Albert, & Zaheer, 1999).

We might ask, "Does the cue ball cause the eight ball to go in the pocket?" If the cue ball moves first and strikes the eight ball, and then the eight ball goes in the pocket, we may be able to argue that the cue ball caused the eight ball to go in the pocket. What if we ask, "Does training in billiards cause the eight ball to go in the pocket?" Training is a process that takes place over weeks, months, and years, whereas the strike of a cue ball takes a fraction of a second. How do we relate training, with its long time frame, with the eight ball going in a pocket in less than a second? Additionally, training effects often

dissipate with time. We face similar challenges when trying to relate something like FAA regulations to aircrew behavior on a flight.

It doesn't seem controversial to suggest that FAA regulations alter flight crew behavior. That's their raison d'etre. But how would we measure and model this relationship? It would be difficult. If we gather measures of the FAA regulations today and relate them to pilot behavior tomorrow, we wouldn't see much, if any, relationship. *X* (FAA regulations) most likely causes *Y* (pilot behavior), and we see and measure *X* preceding *Y*. However, we are unlikely to detect the relationship because *X* is nearly invariant over this time window. This is true even if we measured FAA regulations first and then assessed pilot behavior 6 months later. FAA regulations evolve over months and years, whereas aircrew behavior can take place in minutes, hours, and days. We have a mismatch in timescale associated with levels.

To see the causal connection between FAA regulations and flight crew behavior, we require either molar observations that transcend time (e.g., number of regulations or content of regulations; aircrew safe landings, aircrew errors, flight disruptions over time) or multiple observations over time at the timescale of each of the entities involved (regulation time 0; aircrew safety time 1; regulation time 1; aircrew safety time 2; and so forth). Thus, we need to know the durations, scaling, and time lag (Mitchell & James, 2001; Zaheer et al., 1999). As an illustration, if the idea is that FAA regulations involving pilot training influence aircrew behavior, then we might explore the impact of this intervention using an interrupted time-series design over years or decades wherein we collect data on pilot behavior both before and after changes were made regarding FAA regulations (Shadish et al., 2002). The types of measures of FAA regulations would depend on what is being measured and the rate at which FAA regulations can evolve. The time frame for the subsequent assessments of aircrew behavior would be dependent on whether FAA changes are likely to have immediate impact or whether it takes 6 months or a year to see the changes. Pick the wrong time frame and nothing will be seen (George & Jones, 2000). Another option might be to approach the question qualitatively, talking with pilots and regulators before and after the changes to see what evidence might exist for various causal relationships.

The point being made is that powerful causal influences of higher level variables can be hidden because of differing time frames across levels and the inability to connect the time frames across levels in a meaningful fashion. Governmental laws (e.g., Civil Rights Act of 1964) almost certainly affect human behavior, but getting a quantitative measure of governmental regulation to correlate with specific individual behaviors may prove difficult unless we pay attention to the role of time.

We can't ignore important causal influences of higher level variables simply because they evolve more slowly or because it's difficult to assess causality. Some of the most important variables may be of this sort—slowly changing, difficult to measure, yet powerful as a causal antecedent.

Lagged Outcomes

The second issue relevant to time and temporal precedence in multilevel systems is the potential delay between a cause and the effect. Consider a study that purports to examine whether new CEOs tend to implement a new strategy that affects organizational performance. If we stay at the CEO and organizational level and do not approach this question from a multilevel systems perspective, then it might seem reasonable to assess the outcome a year later. But if we adopt a more sophisticated multilevel perspective, we may realize that CEO changes in strategy unfold at different rates across different levels across the organization. Consider an organization that hires a new CEO. She comes on board and after a few months of talking with executives, employees, and customers, she decides on a new strategic approach for the company. The top-level executives then work on plans, the organization's information technology professionals identify technologies to enable strategic pursuit, human resources decision makers implement changes to organizational structure and culture, training and development specialists train employees, and everyone adjusts to the inevitable bumps along the way. Sometime later the strategy begins to effect changes in relationships with customers, and as word spreads among customers through marketing

and spillover effects, changes in sales and earnings become apparent. How long might it take for the CEO's changes to show their impact on return on equity measures? Is a year enough? Two years? The delayed effect on outcomes can easily mask our ability to detect relationships (covariance) if we don't think about the multilevel system generating those outcomes.

Fragile Homeostasis

The final issue to consider with the idea of seeing the cause preceding the effect in a multilevel system is the notion of fragile homeostasis. This is related to the idea of lagged effects, but it's the result of a distinct process. Lagged effects take time to manifest, but there's a smooth and even connection to the outcome. The challenge is getting the timing right in order to unveil the causal connection. Fragile homeostasis and homeostatic breakdown (or breakout) is different because it is more abrupt.

Complex social structures ranging from teams to governments and societies exist in a state of quasi-homeostasis. Many social collectives (e.g., teams, schools, organizations, professional societies) have a clearly defined structure and set of processes that allow them to exist independent of individual membership. Members may come and go, but the overall structure of the social collective remains largely static. The higher level entities are in homeostasis, but it is a partial homeostasis because they are changing, growing, shrinking, and evolving. When a cause enters the system, it may exhibit no easily discernable effect on the outcome, yet a change may occur. Over time the social system can reach a critical state preceding homeostatic breakdown, when abrupt change can take place. It's similar to conditions in chemistry when water can be superheated yet not boil. Then, when jostled even slightly the water roils and boils over the edge of the cup. Or conversely, liquids can be supercooled but still in a liquid state, but then the smallest perturbation can result in a nearly instant shift to a solid state. George and Jones (2000) referred to this process as *discontinuous change*.

What does it look like when applied to social structures? Consider Rosa Parks. She didn't cause the civil rights movement. Society was pressurized, with inequality, prejudice, and even violence creating differences in quality of life and opportunity for a large segment of the population. As slight was heaped upon slight, the social context became ripened for a phase shift—a change to something new. Then something happens to disturb the system and the structure abruptly changes. For another example, the assassination of the Archduke Ferdinand didn't cause World War I. The world was poised for war and the assassination precipitated the outcome. Teams abruptly coalesce and rally to become something special, and they sometimes explode unexpectedly. Organizations appear fine until a tightly wound spring such as excess leverage precipitates a dramatic fall. Marriages seem fine until financial difficulties rattle the relationship.

Fragile homeostasis and homeostatic breakdowns (or breakouts) make it difficult to determine whether the temporal sequencing of *X* before *Y* is in fact causing *Y*. Perhaps *X* does cause *Y*, over time, creating a state of fragile homeostasis, but then *Z* occurs and creates the homeostatic breakdown and abrupt phase shift in *Y*. As a result, we may ascribe the cause to *Z*, not *X*, but both *Z* and *X* could be the primary causative factor creating the breakdown in fragile homeostasis. Consider a student who has been generally unhappy and disconnected for a large number of years. What causes the student to drop out of school isn't necessarily some specific triggering event, *Z* (e.g., moving to a new school because a parent takes a new job, a failing grade), but rather *Z* creates a context in which the impact of *X* on *Y* became more salient or pronounced. If *X* is educational engagement and *Y* is school satisfaction (or staying in school), then levels of engagement may be correlated with satisfaction. However, even the least engaged student may not plan to quit school unless some triggering event strengthens the relationship between lack of engagement and dissatisfaction (e.g., moving to a new school, failing a class).

Addressing the condition of temporal precedence is not an easy task to accomplish. Levels often operate on a different scaling of time, but fragile homeostasis and homeostatic breakdown can make the situation more complex to understand. Stars evolve over billions of years, but when the conditions are right,

massive stars can become supernovas within minutes. National governments and societies may lumber along, and it's business as usual for decades or centuries, but then when conditions are right, homeostatic breakdown takes place and revolutions occur.

We can and must do a better job of attending to the role of time in the complex open environments we call *multilevel systems* (George & Jones, 2000; Mitchell & James, 2001). We have to think more carefully about time-boundedness, the role of time scaling, lagged effects, and homeostatic breakdowns. As K. H. Roberts et al. (1978) stated:

> Many relations we study are time- and place-bound. That is, a relation observed in an organization today may not be observed in another organization and may not be observed in the same organization next year. Although we are reasonably sensitive to environmental influences on relations, we are generally insensitive to the time boundaries of our data. (p. 22)

We tend to be insensitive to the time boundaries and evolutionary processes of our theories. We haven't even mentioned reciprocal dynamics and the mediation that occurs between any putative cause and effect (cf. James et al., 1982; Mathieu & Taylor, 2007; Zhang, Zyphur, & Preacher, 2009). For every action and reaction, there's a potential intermediating effect that might also deserve examination, and sometimes it is the reciprocal effects of *X* and *Y* over time. Additionally, there can be other variables mediating effects. Indeed, between any two mediators there exists the possibility of a third mediator. If we want to truly understand a system, we may need to capture the micromediation processes taking place over time. Another issue is that in an open system for every 0 point in time, there's a $t-1$ point in time. That is, in open systems there's always the possibility of a time frame that precedes the time frame being investigated, and the preceding time frame could contain a key causative factor. With all this challenge and complexity, what should we do?

Social scientists can begin to address these issues by being more aware of the multifaceted ways that time exerts effects in multilevel contexts. With the long time frame for some effects, we may need to borrow from other disciplines such as historiography, ethnography, and econometrics to consider how best to predict and test potential cause and effect relationships. We may also need to use both qualitative designs and computational modeling or simulations to understand the dynamics involved (Kozlowski, Chao, Grand, Braun, & Kuljanin, 2013). We can borrow ideas from other sciences, even chemistry, astronomy, or physics, to metaphorically (and perhaps analytically) grasp how partially homeostatic systems in a quasi-equilibrium function as causal influences. We have to try new and different approaches to our theoretical and mathematical modeling of causes and effects.

> **Principle 3:** Adopting a holistic, open systems view of social systems and building multilevel theory to account for phenomena will enable scholars to more effectively detect cause and effect patterns over time, with lagged effects, and in the presence of fragile homeostasis.

SUBJECTIVE VERSUS OBJECTIVE EFFECTS

When trying to identify the multilevel effects at play within a system, we should consider the distinction between subjective, or perceived, effects and objective or physically manifested effects. Can effects at the level of cosmological events affect human behavior at work? Perhaps this is a laughable question, because who would care? The skip between levels seems simply too vast. But consider that objective cosmological effects have physically shaped our reality, and they include asteroid strikes that have led to the large-scale extinction of dinosaurs and supernovas that may have disrupted our atmosphere and caused small-scale extinctions. Mass coronal ejections have disrupted our electrical grids and are highly likely to do so again in the future. We can consider the impact of solar flares on communication patterns of individuals and societies. We don't need to turn to astrology to find strong evidence that physical cosmological events can and most likely

will shape human behavior in the future. However, the vast scaling of time influences whether we attend to these issues.

The long time frame for such events makes studying such phenomena uninteresting to most organizational scientists. It's unlikely to happen in their lifetimes, and it's a rare event, so why bother? Yet these questions and issues are germane to human productivity and survival. Planetary-level effects such as climate change alter availability of food and water, and these will later affect geopolitical power and stability among nations. It's difficult to consider these relationships and their impact on individuals, families, and organizations, but it's important to try.

Cosmology or planetary change doesn't have to manifest as a physical event to influence human behavior. As long as people think about and perceive cosmological events as important, then they will exert effects on people. For example, we would never advocate taking astrology seriously as a science. However, in 2012 as much as 42% of the U.S. population thought astrology was either "sort of scientific" or "very scientific" (National Science Foundation, 2014). For believers, it doesn't matter if it's real; it's enough to believe. Thus, we can ask, "do astrological perceptions affect individual behavior?" Most likely they do for a subset of people. Ronald Reagan and his wife Nancy were said to have taken astrology seriously during his presidency; it affected the scheduling of important events (S. V. Roberts, 1988). Do others in work environments take astrology seriously? Might such beliefs affect perceived risk (Sjöberg & af Wåhlberg, 2002)?

We can also ask whether actual organizational policies affect employee behavior or whether perceptions and attributions about organizational policies matter more. We are reminded of a friend who lamented working long work weeks at a law firm when he had a newborn at home. We asked, "Doesn't your firm have paternity leave policies?" He said, "Yes, but everyone knows you'll never make partner if you take it." In many respects, it doesn't matter if the statement is true, because the statement is true for him. If he believes in his perceptions, then he will act accordingly regardless of whether the perceptions are objectively true. If such perceptions are shared, they manifest as collective constructs, shaping individual behavior whether or not they are accurate. There are many instances in human history of behavior being shaped not by reality but by the perception of reality.

Objective impact of lower level factors can exist. People can be exposed to chemical compounds, viruses, or bacteria that alters human chemistry and changes our cognitions and behaviors over time. Genetics (which focuses on the molecular level) can affect people. Using twin studies, researchers have found that as much as 30% of the variance in job satisfaction may be associated with genetic factors (Arvey, Bouchard, Segal, & Abraham, 1989). Changes in brain structure or chemistry over time clearly can shape behavior. Researchers have associated ethical decision processes with neural activation patterns in the brain using fMRI techniques (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Robertson et al., 2007).

Perceptions of lower level phenomena may be important too. If people think there are deadly germs all around them, then this will shape their behavior regardless of whether the germs actually exist. If people believe in a heritable trait theory of intelligence, then they will act and react differently to mistakes and errors than people who believe in malleable, contextually driven intelligence (Dweck & Leggett, 1988; James & LeBreton, 2012). Again, the more shared the perceptions, the more mutually reinforcing they become, eventually manifesting as higher level effects, even if they are higher level effects about perceptions of lower level phenomena.

> **Principle 4:** Scholars must attend to potential objective and subjective effects across levels of analysis to build a more integrated understanding of how a multilevel system influences outcomes.

ON VARIANCE AND VARIATION

Our typical modeling approach for multilevel data is to gather data, set up identifiers for units at two or more levels, partition the variance into higher- and lower level portions, and then determine the significance of predictor variables at a particular

level for predicting the outcome variance at that level (cf. Dansereau et al., 1984; Raudenbush & Bryk, 2002). When modeling group effects, we are testing group predictors against the component of variance in individual-level scores that exists across groups. And when modeling individual level effects, we are testing individual predictors against the component of variance that represents variation of individuals within groups. This approach does well for what it is supposed to do: provide significance tests of specific effects within a given set of data. However, this approach will not tell us where the variance lives, unless we happen to get lucky and sample, measure, and analyze the correct variables at the correct level at the correct point in time.

By thinking of variance as a fixed pie to be sliced up into appropriate tests at various levels, we ignore some very important issues. First, the variance we see in our data may or may not be the important variance that exists. Second, we tend to collect data as a snapshot, yet phenomena unfold over time. Finally, we argue that variance is not a fixed pie. As shown in Figure 1.2, variance can increase and decrease over time, both within and across levels, as part of an adaptive or evolutionary process. Sometimes both individual and team variance can increase over time as in the propagation and adoption of innovations. At other times or in other contexts, individual variance can shrink while variance across teams can increase, such as when there are strong team effects influencing individual behavior. There are probably occasions when both individual variance within teams and variance between teams shrink. Perhaps this could be the result of strong socialization processes within an organization organized by units such as the military (James, Demaree, Mulaik, & Ladd, 1992). All of these patterns have profound effects on how much variance is seen at what level over time and our ability to see what is going on. This attaches to our previous conversation about time: what is increasing or decreasing over time and why?

Variance is not fixed to a particular level over time; it shifts and morphs as social systems unfurl their effects. Consider a brief thought experiment: Imagine for a moment that we have 500 individuals within 100 teams. On a measure of work satisfaction (assessed using a 10-point Likert-type scale), nearly all of the teams have a mean work satisfaction score of 8 (within sampling error). All of the individual-level ratings of work satisfaction (i.e., each employee within each team) also hover around an 8. Thus, there are no significant differences between teams and no significant differences within teams (i.e., between individuals nested in a particular team). At what level in the system does the action exist for the work satisfaction construct? Answer: No one can tell you on the basis of variance. Individual and team effects are indistinguishable, and there's no evidence to suggest there is a particular level effect (or there are both individual and team effects and they can't be separated).

Now, what if we told you that previously, the individual-level scores vary widely from 1 to 10, but there were no specific team-level differences? Assume we've collected data moment by moment for many weeks. Scores are all over the place in the beginning, but there's a small team of three people, with one new charismatic member, all with scores of 7 or 8. Then we see the small group with scores around 8 expand to five. Eventually, the entire work unit averages close to 8. Over time, the initial team solidifies, individual scores drawing tighter around 8. Abruptly a second team rapidly transitions from scores ranging from 1 to 10 to a mean around 8. Then a third team follows the same pattern. This continues until all 100 teams average near 8. What is the level at which the system is operating? Clearly, the initial level was individual within team, with emergence or social contagion processes drawing people together. The next phase, however, operated at the team level, with the contagion "jumping" from one team to another. Perhaps because of the initial team, the organizational leaders realized there was a better way to do things and began to implement new best practices, one team at a time. Or perhaps other teams saw something working for one team and incorporated the practices into their own unit.

We can imagine other possibilities. What if the means of each team vary widely but individuals within each team begin to converge on similar values so that one team averages near 3 while another averages near 7 or 8, with all members within a

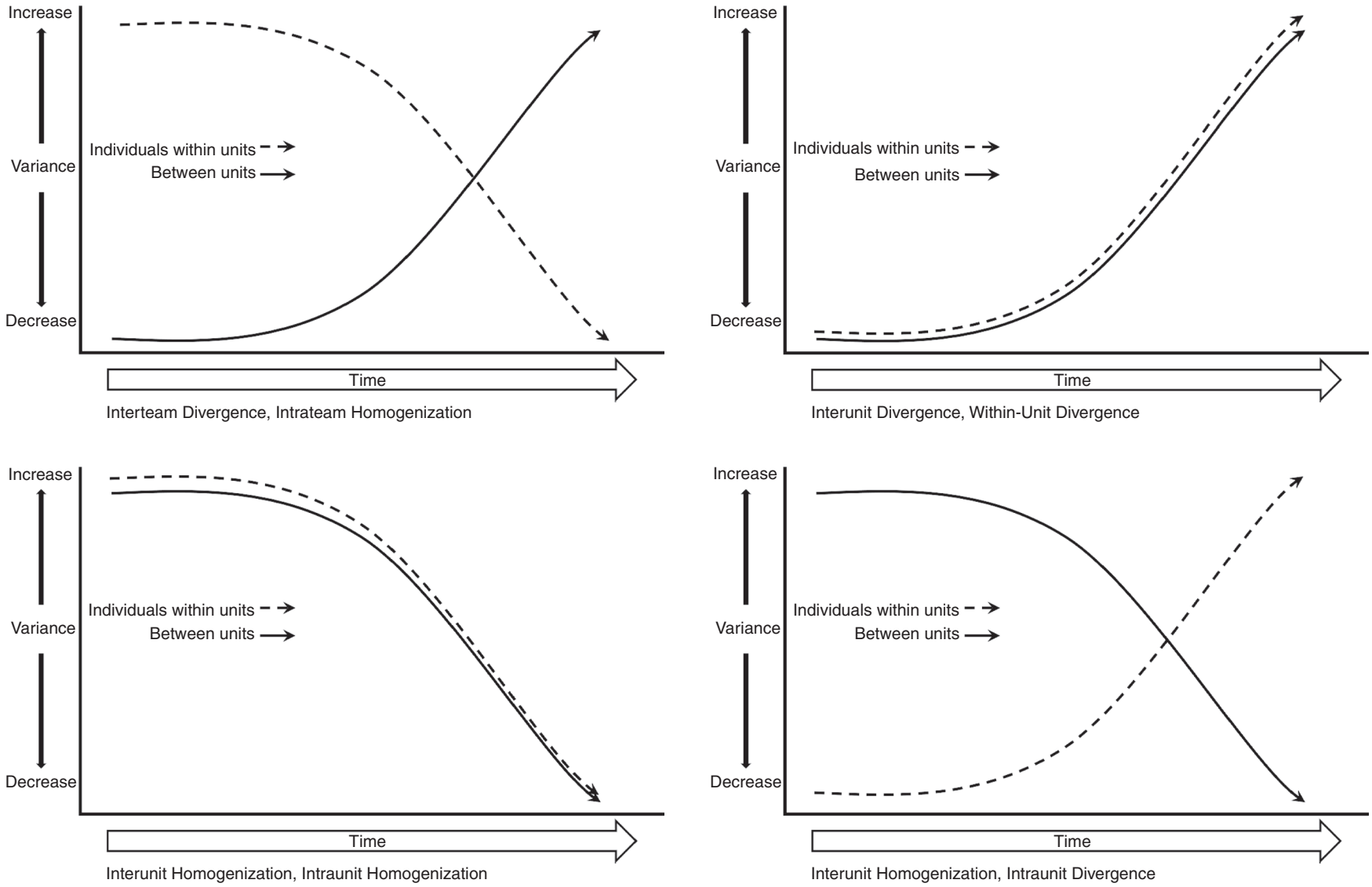


FIGURE 1.2. Variance shifts over time.

team clustering. For a long time the teams vary from one another, but individuals within teams converge around their respective team means. After some time, all teams begin to slowly converge toward 8, so that the final result is that both individuals and teams hover around 8. This is a very different pattern than the preceding one. This situation clearly indicates that team context effects exist, so at least part of the action is at the team level. Over time, however, the teams converge, perhaps through the sharing of team best practices that make all of the teams more effective and therefore more satisfying. This might occur, for example, if innovation diffuses across teams.

There are three key points to consider in this thought experiment. First, by examining how variance changes across levels over time, we went from having no clear idea as to how the multiple-level system was operating to having a pretty good idea of what might have happened in each situation. Second, we see that effects can shift up and down levels over time as the variance components shift. Third, we have to measure processes over time to see the change in variance components.

The important issue to keep in mind is that structures or entities across levels come and go. There was once no Google, but a couple of people got together and began to change the world. Perhaps Google will continue to change the world for centuries, but it's worth remembering that a large-scale survival analysis suggested the average lifespan of a business is about a decade (Daepp, Hamilton, West, & Bettencourt, 2015). Teams form, storm, norm, perform (Chang, Bordia, & Duck, 2003; Mannix & Jehn, 2004) and then many disband, get reassigned or restructured, or become incorporated into some other entity. Members come and go and sometimes they change the character of the teams they join or leave (Mathieu & Chen, 2011). Most social structures have some degree of impermanence.

There was a time when there was little or no variance across large manufacturing companies in the use of assembly line production technologies. Companies didn't use this approach, so the variance was near zero—no organizational effect on assembly line technology. Then certain companies implemented the assembly line system, and variance on this technology increased. Other large manufacturing companies began to adopt the approach, and variance across companies increased in the use of this technology. Now an organizational effect on the use of assembly lines could be detected; there was unit-level variance. However, variance in the types of jobs performed by each individual decreased as the assembly line technology was implemented. Eventually, nearly all large manufacturing companies adopted assembly line production methods, and the variance returned to near zero.

Where does the level exist for the diffusion of innovation of assembly line production techniques? It may exist at the organizational level, but many factors are in operation and we could only detect the organizational effect using organizational level prediction of variance during a small window of time (George & Jones, 2000; Mitchell & James, 2001). Today we won't see much variance in whether or not assembly line technologies are used. But these manufacturing systems have a profound effect on how employees work and feel on the job. Are we to say that there's no organizational effect of assembly line production technologies on employee behavior because we can't see variance? This would be a highly misleading conclusion because everything about the system, which exists at the organizational level, affects what employees do (and cannot do).

As scholars, we must be more sensitive to how variance changes over time and across levels. This means we have to think more carefully about the meaning of time, its scaling, and the types of effects we may see (lagged, homeostatic breakdown). By observing how units at multiple levels shift in variance, we can begin to detect the fundamental processes at play (e.g., social contagion, innovation diffusion). Failure to see variance does not mean that a higher level effect does not exist. The effects may be at play and they may be profound and powerful, but precisely because they are profound and powerful, there's little variance across higher units.

Returning to the example of FAA regulations: They work so everyone uses them, and the variance is not discernible. However, it drives the system, process, and outcomes regardless of the variance components. One merely needs to look at the history of aviation (e.g., Bryson, 2013) to see the reckless and wild early days of flight. Indeed, variance in

flight safety was high in the 1930s. The FAA came to exist to reduce this variance and create standard operating procedures that would lead to flight safety, and it worked. The variance is gone but the system is there, nevertheless.

> **Principle 5:** Developing a deep understanding of complex multilevel systems requires scholars to move beyond snapshots of variance to reframing the relative partitioning of variance within and between units over time.

DUAL PROCESSES IN MULTILEVEL SYSTEMS

A final issue to consider when thinking about identifying the correct level are the dual processes of variance creation and restriction in multilevel systems. As shown in our earlier examples, units within a multilevel system change, adapt, and evolve over time. Families, teams, organizations, societies, and nations are not static entities. For example, a given team has not existed for an eternity. It comes into being for an externally driven purpose or because a collection of individuals shares a common vision and set of goals. Eventually, a group of individuals begin to restrict variance on key attitudes and share core values. This is one example of the process of emergence (Kozlowski et al., 2013). We see variance between groups/teams increasing and variance within teams decreasing on core values. Teams can also experience divergence, cases in which team members increasingly drift apart, eventually leading to the dissolution of the team.

The dual processes of variance creation and variance restriction forge and dissolve higher and lower level entities and their associated effects. The processes of variance creation and restriction occur both within units as well as across units. Exhibit 1.1 introduces some potential factors that can create or restrict variance within or across units.

This list is not meant to be exhaustive. Instead, it is meant to be a point of departure for thinking about the systems and process that drive variance creation and reduction and eventually lead to the existence and dissolution of higher and lower entities and their associated effects. We attempt to highlight how countervailing forces act to create levels effects.

EXHIBIT 1.1

Variance Creation and Restriction Processes

Variance creation	Variance restriction
Exploration	Exploitation
Innovation	Assimilation
Serendipity	Structuration
Adaptation	Elimination
Differentiation	Conformation
Contagion	Truncation
Deregulation	Regulation

March (1991) described *exploration* as the process of experimenting, investigating, and innovating to find new products, markets, technologies, and services. The process of looking for and finding these new innovations necessarily creates variance across work units and organizations as each pursues a distinct path. Eventually, however, once the innovation is created or identified, organizations must become organized to take advantage of, or *exploit*, this new opportunity. Apple created the iPod and organized its marketing and manufacturing units to pursue the common goal of taking advantage of this new product. Initially the systems and processes innovating the iPod created variance, and then the systems and processes geared to take advantage of the iPod reduced certain types of variance (e.g., product goals). It happened at the industry level as well. The entry of the iPod created variance as other organizations rushed to take advantage of the opportunity, and eventually many technology players had some type of mp3 player on the market, reducing variance. We've seen this cycle many times, ranging from electricity, to the Internet, to smartphones and tablets.

There are similar forces with innovation and assimilation. New ideas, technologies, services, and markets can create competitive advantage for certain firms or teams. This creates variance among units as some are "haves" and some are "have nots." As other units see the success of the early adopters, they are more likely to assimilate the innovations into

their own systems and structures. Over time, best practices become institutionalized, and nearly all units become "haves," thus restricting variance. Thus, we see that variance becomes an indicator for the ebb and flow of new approaches toward success. This sometimes results in a physical assimilation such as in an acquisition when a firm acquires another for people, products, or technologies. Assimilation of values or process approaches happens at not only the organizational level but the teams level as well, as some teams observe others engaging in practices that enable success. Over time, variance increases as some teams branch out, then variance decreases as teams become increasingly similar due to consistency in values, practices, and technologies. Most firefighting, hospital, and police units are similar, yet some show variance on key factors. Some hospitals are embracing a professional and open approach to communication, some police units are reaching out to the community, and some firefighting crews are embracing diversity. This creates variance and as we see which approaches work, other units adopt effective approaches, eventually reducing variance.

Serendipity should not be discounted as a force for variance. Random mutations create speciation over time, and random errors can result in fortunate discoveries (R. M. Roberts, 1989). Post-it® notes were the unexpected result of an adhesive experiment. For some time, 3M was the only firm to have this technology but over time other firms obtained it, creating variance. Now it seems that low adhesion adhesives are used in products everywhere, and many firms have it, reducing variance. Through *structuration* organizations become highly structured to take advantage of the new discoveries. Random changes that result in powerful positive outcomes eventually lead to new structures that are designed to take advantage of the serendipity. Teams and other social units can experience this. There has been a long tradition of people who get together to play games socially. Some of these groups began doing it online and today, many social gamers interact virtually. Groups adopt what works for other groups, eventually restricting variance of certain types.

The forces described above create adaptation at the individual, group, and organizational levels. Entities that adapt and evolve continue to exist as the context changes. However, failure to adapt to contextual forces results in *elimination* or *dissolution* of the entity. Adaptation creates variance until the adaptation is generally embraced or assimilated across units. Elimination reduces variance as units that fail to adapt fall out of existence. These forces can have strong influences on our ability to detect higher level effects. Consider, for example, the idea that strategic human resource management practices need to be aligned with organizational strategies. This seems obvious, even intuitive, yet the empirical research supporting this perspective is notably weak (e.g., Huselid, 1995). One possibility is that organizations that have not aligned their talent strategies with their business strategies are eliminated from the competitive landscape. If true, it would mean a key quadrant enabling detection of the alignment hypothesis would be missing. In other words, if the theory that businesses need to align their talent strategies with their business strategies to succeed is correct, then the ability to test that theory would be compromised because the firms that fail to do so would drop out of the variance space. This can play out at the individual or team level as well. We might have a theory that people who are poor fits with an organization and who perform poorly have a different value system from those who are poor fits and perform well or those who are good fits and perform poorly. If we tried to run such a study, how many people do we expect to see remain in an organization who are poor fits, performing poorly? The elimination process would drive most of these people out. Adaptation and elimination processes can have profound effects on the variances we are able to see, both within and between units of interest.

We also see forces for *differentiation* and *conformation*. Some people, groups, and organizations embrace difference and actively try to be different from others. If everyone was the same, the world would be uninteresting. If all organizations are the same, there is no competitive advantage and they would all be subject to the tyranny of mediocrity. But there's a certain inelasticity to difference. If an entity is too different, too out there, then it can struggle, whether the entity is an individual, team, or organization. There are pressures for conformity

that restrict variance. You wouldn't go to a restaurant and expect wait staff to be rude or dressed in pajamas or both. Individuals can't go to stores with pants on their heads and shirts on their legs. Part of the conformance effect is practical. Some things don't work and they are eliminated from the repertoire. However, social constraints driving conformance are powerful but don't necessarily have a pragmatic purpose. There's no reason that certain hand gestures (e.g., the "okay sign") have to mean something positive in some cultures and something negative in others. These are socially constructed interpretations. But conformance in each culture means that it is or is not appropriate to engage in the behavior. It's true for organizational strategy, too. Businesses operate differently in China, Germany, and France than they do in the United States.

Social and physical *contagion* processes have the ability to generate variance through bottom-up emergence effects. An individual or small group of people can start a social movement, create a new product, or develop a better service, and this can cascade out through a team, an organization, an industry, or a nation. Physical contagion can cascade out as well, changing how people interact with each other. As the contagion takes hold, variance is created but over time, either it is assimilated or truncated. Social *truncation* occurs when opposing forces work to squelch, control, or sometimes adopt and assimilate the social contagion. Eventually, either contagion wins and variance is restricted because most units adopt the new "thing," or truncation wins by squeezing the contagion out.

Regulation, including legislation, is designed to reduce variance by fiat, usually to enhance efficiency or increase safety. Regulations can also be derived from social construction. We have regulations intended to reduce risk in the finance industry, but they also restrict variance in the business practices and investment strategies that can be pursued. Some hospitals have regulations for surgical teams to use checklists because they've been shown to reduce surgical errors by ensuring that all teams engage in consistent effective practices. Likewise, the FAA has checklists for aircrews. We also have regulations for what people can do, what they eat, or how they act. The counter force for regulation is *deregulation*. By removing constraints on individual, team, and organizational behavior, we allow variation in behaviors and practices. At times such variation can result in innovation and adaptation, at other times it can result in negative outcomes.

We suggest that the dual forces of variance creation and variance restriction shape the formation, evolution, and dissolution of units across levels over time. Attending to how various types of forces affect variance will help us better understand the system, as a system, rather than thinking about piecemeal slices of the system. As Kozlowski et al. (2013) pointed out, the emergence of levels is complex because it incorporates both process and structure. Process involves dynamic interactions among entities, and structure is the manifestation over time of a collective property. Both process and structure show their effects by transitions over time, and we further suggest that process and structure evolve in response to the dual forces of variance creation and variance restriction over time. As Kerlinger (1973) discussed, the concept of variation, or variance, is essential to all scientific efforts as the research process seems to understand or explain variation.

We agree and take this statement further in two ways. First, we suggest that understanding phenomena in open systems requires an understanding and explanation of variation at multiple levels of analysis. Second, we argue that variance is not static; it shifts and changes over time in response to dual forces of creation and restriction, and this occurs at multiple levels. Third, we need to think about the variance of the system as a whole in order to understand the phenomena of interest. We can begin by examining variance of the component parts, but eventually we need to build a more integrative understanding of the sources and outcomes of variance across levels.

SUMMARY

We suggest that researchers must think in more complex ways about the phenomena being studied. As Hanges and Wang (2012) noted, complex multilevel and adaptive systems have tangled feedback loops among the system's elements. As a result, causal influences flow in multiple directions within

the system and across levels. We have to think more carefully about the system as a system and work toward interdisciplinary understanding of the phenomena we care about.

As we progress, we have to take seriously the notion that even if there are multiple Xs across levels predicting *Y*, understanding which *X* is causing *Y* is challenging because *X* doesn't always cause *Y*, seeing *X* precede *Y* doesn't mean it is the only influence on *Y*, and not seeing *Y* when *X* is present doesn't mean that *X* isn't important. Additionally, we have to be much more serious about the role of time in multilevel phenomena. In particular, we have to be sensitive to scaling of time, lagged effects, and fragile homeostasis. Timescales can mask or enhance our observed of relationships among variables of interest.

Finally, we have to think about variance creation and restriction processes that can influence our ability to see and detect phenomena. Most things are in a process of dynamic change at different timescales. As a result, we should think about how different forces for variance creation and restriction operate over time to influence outcomes. Do laws affect marital behavior? Certainly. But if laws change slowly and we gather observations over a year when they are unchanging, we may not detect the effects unless we adopt qualitative approaches of inquiry.

We began with the question "What is the appropriate level?" The answer is, there is no appropriate level. The level is the system, which is what we should examine if we want to understand multilevel causality. The ability to examine the system depends on timing of observation, and bracketing is not enough. Our efforts to understand such multilevel systems may require new interdisciplinary approaches and a willingness to broaden our own perspectives and conceptualizations of phenomena at hand.

References

Arvey, R. D., Bouchard, T. J., Segal, N. L., & Abraham, L. M. (1989). Job satisfaction: Environmental and genetic components. *Journal of Applied Psychology*, *74*, 187–192. http://dx.doi.org/10.1037/0021-9010.74.2.187

Avolio, B. J., Zhu, W., Koh, W., & Bhatia, P. (2004). Transformational leadership and organizational commitment: Mediating role of psychological empowerment and moderating role of structural distance. *Journal of Organizational Behavior*, *25*, 951–968. http://dx.doi.org/10.1002/job.283

Baltes, B. B., Zhdanova, L. S., & Parker, C. P. (2009). Psychological climate: A comparison of organizational and individual level referents. *Human Relations*, *62*, 669–700. http://dx.doi.org/10.1177/0018726709103454

Barrick, M. R., Stewart, G. L., Neubert, M. J., & Mount, M. K. (1998). Relating member ability and personality to work-team processes and team effectiveness. *Journal of Applied Psychology*, *83*, 377–391. http://dx.doi.org/10.1037/0021-9010.83.3.377

Baysinger, M. A., Scherer, K. T., & LeBreton, J. M. (2014). Exploring the disruptive effects of psychopathy and aggression on group processes and group effectiveness. *Journal of Applied Psychology*, *99*, 48–65. http://dx.doi.org/10.1037/a0034317

Binning, J. F., & Barrett, G. V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology*, *74*, 478–494. http://dx.doi.org/10.1037/0021-9010.74.3.478

Blakely, T. A., & Woodward, A. J. (2000). Ecological effects in multi-level studies. *Journal of Epidemiology and Community Health*, *54*, 367–374. http://dx.doi.org/10.1136/jech.54.5.367

Bragger, J. D., Hantula, D. A., Bragger, D., Kirnan, J., & Kutcher, E. (2003). When success breeds failure: History, hysteresis, and delayed exit decisions. *Journal of Applied Psychology*, *88*, 6–14. http://dx.doi.org/10.1037/0021-9010.88.1.6

Bridoux, F., & Stoelhorst, J. W. (2016). Stakeholder relationships and social welfare: A behavioral theory of contributions to joint value creation. *Academy of Management Review*, *41*, 229–251. http://dx.doi.org/10.5465/amr.2013.0475

Bryson, B. (2013). *One summer: America, 1927*. Toronto, Ontario, Canada: Doubleday Canada.

Bureau of Labor Statistics. (2015a). *Census of fatal occupational injuries charts, 1992–2015 (final data)*. Retrieved from https://www.bls.gov/iif/oshwc/cfoi/cfch0014.pdf

Bureau of Labor Statistics. (2015b). *Employer-reported workplace injuries and illnesses—2015*. Retrieved from https://www.bls.gov/news.release/archives/osh_10272016.pdf

Bureau of Labor Statistics. (2016). *Career outlook: Adrenaline jobs: High-intensity careers*. Retrieved from https://www.bls.gov/careeroutlook/2016/article/adrenaline-jobs.htm

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. http://dx.doi.org/10.1037/0021-9010.83.2.234

Chang, A., Bordia, P., & Duck, J. (2003). Punctuated equilibrium and linear progression: Toward a new understanding of group development. *Academy of Management Journal*, *46*, 106–117. http://dx.doi.org/10.5465/30040680

Chen, G., Mathieu, J. E., & Bliese, P. D. (2004). A framework for conducting multilevel construct validation. In F. J. Yammarino & F. Dansereau (Eds.), *Research in multilevel issues: Multilevel issues in organizational behavior and processes* (Vol. 3, pp. 273–303). Oxford, UK: Elsevier. http://dx.doi.org/10.1016/S1475-9144(04)03013-9

Cheshin, A., Rafaeli, A., & Bos, N. (2011). Anger and happiness in virtual teams: Emotional influences of text and behavior on others' affect in the absence of non-verbal cues. *Organizational Behavior and Human Decision Processes*, *116*, 2–16. http://dx.doi.org/10.1016/j.obhdp.2011.06.002

Civil Rights Act of 1964, Pub. L. No. 88-352, 78 Stat. 241, July 2, 1964.

Colquitt, J. A., & Zipay, K. P. (2015). Justice, fairness, and employee reactions. *Annual Review of Organizational Psychology and Organizational Behavior*, *2*, 75–99. http://dx.doi.org/10.1146/annurev-orgpsych-032414-111457

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Costa, P. T., Jr., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology*, *38*, 668–678. http://dx.doi.org/10.1037/0022-3514.38.4.668

Cox, A., Zagelmeyer, S., & Marchington, M. (2006). Embedding employee involvement and participation at work. *Human Resource Management Journal*, *16*, 250–267. http://dx.doi.org/10.1111/j.1748-8583.2006.00017.x

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302. http://dx.doi.org/10.1037/h0040957

Daepp, M. I. G., Hamilton, M. J., West, G. B., & Bettencourt, L. M. A. (2015). The mortality of companies. *Journal of the Royal Society Interface*, *12*, 20150120. http://dx.doi.org/10.1098/rsif.2015.0120. Published 1 April 2015.

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

Diba, K., Amarasingham, A., Mizuseki, K., & Buzsáki, G. (2014). Millisecond timescale synchrony among hippocampal neurons. *The Journal of Neuroscience*, *34*, 14984–14994. http://dx.doi.org/10.1523/JNEUROSCI.1091-14.2014

Di Tella, R., MacCulloch, R. J., & Oswald, A. J. (2003). The macroeconomics of happiness. *The Review of Economics and Statistics*, *85*, 809–827. http://dx.doi.org/10.1162/003465303772815745

Dul, J. (2016). Necessary condition analysis (NCA): Logic and methodology of "Necessary but Not Sufficient" causality. *Organizational Research Methods*, *19*, 10–52. http://dx.doi.org/10.1177/1094428115584005

Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, *95*, 256–273. http://dx.doi.org/10.1037/0033-295X.95.2.256

George, J. M., & Jones, G. R. (2000). The role of time in theory and theory building. *Journal of Management*, *26*, 657–684. http://dx.doi.org/10.1177/014920630002600404

Gerstein, D. R. (1987). To unpack micro and macro: Link small with large and part with whole. In J. C. Alexander, B. Giesen, R. Münch, & N. J. Smelser (Eds.), *The micro–macro link* (pp. 86–111). Los Angeles: University of California Press.

Grant, A. M., Christianson, M. K., & Price, R. H. (2007). Happiness, health, or relationships? Managerial practices and employee well-being tradeoffs. *The Academy of Management Perspectives*, *21*, 51–63. http://dx.doi.org/10.5465/AMP.2007.26421238

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, *293*, 2105–2108. http://dx.doi.org/10.1126/science.1062872

Groff, E. R., Weisburd, D., & Yang, S. M. (2010). Is it important to examine crime trends at a local "micro" level? A longitudinal analysis of street to street variability in crime trajectories. *Journal of Quantitative Criminology*, *26*, 7–32. http://dx.doi.org/10.1007/s10940-009-9081-y

Groysberg, B. (2010). *Chasing stars: The myth of talent and the portability of performance*. Princeton, NJ: Princeton University Press. http://dx.doi.org/10.1515/9781400834389

Groysberg, B., Nanda, A., & Nohria, N. (2004). The risky business of hiring stars. *Harvard Business Review*, *82*, 92–100, 151.

Gully, S. M., & Phillips, J. M. (2005). A multilevel application of learning and performance orientations to individual, group, and organizational outcomes. In J. Martocchio (Ed.), *Research in personnel and human resources management* (Vol. 24, pp. 1–51). Greenwich, CT: JAI Press/Elsevier Science. http://dx.doi.org/10.1016/S0742-7301(05)24001-X

Hackman, J. R. (1987). The design of work teams. In *Handbook of organizational behavior* (pp. 315–342). Englewood Cliffs, NJ: Prentice-Hall.

Hackman, J. R. (2003). Learning more by crossing levels: Evidence from airplanes, hospitals, and orchestras. *Journal of Organizational Behavior*, *24*, 905–922. http://dx.doi.org/10.1002/job.226

Hanges, P. J., & Wang, M. (2012). Seeking the Holy Grail in organizational science: Uncovering causality through research design. In S. W. J. Kozlowski (Ed.), *The Oxford handbook of organizational psychology* (pp. 79–116). New York, NY: Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780199928309.013.0003

Hitt, M. A., Beamish, P. W., Jackson, S. E., & Mathieu, J. E. (2007). Building theoretical and empirical bridges across levels: Multilevel research in management. *Academy of Management Journal*, *50*, 1385–1399. http://dx.doi.org/10.5465/AMJ.2007.28166219

Hodgson, G. M. (1998). The approach of institutional economics. *Journal of Economic Literature*, *36*, 166–192.

House, R. J., Rousseau, D. M., & Thomas-Hunt, M. (1995). The third paradigm: Meso organizational research comes to age. *Research in Organizational Behavior*, *17*, 71–114.

Huselid, M. A. (1995). The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of Management Journal*, *38*, 635–672. http://dx.doi.org/10.5465/256741

Indik, B. (1968). The scope of the problem and some suggestions toward a solution. In B. P. Indik & F. K. Berrien (Eds.), *People, groups and organizations* (pp. 3–26). New York, NY: Teachers College Press.

James, L. R. (1980). The unmeasured variables problem in path analysis. *Journal of Applied Psychology*, *65*, 415–421. http://dx.doi.org/10.1037/0021-9010.65.4.415

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, *67*, 219–229. http://dx.doi.org/10.1037/0021-9010.67.2.219

James, L. R., Demaree, R. G., Mulaik, S. A., & Ladd, R. T. (1992). Validity generalization in the context of situational models. *Journal of Applied Psychology*, *77*, 3–14. http://dx.doi.org/10.1037/0021-9010.77.1.3

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. http://dx.doi.org/10.1037/0021-9010.69.1.85

James, L. R., & LeBreton, J. M. (2012). *Assessing implicit personality through conditional reasoning*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/13095-000

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). *Causal analysis: Assumptions, models, and data*. Beverly Hills, CA: Sage.

Jepperson, R., & Meyer, J. W. (2011). Multiple levels of analysis and the limitations of methodological individualisms. *Sociological Theory*, *29*, 54–73. http://dx.doi.org/10.1111/j.1467-9558.2010.01387.x

Johnson, S., Cooper, C., Cartwright, S., Donald, I., Taylor, P., & Millet, C. (2005). The experience of work-related stress across occupations. *Journal of Managerial Psychology*, *20*, 178–187. http://dx.doi.org/10.1108/02683940510579803

Katz, D., & Kahn, R. L. (1978). *The social psychology of organizations* (2nd ed.). New York, NY: Wiley.

Keltner, D., & Haidt, J. (1999). Social functions of emotions at four levels of analysis. *Cognition and Emotion*, *13*, 505–521. http://dx.doi.org/10.1080/026999399379168

Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd ed.). New York, NY: Holt, Rinehart and Winston.

Kim, A., Han, K., Blasi, J. R., & Kruse, D. L. (2015). Anti-shirking effects of group incentives and human-capital-enhancing HR practices. In A. Kauhanen (Ed.), *Advances in the economic analysis of participatory & labor-managed firms* (pp. 199–221). Bingley, England: Emerald Group. http://dx.doi.org/10.1108/S0885-333920150000016014

Klein, G. A. (1998). *Sources of power: How people make decision*. Cambridge, MA: MIT Press.

Klein, K. J., Conn, A. B., Smith, D. B., & Sorra, J. S. (2001). Is everyone in agreement? An exploration of within-group agreement in employee perceptions of the work environment. *Journal of Applied Psychology*, *86*, 3–16. http://dx.doi.org/10.1037/0021-9010.86.1.3

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

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.

Lebreton, J. M., Burgess, J. R. D., Kaiser, R. B., Atchley, E. K. P., & James, L. R. (2003). The restriction of

variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, *6*, 80–128. http://dx.doi.org/10.1177/1094428102239427

Lount, R. B., Jr., & Wilk, S. L. (2014). Working harder or hardly working? Posting performance eliminates social loafing and promotes social laboring in workgroups. *Management Science*, *60*, 1098–1106. http://dx.doi.org/10.1287/mnsc.2013.1820

Mannix, E., & Jehn, K. A. (2004). Let's norm and storm, but not right now: Integrating models of group development and performance. In E. Salas (Ed.), *Research on managing groups and teams: Vol. 6. Time in groups* (pp. 11–37). Bingley, England: Emerald Group. http://dx.doi.org/10.1016/S1534-0856(03)06002-X

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, *2*, 71–87. http://dx.doi.org/10.1287/orsc.2.1.71

Mathieu, J. E., & Chen, G. (2011). The etiology of the multilevel paradigm in management research. *Journal of Management*, *37*, 610–641. http://dx.doi.org/10.1177/0149206310364663

Mathieu, J. E., & Taylor, S. R. (2007). A framework for testing meso-mediational relationships in organizational behavior. *Journal of Organizational Behavior*, *28*, 141–172. http://dx.doi.org/10.1002/job.436

McNeely, C. A., Nonnemaker, J. M., & Blum, R. W. (2002). Promoting school connectedness: Evidence from the national longitudinal study of adolescent health. *The Journal of School Health*, *72*, 138–146. http://dx.doi.org/10.1111/j.1746-1561.2002.tb06533.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Mitchell, T. R., & James, L. R. (2001). Building better theory: Time and the specification of when things happen. *Academy of Management Review*, *26*, 530–547.

National Science Foundation. (2014). Science and technology: Public attitudes and understanding. In *Science and engineering indicators 2014* (NSB 14-01). Arlington, VA: National Center for Science and Engineering Statistics. https://www.nsf.gov/statistics/seind14/content/chapter-7/chapter-7.pdf

Nembhard, D. (2014). Cross training efficiency and flexibility with process change. *International Journal of Operations & Production Management*, *34*, 1417–1439. http://dx.doi.org/10.1108/IJOPM-06-2012-0197

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event- and interval-contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin*, *27*, 771–785. http://dx.doi.org/10.1177/0146167201277001

O*Net. (2017). *Occupational Information Network (O*NET)*. Retrieved from https://www.onetcenter.org/overview.html

Piaget, J. (1971). *Biology and knowledge: An essay on the relations between organic regulations and cognitive processes*. Chicago, IL: University of Chicago Press.

Raudenbush, S. W., & Bryk, A. S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. *Review of Research in Education*, *15*, 423–475. http://dx.doi.org/10.3102/0091732X015001423

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

Roberts, K. H., Hulin, C. L., & Rousseau, D. M. (1978). *Developing an interdisciplinary science of organizations*. San Francisco, CA: Jossey-Bass.

Roberts, R. M. (1989). *Serendipity: Accidental discoveries in science*. Hoboken, NJ: Wiley & Sons.

Roberts, S. V. (1988, May 4). White House confirms Reagans follow astrology, up to a point. *New York Times*. http://www.nytimes.com/1988/05/04/us/white-house-confirms-reagans-follow-astrology-up-to-a-point.html.

Robertson, D., Snarey, J., Ousley, O., Harenski, K., DuBois Bowman, F., Gilkey, R., & Kilts, C. (2007). The neural processing of moral sensitivity to issues of justice and care. *Neuropsychologia*, *45*, 755–766. http://dx.doi.org/10.1016/j.neuropsychologia.2006.08.014

Rousseau, D. M. (1985). Issues of level in organizational research: Multilevel and crosslevel perspectives. In L. L. Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 7, pp. 1–37). Greenwich, CT: JAI Press.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*, 262–274. http://dx.doi.org/10.1037/0033-2909.124.2.262

Schramm, V. L. (2011). Enzymatic transition states, transition-state analogs, dynamics, thermodynamics, and lifetimes. *Annual Review of Biochemistry*, *80*, 703–732. http://dx.doi.org/10.1146/annurev-biochem-061809-100742

Schyns, P. (1998). Crossnational differences in happiness: Economic and cultural factors explored. *Social Indicators Research*, *43*, 3–26. http://dx.doi.org/10.1023/A:1006814424293

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Shafer, S. M., Nembhard, D. A., & Uzumeri, M. V. (2001). The effects of worker learning, forgetting, and heterogeneity on assembly line productivity. *Management Science*, *47*, 1639–1653. http://dx.doi.org/10.1287/mnsc.47.12.1639.10236

Shinn, M., & Rapkin, B. D. (2000). Cross-level research without cross-ups in community psychology. In J. Rappaport & E. Seidman (Eds.), *Handbook of community psychology* (pp. 669–695). New York, NY: Kluwer/Springer. http://dx.doi.org/10.1007/978-1-4615-4193-6_28

Simon, H. A. (1973). The organization of complex systems. In H. H. Pattee (Ed.), *Hierarchy theory: The challenge of complex systems* (pp. 1–27). New York, NY: George Braziller.

Singer, J. D. (1961). The level-of-analysis problem in international relations. *World Politics*, *14*, 77–92. http://dx.doi.org/10.2307/2009557

Sjöberg, L., & af Wåhlberg, A. (2002). Risk perception and new age beliefs. *Risk Analysis*, *22*, 751–764. http://dx.doi.org/10.1111/0272-4332.00066

Snijders, T. A., & Kenny, D. A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships*, *6*, 471–486. http://dx.doi.org/10.1111/j.1475-6811.1999.tb00204.x

Takeuchi, R., Chen, G., & Lepak, D. P. (2009). Through the looking glass of a social system: Cross-level effects of high-performance work systems on employees' attitudes. *Personnel Psychology*, *62*, 1–29. http://dx.doi.org/10.1111/j.1744-6570.2008.01127.x

Thompson, P. B., McWilliams, A., & Shanley, M. (2014). Creating competitive advantage: A stakeholder view of employee ownership. *International Journal of Strategic Change Management*, *5*, 262–279. http://dx.doi.org/10.1504/IJSCM.2014.064468

Tsang, L. L. W., Harvey, C. D., Duncan, K. A., & Sommer, R. (2003). The effects of children, dual earner status, sex role traditionalism, and marital structure on marital happiness over time. *Journal of Family and Economic Issues*, *24*, 5–26. http://dx.doi.org/10.1023/A:1022478919443

Van Lange, P. M., Joireman, J., Parks, C. D., & Van Dijk, E. (2013). The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes*, *120*, 125–141. http://dx.doi.org/10.1016/j.obhdp.2012.11.003

Zaheer, S., Albert, S., & Zaheer, A. (1999). Time scales and organizational theory. *Academy of Management Review*, *24*, 725–741.

Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods*, *12*, 695–719. http://dx.doi.org/10.1177/1094428108327450

CHAPTER 1

A Multilevel Approach to Theory and Research in Organizations

Contextual, Temporal, and Emergent Processes

Steve W. J. Kozlowski
Katherine J. Klein

Organizations are multilevel systems. This axiom—the foundation of organizational systems theory—is reflected in the earliest examples of organizational theory, including the Hawthorne Studies (Roethlisberger & Dickson, 1939), Homans's theory of groups (1950), Lewin's field theory (1951), sociotechnical systems theory (Emery & Trist, 1960), Likert's theory of organizational effectiveness (1961), Thompson's (1967) theory of organizational rationality, and Katz and Kahn's (1966) social organizational theory, to name but a few. Further, this axiom continues to provide a foundation for virtually all contemporary theories of organizational behavior. Yet, despite the historical tradition and contemporary relevance of organizational systems theory, its influence is merely metaphorical. The system is sliced into organization, group, and individual levels, each level the province of different disciplines, theories, and approaches. The organization may be an integrated system, but organizational science is not.

There are signs that this is beginning to change, that we are moving toward the development of an integrated conceptual and

methodological paradigm for organizational science. We have witnessed the evolution, over the last two decades, of multilevel frameworks that have well-developed conceptual foundations and associated analytic methodologies. Organizational science is moving toward the development of a paradigm that can bridge the micro-macro gap in theory and research. We are witnessing the maturation of the multilevel paradigm in organizational science.[1]

As with all maturation, however, the process has not proceeded without pain. The roots of the multilevel perspective are spread across different disciplines and literatures, obscured by the barriers of jargon, and confused by competing theoretical frameworks and analytic systems. Although there are some explicit efforts to specify general multilevel frameworks for organizational science (e.g., Dansereau, Alutto, & Yammarino, 1984; House, Rousseau, & Thomas-Hunt, 1995; Klein, Dansereau, & Hall, 1994; Roberts, Hulin, & Rousseau, 1978; Rousseau, 1985), real and apparent differences among the frameworks have created the impression of little common ground (e.g., George & James, 1994; Klein, Dansereau, & Hall, 1995). Further, the best way to evaluate multilevel theories (e.g., George & James, 1993; Yammarino & Markham, 1992) and establish emergent constructs (e.g., James, Demaree, & Wolf, 1993; Kozlowski & Hattrup, 1992; Schmidt & Hunter, 1989) is much contested. No single source exists to cut across these differences and to guide the interested researcher in the application of multilevel concepts. This contributes to confusion and limits the development of multilevel theories. Accordingly, a review of the current literature is likely to leave those who are tempted to test multilevel theories intrigued yet confused—inspired yet wary.

Our goal in this chapter is to help resolve this confusion by synthesizing and extending prior work on the development of multilevel theory and research for organizations. The chapter is organized into three sections. In the first section, we review the theoretical roots of the multilevel perspective as it relates to theory building and research in organizations. The epistemological foundation and several basic assumptions for the levels perspective are rooted in general systems theory (von Bertalanffy, 1968) and related variants. Early and enduring applications of the levels perspective to research on organizational characteristics and organizational climate had a formative impact on the development of the levels perspective and continues to exert considerable influence.

In the second section, we clarify, synthesize, and extend basic principles to guide multilevel theory development and to facilitate empirical research. We first provide principles to guide the development of multilevel theory. We discuss theoretical issues pertaining to the origin and direction of phenomena across levels, unit and entity specification, time, and factors affecting the degree of coupling or linkage of phenomena across levels. With this theoretical foundation in place, we next explain and illustrate how to specify and operationalize multilevel models. Critical issues focus on establishing an alignment among levels of theory, constructs, and measures. We also specify different types of levels models, examine implications for research sampling, and provide an overview of data-analysis issues.

In the third section, we extend multilevel organizational theory by drawing particular attention to relatively neglected *bottom-up* processes. Many organizational theories are implicitly or explicitly *top-down,* addressing the influence of macro levels (for example, organization or group characteristics) on micro levels (for example, individuals). Such models focus on contextual factors at higher levels that constrain and influence lower-level phenomena. Bottom-up models describe phenomena that have their theoretical origin at a lower level but have emergent properties at higher levels (for example, psychological and organizational climate, individual and team effectiveness, individual and organizational learning). Models of emergence have been largely restricted to isomorphic composition processes, which has limited the development of bottom-up multilevel theory and research. We elaborate discontinuous, configural compilation processes and describe how they allow the conceptualization of alternative manifestations of emergence. We use this perspective to extend extant models of emergence. We develop a *typology of emergence* to illustrate and explain several alternative models that range from isomorphic composition to discontinuous compilation. We are hopeful that these alternative models of emergence will stimulate and guide research on these central but neglected multilevel phenomena.

Foundations for Multilevel Theory in Organizations

Conceptual Underpinnings

General Systems Theory

General systems theory (GST) has been among the more dominant intellectual perspectives of the twentieth century and has been shaped by many contributors (e.g., Ashby, 1952; Boulding, 1956; Miller, 1978; von Bertalanffy, 1972). Systems concepts originate in the "holistic" Aristotelian worldview that the whole is greater than the sum of its parts, in contrast with "normal" science, which tends to be insular and reductionistic. The central goal of GST is to establish principles that generalize across phenomena and disciplines—an ambitious effort that is aimed at nothing less than promoting the unity of science.

Systems principles are manifest as analogies or *logical homologies*. Logical homologies represent identical concepts (that is, *isomorphism*), and parallel processes linking different concepts (that is, *homology*), that generalize to very different systems phenomena (von Bertalanffy, 1972). For example, it is noted that open systems counteract the second law of thermodynamics—entropy—by importing energy and information from the external environment, and transforming it, to maintain homeostasis. Feedback and servomechanisms are the basis for the purposive responses of cybernetic systems. Organizational systems are proposed to have analogous structures and processes (e.g., Katz & Kahn, 1966; Miller, 1978).

Whether one takes a more macro (Parsons, 1956, 1960) or micro (Allport, 1954) perspective, the influence of GST on organizational science has been pervasive. Unfortunately, however, that influence has been primarily metaphorical. The bureaucratic–closed systems–machine metaphor is contrasted with a contingent–open systems–living organism metaphor. Although metaphor has important value—virtually all formal theory is rooted in underlying metaphor (Morgan, 1983)—lack of specificity, formal identity, and precise definition can yield truisms that mislead and fail the test of science (Pinder & Bourgeois, 1982; Bourgeois & Pinder, 1983). GST has exhibited heuristic value but has contributed relatively little to the development of *testable principles* in the organizational sciences (Roberts et al., 1978). It is to this latter concern that the multilevel perspective is directed.

As social systems, organizations are qualitatively distinct from living cells and other concrete physical systems. The goal of the multilevel perspective is not to identify principles that generalize to other types of systems. Although laudable, such an effort must often of necessity gloss over differences between qualitatively different systems in order to maintain homology across systems (compare Miller, 1978). *The primary goal of the multilevel perspective in organizational science is to identify principles that enable a more integrated understanding of phenomena that unfold across levels in organizations.*

Macro and Micro Perspectives

Fundamental to the levels perspective is the recognition that micro phenomena are embedded in macro contexts and that macro phenomena often emerge through the interaction and dynamics of lower-level elements. Organizational scholars, however, have tended to emphasize either a micro or a macro perspective. The macro perspective is rooted in its sociological origins. It assumes that there are substantial regularities in social behavior that transcend the apparent differences among social actors. Given a particular set of situational constraints and demographics, people will behave similarly. Therefore, it is possible to focus on aggregate or collective responses and to ignore individual variation. In contrast, the micro perspective is rooted in psychological origins. It assumes that there are variations in individual behavior, and that a focus on aggregates will mask important individual differences that are meaningful in their own right. Its focus is on variations among individual characteristics that affect individual reactions.

Neither single-level perspective can adequately account for organizational behavior. The macro perspective neglects the means by which individual behavior, perceptions, affect, and interactions give rise to higher-level phenomena. There is a danger of superficiality and triviality inherent in anthropomorphization. Organizations do not behave; people do. In contrast, the micro perspective has been guilty of neglecting contextual factors that can significantly constrain the effects of individual differences that lead to collective responses, which ultimately constitute macro phenomena (House et al., 1995; Klein et al., 1994; Roberts et al., 1978; Rousseau, 1985).

Macro researchers tend to deal with global measures or data aggregates that are actual or theoretical representations of lower-level phenomena, but they cannot generalize to those lower levels without committing errors of misspecification. This renders problematic the drawing of meaningful policy or application implications from the findings. For example, assume that we can demonstrate a significant relationship between organizational investments in training and organizational performance. The intuitive generalization—that one could use the magnitude of the aggregate relationship to predict how individual performance would increase as a function of increased organizational investments in training—is not supportable, because of the well-known problem of ecological inference. Relationships among aggregate data tend to be higher than corresponding relationships among individual data elements (Robinson, 1950; Thorndike, 1939). This fact continues to be a significant difficulty for macro-oriented policy disciplines—sociology, political science, economics, education policy, epidemiology—that attempt to draw individual-level inferences from aggregate data.

Micro researchers suffer from an obverse problem, which also makes the desire to influence human resource management policy difficult. We may, for example, be able to show that individual cognitive ability increases individual performance. However, we cannot then assert that selection systems that produce higher aggregate cognitive ability will necessarily yield improved organizational performance. Perhaps they will, but that inference is not directly supported by individual-level analyses. Misspecifications of this sort, however, are not unusual (Schmidt, Hunter, McKenzie, & Muldrow, 1979). Such "atomistic fallacies," in which organizational psychologists suggest team- or organization-level interventions based on individual-level data, are common in our literature.

A levels approach, combining micro and macro perspectives, engenders a more integrated science of organizations. House and colleagues (1995) suggest the term *meso* because it captures this sense that organizational science is both macro and micro. Whatever it is called, we need a more integrated approach. The limitations that the organizational disciplines suffer with respect to influencing policy and applications can be resolved through the development of more complete models of organizational phenomena—models that are system-oriented but do not try to capture the complexity of the entire system. Instead, by focusing on significant and salient phenomena, conceptualizing and assessing at multiple levels, and exhibiting concern about both top-down and bottom-up processes, it is possible to build a science of organizations that is theoretically rich and application-relevant.

Formative Theory Development: The Emergence of a Levels Perspective

Early efforts to conceptualize and study organizations as multilevel systems were based in the interactionist perspective (Lewin, 1951) and focused on the construct of organizational climate.[2] Those early efforts played a significant role in developing a "levels" perspective. Interactionists see behavior as a function of both person and situation, with the nature of the combined effect broadly conceived (as, for example, additive, multiplicative, and reciprocal; see Schneider, 1981; Terborg, 1981). Thus behavior is viewed as a combined result of contextual and individual-difference effects. The interactionist perspective has had a pervasive influence on organizational research. It has played a dominant role in shaping research on climate, first posited by Lewin, Lippitt, and White (1939). It continues to exert influence through research on person-organization fit.

As organizational psychology developed as a distinct subdiscipline in the 1950s, organizational climate emerged as a central construct for understanding organizational effectiveness. Researchers of this era described climate as a representation of "organizational stimuli" or "environmental characteristics" presumed to affect individual behavior and attitudes. Forehand and Gilmer (1964) reviewed the climate literature, highlighting problems of conceptualization and measurement. They criticized researchers' failure to consistently and clearly distinguish whether climate was viewed as an objective property of the organization or as an individual perception, and they bemoaned the resulting confusion regarding whether climate should be assessed at the organizational level, via objective characteristics, or at the individual level, via perceptions.

James and Jones's (1974) subsequent review helped to dispel much of this confusion. They distinguished objective characteristics of the organizational context, which are the antecedents of climate, from individuals' interpretive perceptions, which ascribe

meaning to the context. This conceptualization views climate perceptions as a result of both contextual and individual influences. In addition, James and Jones distinguished psychological (that is, individual-level) climate from organizational climate, arguing that homogeneous perceptions could be aggregated to represent climate as a property of the organization. James and Jones's conclusions influenced the nature of climate research for the next two decades.

There were two critical contributions of this formative research on the development of a levels perspective in organizational science. First, this research made top-down cross-level contextual effects salient, establishing the need to conceptualize and assess organization, subunit, and group factors that had the potential to affect individual perceptions, attitudes, and behavior. This energized a stream of research that linked organizational structure and technology to individual attitudes (e.g., Herman & Hulin, 1972; James & Jones, 1976; Rousseau, 1978b). As this research progressed, models were elaborated to include mediating perceptions. Many studies were conducted that demonstrated that individual-level climate and/or job-characteristics perceptions mediated the linkage between contextual factors at higher levels (group, subunit, or organization) and individual-level outcomes (e.g., Brass, 1981, 1985; Oldham & Hackman, 1981; Kozlowski & Farr, 1988; Rousseau, 1978a). This work emphasized the importance of top-down cross-level contextual effects on lower-level phenomena. *Thus group and organization factors are contexts for individual perceptions, attitudes, and behaviors and need to be explicitly incorporated into meaningful models of organizational behavior.*

The second contribution of this research was to make salient emergent phenomena that manifest at higher levels. Although organizational policies, practices, and procedures are the antecedents of individual-level climate perceptions, individuals in organizations do not exist in a vacuum. People in groups and subunits are exposed to common features, events, and processes. They interact, sharing interpretations, which over time may converge on consensual views of the group or organizational climate (James, 1982; Kozlowski & Hattrup, 1992). Processes such as attraction, selection, and attrition; socialization (Schneider & Reichers, 1983); and leadership (Kozlowski & Doherty, 1989) also operate to reduce

the variability of individual differences and perceptions, facilitating common interpretations of the climate. In such conditions, individual-level perceptions can be averaged to represent higher-level group, subunit, or organizational climates (Jones & James, 1979; Kozlowski & Hults, 1987; Schneider & Bowen, 1985). This work emphasized the importance of bottom-up emergent processes that yield higher-level phenomena. *Thus individual social-psychological processes can be manifest as group, subunit, and organizational phenomena and need to be explicitly incorporated into meaningful models of organizational behavior.*

Multilevel Organizational Theory and Research

Overview

Although interest in the development and testing of multilevel theoretical models has increased dramatically in the past decade, there have been relatively few efforts to provide multilevel theoretical frameworks for organizational researchers (e.g., House et al., 1995; Klein et al., 1994; Rousseau, 1985). Multilevel theory building presents a substantial challenge to organizational scholars trained, for the most part, to "think micro" or to "think macro" but not to "think micro *and* macro"—not, that is, to "think multilevel." Our goal is to explain fundamental issues, synthesize and extend existing frameworks, and identify theoretical principles to guide the development and evaluation of multilevel models.

In the first part of this section, we describe multilevel theoretical processes, providing insights into and principles for "thinking multilevel." The issues we examine are central to the development of multilevel theories and provide conceptual guidance for theorists seeking to develop specific multilevel models. In the second part of this section, we focus on model operationalization. Most of the difficulties of conducting multilevel research have concerned the consequences of incongruent levels among constructs, measures, or analyses (for example, misspecification errors, aggregation biases, ecological correlation; see Burstein, 1980; Firebaugh, 1979; Freeman, 1980; Hannan, 1991; Robinson, 1950; Thorndike, 1939). We provide principles to guide the interested researcher through the problem of model specification.

The principles we derive are intended to be general guidelines applicable to most circumstances; they are not immutable laws. We acknowledge at the onset that the complexity of the issues involved in multilevel theory makes exceptions to the general principles inevitable. In such cases, theory takes precedence—that is the one overarching principle.

Principles for Multilevel Organizational Theory Building

This section describes fundamental theoretical processes that provide the underpinnings for developing multilevel theories. We hope to assist readers in emulating and extending the best of current multilevel thinking. Toward this end, we highlight established principles and consider provocative new possibilities for multilevel theory building and research. For ease of presentation, we present central principles of multilevel theory building and research organized around the *what, how, where, when,* and *why* (and *why not*) of multilevel theoretical models.

What

On what should multilevel theory building and research focus? The possibilities are virtually endless, reflecting the full breadth of organizational processes, behavior, and theory. Nevertheless, a few guidelines regarding the process of choosing a focus for study are possible. First, we urge scholars to begin to fashion their theoretical models by focusing on the endogenous construct(s) of interest: What phenomenon is the theory and research attempting to understand? The endogenous construct, or dependent variable, drives the levels, constructs, and linking processes to be addressed by the theory. Too frequently, researchers begin theory development with the antecedents of interest: "These are interesting constructs; I wonder how well they predict generic outcomes." Such an approach invites the development of a trivial or misspecified theory. Without careful explication of the phenomenon of interest, it is exceedingly difficult to specify a meaningful network of potential antecedents.

PRINCIPLE: *Theory building should begin with the designation and definition of the theoretical phenomenon and the endogeneous construct(s) of interest.*

Second, multilevel theory is neither always needed nor always better than single-level theory. Micro theorists may articulate theoretical models capturing individual-level processes that are invariant across contexts, or they may examine constructs and processes that have no meaningful parallels at higher levels. Similarly, macro theorists may develop theoretical models that describe the characteristics of organizations, distinct from the actions and characteristics of organizational subunits (groups, individuals). Although we think that such phenomena are likely to be rare, in such cases multilevel theory building is not necessary.

Finally, theorists may also find it impractical to develop multilevel models for processes, relationships, and outcomes new to organizational science; that is, when tackling phenomena previously unexplored in the organizational literature, a theorist may find it helpful to initially act as if the phenomena occur at only one level of theory and analysis. In this way, a theorist temporarily restricts his or her focus, putting off consideration of multilevel processes for a period. Huselid's work (1995) on strategic human resource management provides an example. Huselid has documented organization-level relationships among human resource practices, aggregate employee outcomes, and firm financial performance, but what are the cross-level and emergent processes—the linkages of individual responses to human resource practices—that mediate the relationship between organizational human resource practices and organizational performance? The time is now ripe for such multilevel theory building (Ostroff & Bowen, Chapter Five, this volume).

Having acknowledged that there may be instances in which multilevel models may be unnecessary, we also offer the following caveat: given the nature of organizations as hierarchically nested systems, it will be difficult in practice to find single-level relations that are unaffected by other levels. The set of individual-level phenomena that are invariant across contexts is likely to be very small. Similarly, the set of group- or organization-level phenomena that are completely uninfluenced by lower levels is also likely to be

small. Failure to account for such effects when they exist will yield incomplete or misspecified models.

PRINCIPLE: *Multilevel theoretical models are relevant to the vast majority of organizational phenomena. Multilevel models may, however, be unnecessary if the central phenomena of interest (a) are uninfluenced by higher-level organizational units, (b) do not reflect the actions or cognitions of lower-level organizational units, and/or (c) have been little explored in the organizational literature. Caveat: Proceed with caution!*

How

By definition, multilevel models are designed to bridge micro and macro perspectives, specifying relationships between phenomena at higher and at lower levels of analysis (for example, individuals and groups, groups and organizations, and so on). Accordingly, a multilevel theoretical model must specify how phenomena at different levels are linked. Links between phenomena at different levels may be top-down or bottom-up. Many theories will include both top-down and bottom-up processes.

Top-down processes: contextual influences. Each level of an organizational system is embedded or included in a higher-level context. Thus individuals are embedded within groups, groups within organizations, organizations within industries, industrial sectors within environmental niches, and so on. Top-down processes describe the influence of higher-level contextual factors on lower levels of the system. Fundamentally, higher-level units may influence lower-level units in two ways: (1) higher-level units may have a direct effect on lower-level units, and/or (2) higher-level units may shape or moderate relationships and processes in lower-level units.

An organization has a direct effect on the behavior of its individual employees when, for example, its culture determines the accepted patterns of employee interaction and work behavior (for example, how formally employees address each other, or the extent to which employees question their supervisors' directives). An organization has a moderating effect on lower-level relationships when the relationship between two lower-level constructs changes as a function of organizational context. Thus, for example, the relationship between employees' conscientiousness and performance

may vary across organizational contexts. In contexts that provide autonomy and resources, conscientiousness may be associated with performance. However, contexts low on autonomy and resources are likely to constrain the effects of conscientiousness on performance, hence the relationship will be weak.

PRINCIPLE: *Virtually all organizational phenomena are embedded in a higher-level context, which often has either direct or moderating effects on lower-level processes and outcomes. Relevant contextual features and effects from the higher level should be incorporated into theoretical models.*

Bottom-up processes: emergence. Many phenomena in organizations have their theoretical foundation in the cognition, affect, behavior, and characteristics of individuals, which—through social interaction, exchange, and amplification—have emergent properties that manifest at higher levels. In other words, many collective constructs represent the aggregate influence of individuals. For example, the construct of organizational culture—a particularly broad and inclusive construct—summarizes the collective characteristics, behaviors, and values of an organization's members. Organizational cultures differ insofar as the characteristics, behaviors, and values of organizational members differ.[3]

Bottom-up processes describe the manner in which lower-level properties emerge to form collective phenomena. The emergence of phenomena across increasingly higher levels of systems has been a central theme of GST. Formative efforts to apply GST focus on the structure of emergence—that is, on the higher level, collective structure that results from the dynamic interactions among lower-level elements. The broad system typologies of Boulding (1956) and Miller (1978) attempt to capture the increasingly complex collectivities that are based on lower-level building blocks of the system. Thus, for example, interactions among atoms create molecular structure, or interactions among team members yield team effectiveness. This perspective views an emergent phenomenon as unique and holistic; it cannot be reduced to its lower-level elements (e.g., Dansereau et al., 1984).

A more contemporary perspective, one that has its roots in GST, derives from theories of chaos, self-organization, and complexity, and it views emergence as both process and structure. This

perspective attempts to understand how the dynamics and interactions of lower-level elements unfold over time to yield structure or collective phenomena at higher levels (Arthur, 1994; Gell-Mann, 1994; Kauffman, 1994; Nicolis & Prigogine, 1989; Prigogine & Stengers, 1984). This perspective is not a reversion to reductionism; rather, it is an effort to comprehend the full complexity of a system—its elements, their dynamics over time, and the means by which elements in dynamic interaction create collective phenomena (e.g., Cowan, Pines, & Meltzer, 1994). The two perspectives are compatible but different. We draw on this latter perspective and attempt to understand both process and structure in our conceptualization of emergence.

Emergence can be characterized by two qualitatively distinct types—composition and compilation—that may be juxtaposed as anchors for a range of emergence alternatives. To simplify the discussion that follows and make distinctions more apparent, we treat composition and compilation as ideal or pure types. Later in the chapter, we further elaborate their underlying theoretical differences, discuss interaction processes and dynamics that shape emergence, and explore forms of emergence that are more akin to composition or more akin to compilation. *Composition,* based on assumptions of isomorphism, describes phenomena that are essentially the same as they emerge upward across levels. Composition processes describe the coalescence of identical lower-level properties—that is, the convergence of similar lower-level characteristics to yield a higher-level property that is essentially the same as its constituent elements. *Compilation,* based on assumptions of discontinuity, describes phenomena that comprise a common domain but are distinctively different as they emerge across levels. The concepts are functionally equivalent—that is, they occupy essentially the same role in models at different levels, but they are not identical, as in composition. Compilation processes describe the combination of related but different lower-level properties—that is, the configuration of different lower-level characteristics to yield a higher-level property that is functionally equivalent to its constituent elements.

The distinction between composition and compilation forms of emergence is best illustrated with examples. Consider the composition model for psychological and organizational climate

(James, 1982; Kozlowski & Hattrup, 1992). It indicates that both constructs reference the same content, have the same meaning, and share the same nomological network (Jones & James, 1979; Kozlowski & Hults, 1987). For example, an organization's climate for service is a reflection of organizational members' shared perceptions of the extent to which organizational policies, procedures, and practices reward and encourage customer service (Schneider & Bowen, 1985). An organization's climate for service—whether positive or negative—emerges from the shared, homogeneous perceptions of organizational members. Thus individual and organizational climates are essentially the same construct, although there are some qualitative differences at higher levels. Organizational climate is more inclusive and may have some unique antecedents relative to its lower-level origin in psychological climate (Rousseau, 1988). Composition models based on isomorphic assumptions have been the primary means of conceptualizing emergent phenomena (Brown & Kozlowski, 1997; House et al., 1995). We describe collective phenomena that emerge through composition processes as shared properties, and we discuss them in more detail in a subsequent section.

Sometimes lower-level characteristics, behaviors, and perceptions may not coalesce. Instead, lower-level characteristics, behaviors, and/or perceptions may vary within a group or organization, and yet the configuration or pattern of lower-level characteristics, behaviors, and/or perceptions may nevertheless emerge, bottom-up, to characterize the unit as a whole. Consider, for example, individual and team performance. The compilation model for individual and team performance references performance as a functionally equivalent domain but specifies different antecedents and processes at different levels (Kozlowski, Gully, Nason, & Smith, 1999). Individual performance entails task-specific knowledge, skills, and abilities. Dyadic performance entails coordinated role exchanges. Team performance is a complex function of specific individual and dyadic—networked—contributions. Thus, in compilation models, the higher-level phenomenon is a complex combination of diverse lower-level contributions (Kozlowski, 1998, 1999). The form of emergence described by compilation is not widely recognized and yet is inherent in many common phenomena, including the domains of learning, performance, norms, power, conflict, and effectiveness,

among many others. Compilation-based emergent processes are relatively little explored from a multilevel perspective in the organizational literature. We describe collective phenomena that emerge through compilation processes as *configural properties* and discuss them in more detail in a subsequent section.

The type of emergent process is fundamentally affected by the nature of social-psychological interactions and can vary for a given phenomenon; that is, a particular emergent phenomenon may be compositional in some circumstances and compilational in others. Consider team performance once again. Team performance emerges from the behaviors of individual team members. But does team performance emerge as a result of the coalescence of the essentially identical behaviors of individual team members so that team performance simply reflects the sum or average performance of individual team members? Or is team performance the result of the array or pattern of individual team members' performance—the complex culmination of one team member's excellence on one task, another team member's excellence on a second task, and a third team member's fortunately inconsequential performance on yet a third task? The first conceptualization is an example of composition; the second is an example of compilation. Neither conceptualization is "right" in all circumstances. Rather, the determining factors are the dimension of interest for team performance, the nature of the team's work-flow interdependence, and the organizational context in which the team exists, among others. This example hints at the challenges inherent in explicating the precise bottom-up processes that yield many higher-level constructs. Despite the challenges, however, precise explication of these emergent processes lays the groundwork for operationalizing the construct—a point on which we elaborate later in this chapter.

PRINCIPLE: *Many higher-level phenomena emerge from characteristics, cognition, behavior, affect, and interactions among individuals. Conceptualization of emergent phenomena at higher levels should specify, theoretically, the nature and form of these bottom-up emergent processes.*

Where

Virtually inseparable from the question of *how* is the question of *where*—that is, precisely where do top-down and bottom-up processes originate and culminate? The answers to these questions

specify the focal entities—the specific organizational levels, units, or elements—relevant to theory construction. Suppose, for example, that a theorist is interested in the influence of unit climate on individual actions. What is the level of interest? For example, is it group climate? division climate? organizational climate? the climate of the informal friendship network? In the passages that follow, we will first explore the nature of organizational units as evoked by multilevel theory and then describe processes that determine the strength of the ties that link organizational levels or units.

Nature of organizational units. All but the smallest organizations are characterized by differentiation (horizontal divisions) and integration (vertical levels). These factors yield myriad entities, units, or levels. In organizational research, levels of *theoretical interest* focus on humans and social collectivities. Thus individuals, dyads, groups, subunits, and organizations are relevant levels (units, or entities) of conceptual interest. The structure is hierarchically nested so that higher-level units encompass those at lower levels. Many writers (Brown & Kozlowski, 1997; Freeman, 1980; Glick, 1985; Hannan, 1991; Simon, 1973) assert the importance of using formally designated units and levels for specification; for example, leadership research typically defines the "leader" as the formal unit manager. Generally speaking, formal units can be defined with little difficulty, although there can be exceptions, where unit boundaries or memberships are fuzzy.

Yet organizations are social systems in which people define their own informal social entities (Katz & Kahn, 1966). A variety of phenomena may define units or entities that do not correspond with formal unit boundaries. For example, vertical dyad linkage (VDL) theory (Graen, 1976) posits the formation of in-and outgroups as distinctive entities within a formal unit. Rentch (1990) demonstrates that patterns of social interaction across formal units influenced consensus on organizational climate, indicating that informal entities affect sensemaking processes. Often unit specification is based on expedience rather than on careful consideration. This can be problematic when the phenomena of interest are examined within formal units but are driven by informal processes that yield nonuniform patterns of dispersion (Brown & Kozlowski, 1997). Therefore, levels and units should be consistent with the

nature of the phenomenon of interest (Campbell, 1958; Freeman, 1980).

PRINCIPLE: *Unit specification (formal versus informal) should be driven by the theory of the phenomena in question. Specification of informal entities that cut across formal boundaries, or that occur within formal units and lead to differentiation, requires careful consideration.*

Determinants of the strength of ties linking organizational levels or units. One overgeneralization of the systems metaphor is that everything is related to everything. In reality, some levels and units are much more likely than others to be strongly linked, through what Simon (1973) refers to as *bond strength.* The theorist needs to chose appropriate units and levels or risk a misspecified or ineffective theory. Bond strength and related concepts help to explain what is likely to be connected across levels, and why.

Simon (1969, 1973) views social organizations as nearly decomposable systems. In other words, limited aspects of the larger system can be meaningfully addressed without compromising the system's integrity. A social organization can be conceptualized as a set of subsystems composed of more elemental components that are arrayed in a hierarchical structure. The linkage among levels—individual, group, and organizational—and subsystems is determined by their bond strength, which refers to the extent to which characteristics, behaviors, dynamics, and processes of one level or unit influence the characteristics, behaviors, dynamics, and processes of another level or unit (Simon, 1973). The greater the implications of one unit's actions for another unit, the greater the strength of the bond linking the two units. Therefore, meaningful linkages increase in strength with proximity and inclusion, and they decrease in strength with distance and independence.

Other researchers have used similar concepts to express the same basic principle. Weick (1976) uses the concept of *coupling* to reference decomposable subsystems. House and colleagues (1995) describe *inclusion* as the proportion of a lower-level unit's activities that are devoted to a higher level; units that are highly included will be more closely linked to the higher level. Kozlowski and Salas (1997) use the term *embeddedness* to describe how lower-level phenomena are aligned with contextual factors and processes that

originate at higher levels in the organizational system; alignment reflects strong bonds or inclusion across levels. Technostructural factors such as organizational goals, technology, and structure, as well as enabling processes such as leadership, socialization, and culture, influence embeddedness. From an interactionist perspective, Indik (1968) and James and Jones (1976) assert that strong interactions between levels require propinquity of structure and process and alignment of content. Constructs and processes implicated in bond strength, coupling, inclusion, and embeddedness will be more strongly linked across levels for relevant units.

This has obvious implications for models that incorporate multiple levels or units. Proximal, included, embedded, and directly coupled levels and units exhibit more meaningful relations than distal levels or loosely coupled units. Moreover, the content underlying constructs at different levels has to have some meaningful connection. For example, work-unit technology and structure exhibit cross-level effects on individuals because they constrain the characteristics of jobs (Kozlowski & Farr, 1988; Rousseau, 1978a, 1978b). The levels are coupled and the content is meaningfully related in a common network of relations. In contrast, the potential effects of organization-level strategy on individual jobs is likely to be quite small. This does not mean that strategy has no effect; rather, its effects are mediated through so many intervening levels, units, and content domains that direct effects are likely to be very difficult to detect at the individual level because bond strength is weak and the focal content is not meaningfully related. The effects of strategy are likely to be indirect.

PRINCIPLE: *Linkages across levels are more likely to be exhibited for proximal, included, embedded, and/or directly coupled levels and entities.*

PRINCIPLE: *Linkages are more likely to be exhibited for constructs that tap content domains underlying meaningful interactions across levels.*

When

Time is rarely a consideration in either single-level or multilevel organizational models (House et al., 1995), yet it is clearly the case that many if not most organizational phenomena are influenced and shaped by time. Here we explore three ways in which time may

be incorporated into a multilevel model, increasing the rigor, creativity, and effectiveness of multilevel theory building.

Time as a boundary condition or moderator. Many organizational phenomena have a unidirectional effect on higher- or lower-level organizational phenomena, but multilevel relationships are not always so simple; instead, over time the relationship between phenomena at different levels may prove bidirectional or reciprocal. A given phenomenon may appear to originate at a higher or lower level according to the theorist's assumption about the current time point in a stream or cycle of events. The failure, quite common, to make such assumptions explicit can lead to apparently contradictory models of the same phenomenon and to debates about its "true" level.

For example, organizational culture is more likely to be based on emergent processes, either when the organization is at an early point in its life cycle or when the organization is undergoing dramatic change. In effect, individual sensemaking and social construction are more active and have a greater impact when the organizational context is ambiguous or in a state of flux. Therefore development or change in organizational culture will appear to be a bottom-up process. Over time, however, culture becomes stable and institutionalized. Formative events that were salient during emergence become the stuff of myth, legend, and tradition. Founding members move on. New members are socialized and assimilated into enduring contexts that resist change. Therefore, organizational culture appears to have a top-down influence on lower-level units.

The distinction between the two perspectives just sketched does not have to do with which one represents the "true" model of organizational culture; both are veridical. A variety of factors and processes can influence the apparent direction, top-down or bottom-up, of a cross-level process. This illustrates the necessity for the theorist to explicitly specify the temporal assumptions for the phenomenon in question. Thus time may serve as a boundary condition for the model; for example, the theorist states that the model applies only to mature organizations, or only to new ones. Alternatively, in a theoretical model, time may serve as a moderator of the phenomenon; for example, the theorist posits that the direction (top-down or bottom-up) and effects of the phenomenon vary as a function of the organization's maturity.

PRINCIPLE: *The temporal scope, as well as the point in the life cycle of a social entity, affect the apparent origin and direction of many phenomena in such a way that they may appear variously top-down, bottom-up, or both. Theory must explicitly specify its temporal reference points.*

Time-scale variations across levels. Differences in time scales affect the nature of links among levels (Simon, 1973). Lower-level phenomena tend to have more rapid dynamics than higher-level and emergent phenomena, which makes it is easier to detect change in lower-level entities. This is one reason why top-down models predominate in the literature. For example, efforts to improve organizational outcomes (for example, quality) through training (for example, total-quality management, or TQM) assume emergent effects that originate at the individual level. Models of training effectiveness focus on the transfer of trained skills to the performance setting. Higher-level contextual support (for example, a transfer climate; see Rouiller & Goldstein, 1993) enhances transfer in such a way that the effects of TQM training on quality are relatively immediate. However, the effect of individual-level TQM training on organizational outcomes is emergent and requires a much longer time scale. Individual cognition, attitudes, and behaviors must combine through social and work interactions. Depending on the nature of the vertical transfer process, individual outcomes will compose or compile to the group level and, over longer time frames, will yield organizational outcomes (Kozlowski & Salas, 1997; Kozlowski, Brown, Weissbein, & Cannon-Bowers, Chapter Four, this volume). Thus contextual or top-down linkages can be manifest within short time frames, whereas emergent, bottom-up linkages necessitate longer time frames.

PRINCIPLE: *Time-scale differences allow top-down effects on lower levels to manifest quickly. Bottom-up emergent effects manifest over longer periods. Research designs must be sensitive to the temporal requirements of theory.*

One implication of this effect of time scale is that phenomena at different levels may manifest at different points in time. For example, Kozlowski and his colleagues have proposed that team performance compiles and emerges across levels, from individuals to dyads to teams, at different points in the team-development process (Kozlowski et al., 1994, 1999). Others, in related fashion, have noted that level of a relationship in a multilevel model—homogeneous groups, heterogeneous groups, or independent individuals—can be influenced by factors that, over time, change the level of the relationship (Dansereau, Yammarino, & Kohles, 1999).

Entrainment: changing linkages over time. The term *entrainment* refers to the rhythm, pacing, and synchronicity of processes that link different levels (Ancona & Chong, 1997; House et al., 1995). Coupling across levels or units is tightened during periods of greater entrainment. Entrainment is affected by task cycles and work flows, budget cycles, and other temporally structured events that pace organizational life (Ancona & Chong, 1997). For example, the concept of entrainment has been used in the group and team performance literature to capture the idea that work-flow interdependence is not necessarily uniform over time; rather, the degree of interdependence or coupling can vary significantly depending on the timing of events or acts that require a synchronous and coordinated response (e.g., Fleishman & Zaccaro, 1992; Kozlowski, Gully, McHugh, Salas, & Cannon-Bowers, 1996; Kozlowski et al., 1999; McGrath, 1990). Thus levels or units that ordinarily are loosely coupled will be tightly coupled during periods of synchronicity.

Accordingly, entrainment processes must be considered during theory construction. Further, entrainment has rather obvious implications for research designs that intend to capture entrained processes. At some points in the cycle, two entities or levels may be tightly coupled or entrained, whereas at other points they will be decoupled and will appear independent. This variability creates demands for precise theory and measurement in order to capture the coupling; data collection must be sensitive to entrainment cycles and periods.

PRINCIPLE: *Entrainment can tightly couple phenomena that ordinarily are only loosely coupled across levels. Theories that address entrained phenomena must specify appropriate time cycles and must employ those cycles to structure research designs.*

Why and Why Not?

Argument by assertion is invariably a poor strategy for theory building. Argument by logical analysis and persuasion—argument that explains why—is always preferable. In multilevel theory building, explaining why is not merely preferable but essential. A great deal of organizational multilevel theory building spans organizational subdisciplines (industrial/organizational psychology and organizational theory, for example). Therefore, the unstated assumptions in a multilevel theory may be obvious to the members of one subdiscipline but not to the members of another, who are also interested in the new multilevel theory. Furthermore, multilevel theories often incorporate novel constructs (for example, team mental models, or organizational learning). The meaning of such constructs may well be obscured in the absence of thorough explanations concerning why. Finally, multilevel data analysis has been the subject of considerable and continuous debate. Conflicts regarding the best way to analyze multilevel models abate considerably, however, in the presence of carefully and fully explicated theoretical models (Klein et al., 1994) that make the choice of analytical strategy clear (Klein, Bliese et al., Chapter Twelve, this volume). Thus multilevel theorists must not only specify what, how, where, and when but also why: Why are relationships in the model conceptualized as top-down rather than bottom-up? Why are constructs conceptualized as compositional rather than compilational? Why are predictors assumed to have immediate rather than long-term consequences for the outcomes of interest?

Nearly as important as the question of why, and perhaps even more interesting, is the question of why not. Why might bottom-up processes *not* yield a group-level property? That is, why might members' perceptions *not* converge to form a shared unit norm or climate? Why might top-down processes *not* constrain relationships in an organizational subunit? Why might predictors, hypothesized to be influential over time, prove instead to have immediate

consequences? In exploring why not, theorists may refine their models, incorporating important insights and nuances. This adds diversity and depth to theory; it is how a science is built.

PRINCIPLE: *Multilevel theoretical models must provide a detailed explanation of the assumptions undergirding the model. Such explanations should answer not only the question of* why *but also the question of* why not.

In sum, rigorous multilevel theories must carefully consider what, how, where, when, why, and why not. In what follows, we explicate how these basic questions inform the definition and measurement of constructs in multilevel models. We then describe distinctive forms or frameworks that multilevel models may take, the kinds of research designs and samples necessary to test multilevel models, and possible data analytic strategies.

Principles for Model Specification: Aligning Constructs, Measures, Models, Design, and Analyses

Many of the controversies and problems associated with multilevel research are based on misspecifications or misalignments among the theoretical level of constructs, their measurement, and their representation for analysis. Misalignment is a problem for any research design that incorporates mixed levels, but it is also a problem for single-level research that incorporates emergent constructs. The nature of these misalignments is well documented elsewhere (Burstein, 1980; Firebaugh, 1979; Freeman, 1980; Hannan, 1991; Robinson, 1950; Rousseau, 1985; Thorndike, 1939). The following are some common problems: blind aggregation of individual-level measures to represent unit-level constructs, use of unit-level measures to infer lower-level relations (the well-known problems of aggregation bias and ecological fallacies), and use of informants who lack unique knowledge or experience to assess unit-level constructs.

Misalignments degrade construct validity and create concerns about generalizability. To build theoretical models that are clear and persuasive, scholars must explicate the nature of their constructs with real care. Precise explication lays the foundation for sound measurement. Constructs that are conceptualized and measured at

different levels may be combined in a variety of distinctive multilevel models. Research design and analytical strategies need to be aligned with the levels inherent in these models. Principles relevant to these concerns are considered in the remainder of this section.

Constructs in Multilevel Theory

Construct level and origin. Constructs are the building blocks of organizational theory. A construct is an abstraction used to explain an apparent phenomenon. The level of a construct is the level at which it is hypothesized to be manifest in a given theoretical model—the known or predicted level of the phenomenon in question. Although organizational theorists have often discussed "the level of theory," we prefer to use the phrase *level of the construct* because mixed-level models, by definition, include constructs that span multiple levels; that is, generalizations are constrained by the level of the endogenous construct ("the level of the theory"), but other constructs in a model may be at higher or lower levels. Thus, in mixed-level research, the theoretical explanation will span several levels in the effort to understand an endogenous construct at a given focal level.

The first and foremost task in crafting a multilevel theory or study is to define, justify, and explain the level of each focal construct that constitutes the theoretical system. Remarkably, the level of many organizational constructs is unclear. This problem, we have noted, once plagued the climate literature. Researchers and critics asked whether climate was to be conceptualized and measured as an organizational (unit) construct or as a psychological (individual) one. Climate researchers resolved this question, differentiating explicitly between a consensual unit climate and its origins in psychological climate. However, the question of level is often unasked in other research. Consider the familiar construct of worker participation. What is its level? Is worker participation an individual-level phenomenon, describing the influence an individual exerts in unit decisions? Or is worker participation at the unit level, describing a set of formal structures and work practices (for example, quality circles) characteristic of units, not individuals? For the most part, the participation literature reveals neither clear consensus regarding the level of the construct nor explicit discussion of its level (Klein et al., 1994).

PRINCIPLE: *The theorist should explicitly specify the level of each construct in a theoretical system.*

In specifying the level of a construct, the theorist must build a targeted theory, or "minitheory," of the phenomenon, explicating where, when, and how the construct forms and is manifest. Many phenomena we study in organizations have their theoretical origins in the cognition, affect, and behavior of individuals but emerge, through compositional or compilational processes, to manifest as higher-level phenomena. A given construct may be an individual-level construct in some circumstances and a unit-level construct in others. When a theorist specifies that a construct originates at the individual level and manifests at a higher level, the theorist must explicate when, how, and why this process occurs. The theoretical foundation for emergent effects must be at the level of origin. When psychological and social-psychological phenomena are emergent at higher levels, the researcher needs to distinguish the level of theoretical origin and the level at which the focal construct is manifest—the level of the construct. The researcher must also explain the theoretical process that yields higher-level emergence—the conditions in which the higher-level construct exists or does not exist. This is essential to determining an appropriate means of assessing and representing the emergent higher-level construct.

PRINCIPLE: *When higher-level constructs are based on emergent processes, the level of origin, the level of the construct, and the nature of the emergent process must be explicitly specified by the theory.*

We elaborate further in what follows, explaining links between the previously described principles of multilevel theory (what, where, when, how, why, and why not) and the definition, explication, and measurement of theoretical constructs. Our quarrel with much of the existing theoretical literature on organizations is not that authors are too complex in characterizing the multiple, even shifting, levels of their constructs but just the opposite: that, too often, authors' conceptualizations of the theoretical processes and levels of their constructs lack important detail, depth, and complexity. We now consider different types of higher-level constructs and address the implications for measurement.

Types of unit-level constructs. Unit-level constructs describe entities composed of two or more individuals: dyads, groups, functions, divisions, organizations, and so on. In the organizational literature, many problems and controversies revolve around the definition, conceptualization, justification, and measurement of unit-level constructs. The "level" of many higher-level constructs (culture, leadership, or participation, for example) is often debated. The debate is due in part to the potential for these constructs to emerge from lower-level phenomena.

To help resolve the controversies and confusion that often surround the definition, meaning, and operationalization of unit-level constructs, we distinguish three basic types:

1. Global unit properties
2. Shared unit properties
3. Configural unit properties

Global unit properties differ from shared and configural unit properties in their level of origin. Global unit properties originate and are manifest at the unit level. Global unit properties are single-level phenomena. In contrast, shared and configural unit properties originate at lower levels but are manifest as higher-level phenomena. Shared and configural unit properties emerge from the characteristics, behaviors, or cognitions of unit members—and their interactions—to characterize the unit as a whole. Shared and configural unit properties represent phenomena that span two or more levels. Shared unit properties are essentially similar across levels (that is, isomorphic), representing composition forms of emergence. In contrast, configural unit properties are functionally equivalent but different (that is, discontinuous), representing compilation forms of emergence. Configural unit properties capture the variability or pattern of individual characteristics, constructs, or responses across the members of a unit. We elaborate in what follows, and then we discuss how the nature of a unit construct influences its measurement.[4]

Global unit properties. Global constructs pertain to the relatively objective, descriptive, easily observable characteristics of a unit that originate at the unit level. Global unit properties do not originate

in individuals' perceptions, experiences, attitudes, demographics, behaviors, or interactions but are a property of the unit as a whole. They are often dictated by the unit's structure or function. Group size and unit function (marketing, purchasing, human resources) are examples of global properties. There is no possibility of within-unit variation because lower-level properties are irrelevant; indeed, any within-unit variation is most likely the result of a procedure that uses lower-level units to measure the global property. If, for example, group members disagree about the size of their group, someone has simply miscounted. Unit size has an objective standing apart from members' characteristics or social-psychological processes. In contrast, "perceived group membership" is an entirely different type of construct.

Shared unit properties. Constructs of this type describe the characteristics that are common to—that is, shared by—the members of a unit. Organizational climate, collective efficacy, and group norms are examples of shared unit-level properties. Shared unit properties are presumed or hypothesized to originate in individual unit members' experiences, attitudes, perceptions, values, cognitions, or behaviors and to converge among group members as a function of attraction, selection, attrition, socialization, social interaction, leadership, and other psychological processes. In this way, shared unit properties emerge as a consensual, collective aspect of the unit as a whole. Shared unit properties are based on *composition* models of emergence, in which the central assumption is one of isomorphism between manifestations of constructs at different levels; the constructs share the same content, meaning, and construct validity across levels. When researchers describe and study shared unit properties, they need to explain in considerable detail the theoretical processes predicted to yield restricted within-unit variance with respect to the constructs of interest: How does within-unit consensus (agreement) or consistency (reliability) emerge from the individual-level characteristics (experiences, perceptions, attitudes, and so on) and interaction processes among unit members?

Configural unit properties. Constructs of this type capture the array, pattern, or configuration of individuals' characteristics within a unit. Configural unit properties, like the shared properties of a

unit, originate at the individual level. Unlike shared unit properties, however, configural unit properties are not assumed to coalesce and converge among the members of a unit. The individual contributions to configural unit properties are distinctly different. Therefore, configural unit properties have to capture the array of these differential contributions to the whole. Configural unit properties characterize patterns, distribution, and/or variability among members' contributions to the unit-level phenomenon. Configural unit properties do not rest on assumptions of isomorphism and coalescing processes of composition but rather on assumptions of discontinuity and complex nonlinear processes of *compilation*. The resulting constructs are qualitatively different yet functionally equivalent across levels.

Configural unit properties are relatively rare in the organizational literature, but they are not rare in organizations. We can distinguish two types of configural unit properties: *descriptive characteristics,* which reference manifest and observable features, and *latent constructs,* which reference hypothetical and unobserved properties of the unit in question. Descriptive characteristics are straightforward. For example, diversity—the extent to which unit members' demographic characteristics are dissimilar—is a configural descriptive unit property. However, whereas diversity is a manifest unit characteristic, it most likely has effects through latent constructs that tap underlying psychological differences (e.g., Millikin & Martins, 1996). For example, diversity in unit-level sex or age are descriptive characteristics that may be linked to unit-level variability for the constructs of attitudes and values.

Unit-level conceptualizations of constructs are often configural.[5] For example, the combination of team members' abilities or personality characteristics constitutes the configural properties of the unit (Moreland & Levine, 1992). Configural constructs may also capture the pattern of individual perceptions or behavior within a unit. For example, team performance is often regarded as a global property of the team, yet when individual team members perform different but interdependent tasks, team performance may be conceptualized as a configural construct; team members do not engage in identical behaviors (Kozlowski et al., 1999). Finally, network characteristics (for example, network density) are configural insofar as they depict the pattern of the relationships within a unit (or

network) as a whole (Brass, 1995). Configural unit properties are based on compilation models of emergence (e.g., Kozlowski et al., 1999). When studying configural unit properties, researchers need to explain in detail the theoretical processes by which different individual contributions combine to yield the emergent unit property—that is, how are the individual origins represented in the summary, pattern, configuration, or array of the unit-level property?

PRINCIPLE: *Theorists whose models contain unit-level constructs should indicate explicitly whether their constructs are global unit properties, shared unit properties, or configural unit properties. The type of unit-level construct should drive its form of measurement and representation for analyses.*

Levels of Measurement

Basic issues. The level of measurement is the level at which data are collected to assess a given construct. Individual-level constructs should, of course, be assessed with individual-level data. Unit-level constructs, in contrast, may be assessed with either unit-level or individual-level data. When unit-level constructs are assessed with unit-level measures, an expert source (a subject matter expert, for example, or an objective archive) provides a single rating of each unit. When unit-level constructs are assessed with individual-level measures, unit members provide individual-level data (for example, individual ratings of climate, or individuals' reports of their own demographic characteristics), which are subsequently combined in some way to depict the unit as a whole. Rousseau (1985, p. 31) advises researchers to measure unit-level constructs with global (that is, unit-level) data whenever possible: "Use of global data is to be preferred because they are more clearly linked to the level of measurement, avoiding the ambiguity inherent in aggregated data." Klein and colleagues (1994, p. 210) note that when a researcher uses "a global measure to characterize a group, he or she lacks the data needed to test whether members are, indeed, homogeneous within groups on the variables of interest." Accordingly, Klein and colleagues (1994, p. 210) recommend that researchers use global measures to capture unit-level constructs only when the level of the construct is "certain" or "beyond question." Here, we elaborate on Rousseau's (1985) and Klein and colleagues' (1995) admonitions, advising that the level of measurement should be determined by the type of the unit-level construct.

Individual-level constructs. Individual-level constructs should, as already noted, be assessed at the individual level. For example, individuals may complete measures of their own job satisfaction, turnover intentions, self-efficacy, psychological climate, and so forth. In some cases, one or more experts may provide assessments of the characteristics of other individuals. This procedure can be used when the characteristic is observable, or when the informant has unique access to relevant information (Campbell, 1955; Seidler, 1974). A supervisor may describe his or her individual subordinates' performance behavior, an observer may record individual demographic characteristics, or a researcher may use archival records to assess individuals' ages, tenure, or experience. In each case, data are assigned to individuals and are considered individual-level data. Issues of measurement quality are, of course, still relevant.

Global properties. The measurement of unit-level variables is often more complex and more controversial. Least complex and least controversial is the measurement of the global properties of a unit. By definition, global properties are observable, descriptive characteristics of a unit. Global properties do not emerge from individual-level experiences, attitudes, values, or characteristics. Accordingly, there is no need to ask all the individuals within a unit to describe its global properties. A single expert individual may serve as an informant when the characteristic is observable, or when the informant has unique access to relevant information. Thus a vice president for sales may report his or her company's sales volume, a CEO may report a firm's strategy, or a manager may report a unit's function. Although these examples each use an individual respondent, the data are considered global unit-level properties.

Shared properties. In contrast, shared properties of a unit emerge from individual members' shared perceptions, affect, and responses. The theoretical origin of shared properties is the psychological level, and so data to assess these constructs should match the level of origin. This provides an opportunity to evaluate the composition model of emergence underlying the shared property;

that is, the predicted shared property may not in fact be shared, in which case the data cannot be averaged to provide a meaningful representation of the higher-level construct. Therefore, the data to measure shared unit properties should be assessed at the individual level, and sharedness within the unit should be evaluated. Given evidence of restricted within-unit variance, the aggregate (mean) value of the measure should be assigned to the unit. Several empirical examples of this approach to the conceptualization, assessment, and composition of unit-level constructs can be found in the literature (e.g., Campion, Medsker, & Higgs, 1993; Hofmann & Stetzer, 1996; Kozlowski & Hults, 1987). This approach ensures both that the data are congruent with the construct's origin and that they conform to the construct's predicted form of emergence, thereby avoiding misalignment.

Configural properties. When a construct refers to a configural property of a unit, the data to assess the construct derive from the characteristics, cognitions, or behaviors of individual members. Individual-level data are summarized to describe the pattern or configuration of these individual contributions. As before, theory—the conceptual definition of the emergent construct—drives the operationalization of the measure. Configural properties emerge from individuals but do not coalesce as shared properties do. Thus a researcher, in operationalizing the configural properties of a unit, need not evaluate consensus, similarity, or agreement among individual members except to rule out coalescence. The summary value or values used to represent the configural property are based on the theoretical definition of the construct and on the nature of its emergence as a unit-level property. A variety of data-combination techniques may be used to represent, capture, or summarize configural properties, including the minimum or maximum, indices of variation, profile similarity, multidimensional scaling, neural nets, network analyses, systems dynamics and other nonlinear models, among others. The mean of individual members' characteristics is generally not an appropriate summary statistic to depict a configural unit property, although it may be combined with an indicator of variance or dispersion (Brown et al., 1996). In the absence of within-unit consensus, means are equifinal, ambiguous, and questionable representations of higher-level constructs.

PRINCIPLE: *There is no single best way to measure unit-level constructs. The type of a unit-level construct, in addition to its underlying theoretical model, determine how the construct should be assessed and operationalized. As a general rule, global properties should be assessed and represented at the unit level. Shared and configural properties should be assessed at the level of origin, with the form of emergence reflected in the model of data aggregation, combination, and representation.*

Establishing the construct validity of shared properties. The assumption of isomorphism that is central to the conceptualization of shared constructs requires explicit consideration. There are two primary issues relevant to testing models with one or more shared unit properties:

1. Establishing the measurement model
2. Evaluating the substantive theoretical model

The issue of the measurement model addresses the construct validity of aggregated lower-level measures as representations of higher-level constructs. It is generally addressed through examining patterns of within-group variance. Consensus- or agreement-based approaches—for example, $r_{wg(j)}$—evaluate within-group variance against a hypothetical expected-variance (EV) term. Agreement is examined for each shared property measure for each unit: a construct-by-group approach. Consistency- or reliability-based approaches—for example, ICC(1), ICC(2), and within-and-between analysis (WABA)—evaluate between-group variance relative to total (between and within) variance, essentially examining interrater reliability for each shared property across the sample: a construct-by-sample approach (Kozlowski & Hattrup, 1992; Bliese, Chapter Eight, this volume).

These different treatments have been the source of some debate (e.g., George & James, 1993; Yammarino & Markham, 1992). Consensus approaches treat issues 1 and 2 as distinct (e.g., James, Demaree, & Wolf, 1984; James et al., 1993; Kozlowski & Hults, 1987; Kozlowski & Hattrup, 1992). The strength is that construct misspecification, for any construct in any group, is avoided. The disadvantage is that there may be insufficient between-group variance for model evaluation, and this problem will not be revealed

until data analysis. Consistency-based approaches treat the issues as more unitary (e.g., Yammarino & Markham, 1992). The strength is that both within and between variance are considered in the computation of reliability, and so aggregated measures also have adequate between variance for the evaluation of substantive relations. The disadvantage is that some constructs may not actually have restricted variance in some groups, and so there is some potential for construct misspecification, which may be masked in the construct-by-sample approach.

We assert that consideration of both within-group and between-group variance is critical. However, the particular approach chosen is a matter of consistency with one's theory and data. Both approaches have different strengths and drawbacks. In the appropriate circumstances, either of the approaches is acceptable; there is no universally preferable approach.

PRINCIPLE: *The assumption of isomorphism of shared unit properties should be explicitly evaluated to establish the construct validity of the aggregated measure. The selection of a consensus- or consistency-based approach should be dictated by theory and data; no approach is universally preferable.*

Data source, construct, and measurement levels. Individuals as sources of data play different roles in measuring the three different types of unit constructs. This observation highlights the distinction between the data source, on the one hand, and the level of the construct and its measurement, on the other. For example, a knowledgeable individual may act as the data source for a global unit property such as size, function, or strategy, but in such a case the level of measurement is not considered the individual but rather the unit as a global entity.

A single informant may provide the data to measure the configural or distributional properties of a unit when the properties are directly and reliably observable, or when the informant has unique access to relevant information. For example, a supervisor may report the distribution of males and females in a unit. A manager may report unit members' tenure, thus providing the data necessary for the calculation of a unit's variability with respect to tenure. Individual-level performance data may be reported by a team leader to assess the configuration of team performance. In these examples, the configural construct is a unit-level construct even though the source is a single expert.

In contrast, a single individual may rarely if ever serve as the data source regarding a shared property of the construct. For example, it is generally not appropriate to use single informants (for example, a supervisor or a CEO) to assess unit or organizational climate; climate originates as individual interpretations and emerges via social interaction, and single informants are not uniquely situated to know the inner interpretations of multiple perceivers. Thus assessment should model the theory regarding the origin and nature of the construct.

PRINCIPLE: *Individuals may serve as expert informants for higher-level constructs when they can directly observe or have unique knowledge of the properties in question. As a general rule, expert informants are most appropriate for the measurement of global unit-level properties and observable (manifest) configural properties. They are least appropriate for the measurement of shared properties and unobservable (latent) configural properties.*

Item construction. Several authors have provided guidelines for item construction, primarily for the measurement of shared properties. In general, the advice is to focus respondents on description as opposed to evaluation of their feelings (James & Jones, 1974) and to construct items that reference the higher level, not the level of measurement (James, 1982; Klein et al., 1994; Rousseau, 1985). In practice, research has tended to use items framed at both the individual level (data source) and at higher levels. Recently, Chan (1998) distinguished these practices as representing different composition models of the constructs in question. For example, Chan views climate items referencing self-perceptions (for example, "I think my organization . . .") as constructs distinct from items that tap the same content but reference collective perceptions (for example, "We think the organization . . .")—what he refers to as "reference shift consensus."

Research that has tested the merits of this advice is, however, very limited. Klein, Conn, Smith, and Sorra (1998) have found that survey items referencing the unit as a whole (for example, "Employees'

work here is rewarding") do engender less within-group variability and more between-group variability than comparable survey items that reference individual experiences and perceptions (for example, "My work here is rewarding"). However, many climate researchers assessing shared unit properties have used self-referenced items and have demonstrated meaningful within-unit consensus (e.g., Kozlowski & Hults, 1987; Ostroff, 1993; Schneider & Bowen, 1985). It may well be the case that item content is critically important to the unit of reference. Perhaps climate-related content (for example, "I think the reward system . . .") that taps the broader work environment may be more robust to differences between self-reference versus collective reference. The perspective, whether the self or the larger unit, may be largely the same, whereas content that taps more variable properties (for example, "My job is . . .") may be more sensitive to the point of view incorporated in the item.

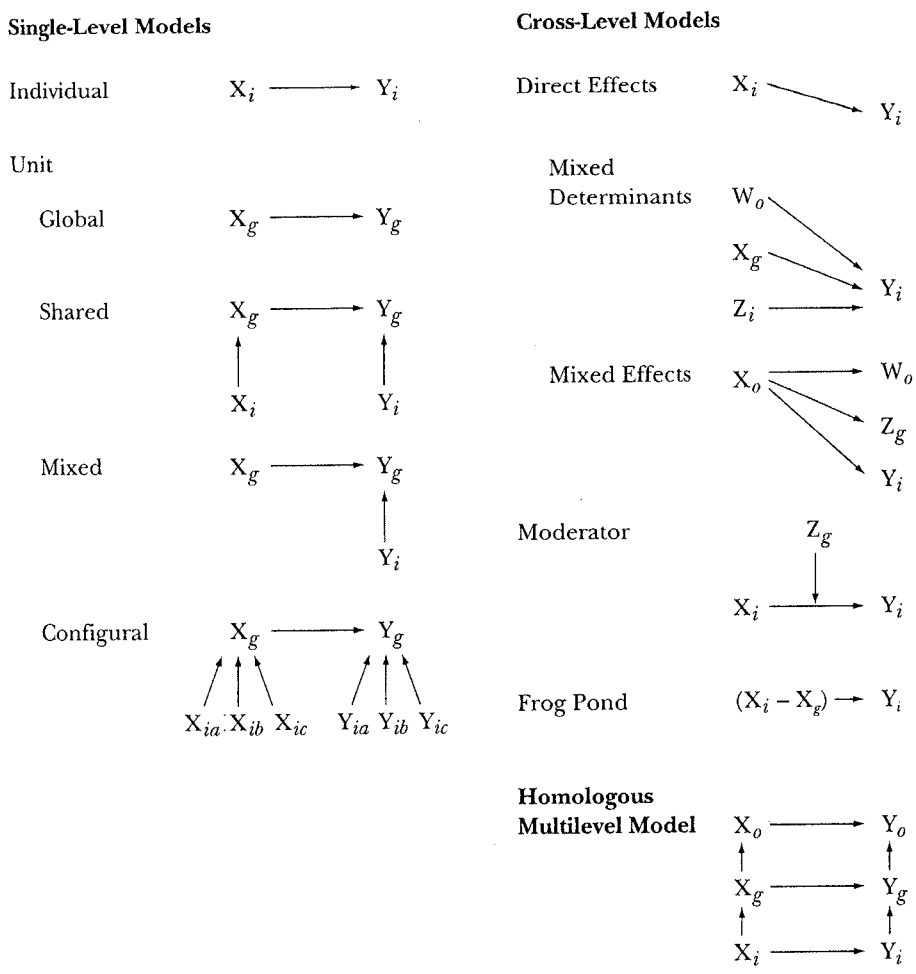
Clearly, more empirical work is needed to establish which item characteristics are critical to construct fidelity and which ones are not essential. In the meantime, we suggest that researchers employ measures consistent with the conceptualization of their constructs, using unit-level referents, if possible, to assess shared unit-level constructs. However, without more definitive empirical evidence, we do not encourage this as a litmus test and do not offer a principle. We do encourage more empirical research on guidelines for the construction of items to assess emergent constructs.

Types of Multilevel Models

Theoretical models describe relationships among constructs. A multilevel perspective invites—indeed, necessitates—special attention to the level of the constructs united within a theoretical model. In this section, we build on the preceding section by describing broad types of models distinguished by the levels of the constructs they encompass, as well as by the links they propose among constructs. Model specifications are illustrated in Figure 1.1. Following our description of basic models, we note further complexities in the creation of multilevel models.

Single-level models. Single-level models, as their name suggests, specify the relationship between constructs at a single level of theory

Figure 1.1. Model Specification.



and analysis. Such models are common in our literature and generally represent particular disciplinary perspectives. Psychologists are likely to find *individual-level* models the most familiar and straightforward type of single-level model. Individual-level models may be conceptually complex, specifying intricate interactional relationships among numerous constructs. However, individual-level models, by definition, ignore the organizational context of individual perceptions, attitudes, and behaviors. Thus the simplicity of individual-level models is in many cases a major limitation. Indeed, ignoring the context when it is relevant will lead to biases in the examination of construct relations (that is, the standard-error estimates of parameters will be biased).

Potentially far more complex are *unit-level* models, for these models may combine the three types of unit constructs in a variety of ways, in some cases necessitating mixed-level conceptualization, data collection, and analysis. Group-level models that depict the relationship of two global constructs are, from a levels perspective, the least complicated. To test these models, a researcher gathers unit-level data, consulting objective sources or experts to operationalize constructs. Tests of the effects of organizations' global human resource practices (for example, the presence or absence of merit pay and quality circles) on objective measures of organizational performance provide an example. But such models are very simple—perhaps too simple, like their individual-level counterparts. We suggest possible elaborations in what follows.

More complex, from a levels perspective, are unit-level models that include shared constructs. Consider a model linking two shared constructs: perhaps, for example, unit climate is hypothesized to predict unit morale. In proposing such a model, a scholar must explicate not only the processes linking the independent and dependent variables but also the processes engendering the emergence of climate perceptions and feelings of morale to the unit level: How do climate perceptions and feelings of morale, respectively, come to be shared by unit members? Further, to test such a model, a researcher must gather data from the level of origin—that is, from unit members—ascertaining the presence of restricted within-unit dispersion prior to aggregating data measuring the independent variable (climate) and the dependent variable (morale)

and conducting unit-level analyses. Thus a seemingly simple unit-level model may, if it includes shared constructs, effectively include a multilevel (compositional) model in the very definition and operationalization of each shared construct.

Unit-level models may also link global and shared constructs in direct and mediated relationships. A researcher may predict, for example, that global organizational human resources practices enhance global organizational performance by increasing the level of (shared) organizational citizenship behavior. In proposing such a model, a theorist moves beyond the simple unit-level model of global constructs (already outlined), offering a richer and more sophisticated analysis of the possible determinants of organizational performance. Ideally, such a theory explicates the influence of human resources practices on organizational citizenship behavior, the emergence of shared organizational citizenship behavior to the organizational level, and the influence of shared organizational citizenship behavior on global measures of organizational performance. Further, to test such a model, a researcher must, as before, collect individual-level data to tap the shared construct of interest.

Unit-level models incorporating configural constructs are also plausible. For example, the variation in cognitive ability within a unit may be predicted to influence global measures of unit performance. Or consider a more complex model: perhaps the personality configuration of a unit is predicted to influence unit creativity; that is, units with more diverse personality types may develop more creative ideas than units with less dissimilarity. Such a model requires not only the careful definition and operationalization of personality configuration but also the careful definition and operationalization of unit creativity. How does unit creativity emerge from the ideas and behaviors of unit members? Is it a shared construct—a unit average—or a configural construct, reflecting a more complex weighing, or configuration, of individual contributions? These questions hint at the rigor that a multilevel perspective may bring to the processes of theory building and theory testing. At first glance, the construct of unit creativity appears straightforward, unremarkable. But a further, multilevel examination indicates much work to be done in defining, explicating, and operationalizing the nature and emergence of unit-level creativity.

Cross-level models. Cross-level theoretical models describe the relationship between different independent and dependent constructs at different levels of analysis (Rousseau, 1985). Typically, organizational cross-level models describe the top-down impact of higher-level constructs on lower-level constructs (outcomes and processes). Although theory often conceptualizes the potential impacts of lower-level constructs on higher levels (the impact of newcomers on group cohesion, for example), bottom-up cross-level modeling is a distinct rarity in the empirical literature because of its analytic limitations. We should note, however, that recent work is beginning to address this problem (Griffin, 1997). Here, we outline three primary types of top-down cross-level models:

1. *Cross-level direct-effect models* predict the direct effect of a higher-level (for example, unit-level) construct on a lower-level (for example, individual-level) construct. Typically, such models predict that the higher-level construct in some way constrains the characteristics (for example, perceptions, values, or behaviors) of lower-level entities. Thus, for example, a cross-level direct-effect model may highlight the influence of unit technology on the nature of the individual job characteristics in each unit. Routine unit technologies are likely to yield jobs that are low in discretion, variety, and challenge. Conversely, uncertain technologies are likely to yield jobs high in discretion, variety, and challenge (e.g., Kozlowski & Farr, 1988; Rousseau, 1978a). Cross-level direct-effect models may, of course, highlight the effects of global, shared, or configural unit properties on lower-level constructs. For example, unit norms (a shared construct) may constrain individual behavior, or the density of a unit's social network (a configural construct) may influence individual satisfaction and turnover within the unit. Finally, cross-level direct-effect models may describe the influence not only of units on individuals but of other, higher-level entities (for example, industries) on lower-level entities (for example, organizations). Variants of cross-level direct-effect models include *mixed-determinant* and *mixed-effect* models (Klein et al., 1994). A mixed-determinant model specifies multilevel determinants (for example, both unit and individual) of a single-level (for example, individual-level) outcome or outcomes. A mixed-effect model specifies multiple-level outcomes of a single-level predictor. Thus, for example, an organization's adoption and implementation of a new computerized technology may engender changes in the image of the organization to outsiders, in the extent to which distinct groups within the organization coordinate their work tasks, and in individual employees' feelings of job security as a function of their technical expertise and trust in the organization. Mixed-determinant and mixed-effect models may be combined to create complex cross-level models of antecedent and outcome networks.

2. *Cross-level moderator models* suggest that the relationship between two lower-level constructs is changed or moderated by a characteristic of the higher-level entity in which they are both embedded. One may also formulate the model so that a cross-level relationship between a higher-level construct and a lower-level construct is moderated by another lower-level construct. These two forms are actually identical because each model specifies direct and interactional effects of the higher- and lower-level constructs on a lower-level outcome measure. As an example, consider the effects of unit technology on the relation between individual cognitive ability and individual job performance. Generally, higher ability is associated with higher performance. However, routine unit technology limits individual discretion, thereby limiting the relevance of cognitive ability to performance. Conversely, uncertain unit technology fosters high individual job discretion, allowing cognitive ability to enhance job performance. Unit technology thus moderates the relationship of individual ability and performance.

3. *Cross-level frog-pond models* highlight the effects of a lower-level entity's relative standing within a higher-level entity. The term *frog pond* captures the comparative or relative effect that is central to theories of this type: depending on the size of the pond, the very same frog may be small (if the pond is large) or large (if the pond is small). Also called *heterogeneous, parts,* or *individual-within-the-group* models (Dansereau et al., 1984; Glick & Roberts, 1984; Klein et al., 1994), theoretical models of this type are cross-level models in that the consequences of some lower-level (typically individual-level) construct depend on the higher-level (typically group-level) average for this construct: where one stands relative to the group average. Consider, for example, the relationship between an individual's amount of education and his or her influence in problem-solving discussions within a group. A college-educated individual

may have a great deal of influence if his or her group members' average amount of education is relatively low (few graduated from high school), or very little influence if his or her group members' average amount of education is relatively high (most have postgraduate degrees). Thus the relationship between an individual's education and his or her influence in a group depends on the individual's relative standing within his or her group's degree of education. Frog-pond models of this type, we should note, may be categorized in different ways in levels typologies. We have classified frog-pond models as cross-level models, but we recognize that frog-pond models do not evoke unit-level constructs in the same way as the other cross-level models already described. The "group average" specified in a frog-pond model is not conceptualized as a shared property of the unit. Indeed, were the construct predicted to be shared within each group, then it would make no conceptual or empirical sense to assess individual standing on the construct relative to the mean—the hallmark of frog-pond models (X_i – the group mean of X). Nor is the "group average" considered a global property of the unit; perhaps the group average, in combination with deviations, may be considered a configural property of the unit. This insight is subtle and complex, but it may help clarify why the frog-pond effect has been classified by some scholars as a distinct phenomenon or even as a distinct level of analysis. Just as we have created a distinct category for configural unit-level properties—unit properties that are characteristics of the unit but are neither global nor shared (isomorphic)—so others (e.g., Klein et al., 1994; Dansereau & Yammarino, Chapter Ten, this volume), in their conceptualizations, have designated frog-pond (heterogeneous or parts) models as a distinctive level.

Homologous multilevel models. These models specify that constructs and the relationships linking them are generalizable across organizational entities. For example, a relationship between two or more variables is hypothesized to hold at the individual, group, and organizational levels. Such models are relative rarities. The most commonly cited example of such a model is Staw, Sandelands, and Dutton's (1981) model of threat rigidity. Staw and his colleagues posit that the way in which individuals, groups, and organizations respond to threat is by rigidly persisting in the current response. By

arguing for parallel constructs and homologous linking processes, they have developed a homologous multilevel model of threat-rigidity effects. However, the model has not been tested empirically, its propositions are open to debate (e.g., House et al., 1995), and its attention to construct composition is limited. Lindsley, Brass, and Thomas's model (1995) of efficacy-performance spirals is an excellent example of a homologous multilevel model that carefully attends to the composition of its constructs. However, we know of no empirical test, in the published organizational literature, of a fully homologous multilevel model.

Given their generalizability across levels, homologous multilevel models are, at their best, uniquely powerful and parsimonious. At their worst, however, multilevel homologies may be trite. A search for parallel and generalizable constructs and processes may so reduce and abstract the phenomenon of interest that the resulting model may have little value at any level. The basic notion that goals influence performance at the individual, group, and organizational levels may be valid but not, at least in its bare-bones formulation, very interesting or useful. A hypothesis that is readily applicable to many levels may be a very basic hypothesis, indeed. In the literature there are examples of efforts to develop and apply homologous multilevel models to organizational behavior (e.g., Kuhn & Beam, 1982; Tracy, 1989), although these models have had little influence on theory or research. Thus the theorist must be aware of the tension inherent in the construction of multilevel models: good ones have the potential to advance and unify our field, but weak ones offer little to our understanding of organizational phenomena.

Sampling in Multilevel Research

Sampling within and across units. When testing individual-level theoretical models, researchers endeavor to ensure that their samples contain sufficient between-individual variability to avoid problems of range restriction. Sampling issues in multilevel research are more complex but comparable. In testing unit-level theoretical models (for example, the relationship between organizational climate and organizational performance) and mixed-level models containing unit- and individual-level variables (for example, the relationship of organizational human resources practices and individual organizational commitment), researchers must endeavor to

ensure that their samples show adequate variability on the constructs of interest, *at all relevant levels in the model.* Thus, for example, it may be inappropriate to test a cross-level model linking a group construct to an individual outcome in a single-organization sample. If a higher-level organizational characteristic constrains between-group variability, it will yield range restriction on the measure of the group construct and preclude a fair test of the model. Unfortunately, this problem is all too common in levels research.

In testing models containing shared unit-level constructs, researchers must endeavor to obtain samples showing within-unit homogeneity *and* between-unit variability on the shared constructs. Thus, for example, if a theoretical model asserts that units develop shared norms over time and that these norms influence unit-level or individual-level outcomes, then a test of the model requires units in which individuals have worked together for a considerable period; newly formed task groups, for example, would provide an inappropriate sample for the study. The researcher's sampling goal, then, is to obtain experienced units showing shared norms that differ between the units. Alternatively, a researcher may explicitly model and gather data to test the hypothesis that the length of time unit members have worked together predicts the emergence of shared norms, which in turn influence unit-level or individual-level outcomes. In this scenario, the researcher's sample should contain units showing substantial variability in the length of time that unit members have worked together. This strategy allows a researcher to test the variable (time that unit members have worked together) hypothesized to engender the emergence of shared norms. The outcome measure for this hypothesis, then, is not the level or nature of a shared norm but the extent to which the norm is shared (or, conversely, its dispersion across group members).

The collection of data to test a multilevel model, or even a single unit-level model, is thus likely to be labor-intensive and time-consuming. It is not enough to sample many people in one organization. The multilevel researcher, whose variables include measures of shared and configural constructs, must sample many people in many units that are nested in many higher-level units. In other words, multilevel research generally necessitates sampling several organizations, units within these organizations, and individuals within these units. To be forewarned is to be forearmed: it is not reasonable to whine about range restriction in mixed-level data after the fact!

PRINCIPLE: *In the evaluation of unit-level or mixed unit-level and individual-level theoretical models, the sampling strategy must allow for between-unit variability at all relevant levels in the model. Appropriate sampling design is essential to an adequate test of such models.*

Sampling across time. In the section on theoretical principles (see "Principles for Multilevel Organizational Theory Building," pp. 21–25), we highlighted the importance of time, as well as its general neglect in theory construction for processes that link different levels. However, temporal considerations are important not only for theory; they are also essential to research design. Two issues are central: differential time scales across levels, and entrainment.

The first issue, differential time scales across levels, concerns the fact that higher-level and lower-level phenomena operate on different time scales. In general, lower-level phenomena change more quickly, whereas higher-level phenomena tend to change more slowly, and so it is easier to detect change in lower-level entities. This means that top-down cross-level relations, if present, can be readily detected with cross-sectional and short-term longitudinal designs. In related fashion, emergent phenomena generally need longer time frames to unfold and manifest at higher levels, and so bottom-up emergent effects require longitudinal designs.

PRINCIPLE: *Time-scale differences allow top-down cross-level effects to be meaningfully examined with cross-sectional and short-term longitudinal designs. Bottom-up emergent effects necessitate long-term longitudinal or time-series designs.*

The second issue, entrainment, concerns the fact that the links between some phenomena are cyclical; that is, the strength of a link may vary over time and will be detectable only during periods of entrainment. Therefore, a theory that includes entrained phenomena necessitates a very carefully timed research design that can sample relevant data during periods of entrainment. To the extent that such a theory represents an effort to evaluate entrainment as a

process, the design must also be capable of sampling relevant data during periods when the phenomena are not entrained.

PRINCIPLE: *Entrainment tightly links phenomena that are ordinarily only loosely connected across levels. Sampling designs for the evaluation of theories that propose entrained phenomena must be guided by theoretically specified time cycles, to capture entrainment and its absence.*

Analytic Strategies

Several techniques are available for the analysis of multilevel data: analysis of covariance (ANCOVA) and contextual analysis using ordinary least squares (OLS) regression (e.g., Mossholder & Bedeian, 1983); cross-level and multilevel OLS regression; WABA (Dansereau et al., 1984); multilevel random-coefficient models (MRCM), such as hierarchical linear modeling (HLM; Bryk & Raudenbush, 1992); and multilevel covariance structure analysis (MCSA; Muthen, 1994). The techniques differ in their underlying theoretical assumptions and are designed to answer somewhat different research questions. Therefore, no single technique is invariably superior in all circumstances; rather, the choice of an analysis strategy is dependent on the nature of the researcher's questions and hypotheses. Here we see again the primacy of theory in dictating the resolution of levels issues. The best way to collect and the best way to test multilevel data will depend on the guiding theory. The more explicit and thorough the guiding theory, the more effective data collection and analysis are likely to be. We provide a brief overview of these analytic approaches here but direct the reader to later chapters in this volume for in-depth consideration of contextual and regression analysis (James & Williams, Chapter Nine), WABA (Dansereau & Yammarino, Chapter Ten), and multilevel random-coefficient models (Hofmann, Griffin, & Gavin, Chapter Eleven).

ANCOVA and contextual analysis. Among the earliest approaches to the analysis of cross-level data were adaptations of ANCOVA and the use of OLS regression to conduct contextual analysis (Firebaugh, 1979; Mossholder & Bedeian, 1983). The ANCOVA approach is used to determine whether there is any effect on an individual-level dependent variable that is attributable to the unit, beyond the effect accounted for by individual differences. Essentially, this approach treats the individual-level variables as covariates and then uses unit membership as an independent variable to determine how much variance is attributable to the unit. Unit membership as a variable accounts for all possible remaining differences across units. Therefore, this approach cannot identify the specific constructs relevant to unit membership that are actually responsible for observed differences among groups; such effects are unexplained. Nevertheless, to the extent that there are any differences attributable to the grouping characteristic, this approach will capture it (Firebaugh, 1979).

The regression approach to contextual analysis typically uses aggregation and/or disaggregation to specify contextual constructs of interest. Although it is typically used to determine the effects of one or more higher-level contextual constructs on an individual-level dependent variable, it is actually flexible with respect to level. "Classic" contextual analysis includes individual-level predictors and unit means on the same predictors, to assess the relative amounts of variance attributable to the unit (Firebaugh, 1979). To the extent that unit means on the variables of interest account for variance beyond that explained by their individual-level counterparts, a contextual effect is demonstrated. This approach generally explains less variance than ANCOVA because the substantive unit variables are usually a subset of the total group composite effect, but it does identify the unit characteristic responsible for differences. Note that the aggregation process in classic contextual analysis is typically atheoretical (that is, no theoretical model of emergence is modeled), and isomorphism is not evaluated.

Cross-level and multilevel regression. In the organizational literature, OLS regression has been adapted to examine cross-level and multilevel effects and is quite flexible with respect to the type of model it can evaluate. Contemporary uses of this approach treat aggregation as an issue of construct validity (James, 1982; Kozlowski & Hattrup, 1992) so that a model of emergence is first evaluated before individual-level data are aggregated to the group level (e.g., Kozlowski & Hults, 1987; Ostroff, 1993). Therefore, with respect to the specification and measurement of construct types, this approach is relevant to the issues we have discussed in this chapter. Once the measurement model of the higher-level (aggregated)

constructs is established, the analysis proceeds to test substantive hypotheses. For example, if the theory assumes shared perceptions of unit climate as predictors of individual satisfaction, then one establishes restricted within-unit variance on climate, aggregates the data to the unit level (that is, computes means), and then disaggregates to the individual level of analysis (that is, assigns the means to individuals in the unit). The analysis then estimates the amount of variance in individual satisfaction that is attributable to unit climate. Individual-level analogues of the contextual construct are not necessarily controlled (as in contextual analysis) unless the question is of substantive interest (James & Williams, Chapter Nine, this volume).

Within-and-between analysis. The basic WABA equation (Dansereau et al., 1984) is modeled on the classic decomposition of within-and-between variance terms formulated by Robinson (1950) to model individual-level and aggregate group-level correlations. The "classic" WABA analysis examines bivariate relationships, assumes measures at the lowest level of analysis for all constructs, and proceeds in two phases. The first phase, WABA I, establishes the level of the variables. The second phase, WABA II, evaluates the level of relations between all the variables in the analysis (Dansereau et al., 1984). WABA I is designed to assess whether measures, treated one at a time, show variability in the following ways: both within and across units (as typically with an individual-level construct), primarily between units (as typically with a unit-level construct), and primarily within units (as with a frog-pond, parts, or heterogeneous construct). WABA II is designed to assess whether two measures covary in the following ways: both within and across units (as typically with individual-level relationships), primarily between units (as typically with unit-level relationships), and primarily within units (as typically with a frog-pond, parts, or heterogeneous relationship; see Klein et al., 1994). Although WABA was originally developed to examine bivariate relations at multiple levels, it has been extended to address multivariate relations (Schriesheim, 1995; Dansereau & Yammarino, Chapter Ten, this volume).

Multilevel random-coefficient modeling. The MRCM analysis strategy is represented by several packages of statistical software (for example, PROC MIXED in SAS; MLn; lme in S-PLUS), of which

HLM is probably the most familiar. HLM analysis assumes hierarchically organized, or nested, data structures of the sort that are typically encountered in organizations: individuals nested in units, units nested in organizations, and organizations nested in environments. Models of theoretical interest typically represent multiple levels of data. For instance, many cross-level models involve an outcome variable at the lowest level of analysis, with multiple predictors at the same and higher levels. HLM is well suited to the handling of such data structures.

The logic of HLM involves a simultaneous two-stage procedure. Level 1 analyses estimate within-unit intercepts (means) and slopes (relations). To the extent that unit intercepts and/or slopes vary significantly across units, Level 2 analyses treat them as outcomes. Thus Level 2 analyses model the effects of unit-level predictors on unit intercepts and slopes so that effects on intercepts are indicative of direct cross-level relations, and effects on slopes are indicative of cross-level moderation. HLM relies on a generalized least squares (GLS) regression procedure to estimate fixed parameters, and on the EM algorithm to generate maximum-likelihood estimates of variance components. This provides many statistical advantages over analogous OLS regression–based approaches (Hofmann et al., Chapter Eleven, this volume).

An in-depth description of these techniques is beyond the scope of this chapter; assumptions, applications, and differences among the techniques are addressed elsewhere in this volume. However, we will note here that all these techniques have the potential to be misused in an atheoretical attempt to establish "the" level at which effects occur. We reiterate that the conceptual meaning of higher-level aggregations (however they are statistically determined) must have an a priori theoretical foundation.

PRINCIPLE: *There is no one, all-encompassing multilevel data-analytic strategy that is appropriate to all research questions. Particular techniques are based on different statistical and data-structure assumptions, are better suited to particular types of research questions, and have different strengths and weaknesses. Selection of an analytic strategy should be based on (a) consistency between the type of constructs, the sampling and data, and the research question; and (b) the assumptions, strengths, and limitations of the analytic technique.*

Extending Models of Emergent Phenomena

Some of the most engaging and perplexing natural phenomena are those in which highly structured collective behavior emerges over time from the interaction of simple subsystems [Crutchfield, 1994, p. 516].

A central theme woven throughout this chapter is the need for a more extended understanding of emergence as a critical multilevel process in organizational behavior. There is evident dissatisfaction with the overreliance on isomorphism-based composition as the primary model for conceptualizing collective constructs (House et al., 1995; Rousseau, 1985). Indeed, there is increasing recognition that emergence based on isomorphism may well be the exception rather than the rule. Although isomorphic emergence is a very powerful conceptual model, it is but one possible model. Emergent phenomena are not necessarily shared, uniform, and convergent. In their discussion of dispersion theory, a precursor to our typology, Brown and Kozlowski (1997, p. 7) note that nonuniform "phenomena marked by differentiation, conflict, competition, coalition formation, and disagreement are common" in organizations.

There are many theories, in our literature and others, that implicitly or explicitly address alternative forms of emergence. Power, conflict, and competition all involve compilational, discontinuous forms of emergence. The *varient paradigm* (Dansereau & Yammarino, Chapter Ten, this volume), with its interest in "parts" relationships, shows a recognition of the plausibility of compilation. This is a good beginning, but the "parts" perspective captures but one form of compilation among many. We argue that there is a need to extend the conceptualization of emergence, to make it more inclusive, so that our theories and research can encompass more varied and diverse emergent phenomena. We need to elaborate compilation forms of emergence.

Conceptual Goals

Purpose

Our purpose is to take a step toward this elaboration, describing forms of emergence that until now have received little attention in the organizational literature on levels of analysis. In preceding

sections of this chapter, we contrasted composition (shared unit properties) and compilation (configural unit properties) as distinctive, ideal types of emergence. This contrast was useful in making salient the important differences that affect conceptualization, measurement, and sampling. However, composition and compilation are not necessarily clear-cut dichotomous categories; rather, they are end points for a diverse set of emergence alternatives, with some forms of emergence being more akin to composition and some forms being more akin to compilation.

We now explore varying forms of emergence, hoping to foster increased attention to the structures and processes underlying emergent organizational phenomena. We undertake this exploration here by elaborating the theoretical underpinnings of emergence. First we consider, in greater depth, the theoretical foundation for emergence. A primary focus of our attention is the central role that interaction processes and dynamics among individuals play in shaping the form of the emergent phenomenon. Next, with this foundation in place, we identify more specific theoretical assumptions that distinguish the ideal or pure types of composition and compilation forms of emergence. We describe and illustrate how the assumptions change when one is considering discontinuous compilation relative to isomorphic composition. Finally, we develop a typology, posing a set of emergence exemplars that range between the ideal types of composition and compilation. We discuss each exemplar, providing examples from the literature that consider unit performance, unit learning–cognition–knowledge, and other unit phenomena, to illustrate how the theoretical assumptions help to explicate the nature of emergence for that exemplar. Our use of the typology is intended to help elaborate the theoretical underpinnings that shape the conceptualization of alternative forms of emergence.

Contributions

There are three primary conceptual contributions of this effort. First, our intent is to be inclusive, encompassing multiple perspectives. Several recent theoretical efforts have started to explore emergence and the ways in which it may be manifest (Brown & Kozlowski, 1997, 1999; Brown et al., 1996; Chan, 1998; Kozlowski, 1998, 1999; Morgeson & Hofmann, 1999a, 1999b). Although these efforts are for the most part compatible, they have also chosen different points of theoretical

departure, different language, and different organizing structures. It is not our goal to explicitly integrate these efforts, but we believe our framework makes their compatibilities more explicit. We build on the strong theoretical and research foundation provided by isomorphism-based composition and elaborate it to embrace different, alternative, and neglected forms of emergent organizational phenomena that follow from a consideration of discontinuity-based compilation. Because compilation entails less restrictive assumptions, it allows for many more possible emergent forms relative to composition. We argue that a broader range of alternatives, from composition to compilation, is necessary to more fully capture complex emergence.

Second, an important contribution of our perspective is the recognition that higher-level phenomena do not necessarily exhibit universal forms of emergence; that is, a given phenomenon may emerge in different ways depending on the context and the nature of lower-level interaction processes. We need to attend to the ways in which interaction processes and dynamics shape the form of emergence. Therefore, the search for universal models of emergence, to be applied in each and every instance, may be misguided. Our perspective emphasizes that a collective phenomenon—unit performance—may emerge in a variety of different ways in different units. We need flexible conceptual tools that allow us to seek out, explore, and characterize variation in forms of emergence.

Third, our intent is to stimulate a more extended conceptualization of the theoretical mechanisms that characterize different forms of emergence. We develop a typology of emergence that explicitly links exemplars of different emergent forms to key theoretical underpinnings. Our focus is on theory development, not on mere classification. We are not advocating simple reductionist explanations for higher-level phenomena. We recognize that many organizational phenomena are top-down rather than bottom-up. Further, as we have already explained, many phenomena reflect *both* top-down and bottom-up processes unfolding over time. Moreover, we are not rejecting macro single-level approaches that do not explicitly address the emergent origins of the higher-level phenomena. Rather, we seek to promote more inclusive, extensive, and coherent explanations of collective phenomena. We are interested in both structure *and* process. We wish both to understand the whole *and* keep an eye on the parts.

The issues we address go to the conceptual meaning of higher-level phenomena that are rooted in individual characteristics and actions. Consider, for example, the global outcome of a baseball game score. One can examine a global predictor of this outcome (for example, average ability of team members), but this predictor can only provide a limited understanding of the team's performance. Baseball team scores are equifinal. True fans know this. They follow box scores so that they can understand how individual team members, in *dynamic interaction*, compiled the team score. We believe that a similar degree of conceptual understanding can pay big dividends in our effort to comprehend meso organizational behavior.

Theoretical Underpinnings of Emergence

What Is Emergence?

Emergence is bottom-up and interactive. The concepts undergirding emergence have broad expression in the biological, social, and physical sciences and are represented in theories of chaos, self-organization, and complexity (Arthur, 1994; Gell-Mann, 1994; Kauffman, 1994; Nicolis & Prigogine, 1989; Prigogine & Stengers, 1984) which address the dynamics of emergence. Our focus is on emergent phenomena that occur within the boundaries and constraints of organizational systems. Emergence is particularly relevant in the continuing effort of our science to understand how individuals contribute to organizational effectiveness. This is a central theme in several of the chapters of this book, including those focused on selection (Schneider, Smith, & Sipe, Chapter Two), performance appraisal (DeNisi, Chapter Three), training effectiveness (Kozlowski et al., Chapter Four), and human resources management (Ostroff & Bowen, Chapter Five). Emergence plays an important role in the linkages involved in interorganizational relationships (Klein, Palmer, & Conn, Chapter Six) and cross-cultural relations (Chao, Chapter Seven).

A phenomenon is emergent when it originates in the cognition, affect, behaviors, or other characteristics of individuals, is amplified by their interactions, and manifests as a higher-level, collective phenomenon (Allport, 1954; Katz & Kahn, 1966). Individual cognition, affect, behavior, and other characteristics denote *elemental content.* Elemental content is the raw material of emergence. Team mental models (cognition), group mood (affect), team performance (behavior), and

group diversity (other characteristics) all represent emergent group properties that have their origins in the elemental content provided by individuals. *Interaction* denotes process. Individuals communicate and exchange information, affect, and valued resources. They share ideas. They communicate mood and feelings. They perform acts and exchange work products. Communication and exchanges may be direct, as in face-to-face interaction, or indirect, as when information or other resource exchange is mediated via some form of technology. The form of the interaction process, in combination with the elemental content, comprises the emergent phenomenon.

Emergence is shaped and constrained. Although emergent phenomena have their origins in lower levels, the process of emergence is shaped, constrained, and influenced by higher-level contextual factors. Interaction in organizations is constrained by a hierarchical structure that defines unit boundaries. The individuals in a unit tend to interact more dynamically and intensely with each other than with individuals outside their unit (Simon, 1973). Moreover, work-flow transactions—the ways in which people are linked to accomplish the work of the unit (Thompson, 1967)—pattern interactions and exchanges. Individuals directly linked by the work flow tend to interact more with each other than with individuals who are only linked indirectly (Brass, 1995). Thus, for example, professors tend to interact more intensely with the students who are involved in their research than with the other students in their programs, and they interact more with students in their programs than with students in other programs. This patterning of interaction by formal structure and work flow shapes emergence.

In addition, informal patterns of interaction—social interaction that transcends formal boundaries and work flows—also shape emergence. People who cross unit boundaries to bond socially are more likely to communicate common perspectives. For example, Rentch (1990) shows that individuals from different organizational units who met informally developed a shared conception of the organization's culture. In organizations, emergent phenomena are shaped by a combination of formal structure and work flows, and by informal social-interaction processes, with the relative importance of one, the other, or both dependent on the phenomenon of interest.

There are also a variety of other forces—such as attraction, selection, and attrition (ASA); common stimuli; socialization; and sensemaking—that affect interaction processes and dynamics. These forces, in combination with formal structures, work flows, and social structures, as already described, shape the nature of emergent phenomena. Generally, these forces have been conceptualized as constraining either the range of elemental content or the interaction process. Given these assumptions, the forces have been used to explain composition-based emergence, but they can also explain compilation. For example, the result of the ASA process is a workforce that is relatively more homogeneous in terms of ability, personality, attitudes, and values (Schneider & Reichers, 1983) and therefore more likely to have viewpoints in common. Organizational environments tend to expose employees to common stimuli—policies, practices, and procedures—that shape common perceptions (Kozlowski & Hults, 1987). Socialization can operate as a powerful force that shapes shared sensemaking (Louis, 1980). In these ways, the forces act as constraints shaping composition forms of emergence that are characterized by stability, uniformity, and convergence.

Sometimes the forces operate to expand rather than limit the range of elemental content or the nature of the interaction process. Compilation is based on the assumption that ASA, socialization, and related processes are not so powerful as to eliminate all meaningful differences in individual organizational members' elemental characteristics. Indeed, these processes may preserve or even engender variability within organizations, at least with respect to many important elemental qualities. For example, selection, attrition, and reward processes are unlikely to eliminate all variability in individual performance. Moreover, some organizations may well select individuals for their varying and idiosyncratic strengths, much as a sports team needs some players who are good on offense and other (typically different) players who are good on defense. Further, interactions among organizational members may engender similarity or dissimilarity; social interactions may unite or polarize employees. Finally, a variety of contextual factors limit an organization's ability (and often its desire) to build an organization of perfectly homogeneous individuals. Some measure of demographic variability is inevitable in most organizations, for example.

Further, diversity in an organization—with respect to organizational members' demographic characteristics, work experiences, education, and so on—may foster organizational creativity and innovation. In these ways, the forces create differences and discontinuities, shaping compilation forms of emergence that are characterized by irregularity, nonuniformity, and configuration.

Emergence varies in process and form. As already noted, interaction dynamics can lead to variation in the ways in which a higher-level phenomenon emerges; that is, a given phenomenon, such as team performance, can arise in a variety of different ways, even in the same organization. Individual characteristics, cognition, affect, and behavior are constrained by their context. Over time, interaction dynamics acquire certain stable properties; stable structure emerges from a dynamic process. Katz and Kahn (1966) describe this as recurrent patterns of interaction. Thus the emergence of a collective phenomenon is the result of a dynamic unfolding of *role exchanges* (Katz & Kahn, 1966), *ongoings* (Allport, 1954), or *compilation processes* (Kozlowski et al., 1999) among individuals. It is from these dynamics that a stable collective pattern emerges.

Morgeson and Hofmann (1999a) describe Allport's notion of *ongoing* as a recurrent pattern representing the intersection of individual action in its context. Individual ongoings encounter one another, creating interaction *events*. Subsequent interactions solidify a recurrent *event cycle*, which represents the emergence of a stable collective phenomenon. Similarly, Kozlowski and colleagues (1999) describe how team performance *compiles* upward from individual behaviors and work-flow transactions: individuals work out transaction patterns that regulate dyadic work flows, and as these dyadic exchanges stabilize, team members develop extended work-flow networks that stabilize around routine task demands. Gersick and Hackman (1990) characterize these stable patterns in teamwork as *habitual routines*.

However, because emergent phenomena are based on patterns of interaction, even small changes in individual behavior or dyadic interaction can yield big changes in the nature of emergence. For example, Kozlowski and colleagues (1999) also propose that task environments can change dramatically and unpredictably. Unexpected shifts, and the novel tasks they present, necessitate adaptation of team networks, an adaptation that is based on individuals and dyads developing alternative work flows. In this model, team performance and adaptability emerge across levels from individual action and dyadic transactions, creating enormous flexibility in the formation of adaptive work-flow networks that may resolve the novel situation. *The implication is that collective phenomena may emerge in different ways under different contextual constraints and patterns of interaction. Emergence is often equifinal rather than universal in form.*

This important implication of our conceptualization of emergence sets our framework apart from most others: a given phenomenon or construct domain does not necessarily have to exhibit a universal form of emergence;[6] that is, a given emergent phenomenon may be the result of composition processes in one situation and of compilation processes in another. A consideration of the examples shown in Figure 1.2 illustrates this point. Consider, for example, how personality makeup can differ across teams (Jackson, May, & Whitney, 1995; Moreland & Levine, 1992). Teams may be characterized by the high homogeneity indicative of personality composition, or by the heterogeneity indicative of personality compilation. There is no a priori theoretical reason to suppose that one or the other is a universal form for the way in which team personality emerges.

Consider collective cognition, for example. The construct of shared mental models (Klimoski & Mohammed, 1995) assumes that team members hold identical mental representations of their collective task. In contrast, alternative conceptualizations assume that team members' mental models have compatible configurations but are not necessarily identical. Group members have somewhat different mental representations of their collective task, based on their specific roles within the team. Members' different mental representations fit together in a complementary way, like the pieces of a puzzle, to create a whole that is greater than the sum of its parts (Kozlowski, Gully, Salas, & Cannon-Bowers, 1996). Similarly, collective knowledge may be conceptualized as the sum of individual knowledge; more nonredundant information is better, and collective knowledge is the sum of the parts. Alternatively, collective knowledge may be conceptualized as configural spirals: some individual knowledge is more useful than other knowledge; useful knowledge is selected and crystallized, and it then attracts

Figure 1.2. Theoretical Underpinnings of Emergence.

Emergent Process	**Composition** ◄·····► **Compilation**
Variation in Emergence	
	• Personality similarity	• Personality diversity
	• Shared mental models	• Compatible mental models
	• Classical decision making (single optimal solution)	• Naturalistic decision making (multiple solutions)
	• Pooled team performance	• Adaptive team networks
	• Organizational learning (sum of individual knowledge)	• Organizational learning (knowledge spirals)
Theoretical Assumptions		
Model	Isomorphism	Discontinuity
Elemental contribution		
Type	Similar	Dissimilar
Amount	Similar	Dissimilar
Interaction process and dynamics	Stable	Irregular
	Low dispersion	High dispersion
	Uniform	Nonuniform
Combination	Linear	Nonlinear
Emergent representation	Convergent point	Pattern

and amplifies related knowledge, in a spiral of collective knowledge acquisition (Nonaka, 1994).

The point of these examples is that given phenomena may emerge in different ways. A variety of contextual and temporal constraints operate to influence interaction dynamics among individuals, which in turn shape the emergent form, yet the dominance of composition models based on isomorphism has tended to limit consideration to shared models of emergence, and to the dichotomous presence or absence of emergence (Brown & Kozlowski, 1997). Theory needs to be able to capture the rich complexity of emergence rather than limiting emergence to universal conceptualizations that often do not exist.

Theoretical Assumptions

Our framework is formulated around theoretical distinctions between ideal forms of composition and compilation, considered in earlier sections of this chapter. Here we turn our attention to three sets of overlapping assumptions, shown in Figure 1.2, that are useful for more finely distinguishing these alternative forms of emergence. The assumptions include the following elements:

1. The theoretical model of emergence, and the type and amount of elemental contribution implicated by the model
2. The interaction process and dynamics that shape the form of emergence
3. The resulting combination rules for representing the emergent form.

At the risk of some redundancy, we will outline these assumptions and apply them to the contrasting of composition and compilation forms of emergence. We will then present a typology, using the assumptions to distinguish alternative forms of emergence ranging between composition and compilation ideals.

Model and elemental contribution. Composition and compilation are distinguished by their underlying theoretical models. Composition is based on a model of isomorphism, whereas compilation is based on a model of discontinuity.[7] Isomorphism and discontinuity represent

differing conceptualizations with respect to the nature and combination of the constituent elements that constitute the higher-level phenomenon.

Isomorphism essentially means that the type and amount of elemental content—the raw material of emergence—are similar for all individuals in the collective. In other words, the notion of isomorphism is based on an assumption that all individuals perceive climate, for example, along the same set of dimensions, or that all team members possess mental models organized around the same content. In addition, isomorphism means that the amount of elemental content is essentially the same for all individuals in the collective. In other words, the climate or mental model is shared. Hence, within-unit convergence (that is, consensus, consistency, homogeneity) is central to composition. Morgeson and Hofmann (1999a, 1999b) describe this similarity in the type and amount of elemental content as *structural equivalence.* Thus isomorphism allows the theorist to treat a phenomenon as essentially the same construct at different levels (Rousseau, 1985). Note that isomorphic constructs are also *functionally equivalent.* That is, they occupy the same roles in multilevel models of the phenomenon; they perform the same theoretical function (Rousseau, 1985).

Discontinuity means that either the amount or type of elemental content is different, or both the amount and type are different. The notion of discontinuity is based on an assumption that the kinds of contributions that individuals make to the collective are variable, not shared and consistent. Essentially, there is an absence of structural equivalence in the nature of the elemental content and in the ways in which it combines (Kozlowski, 1998, 1999; Morgeson & Hofmann, 1999a, 1999b). Nevertheless, there is functional equivalence because the constructs perform the same role and function in models at different levels (Rousseau, 1985), as we shall explain.

The elemental content comes from a common domain—performance, personality, cognition—but the nature of individual contributions can be quite different. For example, baseball players contribute qualitatively different types and amounts of individual performance to accomplish team performance. The pitcher pitches, fielders field, and batters hit. In any given game, some will excel and others will make errors. Different dominant personality traits characterize each team member. Team members possess different but compatible mental models of the game. Therefore, variability and pattern are central to compilation. Because the diverse elemental content is drawn from a common domain and contributes to a similar collective property, there is functional equivalence across levels. This functional equivalence allows the theorist to treat compilational properties as qualitatively different but related manifestations of the phenomenon across levels (Kozlowski, 1998, 1999; Morgeson & Hofmann, 1999a, 1999b).

Interaction process and dynamics. The hallmark of composition forms of emergence is convergence and sharing. In climate theory, for example, a variety of constraining forces have been proposed that are thought to shape the emergence of a shared collective climate. Individuals are exposed to homogeneous contextual constraints—common organizational features, events, and processes (James & Jones, 1974). They develop individual interpretations of these characteristics, yielding psychological climate. ASA processes operate to narrow variation in psychological climate (Schneider & Reichers, 1983). Interpretations are filtered and shaped by leaders (Kozlowski & Doherty, 1989). Individuals interact, communicate perspectives, and iteratively construct a common interpretation. Variations in individual interpretations dissipate as a collective interpretation converges. This is an incremental process that, over time, promotes stability, characterized by reduced dispersion as outliers are trimmed and by increased uniformity as perceptions are pushed to a convergent point. An equilibrium is achieved.

The hallmark of compilation forms of emergence is variability and configuration. Team performance requires that individuals coordinate and dynamically combine distinct individual knowledge and actions. The emergence of team performance is largely shaped by work-flow interdependencies—that is, the linkages that connect individual performance in the team work system (Brass, 1981). Consider once again the performance of a baseball team. There are any number of ways in which team members, working together, can achieve a particular score. They may excel because power hitters recurrently hit home runs. They may have a stable of good but not exceptional hitters; by consistently getting players on base the team is able to accumulate good scores. They may excel by

limiting the success of the opposing team; exceptional pitching, for example, will keep opposing scores low, and good defensive fielding, along with solid teamwork, will be needed to support the pitcher. Each player on the team will make distinctive individual contributions that combine in myriad ways to yield the team's performance. The *score* may be no more than the sum of its parts (that is, runs), but *team performance* is more than a simple sum of parts. Decomposing team performance necessitates an understanding of who did what, when, and of how it all fits together. This is an irregular process rather than incremental, stable interaction. There will be considerable dispersion and nonuniformity in the ways in which individual contributions are coordinated and combined to yield the compiled team performance (Kozlowski et al., 1999).

Combination rules and representation. The representation of an emergent construct is an effort to capture or freeze the result of a dynamic process. The assumptions identified earlier provide the basis for different combination rules—guidelines for summarizing or capturing a collective representation from the elemental content. For composition, similar types and amounts of elemental content that evidences relative stability, uniformity, and low dispersion will generally be summarized with linear additive or averaging rules. This procedure will yield a single indicator—a convergent point capturing the shared unit property. Collective climate, based on composition assumptions, is generally represented by unit means (Kozlowski & Hattrup, 1992). Homogeneous perceptions of worker participation are likewise represented as unit means (Klein et al., 1994).

For compilation, a variety of different nonlinear combination rules may be used to combine the different types and amounts of elemental content. Compilation interaction processes are irregular, high in dispersion, and nonuniform. Elemental content may vary in amount, kind, or both. Therefore, the combination rules for compilation are more varied and complex than those used to characterize composition. A sampling of potential combination rules includes disjunctive, conjunctive, and multiplicative combination models, and indices of variance, proportion, configural fit, and network characteristics, among others (Levine & Fitzgerald, 1992; Meyer, Tsui, & Hinings, 1993). The key issue is that the combination rules should be consistent with the conceptualization of emergence. For example, if the compilation theory emphasizes team networks (Kozlowski et al., 1999), then the representation should capture such meaningful variation in network characteristics as centrality, transaction alternatives, and substitutability (Brass, 1981). If the theory emphasizes the formation of dyadic relationships, as in leader-member exchange (Graen, 1976), then the representation should capture relative standing on the basis of differences between leader-member pairs (Dansereau & Dumas, 1977). If the theory focuses on the formation of in-groups and out-groups (Kozlowski & Doherty, 1989), then the representation should capture in-and out-group standing and differences (Brown & Kozlowski, 1997, 1999).

Summary of distinctions between composition and compilation. The key assumptions that distinguish composition and compilation, respectively, involve the question of whether the following elements are present:

1. Elemental (that is, individual) contributions to the higher-level phenomenon are similar (isomorphism) or dissimilar (discontinuity) in type, amount, or both
2. Interaction processes and dynamics are incremental and stable, exhibit low dispersion, and are uniform in pattern, or interaction processes and dynamics are irregular, high in dispersion, and exhibit nonuniform patterns
3. The emergent phenomenon is consequently represented by a linear convergent point (composition), or the emergent phenomenon is represented as a nonlinear pattern or configuration (compilation)

A Typology of Emergence

The purpose of our typology is to promote a more expansive conceptualization of the theoretical mechanisms that characterize different forms of emergence. Our typology of emergence, shown in Figure 1.3, juxtaposes composition and compilation. The theoretical underpinnings derived previously are used to distinguish a variety of exemplars—specific emergence models. We discuss each exemplar, illustrating the exemplars with examples regarding

Figure 1.3. Typology of Emergence.

Emergence Theory and Features	**Isomorphic Composition** ◄			► **Discontinuous Compilation**		
	↓			↓		
Exemplars	*Convergent*	*Pooled Constrained*	*Pooled Unconstrained*	*Minimum/ Maximum*	*Variance*	*Patterned*
• Performance →	Rowing crew Synchronized swimming	Tug of war Group sales	Social loafing Free riding	Climbing team Jury decision making	Jazz improvisation Dance	Adaptive team performance Performance spirals
• Learning/ Knowledge →	Shared mental models/ knowledge	Group information exchange	Organizational learning/ knowledge	Crew ability/ knowledge	Creativity Knowledge/ diversity	Knowledge spirals Compatible mental models/knowledge Transactive memory
• Others →	Collective Climate Efficacy		Unit rates Absence Turnover Accidents		Norm cystallization Culture strength Personality diversity	LMZ Intragroup conflict
Elemental Contribution						
Type	Similar	Similar	Similar	Similar	Variable	Dissimilar
Amount	Similar	Moderately similar	Similar to dissimilar	Dissimilar	Variable	Dissimilar
Interaction Process, Combination Rules, and Representation	Low dispersion Uniform Sum or mean **(linear)**	Moderate dispersion Uniform Sum or mean **(linear)**	Moderate to high dispersion Uniform Sum or mean **(linear)**	Dispersion NA Uniform NA Minimum/Maximum **(nonlinear)**	Variable dispersion Uniform Variance **(nonlinear)**	High dispersion Nonuniform Patterns Profiles Networks Proportion **(nonlinear)**

collective performance, learning–cognition–knowledge, and other phenomena. We include exemplars for the following types of emergence: convergent, pooled constrained, pooled unconstrained, minimum/maximum, variant, and patterned. Each exemplar describes a different emergence process, based on contextual constraints and interaction processes, for how a lower-level phenomenon is manifested at a higher level. The nature of elemental contributions, in type and amount, and the combination rules applicable to each exemplar are indicated. Although we have used the individual and group levels to make the examples easier to explain, the models are applicable to higher levels as well. The typology is intended to help elaborate the theoretical underpinnings that shape the conceptualization of alternative forms of emergence.

Convergent Emergence

The exemplar for this type of emergence represents the ideal form of composition that we have discussed throughout this chapter. The model is based on the assumption that contextual factors and interaction processes constrain emergence in such a way that individuals contribute the same type and amount of elemental content. Therefore, the phenomenon converges around a common point that can be represented as a mean or a sum. For example, the performance of a crew rowing a scull is dependent on each individual providing the same amount and type of physical thrust at precisely the same time. Synchronized swimmers must execute the same movements, in the same amount, at the same time. Similarly, the notion of team mental models is predicated on all team members sharing the same amount and type of knowledge (Klimoski & Mohammed, 1995). Ideal composition is also illustrated by theory and research on collective climate and collective efficacy. Group members' perceptions converge on the referent construct. Sharing is evaluated on the basis of consensus or consistency. Variability in elemental content and individual contributions is very low and uniform in distribution across members. Therefore, aggregation to the group mean eliminates the small amount of error variance and effectively represents the group on the higher-level construct.

Alternative subforms of this exemplar can be distinguished on the basis of the item referent used to create the emergent construct (Chan, 1998; Klein et al., 1998); that is, individual-level measures

may reference the self ("how I perceive") or the group ("how I believe the group perceives"). The self-referenced-item form is described by Chan (1998) as "direct consensus," and the group-referenced form is described as "referent shift consensus." This latter form is regarded as being more consistent with the conceptual underpinnings of the higher-level construct (James, 1982; Klein et al., 1994; Rousseau, 1985). Some research suggests that the referent-shift form may enhance within-group agreement and between-group variability (Klein et al., 1998). In related fashion, DeShon et al. (1999) indicate that aggregated group-referenced measures are better predictors of group performance than aggregated individual-referenced measures of the same construct. Empirical findings are preliminary at this point. Sometimes the item referent (self or group) makes a difference; at other times it does not. Clearly, this is an important issue that can be resolved only with systematic research.

Pooled Constrained Emergence

This exemplar relaxes the assumptions for the amount of elemental contribution, but the type of content remains similar. The underlying model is based on the assumption that contextual factors and interaction processes shape emergence in such a way that some minimum amount of contribution is required of each individual. Therefore, there will be restricted variability within the group, yielding a pattern across individuals that is relatively uniform and moderate in dispersion. An additive or averaging model combines the elemental contributions.

Consider, for example, group sales performance for a district. Each salesperson makes an incremental, pooled contribution to group performance. The elemental contributions are similar in type but can vary in amount to some extent. Contextual constraints—such as incentives, competitiveness, leadership, and dismissal—are likely to restrict just how little can be contributed. All salespeople are not expected to contribute the same amount, but contributing too little will likely lead to turnover. Therefore, individual and group performance are not identical, but they are closely related.

Wittenbaum and Stasser (1996) provide a model of group discussion and consensus decision making consistent with this form of emergence. In their model, group members possess both

unique and common information that must be discussed and combined to yield a group decision. Although individuals possess both similar and dissimilar types of elemental content (that is, common and unique information), groups have been found to focus virtually all of their discussion on sharing the common information. In effect, the nature of social interaction processes constrains emergence so that only common information is discussed and used for the decision. Although there is some variation in individual contribution, the dissimilar information plays no role in the team product. The group decision is essentially an average of the shared information.

Pooled Unconstrained Emergence

This exemplar fully relaxes the requirement on the amount of elemental contribution, but, as before, the type of content remains similar. Here, variation in the amount of elemental contribution can be quite high. For example, research demonstrates that performance in pooled tasks can be plagued by social loafing and free riding: some individuals contribute far less to the collective when the amount of their contributions cannot be identified (Harkins, Latane, & Williams, 1980). In such circumstances, the group product may be represented as a sum or mean. However, in contrast with the previous exemplar, the group representation and the individual contribution may be dramatically different. Similarly, one conceptualization of organizational climate is based on the assumption that within-group variation in climate perceptions is random measurement error (Glick, 1985, 1988). No restriction is placed on how much variability can be eliminated through averaging.

This exemplar is also frequently used for such group descriptive variables as absence, turnover, and accidents (e.g., Hofmann & Stetzer, 1996; Mathieu & Kohler, 1990). Unit rates are typically counts of the dichotomous presence or absence of some event: additive frequency counts, although sometimes these characteristics are summarized by means. Bliese (Chapter Eight, this volume) labels phenomena of this sort *fuzzy composition* because they lack the sharing that is the hallmark of composition. Other theorists have used group rates as examples of discontinuity (Rousseau, 1985), which is indicative of compilation. Therefore, these phenomena certainly represent fuzzy *something;* whether they are fuzzy composition or fuzzy compilation is not necessarily an important issue unless one is highly interested in classification. However, the fuzziness suggests that this exemplar captures a transition zone between the ideal types. Deeper conceptual digging may be useful for surfacing theoretical nuances that may help us better understand these differing forms of emergence.

One factor to consider in this deeper digging may be the base rate. In some instances, the elemental contribution can be spread across many (though not all) members of a unit—the incidence of stress, for example. In other instances, the rate is often predominantly influenced by the acts of just a few individuals—for example, serious accidents. Perhaps the first group of instances is more akin to fuzzy composition, and the second more akin to fuzzy compilation.

Minimum/Maximum Emergence

This exemplar represents a shift from linear combination rules (that is, additive models) to nonlinear rules. Elemental contribution is based on similar content, but the amount of contribution is qualitatively distinct. Contextual factors and interaction processes constrain emergence so that the pattern across individuals is discontinuous. The standing of one individual on the phenomenon in question determines the standing of the collective. Therefore, dispersion and uniformity are not directly applicable to the conceptualization of this exemplar.

This is a conjunctive (minimum) or disjunctive (maximum) model, in which the highest or lowest value for an individual in the group sets the value of the collective attribute (Steiner, 1972). Consider, for example, group cognitive ability for a tank crew (Tziner & Eden, 1985) or a football team. It is not the average level or dispersion of cognitive ability that is important, because the same sort of cognitive contribution may not be necessary for all members; as long as one person is high on cognitive ability and the rest of the team will take direction, the group as a whole can effectively assess the situation and execute the appropriate strategy. Therefore, the maximum individual-level standing on the attribute determines the standing of the collective. This emergence process is similar to the jury decision-making model, in which a lone holdout (minimum) can yield a hung jury and a mistrial (Davis, 1992), or to a mountain climbing team whose performance is determined by the slowest and

weakest member of the team (e.g., Krakauer, 1997). Therefore, one individual can effectively determine the group-level outcome because the combination rule is nonlinear.

Variance Form of Emergence

Unlike the other exemplars, which focus on representative values to capture the emergent characteristic of the collective, this form of emergence represents the phenomenon as variability within the group. Conceptually, this form of emergence is related to heterogeneity (Klein et al., 1994), parts (Dansereau & Yammarino, Chapter Ten, this volume), and uniform dispersion (Brown & Kozlowski, 1997; Brown et al., 1996; Chan, 1998). The elemental contribution may be similar in type and amount (for example, norm crystallization) or different in type and amount (for example, demographic diversity). Therefore, individuals may make contributions that are similar or different, but the substantive focus is on the variance of contribution (Roberts et al., 1978). It is important to emphasize that this one form captures different types of emergence that may range from low dispersion to high dispersion.

For example, one form of creativity can be characterized by the diversity, or lack thereof, of the knowledge or perspectives that are brought to bear on a problem (Wiersema & Bantel, 1992). Demographic diversity captures the extent to which individual members of a unit differ in their demographic characteristics (Tsui, Egan, & O'Reilly, 1992; Jackson et al., 1995). Homogeneity of charisma (that is, the extent to which a leader has equally charismatic relationships with all of his or her subordinates; see Klein & House, 1995), norm crystallization (Jackson, 1975), and culture strength (Koene, Boone, & Soeters, 1997) are based on variability within a collective. Homogeneity, crystallization, and strength are predicated on low variance, whereas the absence of homogeneity, crystallization, and strength is indicated by high variance. Klein and colleagues (Chapter Six, this volume) explore the antecedents and consequences of variability in organizational boundary spanners' trust in and commitment to their organization's interorganizational partner. Variance, of course, is a key operationalization of variability. Variance can capture emergence that differs across groups, contexts, and time. Therefore, it represents a shift in conceptual focus, from the content of the phenomenon to the nature of emergence itself.

Patterned Emergence

This model is based on the widest variability in the type and amount of elemental contribution, and in the patterns by which those differences combine to represent emergent phenomena. This model incorporates the assumption that emergence may manifest itself as different forms, and it views nonuniform patterns of dispersion as meaningful substantive phenomena.

The variance form of emergence is based on uniform distributions of within-group dispersion, whereas the patterned or configural form is based on nonuniform distributions of within-group dispersion. The term *uniformity* refers to the pattern of the distribution. A uniform distribution is single-modal, indicating strong or weak *agreement*. A nonuniform distribution is highly skewed or multimodal, indicating strong or weak *disagreement* (that is, the formation of subgroup clusters). Indeed, this form is generally indicated by within-unit variance that exceeds what would be expected from purely random responding. Therefore, very high variance within a group may be indicative of polarized factions, or "faultlines," Lau and Murhighan's (1998) metaphor for the divisions that may erupt and split a group. In this sense, disagreement goes beyond lack of agreement; it is indicative of conflict or of opposing perspectives within the collective unit. It is in this respect that dispersion theory uses nonuniform patterns of subgroup bifurcation to capture such complex phenomena as conflict, polarization, competition, and coalition formation (Brown & Kozlowski, 1997, 1999).

In addition to patterns of subgroup bifurcation, this form of emergence includes configurations that attempt to capture networks of linkages. Consider, for example, the model of team compilation proposed by Kozlowski and colleagues (1999). The model specifies different types, amounts, and linking mechanisms to characterize performance contributions at the individual, dyad, and team levels. Adaptive team performance is represented as a configuration of compatible knowledge and actions across team members at different levels of analysis. Or consider notions of team mental models and transactive memory. Early notions of the team mental model concept assumed that all team members shared the same knowledge (e.g., Cannon-Bowers, Salas, & Converse, 1993; Klimoski & Mohammed, 1995). Therefore, early versions of this construct assumed isomorphic composition. As this concept has evolved in the literature, it has been reconceptualized as entailing different *compatible* knowledge

(Kozlowski, Gully, Salas, et al., 1996)—different knowledge across individuals that forms a congruent whole.

Similarly, Wegner (1995) proposes that individual group members may each have unique information essential to performing the group task. It is not necessary for individuals to share the same knowledge (that is, isomorphic assumptions); rather, one or more individuals simply need to know who possesses the unique information. The essential information can then be accessed, as necessary. In this model, group memory is a complex configuration of individual memory, distributed knowledge of the contents of individual memory, and the interaction process that links that information into an emergent whole.

Implications

We introduced this third and last section of the chapter with three intentions: to be inclusive and expansive in our consideration of alternative forms of emergence, to focus on building a theoretical foundation for different forms of emergence, and to use typology as a vehicle for explicating and elaborating on the theoretical underpinnings of emergence. We hope that we have, in some measure, accomplished these goals. We believe, as we shall describe, that our framework is largely consistent with other efforts to explore emergence. We also believe that our particular attention to the underlying processes and dynamics that shape different forms of emergence can enhance understanding of the moderator effects and boundary conditions affecting emergence. An appreciation of the influence of these processes will lead to more precise specification of the theory addressing emergent phenomena. We see our effort as a point of departure for guiding and pushing further theoretical elaboration.

It is interesting to us that when our effort was originally conceived, we viewed our focus on different forms of emergence, and on the processes that shape those forms, as novel. However, a number of other researchers, contemporaneous with the development of this chapter, have also started to explore emergence (Brown & Kozlowski, 1997, 1999; Brown et al., 1996; Chan, 1998; Kozlowski, 1998, 1999; Morgeson & Hofmann, 1999a, 1999b). Although this chapter is not intended as an integration of these efforts, we believe that our framework helps to make explicit the compatibilities

across these apparently disparate efforts to explore emergence. For example, Brown and Kozlowski (1997, 1999) posit *dispersion theory*, which focuses on patterns of within-group variability or the dispersion of phenomena, as opposed to the more common focus on means or convergent points. In dispersion theory, uniform patterns that evidence low dispersion are consistent with composition processes, whereas subgroup bifurcation that creates nonuniform patterns of dispersion are consistent with compilation processes. Similarly, Morgeson and Hofmann (1999a, 1999b) have made a strong case for distinguishing construct *structure* and *function*. Structural and functional identity across levels is consistent with composition processes, and functional but not structural identity across levels is consistent with compilation processes.

Using examples from the literature, Chan (1998) has developed a typology to distinguish different types of "composition" or data-aggregation models. The typology includes additive models (e.g., Glick, 1985), direct-consensus models (e.g., James et al., 1984, 1993; Kozlowski & Hattrup, 1992), referent-shift-consensus models (e.g., James, 1982; Klein et al., 1994; Rousseau, 1985), dispersion models (e.g., Brown et al., 1996; Brown & Kozlowski, 1997), and process models (e.g., Kozlowski et al., 1994, 1999; Kozlowski, Gully, McHugh, et al., 1996). The direct-consensus, additive, and referent-shift-consensus models are consistent with composition processes, whereas the dispersion and process models are consistent with compilation processes.[8] Finally, our typology is also consistent with Steiner's (1972) typology of group performance. In many ways, Steiner's work is a precursor of all such typologies because it captures many of the basic combination rules that determine how individual characteristics, cognition, affect, and behavior can aggregate to represent higher-level, collective phenomena. We believe, as just discussed, that our framework is largely consistent with these other efforts. We also believe that our particular attention to the underlying processes and dynamics that shape different emergent forms enhances understanding of the moderator effects and boundary conditions affecting emergence. An appreciation of the influence of these processes will lead to more precise specification of theory addressing emergent phenomena.

We would be remiss if we did not note that there are also apparent inconsistencies between the contemporary treatments of

emergence (just noted) and other treatments with a tradition in the literature. We see the treatments as compatible yet different efforts to understand the same general class of phenomena. For example, the varient paradigm (Dansereau et al., 1984) treats emergence as a relationship between variables that exists at a higher, collective level but that does not hold between similar variables at a lower level. Thus, for example, a relationship between two variables is said to emerge at the group level of analysis if the two variables are significantly related (both statistically and practically) at the group level of analysis but the relationship between the two variables is not significant at the individual level of analysis. The varient perspective on emergence and our perspective are related but distinct. Dansereau and his colleagues focus on the emergence of relationships between variables at higher unit levels and on the statistical detection of such relationships. In contrast, we have focused primarily on the emergence of higher-level constructs, endeavoring to show the variety of ways in which a higher-level construct may emerge from lower-level entities and interaction processes. Measurement and analysis are important but separable issues. Ultimately, specific theories that assume particular emergent forms will need to be tested empirically. The varient paradigm, other analytic approaches, and even new techniques will be useful in this process.

We believe that the theoretical issues surrounding emergence that we have explored here are critical to the development of our science. How individual cognition, affect, behavior, and other characteristics emerge to make contributions to group and organizational outcomes is largely an uncharted frontier. How theories, interventions, and tools from the fields of industrial/organizational (I/O) psychology and organizational behavior (OB) can enhance these contributions is largely an unanswered question. Like most researchers and practitioners in the field, we believe that I/O-OB theories and techniques make contributions to organizational effectiveness, but we cannot really substantiate that belief (Rousseau, Chapter Fourteen, this volume). The chapters in this volume that deal with theory begin to explore this missing link. The chapter on training effectiveness (Kozlowski et al., Chapter Four), in particular, uses the distinction between composition and compilation to draw implications for how training can influence higher-level outcomes. We are beginning to probe a critical issue, but there is much more to do.

We make no claim that our framework is all-encompassing and complete; it is a work in progress. Although our focus has been primarily conceptual, the alternative forms of emergence have implications for measurement and analysis. We have endeavored to address measurement and data representation where possible, but we readily admit that the more complex compilation forms of emergence do not have well-developed measurement methods and analytic models. We hope that our pushing theorists to consider more complex phenomena will lead to new developments in methods and analytic systems. We hope the theoretical framework and typology presented here will stimulate further efforts to expand the conceptualization of emergent phenomena in organizations.

Conclusion

As the next millennium approaches, we are poised to witness a renaissance in organizational theory and research. There is increasing recognition that the confines of single-level models—a legacy of primary disciplines that undergird organizational science—need to be broken. A meaningful understanding of the phenomena that comprise organizational behavior necessitates approaches that are more integrative, that cut across multiple levels, and that seek to understand phenomena from a combination of perspectives. There is a solid theoretical foundation for a broadly applicable levels perspective, for an expanding, empirically based research literature, and for progress toward the development of new and more powerful analytic tools. A levels perspective offers a paradigm that is distinctly organizational.

Our purposes in this chapter have been to review the conceptual foundations of the levels perspective in organizations, to synthesize principles for guiding theory development and research, and to elaborate neglected models of emergent phenomena. Our goal is to convince researchers that levels issues should be considered in the study of a broad range of phenomena that occur in organizations. We hope that this chapter will, in a small way, push researchers to use established frameworks and to explore new alternatives in their work.

The remaining chapters in this book apply a levels perspective to substantive topics, consider analytic methods, and reflect on the implications of the levels perspective for organizational science.

Several of the substantive topics were selected primarily because typical treatments of these topics in the industrial and organizational literature rarely consider the implications of levels, and yet levels issues are central. When the implications of a multilevel theory are considered, new and unexplored issues are surfaced. Prime examples of such topics include selection (Schneider et al., Chapter Two, this volume), performance appraisal (DeNisi, Chapter Three, this volume), training effectiveness (Kozlowski et al., Chapter Four, this volume), and human resource management (Ostroff & Bowen, Chapter Five, this volume).

Other topics were selected because they naturally embody a levels perspective, but a perspective that forces us to think beyond our current frameworks. Prime examples include cross-cultural (Chao, Chapter Seven, this volume) and interorganizational linkages (Klein et al., Chapter Six, this volume). Both chapters focus on the implications of individuals being representatives of the higher level collectivities to which they belong.

Next, there are chapters addressing each of the primary multilevel analytic methods and issues, including within-group agreement, non-independence, and reliability (Bliese, Chapter Eight, this volume); the cross-level operator and contextual analysis (James & Williams, Chapter Nine, this volume), within-and-between analysis (Dansereau & Yammarino, Chapter Ten, this volume); and hierarchical linear modeling (Hofmann et al., Chapter Eleven, this volume). In addition, we have endeavored to cut through to the heart of the assumptions, differences, and appropriate applications of these multilevel analytic techniques with a collaborative effort that combines our disparate knowledge and perspectives (Klein, Bliese et al., Chapter Twelve, this volume).

Finally, we close the book with reflective comments pertaining to the importance of the levels perspective to the deep historical roots of our science, and to the increasing centrality of levels theory in mainstream organizational theory and research (Brass, Chapter Thirteen, this volume). The multilevel perspective provides a means for us to unify our science, and creates a foundation for enhancing policy impact for the disciplines that study organizations (Rousseau, Chapter Fourteen, this volume). The authors of all these chapters have provided a wealth of ideas and actionable knowledge. We hope that these ideas, and this book, stimulates those, who like us, seek a more unified and impactful science of organizations.

Notes

1. Throughout this chapter, we use the term *multilevel* in a generic sense, to reference all types of models that entail more than one level of conceptualization and analysis. Therefore, our use of the term *multilevel* references composition and compilation forms of bottom-up emergence, cross-level models that address top-down contextual effects, and homologous multilevel models that address parallel constructs and processes occurring at multiple levels.
2. Any effort to briefly characterize the many and myriad contributions to multilevel theory in organizations is doomed from the outset to be incomplete. We recognize that there are other lines of theory and research that have contributed to multilevel theory; many are mentioned throughout this chapter. We have chosen, however, to focus on a very early, sustained, and reasonably coherent effort that spanned many decades and many contributors. Our apologies to all others.
3. We recognize that there are alternative perspectives on organizational culture that view it as a collective construct, one that cannot be decomposed to the individual level. However, research on organizational culture has become increasingly consistent with an emergent perspective (Denison, 1996).
4. Insofar as global, shared, and configural unit properties each describe a unit as a whole, they are "homogeneous constructs," as Klein and colleagues (1994) use the term; here, we elaborate on their typology, illuminating the variety of forms that homogeneous unit-level constructs may take.
5. Unit-level constructs may of course be compositional, as in situations where group members share identical values or the same attitudes, but we expect some characteristics, such as abilities and personality, to be more likely configural than shared.
6. We acknowledge that the conceptualization of phenomena *may* entail a universal form; for example, unit climate is often conceptualized as a unit property when it is shared and as an individual property when it is not (James, 1982).
7. Our definition of discontinuous phenomena is consistent with House and colleagues (1995). Note also that these authors propose three models of *relational* discontinuity, involving (a) magnitude, (b) relational patterns, and (c) behavior-outcome relations. We would characterize these models as top-down contextual models, not bottom-up emergent processes. These three models illustrate (a) cross-level direct effects,

(b) cross-level frog-pond relations, and (c) cross-level moderation, respectively. Our typology focuses on *discontinuity in emergence.*

8. We should clarify that Chan (1998) indicates that his additive, direct consensus, referent-shift consensus, and dispersion models are static, whereas the process model in his typology is more directly interested in the dynamics of emergence. We would argue that emergent process dynamics are relevant to *all* the categories in that such processes shape the emergent form and, therefore, should be an explicit part of the conceptualization.

References

Allport, F. H. (1954). The structuring of events: Outline of a general theory with applications to psychology. *Psychological Review, 61,* 281–303.

Ancona, D., & Chong, C. (1997). Entrainment: Pace, cycle, and rhythm in organizational behavior. In L. L. Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 18, pp. 251–284). Greenwich, CT: JAI Press.

Arthur, W. B. (1994). On the evolution of complexity. In G. Cowan, D. Pines, & D. Meltzer (Eds.), *Complexity: Metaphors, models, and reality* (pp. 65–77). Reading, MA: Addison-Wesley.

Ashby, W. R. (1952). *Design for a brain.* New York: Wiley.

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.

Boulding, K. E. (1956). General systems theory: The skeleton of science. *General Systems, 1,* 1–17.

Bourgeois, V. W., & Pinder, C. C. (1983). Contrasting philosophical perspectives in administrative science: A reply to Morgan. *Administrative Science Quarterly, 28,* 608–613.

Brass, D. J. (1981). Structural relationships, job characteristics, and worker satisfaction and performance. *Administrative Science Quarterly, 26,* 331–348.

Brass, D. J. (1985). Technology and the structuring of jobs: Employee satisfaction, performance, and influence. *Organizational Behavior and Human Decision Processes, 35,* 216–240.

Brass, D. J. (1995). A social network perspective on human resources management. In G. R. Ferris (Ed.), *Research in personnel and human resources management* (Vol. 13, pp. 39–79). Greenwich, CT: JAI Press.

Brass, D. J. (2000). Networks and frog ponds: Trends in multilevel research. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 557–571). San Francisco: Jossey-Bass.

Brown, K. G., & Kozlowski, S.W.J. (1997). Dispersion theory: A framework for emergent organizational phenomena. Unpublished paper, Department of Psychology, Michigan State University.

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 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.* Thousand Oaks, CA: Sage.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education, 8,* 158–233.

Campbell, D. T. (1955). The informant in qualitative research. *American Journal of Sociology, 60,* 339–342.

Campbell, D. T. (1958). Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral Science, 3,* 14–25.

Campion, M. A., Medsker, G. J., & Higgs, A. C. (1993). Relations between group characteristics and effectiveness: Implications for designing effective work groups. *Personnel Psychology, 46,* 823–850.

Cannon-Bowers, J. A., Salas, E., & Converse, S. A. (1993). Shared mental models in expert team decision-making. In N. J. Castellan, Jr. (Ed.), *Current issues in individual and group decision making* (pp. 221–246). Mahwah, NJ: Erlbaum.

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.

Chao, G. T. (2000). Multilevel issues and culture: An integrative view. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 308–347). San Francisco: Jossey-Bass.

Cowan, G., Pines, D. & Meltzer, D. (1994). *Complexity: Metaphors, models, and reality.* Reading, MA: Addison-Wesley.

Crutchfield, J. P. (1994). Is anything ever new? Considering emergence. In G. Cowan, D. Pines, & D. Meltzer (Eds.), *Complexity: Metaphors, models, and reality* (pp. 515–533). Reading, MA: Addison-Wesley.

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

Dansereau, F., & Dumas, M. (1977). Pratfalls and pitfalls in drawing inferences about leadership behavior in organizations. In J. G. Hunt & L. L. Larson (Eds.), *Leadership: The cutting edge* (pp. 68–83). Carbondale: Southern Illinois University Press.

Dansereau, F., & Yammarino, F. J. (2000). Within and Between Analysis: The varient 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.

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

Davis, J. H. (1992). Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples, 1950–1990. *Organizational Behavior and Human Decision Processes, 52,* 3–38.

DeNisi, A. S. (2000). Performance appraisal and performance management: A multilevel analysis. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 121–156). San Francisco: Jossey-Bass.

Denison, D. R. (1996). What IS the difference between organizational culture and organizational climate? A native's point of view on a decade of paradigm wars. *Academy of Management Review, 21,* 619–654.

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 fourteen annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Emery, F. E., & Trist, E. L. (1960). Socio-technical systems. In *Management science models and techniques* (Vol. 2). London: Pergamon.

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

Fleishman, E. A., & Zaccaro, S. J. (1992). Toward a taxonomy of team performance functions. In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 31–56). Norwood, NJ: Ablex.

Forehand, G. A., & Gilmer, B. H. (1964). Environmental variation in studies of organizational behavior. *Psychological Bulletin, 62,* 361–382.

Freeman, J. (1980). The unit problem in organizational research. In W. M. Evan (Ed.), *Frontiers in organization and management* (pp. 59–68). New York: Praeger.

Gell-Mann, M. (1994). Complex adaptive systems. In G. Cowan, D. Pines, & D. Meltzer (Eds.), *Complexity: Metaphors, models, and reality* (pp. 17–29). Reading, MA: Addison-Wesley.

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

George, J. M., & James, L. R. (1994). Levels issues in theory development. *Academy of Management Review, 19,* 636–640.

Gersick, C.J.G., & Hackman, J. R. (1990). Habitual routines in task-performing groups. *Organizational Behavior and Human Decision Processes, 47,* 65–97.

Glick, W. H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review, 10,* 601–616.

Glick, W. H. (1988). Response: Organizations are not central tendencies: Shadowboxing in the dark, round 2. *Academy of Management Journal, 13,* 133–137.

Glick, W. H., & Roberts, K. (1984). Hypothesized interdependence, assumed independence. *Academy of Management Review, 9,* 722–735.

Graen, G. (1976). Role making processes within complex organizations. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 1201–1245). Skolie, IL: Rand McNally.

Griffin, M. A. (1997). Interaction between individuals and situations: Using HLM procedures to estimate reciprocal relationships. *Journal of Management, 23,* 759–773.

Hannan, M. T. (1991). *Aggregation and disaggregation in the social sciences.* San Francisco: New Lexington Press.

Harkins, S. G., Latane, B., & Williams, K. (1980). Social loafing: Allocating effort or taking it easy. *Journal of Experimental Social Psychology, 16,* 457–465.

Herman, J. B., & Hulin, C. L. (1972). Studying organizational attitudes from individual and organizational frames of reference. *Organizational Behavior and Human Performance, 8,* 81–108.

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 behaviors and accidents. *Personnel Psychology, 49,* 301–339.

Homans, G. C. (1950). *The human group.* Orlando, FL: Harcourt Brace.

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 Press.

Huselid, M. A. (1995). The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of Management Journal, 38,* 635–672.

Indik, B. P. (1968). The scope of the problem and some suggestions toward a solution. In B. P. Indik & F. K. Berren (Eds.), *People, groups, and organizations* (pp. 3–30). New York: Teachers College Press.

Jackson, J. (1975). Normative power and conflict potential. *Sociological Methods and Research, 4,* 237–263.

Jackson, S. E., May, K. E., & Whitney, K. (1995). Understanding the dynamics of diversity in decision-making teams. In R. Guzzo, E. Salas, & Associates, *Team effectiveness and decision making in organizations* (pp. 204–261). San Francisco: Jossey-Bass.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67,* 219–229.

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., Demaree, R. G., & Wolf, G. (1993). r_{wg}: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78,* 306–309.

James, L. R., & Jones, A. P. (1974). Organizational climate: A review of theory and research. *Psychological Bulletin, 81,* 1096–1112.

James, L. R., & Jones, A. P. (1976). Organizational structure: A review of structural dimensions and their conceptual relationships with individual attitudes and behavior. *Organizational Behavior and Human Performance, 16,* 74–113.

James, L. R., & Williams, L. J. (2000). The cross-level operator in regression, ANCOVA, and contextual analysis. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 382–424). San Francisco: Jossey-Bass.

Jones, A. P., & James, L. R. (1979). Psychological climate: Dimensions and relationships of individual and aggregated work environment perceptions. *Organizational Behavior and Human Performance, 23,* 201–250.

Katz, D., & Kahn, R. L. (1966). *The social psychology of organizations.* New York: Wiley.

Kauffman, S. A. (1994). Whispers from Carnot: The origins of order and principles of adaptation in complex nonequilibrium systems. In

G. Cowan, D. Pines, & D. Meltzer (Eds.), *Complexity: Metaphors, models, and reality* (pp. 83–136). Reading, MA: Addison-Wesley.

Klein, K. J., Bliese, P. D., Kozlowski, S.W.J., Dansereau, F., Gavin, M. B., Griffin, M. A., Hofmann, D. A., James, L. R., Williams, L. J., 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* (pp. 512–553). San Francisco: Jossey-Bass.

Klein, K. J., Conn, A. L., Smith, D. B, & Sorra, J. S. (1998). Is everyone in agreement? Exploring the determinants of within-group agreement in survey responses. Unpublished manuscript.

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.

Klein, K. J., Dansereau, F., & Hall, R. J. (1995). On the level: Homogeneity, independence, heterogeneity, and interactions in organizational theory. *Academy of Management Review, 20,* 7–9.

Klein, K. J., & House, R. J. (1995). On fire: Charismatic leadership and levels of analysis. *Leadership Quarterly, 6,* 183–198.

Klein, K. J., Palmer, S. L., & Conn, A. B. (2000). Interorganizational relationships: A multilevel perspective. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 267–307). San Francisco: Jossey-Bass.

Klimoski, R., & Mohammed, S. (1995). Team mental model: Construct or metaphor? *Journal of Management, 20,* 403–437.

Koene, B. A., Boone, C.A.J.J., & Soeters, J. L. (1997). Organizational factors influencing homogeneity and heterogeneity of organizational cultures. In S. A. Sackmann (Ed.), *Cultural complexity in organizations: Inherent contrasts and contradictions* (pp. 273–293). Thousand Oaks, CA: Sage.

Kozlowski, S.W.J. (1998, March). Extending and elaborating models of emergent phenomena. Presentation at MESO Organization Studies Group, Arizona State University, Tempe.

Kozlowski, S.W.J. (1999, April). A typology of emergence: Theoretical mechanisms undergirding bottom-up phenomena in organizations. In F. P. Morgeson & D. A. Hofmann (Chairs), *New perspectives on higher-level phenomenon in industrial/organizational psychology.* Symposium conducted at the fourteenth annual conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

Kozlowski, S.W.J., Brown, K. G., Weissbein, D. A., Cannon-Bowers, J. A., & Salas, E. (2000). A multilevel approach to training effectiveness:

Enhancing horizontal and vertical transfer. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 157–210). San Francisco: Jossey-Bass.

Kozlowski, S.W.J., & Doherty, M. L. (1989). An integration of climate and leadership: Examination of a neglected issue. *Journal of Applied Psychology, 74,* 546–553.

Kozlowski, S.W.J., & Farr, J. L. (1988). An integrative model of updating and performance. *Human Performance, 1,* 5–29.

Kozlowski, S.W.J., Gully, S. M., McHugh, P. P., Salas, E., & Cannon-Bowers, J. A. (1996). A dynamic theory of leadership and team effectiveness: Developmental and task contingent leader roles. In G. R. Ferris (Ed.), *Research in personnel and human resource management* (Vol. 14, pp. 253–305). Greenwich, CT: JAI Press.

Kozlowski, S.W.J., Gully, S. M., Nason, E. R., Ford, J. K., Smith, E. M., Smith, M. R., & Futch, C. J. (1994, April). A composition theory of team development: Levels, content, process, and learning outcomes. In J. E. Mathieu (Chair), *Developmental views of team process and performance.* Symposium conducted at the ninth annual conference of the Society for Industrial and Organizational Psychology, Nashville, TN.

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.* San Francisco: Jossey-Bass.

Kozlowski, S.W.J., Gully, S. M., Salas, E., & Cannon-Bowers, J. A. (1996). Team leadership and development: Theory, principles, and guidelines for training leaders and teams. In M. Beyerlein, D. Johnson, & S. Beyerlein (Eds.), *Advances in interdisciplinary studies of work teams: Team leadership* (Vol. 3, pp. 251–289). Greenwich, CT: JAI Press.

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. (1987). An exploration of climates for technical updating and performance. *Personnel Psychology, 40,* 539–563.

Kozlowski, S.W.J., & Salas, E. (1997). An organizational systems approach for the implementation and transfer of training. In J. K. Ford, S.W.J. Kozlowski, K. Kraiger, E. Salas, & M. Teachout (Eds.), *Improving training effectiveness in work organizations* (pp. 247–287). Mahwah, NJ: Erlbaum.

Krakauer, J. (1997). *Into thin air.* New York: Villard.

Kuhn, A., & Beam, R. D. (1982). *The logic of organization.* San Francisco: Jossey-Bass.

Lau, D. C., & Murhighan, J. K. (1998). Demographic diversity and faultlines: The compositional dynamics of organizational groups. *Academy of Management Review, 23,* 325–340.

Levine, R. L., & Fitzgerald, H. E. (1992). *Analysis of dynamic psychological systems* (Vols. 1, 2). New York: Plenum.

Lewin, K. (1951). *Field theory in the social sciences.* New York: HarperCollins.

Lewin, K., Lippitt, R., & White, R. K. (1939). Patterns of aggressive behavior in experimentally created "social climates." *Journal of Social Psychology, 10,* 271–299.

Likert, R. (1961). *The human organization: Its management and value.* New York: McGraw-Hill.

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

Louis, M. R. (1980). Surprise and sense-making: What newcomers experience in entering unfamiliar organizational settings. *Administrative Science Quarterly, 25,* 226–251.

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

McGrath, J. E. (1990). Time matters in groups. In J. Galegher, R. Krout, & C. C. Egido (Eds.), *Intellectual teamwork* (pp. 23–61). Mahwah, NJ: Erlbaum.

Meyer, A. D., Tsui, A. S., & Hinings, C. R. (1993). Guest co-editors' introduction: Configural approaches to organizational analysis. *Academy of Management Journal, 36,* 1175–1195.

Miller, J. G. (1978). *Living systems.* New York: McGraw-Hill.

Millikin, F. J., & Martins, L. L. (1996). Searching for common threads: Understanding the multiple effects of diversity in organizational groups. *Academy of Management Review, 21,* 402–433.

Moreland, R. L., & Levine, J. M. (1992). The composition of small groups. In E. E. Lawler III., B. Markovsky, C. Ridgeway, & H. Walker (Eds.), *Advances in group processes* (Vol. 9, pp. 237–280). Greenwich, CT: JAI Press.

Morgan, G. (1983). More on metaphor: Why we cannot control tropes in administrative science. *Administrative Science Quarterly, 28,* 601–607.

Morgeson, F. P., & Hofmann, D. A. (1999a, April). The structure and function of collective constructs. In F. P. Morgeson & D. A. Hofmann, (Chairs), *New perspectives on higher-level phenomena in industrial/organizational psychology.* Symposium conducted at the Fourteenth

Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

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

Mossholder, K. W., & Bedeian, A. G. (1983). Cross-level inference and organizational research: Perspectives on interpretation and application. *Academy of Management Review, 8,* 547–558.

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

Nicolis, G., & Prigogine, I. (1989). *Exploring complexity.* New York: Freeman.

Nonaka, I. (1994). A dynamic theory of knowledge creation. *Organizational Science, 5,* 14–37.

Oldham, G. R., & Hackman, J. R. (1981). Relationships between organizational structure and employee reactions: Comparing alternative frameworks. *Administrative Science Quarterly, 26,* 66–81.

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

Ostroff, C., & Bowen, D. E. (2000). Moving HR to a higher level: HR practices and organizational effectiveness. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 211–266). San Francisco: Jossey-Bass.

Parsons, T. (1956). Suggestions for a sociological approach to the theory of organizations, I and II. *Administrative Science Quarterly, 1,* 225–239.

Parsons, T. (1960). *Structure and process in modern societies.* New York: Free Press.

Pinder, C. C., & Bourgeois, V. W. (1982). Controlling tropes in administrative science. *Administrative Science Quarterly, 27,* 641–652.

Prigogine, I., & Stengers, I. (1984). *Order out of chaos: Man's new dialogue with nature.* New York: Bantam.

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

Roberts, K. H., Hulin, C. L., & Rousseau, D. M. (1978). *Developing an interdisciplinary science of organizations.* San Francisco: Jossey-Bass.

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

Roethlisberger, F. J., & Dickson, W. J. (1939). Human relations. In *Management and the worker.* Cambridge, MA: Harvard University Press.

Rouiller, J. Z., & Goldstein, I. L. (1993). The relationship between organizational transfer climate and positive transfer of training. *Human Resource Development Quarterly, 4,* 377–390.

Rousseau, D. M. (1978a). Characteristics of departments, positions, and individuals: Contexts for attitudes and behavior. *Administrative Science Quarterly, 23,* 521–540.

Rousseau, D. M. (1978b). Measures of technology as predictors of employee attitude. *Journal of Applied Psychology, 63,* 213–218.

Rousseau, D. M. (1985). Issues of level in organizational research: Multilevel and cross-level perspectives. In L. L. Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 7, pp. 1–37). Greenwich, CT: JAI Press.

Rousseau, D. M. (1988). The construction of climate in organizational research. In C. L. Cooper & I. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 139–158). New York: Wiley.

Rousseau, D. M. (2000). Multilevel competencies and missing linkages. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 572–582). San Francisco: Jossey-Bass.

Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology, 74,* 368–370.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). The impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64,* 609–626.

Schneider, B. (1981). Work climates: An interactionist perspective. In N. Feimer & E. Geller (Eds.), *Environmental psychology: Directions and perspectives.* New York: Praeger.

Schneider, B., & Bowen, D. E. (1985). Employee and customer perceptions of service in banks: Replication and extension. *Journal of Applied Psychology, 70,* 423–433.

Schneider, B., & Reichers, A. E. (1983). On the etiology of climates. *Personnel Psychology, 36,* 19–39.

Schneider, B., Smith, D. B., & Sipe, W. P. Personnel selection psychology: Multilevel considerations. In K. J. Klein & S.W.J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 91–120). San Francisco: Jossey-Bass.

Schriesheim, C. A. (1995). Multivariate and moderated within- and between-entity analysis (WABA) using hierarchical linear multiple regression. *Leadership Quarterly, 6,* 1–18.

Seidler, J. (1974). On using informants: A technique for collecting quantitative data and controlling measurement error in organizational analysis. *American Sociological Review, 39,* 816–831.

Simon, H. A. (1969). *The sciences of the artificial.* Cambridge, MA: MIT Press.

Simon, H. A. (1973). The organization of complex systems. In H. H. Pattee (Ed.), *Hierarchy theory* (pp. 1–27). New York: Braziller.

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

Steiner, I. D. (1972). *Group process and productivity.* Orlando, FL: Academic Press.

Terborg, J. R. (1981). Interactional psychology and research on behavior in organizations. *Academy of Management Review, 6,* 569–576.

Thompson, J. (1967). *Organizations in action.* New York: McGraw-Hill.

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.

Tracy, L. (1989). *The living organization.* New York: Praeger.

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

Tziner, A., & Eden, D. (1985). Effects of crew composition on crew performance: Does the whole equal the sum of its parts? *Journal of Applied Psychology, 70,* 85–93.

von Bertalanffy, L. (1968). *General systems theory.* New York: Braziller.

von Bertalanffy, L. (1972). The history and status of general systems theory. In G. J. Klir (Ed.), *Trends in general systems theory* (pp. 21–41). New York: Wiley.

Wegner, D. M. (1995). A computer network model of transactive memory. *Social Cognition, 13,* 319–339.

Weick, K. E. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly, 21,* 1–19.

Wiersema, M. F., & Bantel, K. A. (1992). Top management team demography and corporate strategic change. *Academy of Management Journal, 35,* 91–121.

Wittenbaum, G. M., & Stasser, G. (1996). Management of information in small groups. In J. L. Nye & A. M. Brower (Eds.), *What's social about social cognition: Research on socially shared cognition in small groups* (pp. 3–28). Thousand Oaks, CA: Sage.

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.

CHAPTER 2

Personnel Selection Psychology

Multilevel Considerations

Benjamin Schneider
D. Brent Smith
William P. Sipe

History shows that great . . . forces flow like a tide over communities only half-conscious of that which is befalling them. Wise statesmen foresee what time is thus bringing, and try to shape institutions and model men's thoughts and purposes in accordance with the change that is silently coming on.
JOHN STUART MILL

Industrial and organizational (I/O) psychology in general, and the subfield of personnel selection in particular, are experiencing a paradigm shift of which they may not be fully aware. The paradigm shift is from a focus on what Nord and Fox (1997) called the "essentialist individual" model of behavior to a newer focus on the organizational implications of personnel selection practices. From

Note: This chapter was prepared with the financial assistance of the Army Research Institute for the Behavioral and Social Sciences. The authors bear total responsibility for the content of the chapter; nothing in the chapter should be construed to represent positions of the U.S. Army and/or the Department of Defense.

Population Association of America, New Orleans.
Ryder, N. B. and C. F. Westoff
1971 Reproduction in the United States 1965. Princeton: Princeton University Press.
Sastry, K. R.
1975 "Female work participation and fertility." Journal of Population Research 2:16–32.
Simon, J. L.
1974 The Effects of Income on Fertility. Monograph 19. Carolina Population Center, University of North Carolina, Chapel Hill.
Stolzenberg, R. and L. Waite
1977 "Age, fertility expectations and employment plans." American Sociological Review 42:769–83.
Stycos, J. M. and R. H. Weller
1967 "Female working roles and fertility." Demography 4:210–7.
Sweet, J.
1973 Women in the Labor Force. New York: Seminar Press.
Terry, G.
1974 "A theoretical examination of the relationship between fertility and female employment." Paper presented at the Population Association of America, New York.
Tittle, C. R. and R. J. Hill
1967 "Attitude measurement and the prediction of behavior: an evaluation of conditions and measurement techniques." Sociometry 30:199–213.
Turchi, B.
1975 The Demand for Children: The Economics of Fertility in the U.S. Cambridge, Ma.: Ballinger.
Waite, L. J.
1976 "Working wives: 1940–1960." American Sociological Review 41:65–80.
Waite, L. J. and R. M. Stolzenberg
1976 "Intended childbearing and labor force participation of young women: insights from nonrecursive models." American Sociological Review 41:235–51.
Westoff, C. F., E. G. Mishler, and E. L. Kelly
1957 "Preferences in size of family and eventual fertility twenty years after." American Journal of Sociology 62:491–7.

A RULE FOR INFERRING INDIVIDUAL-LEVEL RELATIONSHIPS FROM AGGREGATE DATA*

GLENN FIREBAUGH
Vanderbilt University

American Sociological Review 1978, Vol. 43 (August):557–572

Under certain conditions aggregate-level data provide unbiased estimates of individual-level relationships. Here I present these conditions in the form of a single theoretical decision rule: bias is absent when, and only when, the group mean of the independent variable (X) has no effect on Y, with X controlled. This paper introduces this rule, demonstrates it for the general n-variable case, compares it with prior discussions of cross-level inference, and illustrates it with the 1930 census data used by Robinson (1950). The final section discusses the implications of this rule for the converse type of cross-level inference: the use of individual-level data to estimate aggregate-level relationships.

Almost three decades ago Robinson (1950) warned sociologists of the dangers of using aggregate data to study individuals. In his seminal paper, Robinson showed that correlations between variables at the aggregate level differ from correlations between the same variables at the individual level. From this finding Robinson concluded that researchers should not use aggregate data to study individuals;[1] those who did were said to be guilty of the ecological fallacy (Selvin, 1958).

However, as Hammond (1973:765) noted, sociologists (and other social scientists) have continued to use aggregate data

* This rule was first presented, in a rudimentary form, in a fall 1974 methods seminar taught by Karl Schuessler. In refining and extending the ideas of that paper, I received helpful counsel from Leigh Burstein, John Cardascia, Lee Cronbach, Michael Hannan, and Elton Jackson, as well as from Professor Schuessler. Carol Meyer expertly typed and assembled the final draft. Of course, none of the above bears any onus for shortcomings in the final product.

[1] The terms "individual" and "aggregate" refer to units of analysis; an individual need not be a person.

to make inferences about individuals, because appropriate individual-level data are often unavailable. In criminology, for example, deterrence theory suggests that the individual (not the aggregate) is deterred, yet empirical investigations generally have used aggregate data (for example, Tittle, 1969; Chiricos and Waldo, 1970; Logan, 1972; Ehrlich, 1973; 1975). Inference across levels of aggregation (hereafter called cross-level inference)[2] also can be illustrated by studies of voting behavior, where the use of areal data for studying hypotheses about individuals has a venerable history (see, for example, Burnham, 1965; 1971; and the critique by Cowart, 1974). Other examples—from history, political science, and economics, as well as from sociology—could be given of the substitution of aggregate-level data for unavailable individual-level data.

Given the unavailability of individual-level data for many areas of interest to social scientists, one is not surprised to find that most methodological discussions since Robinson have sought to modify the strict prohibition against downward cross-level inference. The most important conclusion of these discussions has been that aggregate data do not always yield biased estimates of individual-level unstandardized *regression* coefficients (Goodman, 1953; 1959). This conclusion has two implications: first, there are certain cases where, except for possible loss of efficiency (i.e., the variance of $b_{\bar{Y}\bar{X}}$ is usually greater than the variance of b_{YX}; see Hannan and Burstein [1974:382]), downward cross-level inference can be made with impunity; second, if downward cross-level inference is made, regression coefficients should be used instead of correlation coefficients.

After the demonstration that an aggregate-level regression coefficient need not differ from its individual-level counterpart, most studies sought to determine the *conditions* under which the regression coefficients do not differ—i.e., the conditions under which cross-level bias is absent. Two approaches are discernible. The first approach could be called the contextual effects approach, since it sees contextual effects as the major source of bias (Hammond, 1973; Przeworski, 1974). The second approach is the structural equations approach or causal models approach (Blalock, 1964: 97–114; Hannan, 1971a; 1971b; Hannan and Burstein, 1974; Burstein, 1974; 1975a; 1975b; 1976; Hannan and Young, 1976); this approach formulates bias in terms of path models and uses econometric techniques to determine the expected value of the parameters.[3]

In this paper I combine the two approaches by employing contextual effects models in a structural equation framework. Two papers in particular stimulated this project: Hammond (1973) and Hannan and Burstein (1974). Hammond's paper is important in that it suggests the link between contextual effects theory and cross-level inference. Hannan and Burstein's paper is important in that it provides a summary of the issues involved in cross-level inference as well as a compact statement of the logic and conclusions of the structural equations approach. Briefly stated, Hannan and Burstein counsel researchers faced with the question of cross-level inference to consider the effects of the variable by which the data are grouped (schools, states, etc.). They show that, in the bivariate case, aggregate data give unbiased estimates of the individual-level relationships when any of the following is true: (1) the grouping variable (A) is uncorrelated with the dependent variable (Y) with the independent variable (X) controlled; (2) A and

[2] Cross-level inference can be either *downward* (the ecological fallacy) or *upward* (the individualistic fallacy; Alker, 1969: the use of individual-level data to make inferences about aggregate-level effects). Except where otherwise noted, by cross-level inference I mean downward cross-level inference.

[3] This literature summary refers mainly to the sociological literature. The econometrics literature most often assumes that the researcher can choose the method of grouping, and discusses the relative efficiencies of different grouping methods (for example, Johnston, 1972: Chap. 7). Recent discussions in political science (Hanushek et al., 1974; Irwin and Lichtman, 1976) have contended that cross-level bias arises from specification error. While the rule introduced in this paper is consistent with this contention, it identifies the source of bias much more specifically than does the generic term "specification error," and thus I judge it to be more useful to the sociologist.

X are uncorrelated; or (3) the variance of X equals the variance of $\bar{X}$, where $\bar{X}$ is the group mean of X. Hannan and Young (1976) confirm these findings in a Monte Carlo simulation containing two regressors. Burstein (1975a; 1975b; 1976) applies these findings to other empirical examples, and compares the results to those obtained using an approach suggested by Feige and Watts (1972).

The utility of combining the contextual effects and structural equations approaches is shown in this paper. First, the approach adopted here generates a parsimonious rule—the $\bar{X}$-rule—for making inferences about individual-level relationships from aggregate data. The $\bar{X}$-rule links cross-level bias to theory on group effects. Since group effects theory is well-known in sociology (e.g., Blau, 1960), this rule often provides theoretical leverage to the sociologist who must determine whether cross-level inference is legitimate in a particular case. Second, this rule is easily generalizable analytically to the n-variable case. This is important since, as Hannan and Young (1976:2) noted, there are formidable obstacles to analytical investigations of the effects of grouping in regression models containing two or more regressors. Finally, unlike most previous approaches, this approach focuses on the difference between the aggregate-level coefficient and the individual-level coefficient of interest (explained in Section III, below).

Like most previous studies, this paper focuses on the conditions for avoiding bias[4] where the variables of interest are interval scale (for the nominal scale case, see Duncan and Davis, 1953; Shively, 1969; Iversen, 1973). Section I introduces the $\bar{X}$-rule; first the bivariate case and then the multivariate case are examined. Section II discusses the theoretical interpretation of the $\bar{X}$-rule, and its implications for research. Section III compares the approach of this paper with previous approaches to cross-level inference. Section IV provides an empirical illustration. Finally, Section V discusses the implications of this paper for analyses using only individual-level data, as well as summarizing its implications for analyses using only aggregate-level data.

I. THE $\bar{X}$-RULE: BIVARIATE CASE AND MULTIVARIATE CASE

Introduction of Basic Ideas

We begin with the simplest case: the relationship (slope) between a dependent variable (Y) and a single independent variable (X) in a population. If data on X and Y are available for all individuals in the population, the unstandardized regression coefficient, β_{YX}, is obtained when Y is regressed on X (here, and throughout this paper, the greek letter "β" is used to refer to population parameters; it does *not* refer to standardized regression coefficients). If, on the other hand, all the individuals in the population are placed into mutually exclusive groups (precincts, for example), and an average (usually the mean) for X and Y is computed for each group, the regression of the dependent variable means on the independent variable means yields $\beta_{\bar{Y}\bar{X}}$.

That $\beta_{\bar{Y}\bar{X}}$ is not necessarily equivalent to β_{YX} is well-known to sociologists; the literature is replete with allusions to the danger of inferring individual-level relationships from aggregate-level relationships. *Why* $\beta_{\bar{Y}\bar{X}}$ may give a biased estimate of β_{YX}, however, is not as well-known. Indeed, a discrepancy between $\beta_{\bar{Y}\bar{X}}$ and β_{YX} may seem counterintuitive, since (1) this discrepancy is not due to sampling error ($\beta_{\bar{Y}\bar{X}}$ and β_{YX} are both population parameters),[5] and (2) the variables are based on data from the same source (i.e., $\bar{X}$ is computed from data on X, and $\bar{Y}$ is computed from data on Y).

[4] The question of efficiency is beyond the scope of this paper. The reader should not assume, however, that efficiency is unimportant; biased but efficient estimators often are preferable to unbiased but inefficient estimators. This paper specifies the conditions for unbiased estimation; future papers will want to attend to the issue of efficient estimation.

[5] As Cronbach (1976:1.9) noted, statistics texts sometimes give the misimpression that aggregation problems involve the issue of inference from sample to population. This is a dangerous misimpression; the reader should clearly distinguish cross-level bias, which involves discrepancies between *population* parameters, from biases which involve discrepancies between sample statistics and population parameters (see also Duncan et al., 1961:62).

The demystification of cross-level bias begins with the recognition that an aggregate variable often measures a different construct than its namesake at the individual level. Often the aggregate-level variable taps more constructs than the individual-level variable. College education is one example (Cronbach, 1976:1.11):

> That an individual is college-educated indicates a good deal about what he would be inclined to purchase or what jobs he would be capable of holding. The aggregate college education in the community not only describes an aggregate market and an aggregate employee pool, it says a good deal about what goods and services probably are well-supplied in the community (pediatricians? art movies? books? brokerage offices? etc.), and a good deal about the kinds of jobs offered.

Race is another example: percent black in a community indicates characteristics which are as relevant to the nonblack members as to the black members—their SES, location (urban vs. rural, South vs. non-South), etc. It is this shift in constructs, as one shifts levels of aggregation, which provides the basis for cross-level bias.

When X and $\bar{X}$ measure different constructs, bias is possible. Consider, for example, the finding that, in the 1968 presidential election, percent black was related positively (r = +.55) to the Wallace vote for the Congressional districts in the South (Schoenberger and Segal, 1971). In this case, $\bar{X}$ (percent black in Congressional district) no doubt measures extraracial characteristics of the Congressional district which affected the Wallace vote (proximity to Alabama, extensiveness of busing, etc.), thus giving rise to cross-level bias.

On the other hand, consider a hypothetical situation where *all* blacks voted for candidate C, and *all* nonblacks voted for some other candidate. In this case, β_{YX} = 1 and $\beta_{\bar{Y}\bar{X}}$ = 1 (where $\bar{Y}$ = percent of vote for candidate C, and $\bar{X}$ = percent black); hence, $\beta_{YX} = \beta_{\bar{Y}\bar{X}}$. Even though X and $\bar{X}$ measure different constructs—i.e., $\bar{X}$ measures extraracial district characteristics—no cross-level bias results, since the extraracial characteristics measured by $\bar{X}$ have no effect on Y, with X controlled ("Y, with X controlled," is hereafter written "Y·X").

We can now state a rule for making downward cross-level inference in the bivariate case:

> Cross-level bias is absent when, and only when, $\beta_2 = 0$ in the structural equation $Y = a + \beta_1 X_1 + \beta_2 \bar{X}_1 + e$ (see Figure 1).

This rule is useful to the researcher whether or not s/he can choose the method of grouping. When the researcher has no choice, the question is: is $\bar{X}$ unrelated to Y·X? If the researcher can choose between methods of grouping, the question becomes: is $\bar{X}$ unrelated to Y·X under any of the methods of grouping? (It is important to note that $\bar{X}$ may be related to Y·X under one method of grouping but unrelated to Y·X under another method.)

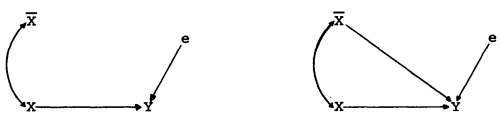
The theoretical interpretation of the rule will be spelled out below. First, however, I give a more rigorous statistical demonstration of the rule; readers who are interested only in its interpretation and application may wish to go directly to Section II.

Bivariate Case

Cross-level bias is the difference, in a population, between the aggregate-level regression coefficient obtained and the individual-level coefficient of interest. In the bivariate case, then, downward cross-level bias (δ) is formally defined as follows:

$$\delta = \beta_{\bar{Y}\bar{X}} - \beta_{yx}, \qquad (1)$$

where $\beta_{\bar{Y}\bar{X}}$ is the regression coefficient in the regression of $\bar{Y}$ on $\bar{X}$ (also called the



No cross-level bias. Cross-level bias.

Figure 1. Representation of the Rule for Cross-Level Inference: Bivariate Case.

between-group slope)[6] and β_{yx} is the regression coefficient in the regression of Y on X within groups (also called the common[7] within-group slope). (Some writers focus on β_{YX} instead of β_{yx}. However, as we will see in Section III, β_{YX} reflects aggregate-level, as well as individual-level effects.)

Next, consider the following structural model:

$$Y_{ij} = a + \beta_1 X_{ij} + \beta_2 \bar{X}_j + e \quad (2)$$
$$(i = 1,2 \ldots N; j = 1,2 \ldots M),$$

where X_{ij} refers to the score on X for the i^{th} person in the j^{th} group, and $\bar{X}_j$ is the group mean of the j^{th} group. This model states that Y is a linear function of X, $\bar{X}$, and a random disturbance, e, where X and $\bar{X}$ are causal variables (in the theoretical section which follows, we relax the assumption that $\bar{X}$ is causal and consider the case where the effect of $\bar{X}$ is spurious). We make the usual assumptions about e: that it has zero mean, constant variance, is uncorrelated with the independent variables, and that the values of e are mutually uncorrelated. Werts and Linn (1971; also Alwin, 1976) showed that the structural parameters β_1 and β_2 are:

$$\beta_1 = \beta_{yx}; \quad (3)$$
$$\beta_2 = \beta_{\bar{Y}\bar{X}} - \beta_{yx}. \quad (4)$$

Equations (3) and (4) state that, in the bivariate case (i.e., in the case of a single X), the individual-level effect of X on Y is β_{yx}, while the aggregate-level effect of X on Y is $\beta_{\bar{Y}\bar{X}} - \beta_{yx}$. But $\beta_{\bar{Y}\bar{X}} - \beta_{yx}$ also measures downward cross-level bias. Hence, in the bivariate case, downward cross-level bias is absent when, and only when, X has no aggregate-level effect on Y. Differently stated: the aggregate-level coefficient ($\beta_{\bar{Y}\bar{X}}$) provides an unbiased estimate of the individual-level effect of X on Y (β_{yx}) when, and only when, $\bar{X}$ has no effect on Y·X.

This result also can be derived by beginning with the analysis of covariance equation. (Covariance analysis is introduced here since it is easier to generalize to the multivariate case.) The standard covariance equation, with a single covariate, is as follows (see Burke and Schuessler, 1974:165):

$$Y_{ij} = \mu + A_j + \beta(X_{ij} - \bar{\bar{X}} + e_{ij} \quad (5)$$
$$(i = 1,2, \ldots N; j = 1,2, \ldots M),$$

where μ is common to all cases, A_j is common to all cases in the j^{th} group, X_{ij} is defined as before, $\bar{\bar{X}}$ is the grand mean of X, and e_{ij} is specific to the i^{th} individual in the j^{th} group. The least-squares solution for the normal equations derived from (5), subject to the constraint that $\sum_{j=1}^{M} n_j A_j = 0$, yields the following for the population parameters in (5):

$$\begin{aligned} \mu &= \bar{\bar{Y}} \text{ (the grand mean of Y)}; \\ \beta &= \beta_{yx}; \\ A_j &= \bar{Y}_j - \beta_{yx}(\bar{X}_j - \bar{\bar{X}}) - \bar{\bar{Y}}. \end{aligned} \quad (6)$$

Noting that $\bar{Y}_j = a_{\bar{Y}\bar{X}} + \beta_{\bar{Y}\bar{X}}\bar{X}_j + e_{\bar{Y}\bar{X}}$, and substituting (6) into (5), we obtain:

$$\begin{aligned} Y_{ij} &= \bar{\bar{Y}} + \{(a_{\bar{Y}\bar{X}} + \beta_{\bar{Y}\bar{X}}\bar{X}_j + e_{\bar{Y}\bar{X}}) \\ &\quad - \beta_{yx}(\bar{X}_j - \bar{\bar{X}}) - \bar{\bar{Y}}\} \\ &\quad + \beta_{yx}(X_{ij} - \bar{\bar{X}}) + e_{ij} \\ &= a_{\bar{Y}\bar{X}} + \beta_{yx}X_{ij} + (\beta_{\bar{Y}\bar{X}} \\ &\quad - \beta_{yx})\bar{X}_j + e. \end{aligned} \quad (7)$$

Hence, whether we begin with equation (2) or with the covariance equation (equation (5)), we conclude that the effect (slope) of $\bar{X}$ on Y·X is $\beta_{\bar{Y}\bar{X}} - \beta_{yx}$ (equation (7)).

To summarize: cross-level bias is possible because $\bar{X}$ and X may measure different constructs—an obvious point, but one which has usually been overlooked in the burgeoning literature on cross-level bias. When $\bar{X}$ and X measure different constructs, $\bar{X}$ may affect Y · X. In the bivariate case, an effect of X on Y·X

---

[6] The computation of $\beta_{\bar{Y}\bar{X}}$ involves weighting by the size of the group:

$$\beta_{\bar{Y}\bar{X}} = \frac{\sum_j n_j (\bar{X}_j - \bar{\bar{X}})(\bar{Y}_j - \bar{\bar{Y}})}{\sum_j n_j (\bar{X}_j - \bar{\bar{X}})^2},$$

where $\bar{X}_j$ is the group mean of the j^{th} group, $\bar{\bar{X}}$ is the grand mean of X, and n_j is the number of individuals in the j^{th} group.

[7] Here, and throughout this paper, effects are assumed to be linear and additive. In the case of nonadditive relationships, β_{yx} is an inappropriate statistic, since each group should be examined separately (Slatin, 1969).

results in cross-level bias. We will now see that the same principles hold in the multivariate case.

Multivariate Case

Cross-level bias is defined as the difference between the obtained aggregate-level regression coefficient(s) and the individual-level regression coefficient(s) of interest. In the bivariate case, examining cross-level bias involves only one comparison: the comparison of $\beta_{\bar{Y}\bar{X}}$ and β_{yx}. However, if there are n independent variables, n comparisons are involved. With two independent variables (X_1 and X_2), for example, $\beta_{\bar{Y}\bar{X}_1\cdot\bar{X}_2}$ is compared with $\beta_{yx_1\cdot x_2}$, and $\beta_{\bar{Y}\bar{X}_2\cdot X_1}$ is compared with $\beta_{yx_2\cdot x_1}$.

As in the bivariate case, cross-level bias is absent in the multivariate case only when $\bar{X}$-effects are absent. This condition is represented in Figure 2: $\bar{X}_1$ has no effect on $Y\cdot X_1, \ldots X_n, \bar{X}_2, \ldots \bar{X}_n$; $\bar{X}_2$ has no effect on $Y\cdot X_1, \ldots X_n, \bar{X}_1, \bar{X}_3, \bar{X}_n$; etc. The rule for cross-level inference in the multivariate case is as follows:

> Cross-level bias is absent when, and only when, $\bar{X}_1, \ldots \bar{X}_n$ have no independent effects in the structural equation $Y = a + \beta_1 X_1 + \ldots + \beta_n X_n + \beta_{n+1} \bar{X}_1 + \ldots + \beta_{2n} \bar{X}_n + e$.

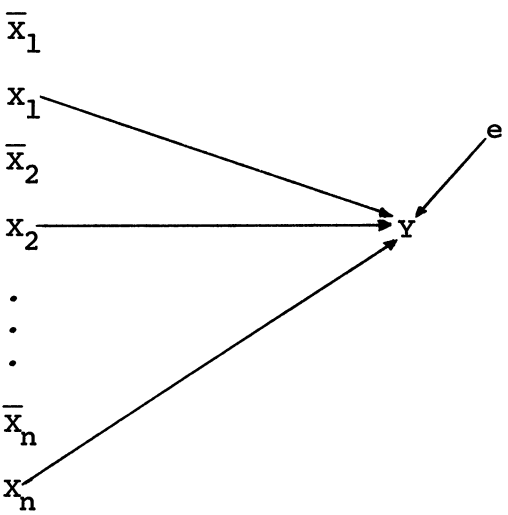


Figure 2. Representation of the Rule for Cross-Level Inference: Multivariate case. (Representation of Zero-Order Correlations between Exogenous Variables Omitted)

I now give a statistical demonstration of this rule.

To simplify the exposition, we first consider two independent variables (X_1 and X_2); the method then easily generalizes to n independent variables. We begin with the following structural model:

$$Y_{ij} = a + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 \bar{X}_{1j} + \beta_4 \bar{X}_{2j} + e, \quad (8)$$

where the variables are defined as before (note that an additional subscript is required to distinguish X_1 and X_2), and the assumptions about the error term are the same as in equation (2).[8] The structural parameters for equation (8) are:

$$\begin{aligned} \beta_1 &= \beta_{yx_1\cdot x_2}; \\ \beta_2 &= \beta_{yx_2\cdot x_1}; \\ \beta_3 &= \beta_{\bar{Y}\bar{X}_1\cdot\bar{X}_2} - \beta_{yx_1\cdot x_2}; \\ \beta_4 &= \beta_{\bar{Y}\bar{X}_2\cdot\bar{X}_1} - \beta_{yx_2\cdot x_1}. \end{aligned} \quad (9)$$

The parameters in (9) can be generated by repeating the procedure used in the bivariate case: (1) begin with the covariance equation; (2) derive the least-squares solution for the parameters (μ, A_j, and β's); (3) state $\bar{Y}$ as a function of the $\bar{X}$'s, and substitute the latter for $\bar{Y}$; (4) rearrange the terms so that Y is given in terms of the X's and $\bar{X}$'s (see the appendix).

Equation (9) shows that $\beta_{\bar{Y}\bar{X}_1\cdot\bar{X}_2} - \beta_{yx_1\cdot x_2}$ gives the effect of $\bar{X}_1$ on $Y\cdot X_1, X_2, \bar{X}_2$, and $\beta_{\bar{Y}\bar{X}_2\cdot\bar{X}_1} - \beta_{yx_2\cdot x_1}$ gives the effect of $\bar{X}_2$ on $Y\cdot X_1, X_2, \bar{X}_1$. Hence, as in the bivariate case, the terms attached to the $\bar{X}$-variables are the cross-level bias terms. Cross-level bias, then, is absent when, and only when, $\bar{X}_1$ and $\bar{X}_2$ have no structural effects.

The generalization of this result to n

[8] Note that mixed models—i.e., models where the $\bar{X}$-effects involve different variables than the individual-level effects—are possible in the multivariate case if we allow individual-level parameters to be zero. For example, if $\beta_2 = \beta_3 = 0$ in equation (8), the model reduces to $Y = a + \beta_1 X_1 + \beta_4 \bar{X}_2 + e$. Mixed models change the cross-level bias terms; in the current example, the individual-level effect of interest is $\beta_{yx_1\cdot\bar{x}_2}$, not $\beta_{yx_1\cdot x_2}$. Except where otherwise noted, the discussion throughout assumes nonzero individual-level parameters.

independent variables is straightforward, though notationally cumbersome. We begin with the following structural model:

$$Y_{ij} = a + \beta_1 X_{1ij} + \ldots + \beta_n X_{nij} + \beta_{n+1} \bar{X}_{1j} + \ldots + \beta_{2n} \bar{X}_{nj} + e, \quad (10)$$

where the variables are defined as before, and the assumptions about the error term and measurement error are the same as in equation (2). The structural parameters for equation (10) are (see the Appendix):

$$\begin{aligned} \beta_1 &= \beta_{yx_1 \cdot x_2, \ldots x_n} \\ &\vdots \\ \beta_n &= \beta_{yx_n \cdot x_1, \ldots x_{n-1}} \\ \beta_{n+1} &= \beta_{\bar{Y}\bar{X}_1 \cdot \bar{X}_2, \ldots \bar{X}_n} - \beta_{yx_1 \cdot x_2, \ldots x_n} \\ &\vdots \\ \beta_{2n} &= \beta_{\bar{Y}\bar{X}_n \cdot \bar{X}_1, \ldots \bar{X}_{n-1}} - \beta_{yx_n \cdot x_1, \ldots x_{n-1}}. \end{aligned} \quad (11)$$

As before, the coefficients of the $\bar{X}$-terms are the bias terms. Hence, equation (11) demonstrates the rule for cross-level inference in the n-variable case: downward cross-level bias is absent when, and only when, $\bar{X}_1, \ldots \bar{X}_n$ have no structural effects on Y.

II. THE $\bar{X}$-RULE: THEORETICAL INTERPRETATION AND IMPLICATIONS FOR RESEARCH

When deriving a general rule for cross-level inference, it is necessary to focus on the structural (i.e., "true") relationships between variables. But the researcher of course deals with *observed* relationships; these relationships may be misspecified. Hence, in applying the $\bar{X}$-rule, the researcher must not only determine if $\bar{X}$-effects are present, but must also assess whether they are true effects. In this process, the researcher making downward cross-level inference is doubly handicapped: not only must s/he deal with the issue of whether $\bar{X}$-effects are structural, but s/he must make this assessment without being able to estimate $\bar{X}$-effects empirically.

This section is intended to help the researcher in applying the $\bar{X}$-rule in empirical analysis. Applying the $\bar{X}$-rule involves asking two questions. (1) Are $\bar{X}$-effects present? (2) If $\bar{X}$-effects are present, can they be eliminated by respecification of the (aggregate) equation? These issues must be determined by theory. First, then, I discuss the general theoretical underpinnings of $\bar{X}$-effects.

Group Effects Theory

Sociologists have often argued that groups can (and do) have effects over and beyond those of the attributes of the group members (see, for example, Durkheim, 1897; Merton and Kitt, 1950; Blau, 1957; 1960). Blau (1960:179) expressed this issue as follows:

> The individual's orientation undoubtedly influences his behavior; the question is whether the prevalence of social values in a community also exerts social constraints upon patterns of conduct that are independent of the influences exerted by the internalized orientations.

Stouffer et al. (1949), for example, found that inexperienced soldiers in veteran units were less likely to say that they were ready for combat than inexperienced soldiers in inexperienced units—an indication of the influence of the veterans (who generally said they were not ready for combat). Sociologists also have found evidence of group effects in public assistance agencies (Blau, 1960), book discussion groups (Davis et al., 1961) and high schools (Alexander and Eckland, 1975; but see Hauser et al., 1976), to name a few.

Some discussions give the impression that groups are unidimensional and that there is, at most, a single group effect.[9] But groups can have numerous distinguishable properties even as individuals can have numerous distinguishable properties. Hence I prefer to speak in terms of macroproperties and microproperties. Macroproperties can be divided into two

[9] This misimpression probably has arisen because the effect of the nominal-scale variable in covariance analysis is often called a group effect. This effect is probably better called the composite group effect, since it measures the total impact of all the group properties affecting Y (Firebaugh, 1977a).

Table 1. Three Cases Where X_1 Would Be Related to $Y \cdot X_1$

Case	Structural Equation	Further Conditions.
Case 1. Spurious Group Effect	$Y = a + \beta_1 X_1 + \beta_2 X_2 + \epsilon$	$\bar{X}_1$ correlated with X_2
Case 2. Non-$\bar{X}_1$ Group Effect	$Y = a + \beta_1 X_1 + \beta_2 \bar{X}_2 + \epsilon$	$\bar{X}_1$ correlated with $\bar{X}_2$
	or[a] $Y = a + \beta_1 X_1 + \beta_2 I + \epsilon$	$\bar{X}_1$ correlated with I
Case 3. Emergent $\bar{X}_1$-Effect	Disputed[b]	———

[a] The symbol "I" represents an integral macroproperty.
[b] See discussion in text.

classes, according to whether or not they are measured by aggregating microproperties. Macroproperties which are not measured by aggregation—form of government, for example—have been called integral properties (Selvin and Hagstrom, 1963).

Recall the earlier observation that a correlation between $\bar{X}$ and $Y \cdot X$ is possible because $\bar{X}$ can measure more constructs than X.[10] Note three possible reasons for a correlation between $\bar{X}$ and $Y \cdot X$: (1) $\bar{X}$ can reflect microproperties other than X; (2) $\bar{X}$ can reflect macroproperties other than $\bar{X}$; and (3) $\bar{X}$ itself can have a causal effect on Y. We will now see that these cases are central in the discussion of eliminating cross-level bias through respecification.

Eliminating $\bar{X}$-Effects

Table 1 presents the three cases where $\bar{X}$ would be correlated with $Y \cdot X$. When the $\bar{X}$-effect is due to uncontrolled individual-level variables (Case 1), cross-level bias can be avoided by respecification of the aggregate equation. On the other hand, the cross-level biases involved in Cases 2 and 3 ordinarily cannot be eliminated by respecification, since they involve causal macroproperties. We now examine each of these cases in more detail.

Case 1. spurious group effect. As Hauser (1970; 1974) has argued convincingly, group effects may be merely individual-level effects in disguise. Spurious $\bar{X}$-effects occur when $\bar{X}$ is correlated with uncontrolled causal individual-level variables. Table 1 gives the prototypical form of this case: the causal variables, X_1 and X_2, are both individual-level variables, and X_2 is correlated with $\bar{X}_1$. In this situation, $\bar{X}_1$ will be correlated with $Y \cdot X_1$ through its correlation with X_2. This problem can be remedied by respecifying the aggregate equation. Observe that $\bar{X}_1$ and $\bar{X}_2$ have no independent effects in a regression of Y on X_1, X_2, $\bar{X}_1$, and $\bar{X}_2$ (note the structural equation in Table 1). Now apply the multivariate form of the $\bar{X}$-rule: if $\bar{X}_1$ and $\bar{X}_2$ have no independent effects, then the equation $\bar{Y} = a + \beta_1 \bar{X}_1 + \beta_2 \bar{X} + \epsilon$ provides unbiased estimates of the parameters of the structural equation $Y = a + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.

In short, in the case of spurious group effects, unbiased estimates are possible even when the researcher does not have recourse to individual-level data. How does one know whether group effects are causal or spurious? In the final analysis, the question of whether the effect of a given macroproperty is spurious—like the question of whether the effect of a given microproperty is spurious—must be resolved by theory. As a general rule, causal group effects seem most likely in groups where group members interact and share relevant life experiences; hence, the macroproperties of "natural" groups (neighborhoods, for example) seem more likely to have causal effects than the macroproperties of arbitrarily-created regions (census tracts, for example).

Case 2. non-X_1 group effect. In Cases 1 and 2, $\bar{X}_1$ is correlated with $Y \cdot X_1$ through its relationship with variables, other than X_1, which cause Y. In Case 1, these causal variables are microproperties; in Case 2 they are macroprop-

[10] I am grateful to an *ASR* reviewer for suggesting that I link this observation more directly to group effects theory.

erties.[11] Unlike Case 1, then, Case 2 involves causal group effects. The causal macroproperties may be either $\bar{X}$-properties or integral properties. Table 1 gives the prototypical forms: in the first equation for Case 2, $\bar{X}_2$ has a causal effect on Y; in the second equation, I (denoting an integral macroproperty) has a causal effect on Y.[12]

Consider again Cronbach's (1976) example of mean education of communities. To provide a focus for our consideration, suppose that we wish to study opera attendance (Y). Certainly education (X_1) causally affects opera attendance. Other microproperties, such as income, probably also affect opera attendance; let's denote these variables $X_2, \ldots X_n$. But community macroproperties also could affect opera attendance; for example, community facilities for opera performances (I) likely affects Y. If mean community education ($\bar{X}_1$), or any other community mean $\bar{X}_2$ to $\bar{X}_n$, is correlated with I, the aggregate equation $\bar{Y} = a + \beta_1 \bar{X}_1 + \ldots + \beta_n \bar{X}_n + \epsilon$ will give biased estimates of the individual-level effects.

At first blush, the solution to this problem seems straightforward: control the causal macroproperties. However, this approach does not necessarily work. Consider the simplest case: two causal variables, X_1 and $\bar{X}_2$ (Table 1). Since $\bar{X}_1$ has no independent effect on $Y \cdot X_1 X_2, \bar{X}_2$, we know that β_1 in the equation $\bar{Y} = a + \beta_1 \bar{X}_1 + \beta_2 \bar{X}_2 + \epsilon$ is an unbiased estimate of $\beta_{yx_1 \cdot x_2}$ (see equation (9), above). However, $\beta_{yx_1 \cdot x_2}$ is not the parameter of interest; we want the effect of X_1 on $Y \cdot \bar{X}_2$, *not* the effect of X_1 on $Y \cdot X_2$. Unless the effect of X_1 on $Y \cdot \bar{X}_2$ is equivalent to the effect of X_1 on $Y \cdot X_2$, then, the regression of $\bar{Y}$ on $\bar{X}_1$ and $\bar{X}_2$ will give a biased estimate of the effect of X_1.

Case 3. emergent $\bar{X}_1$-effect. Consider the following question: since $\bar{X}$ is a sum of X, how can $\bar{X}$ be related to Y once X is controlled? We have already examined two possibilities. $\bar{X}$ could be related to uncontrolled causal microproperties or it could be related to causal macroproperties. Now we consider a third possibility: that $\bar{X}$ itself could give rise to properties which causally affect $Y \cdot X$. That is, Case 3—unlike Cases 1 and 2—involves emergent properties implicit in the aggregation of X.

I am referring in particular to emergent group "atmospheres"; effects of such atmospheres are variously termed "contextual" (Farkas, 1974), "structural" (Blau, 1960), or "compositional" (Davis et al., 1961; Werts and Linn, 1971). We can illustrate such effects with an example from Boudon (1963): assume X= income and Y= voting behavior (conservatism). Boudon suggested that, in France, the mean income of a neighborhood has a positive effect on conservatism, net of individual income. A person living in a high-income neighborhood, then, is expected to be more conservative politically than a person with the same income living in a low-income neighborhood. In such a case, $\bar{X}$ (mean neighborhood income) would be related to $Y \cdot X$.

Social scientists dispute whether $\bar{X}$ should be considered the structural (i.e., true) variable in such a case.[13] This dispute turns on the question of whether the emergent property generated by $\bar{X}$ should be viewed as a separate variable or as an inherent part of $\bar{X}$ (Cronbach, 1976:1.27: "The properties of what the physicists call a critical mass arise from the aggregate itself, not some 'additional variable'. The whole in this case is more than the sum of the parts").

Consider again the effect of mean neighborhood income on political conservatism. How could such an effect come about? The implicit argument apparently

[11] Davis (1966) discussed an effect which is difficult to classify either as micro or macro. This is the "frog pond" effect. In the frog pond effect, group members use some group property as a point of comparison; the individual's position relative to that group standard then affects Y. When frog pond effects involve $\bar{X}$, $\bar{X}$ will be related to $Y \cdot X$ (Firebaugh, 1977b).

[12] In the case where more than one macroproperty causally affects Y, the correlation between X_1 and $Y \cdot X_1$ is determined by the correlation of $\bar{X}_1$ with the *composite* effect of the macroproperties (Werts and Linn, 1971; Firebaugh, 1977a).

[13] Indeed, some social scientists question the possibility of emergent properties. I do not care to enter this debate here; the interested reader should consult Hannan (1971a:Chap. 1).

is that a person is influenced by those with whom s/he interacts, and that people in affluent neighborhoods are more likely than those in poor neighborhoods to interact with conservatives.[14] Hence one might be tempted to say that the causal macroproperty is the neighborhood's level of conservatism, not its level of income; that is, one would propose that the following equation is the structural equation: $Y = a + \beta_1 X + \beta_2 \bar{Y} + \epsilon$. But, under this specification, the mean conservatism of a neighborhood is caused by the mean income of the neighborhood (this easily can be shown by taking the within-group expected value of Y in the equation just given). Which is the structural variable, then, mean income or mean conservatism?

I am inclined to view mean income as the structural variable; others may disagree. But this is a moot point relative to cross-level bias: whether Y is caused by X and $\bar{X}$, or caused by X and $\bar{Y}$, the aggregate-level equations ($\bar{Y} = a + \beta_1 \bar{X} + \beta_2 \bar{X} + \epsilon$ and $\bar{Y} = a + \beta_1 \bar{X} + \beta_2 \bar{Y} + \epsilon$, respectively) obviously cannot be estimated. In short, in the case of emergent $\bar{X}$-effects, individual-level data are required for unbiased estimation.

This section concludes the exegesis of the $\bar{X}$-rule. I will use Robinson's (1950) classic example of the relationship between race and illiteracy to illustrate the major principles set forth by the rule. First, however, I address a final issue: why I have chosen to focus on the within-group slope (β_{yx}) instead of the total individual-level slope (β_{YX}).

III. FOCUSING ON β_{yx}

The focus of this paper differs from the focus of prior discussions of cross-level bias. In discussing the effect of X on Y at the individual level, prior discussions in sociology typically have focused on the following equation:

$$Y_{ij} = a + \beta_{YX} X_{ij} + e. \qquad (12)$$

However, the individual-level effect of X on Y is represented by β_{yx}, not β_{YX}; β_{YX} is a combination of individual-level and aggregate-level effects of X (Duncan et al., 1961:66):

$$\beta_{YX} = \beta_{yx} + E^2_{\bar{X}A} (\beta_{\bar{Y}\bar{X}} - \beta_{yx}), \qquad (13)$$

where $E^2_{\bar{X}A}$ is the correlation between X and A (the grouping variable). Therefore, the researcher should be interested in estimating β_{yx}, not β_{YX} (Cronbach, 1976).

In discussing the rule for cross-level inference, I have found it convenient to refer to the difference between $\beta_{\bar{Y}\bar{X}}$ and β_{yx} as cross-level bias. Many discussions, however, split this difference into two parts ($\beta_{\bar{Y}\bar{X}} - \beta_{YX}$ and $\beta_{YX} - \beta_{yx}$) and focus on the former. Following Hannan and Burstein (1974:387), I will call $\beta_{\bar{Y}\bar{X}} - \beta_{YX}$ "aggregation bias."

From equation (13) one can derive the relationship between cross-level bias (δ) and aggregation bias (θ):

$$\theta/\delta + E^2_{\bar{X}A} = 1, \; E^2_{\bar{X}A} \neq 0 \text{ or } 1, \; \delta \neq 0. \qquad (14)$$

From this identity note, first, that cross-level bias is zero when, and only when, aggregation bias is zero, since $\theta = (1 - E^2_{\bar{X}A})\,\delta$, and $0 < E^2_{\bar{X}A} < 1$ (cross-level bias is indeterminate when $E^2_{\bar{X}A} = 0$ or 1). Hence, in the bivariate case, rules for avoiding cross-level bias apply to aggregation bias, and conversely. Second, note that aggregation bias is always less than cross-level bias. Third, the size of aggregation bias relative to cross-level bias is a function of the correlation between the independent variable and the grouping variable: as $E^2_{\bar{X}A}$ increases, the proportion of cross-level bias that is aggregation bias decreases.

An examination of the empirical example given by Hannan and Burstein (1974:Table 2) illustrates these points. Hannan and Burstein used data for 2,676 incoming university freshmen to assess the likely consequences of grouping under various types of grouping variables. Their purpose was to identify those grouping variables which result in the least aggrega-

[14] There is another possibility: perhaps conservatives select housing in wealthier neighborhoods (and liberals select housing in poorer neighborhoods) than expected on the basis of their income. This is an example of what has been termed "grouping by Y" (Blalock, 1964) or "selection by the dependent variable" (Hammond, 1973). The case of selection by the dependent variable, like the case of emergent $\bar{X}$-effects, results in cross-level bias.

Table 2. Aggregation Bias and Cross-Level Bias under Four Grouping Variables[a]

Grouping Variable (A)	E^2_{XA}	Aggregation Bias (Θ)[b]	Cross-level Bias (δ)	Θ/δ
S.A.T.	.98	-.001	-.037	.02
Father's education	.03	.039	.040	.97
Self-opinion of academic abilities	.28	.060	.084	.72
Achievement test score	.70	.329	1.080	.30

[a] See Hannan and Burstein (1974) for a more complete discussion of these data.
[b] Variables were standardized before grouping; hence these numbers differ from those reported in Hannan and Burstein's Table 2.

tion bias, and those which result in the most; as expected, they found that grouping by the independent variable (aptitude score on S.A.T. test) was the best method, while grouping by the dependent variable (score on an achievement test) was the worst.

By contrast, our purpose is to compare aggregation bias and cross-level bias. Table 2 compares aggregation bias and cross-level bias for four of Hannan and Burstein's grouping variables.[15] These variables were chosen since they cover the range of values for E^2_{XA}, from a very large correlation (grouping by S.A.T.) to a very small correlation (grouping by father's education). As expected, aggregation bias is less than cross-level bias. Further, θ/δ is inversely related to E^2_{XA}, as expected. A comparison of the two extremes on θ/δ—S.A.T. and father's education—underscores the differences in focusing on aggregation bias instead of cross-level bias: while aggregation bias is quite different under the two methods of grouping, cross-level bias is about the same.

IV. ILLUSTRATION: ROBINSON REVISITED

An empirical illustration should crystallize the ideas presented in this paper. A reanalysis of Robinson's (1950) classic illustration of the relationship between race and illiteracy seems fitting, since Robinson's paper is the seminal paper for discussions of cross-level bias in sociology. Using 1930 U.S. census data, Robinson computed the correlation between race (black/nonblack) at the individual level and at the regional level. The correlations were .20 and .95, respectively—graphic confirmation of Robinson's contention that aggregate data misestimate individual-level correlations.

Tables 3 and 4 present the data for Robinson's computations. These data are of course nominal scale, but we can still use them to illustrate the principles set forth in this paper. Unlike Robinson, who used measures of correlation, we employ regression coefficients. We first compute β_{YX}. According to these data (Table 3), 16.3% of blacks and 3.1% of the remainder of the population were illiterate in 1930. Letting I=probability of being illiterate, and letting R be a dummy variable for race (1 if black, 0 otherwise), we can write this individual-level relationship between race and illiteracy as follows:

$$I = .031 + .132R. \qquad (15)$$

This equation states that the probability of being illiterate is .163 (=.031+[.132] [1]) for a black, and .031 (=.031+[.132] [0]) for a nonblack. The coefficient for R is analogous to a regression coefficient; indeed,

[15] Professor Leigh Burstein generously supplied the information needed to construct Table 2.

Table 3. Race and Illiteracy (000's, Population Ten Years and Older): 1930 U.S.[a]

	Black		Nonblack		Total	
	N	%	N	%	N	%
Illiterate	1,514	16.3	2,770	3.1	4,284	4.3
Literate	7,779	83.7	86,661	96.9	94,440	95.7

[a] Source: U.S. Census, 1930.

Table 4. Percent Black and Percent Illiterate, by Nine Regions: 1930 U.S.[a]

Region	% Black[b]	% Illiterate[b]
1. New England	1.1	3.7
2. Middle Atlantic	4.0	3.5
3. East North Central	3.7	2.1
4. West North Central	2.6	1.4
5. South Atlantic	27.6	8.3
6. East South Central	27.2	9.6
7. West South Central	18.8	7.2
8. Mountain	.9	4.2
9. Pacific	1.1	2.1

[a] Source: U.S. Census, 1930.
[b] Population ten years and older.

regressing I (as a dummy variable: 1 if illiterate, 0 otherwise) on R yields $\beta_{IR} = .132$.

Next we compute $\beta_{\bar{Y}\bar{X}}$. Following Robinson, we group by region. To obtain $\beta_{\bar{Y}\bar{X}}$, however, we note that the important regional classification is South/non-South (see Table 4; by "South" I mean regions 5–7). Table 5 presents the aggregate data for race and illiteracy, grouped by region (South/non-South). From Table 5 we can compute the aggregate-level equation analogous to equation (15):

$$\bar{I} = .019 + .26\bar{R}, \qquad (16)$$

where $\bar{I}$ = proportion illiterate and $\bar{R}$ = proportion black. Hence, the relationship between proportion black and proportion illiterate by region overstates the zero-order individual-level relationship between race and illiteracy by .128 (= .26 − .132).

From the finding that $\beta_{\bar{Y}\bar{X}} \neq \beta_{YX}$ we can make two inferences. First, we can infer that percent black by regions has an $\bar{X}$-effect on illiteracy. This $\bar{X}$-effect is no doubt indirect; percent black of a region is probably correlated with more direct causes of illiteracy, such as inferior schools. Second, we can infer that the zero-order individual-level relationship between race and illiteracy is not the relationship of interest; i.e., β_{YX}, like $\beta_{\bar{Y}\bar{X}}$, misestimates the individual-level effect of race on illiteracy.

Table 5. Percent Black and Percent Illiterate, South/Non-South: 1930 U.S.[a]

Region	Percent Black[b]	Percent Illiterate[b]
South	24.7	8.3
Non-South	3.0	2.7

[a] Source: U.S. Census, 1930.
[b] Population ten years and older.

A complete specification of the individual-level and aggregate-level determinants of illiteracy in the U.S. in 1930 is beyond the scope of this paper (but see Hanushek et al., 1974). Nevertheless, the finding that percent black by region has an effect on illiteracy suggests the examination of race and illiteracy, with region controlled (Alker, 1969: 84–5, also suggests this control). Table 6 presents these data; 19.7% of blacks living in the South in 1930 were illiterate, while only 4.6% of blacks living outside the South were illiterate. This suggests that region, not race, was the major determinant of illiteracy. However, illiteracy among nonblacks differed by only 2% between regions (Table 6). Apparently, neither being black, nor living in the South, *in itself* significantly raised the probability of being illiterate. However, being black in the South did. We can see this interaction effect very clearly by translating the above percentages into an equation for illiteracy:

$$I = .026 + .020R + .020S + .131 (R*S), \qquad (17)$$

where I and R are defined as in equation (15), S is a dummy variable for region (1 for South, 0 otherwise), and R*S is a dummy variable for the interaction between race and region (1 for blacks living in the South, 0 otherwise). The effect of being a black in the South is striking: it raises the probability of being illiterate by .131, net of the additive effects of race and region. Without further data, the interpretation of this interaction effect is ambiguous (discrimination? school segregation? etc.). Note, however, that controlling this interaction reduces the independent effect of race to .02; i.e., net of region and the region/race interaction, being black raises the probability of being illiterate by only .02. In this analysis, in sum, the net individual-level effect of race on illiteracy is .02; the zero-order individual-level relationship between race and illiteracy is

Table 6. Race and Illiteracy, South/Non-South (000's, Population Ten Years and Older): 1930 U.S.[a]

	South					
	Black		Nonblack		Total	
	N	%	N	%	N	%
Illiterate	1,416	19.7	1,001	4.6	2,417	8.3
Literate	5,779	80.3	20,972	95.4	26,751	91.7
	Non-South					
Illiterate	96	4.6	1,771	2.6	1,867	2.7
Literate	2,003	95.4	65,685	97.4	67,688	97.3

[a] Source: U.S. Census, 1930.

.132; and the aggregate-level relationship between race and illiteracy is .26. As expected, then, β_{YX} misestimates the net individual-level effect of X on Y when $\beta_{\bar{Y}\bar{X}}$ misestimates β_{YX}.

V. IMPLICATIONS AND CONCLUSION

Implications

The major conclusion of this paper is that downward cross-level inference can be made without bias when, and only when, $\bar{X}$-effects are absent. This conclusion has different implications for upward cross-level inference than it does for downward cross-level inference. We now consider its implications for upward cross-level inference; i.e., for the case where the researcher makes inferences about aggregate-level effects from individual-level data.

The conclusions for downward cross-level inference do not apply in a straightforward way to upward cross-level inference. Consider the bivariate case, for example. In downward inference, at issue is the use of an obtained coefficient ($\beta_{\bar{Y}\bar{X}}$) to estimate the coefficient of interest at a lower level of aggregation (β_{yx}); in upward inference, at issue is the use of an obtained coefficient (β_{YX}) to estimate the coefficient of interest at a higher level of aggregation ($\beta_{\bar{Y}\bar{X}\cdot X}$). (The coefficient of interest at the aggregate level is $\beta_{\bar{Y}\bar{X}\cdot X}$, not $\beta_{\bar{Y}\bar{X}}$, since the latter is a function of individual-level as well as aggregate-level effects.) Upward cross-level inference can be made with impunity, then, only when $\beta_{YX} = \beta_{\bar{Y}\bar{X}\cdot X}$ (note that, since $\beta_{\bar{Y}\bar{X}\cdot X} = \beta_{Y\bar{X}\cdot X}$, $\beta_{\bar{Y}\bar{X}\cdot X} = \beta_{\bar{Y}\bar{X}} - \beta_{yx}$; see equations (2) and (4), above). That the rules for downward and upward cross-level inference are not equivalent can be seen as follows: if $\beta_{\bar{Y}\bar{X}} = \beta_{yx}$, then $\beta_{\bar{Y}\bar{X}} = \beta_{yx} = \beta_{YX}$ (this is easily seen by examining the formulas for these β's); if $\beta_{\bar{Y}\bar{X}} = \beta_{yx} = \beta_{YX}$, then $\beta_{YX} \neq \beta_{\bar{Y}\bar{X}\cdot X}$ ($=\beta_{\bar{Y}\bar{X}} - \beta_{yx}$), except in the uninteresting case that $\beta_{\bar{Y}\bar{X}} = \beta_{yx} = \beta_{YX} = 0$. In short, it is not true, in general, that upward cross-level inference yields unbiased estimates when downward cross-level inference does (or conversely).

A separate discussion of the conditions under which individual-level data provide unbiased estimates of aggregate-level effects would have little utility, since the researcher with individual-level data can estimate directly the individual-level and aggregate-level effects of X on Y (assuming, of course, that the researcher can group the data as desired: for example, computing $\bar{X}$ for nations requires knowing the nationality of the individuals). The researcher with individual-level data, then, most often can include both X and $\bar{X}$ in the equation; through such a multilevel analysis the individual-level and aggregate-level effects of X can be separated (Alwin, 1976; Cronbach, 1976).

The researcher with purely aggregate-level data has fewer options. As we have seen, if $\bar{X}$-effects are present, aggregate-level regression coefficients give biased estimates of individual-level X-effects. Hence, the researcher who desires to obtain unbiased estimates of individual-level effects from aggregate data must respecify the equation so that $\bar{X}$-effects are eliminated. This is possible if the $\bar{X}$-effects are not structural. Hence, if true group effects are so rare or so small that they almost always can be ignored (see Hauser, 1970;

1974), then Hanushek et al. (1974) are correct in their insistence that the researcher restricted to aggregate data should worry primarily about proper specification; the ecological fallacy is itself a near fallacy. If, on the other hand, group effects cannot be dismissed (see Barton, 1970; Farkas, 1974), unbiased estimates of individual-level relationships often will be unobtainable from aggregate data; in such cases, unbiased estimates can be obtained only from multilevel analysis.[16]

Summary and Conclusion

This paper has been directed to the sociologist who faces the question of using aggregate-level data to infer individual-level relationships. In this situation, the crucial question is the following: are the means of the independent variables related to the dependent variable, net of the effects of independent variables? I do not wish to claim too much for my $\bar{X}$-rule; certainly this paper does not solve all the issues in aggregation. Nevertheless, this approach has advantages. First, it links aggregation problems to group effects theory. This not only provides possible theoretical leverage to the researcher puzzling over the legitimacy of downward cross-level inference in a particular case, but it also may demystify the ecological fallacy for some sociologists. Second, the $\bar{X}$-rule is easily generalizable to the n-variable case. Finally, this approach focuses explicitly on the difference between the aggregate-level coefficient and the individual-level coefficient of interest, β_{yx}^{x}.

In conclusion, this paper has dealt with cross-level bias by introducing a general rule for making downward cross-level inference. Of course, the researcher concerned about cross-level bias most often does not have individual-level data and thus cannot determine, empirically, whether the data conform to the rule. This is a problem with downward cross-level inference, to be sure; however, it does not differ in principle from the specification problem faced in all causal analyses. In regression analysis the researcher always makes assumptions about the data used. The validity of these assumptions most often is determined on theoretical grounds; rarely can all the assumptions be tested empirically. This paper suggests that, in an analysis which uses aggregate data to study individual-level relationships, an additional assumption is needed: that there are no $\bar{X}$-effects. When this assumption is met, aggregate data can provide unbiased estimates of individual-level effects.

[16] Fortunately, multilevel analysis does not require individual-level data for all groups. If individual-level data are available for randomly selected groups, estimates can be obtained for β_{yx} and $\beta_{yx.x}$. Indeed, such a strategy often may be necessary since, in many cases, collecting individual-level data for all groups is prohibitively costly.

APPENDIX

Derivation of Equation (9)

The standard covariance equation, with two covariates, is as follows:

$$Y_{ij} = \mu + A_j + \beta_1 (X_{1ij} - \bar{\bar{X}}_1) + \beta_2(X_{2ij} - \bar{\bar{X}}_2) + e_{ij}, \quad (1A)$$

where the variables are defined as before. The least-squares solution for the normal equations derived from (1A), subject to the constraint that $\sum_{j=1}^{M} n_j A_j = 0$, yields the following for the population parameters in (1A):

$$\begin{aligned} \mu &= \bar{\bar{Y}}, \\ \beta_1 &= \beta_{yx_1 \cdot x_2}, \\ \beta_2 &= \beta_{yx_2 \cdot x_1}, \\ A_j &= \bar{Y}_j - \beta_1 (\bar{X}_{1j} - \bar{\bar{X}}_1) \\ &\quad - \beta_2 (\bar{X}_{2j} - \bar{\bar{X}}_2) - \bar{\bar{Y}}, \end{aligned} \quad (2A)$$

where $\beta_{yx_1 \cdot x_2}$ and $\beta_{yx_2 \cdot x_1}$ are the within-group regression coefficients for X_1 and X_2, respectively; these coefficients give the individual-level effect of X_1 and X_2 on Y. Noting that $\bar{Y}_j = a_{\bar{Y}\bar{x}_{1,2}} + \beta_{\bar{Y}\bar{x}_1 \cdot \bar{x}_2} \bar{X}_{1j} + \beta_{\bar{Y}\bar{x}_2 \cdot \bar{x}_1} \bar{X}_{2j} + e_{\bar{Y}\bar{x}_{1,2}}$, and substituting (2A) into (1A), we obtain:

$$\begin{aligned} Y_{ij} &= \bar{\bar{Y}} + \{(a_{\bar{Y}\bar{x}_{1,2}} + \beta_{\bar{Y}\bar{x}_1 \cdot \bar{x}_2} \bar{X}_{1j} \\ &\quad + \beta_{\bar{Y}\bar{x}_2 \cdot \bar{x}_1} \bar{X}_{2j} + e_{\bar{Y}\bar{x}_{1,2}}) \\ &\quad - \beta_1 (\bar{X}_{1j} - \bar{\bar{X}}_1) \\ &\quad - \beta_2 (\bar{X}_{2j} - \bar{\bar{X}}_2) - \bar{\bar{Y}}\} \\ &\quad + \beta_1 (X_{1ij} - \bar{\bar{X}}_1) \\ &\quad + \beta_2 (X_{2ij} - \bar{\bar{X}}_2) + e_{ij} \\ &= a_{\bar{Y}\bar{x}_{1,2}} + \beta_1 X_{1ij} + \beta_2 X_{2ij} \\ &\quad + (\beta_{\bar{Y}\bar{x}_1 \cdot \bar{x}_2} - \beta_1) \bar{X}_{1j} \\ &\quad + (\beta_{\bar{Y}\bar{x}_2 \cdot \bar{x}_1} - \beta_2) \bar{X}_{2j} + e. \end{aligned} \quad (3A)$$

Derivation of Equation (11)

The demonstration of (11) is a straightforward extension of the previous case; I indicate here only the key equations. Beginning with the covariance equation with n covariates, we find that the least-squares solution for A_j is:

$$A_j = \bar{Y}_j - \beta_{yx_1 \cdot x_2, \ldots x_n}(\bar{X}_{1j} - \bar{\bar{X}}_1) - \ldots - \beta_{yx_n \cdot x_1, \ldots x_{n-1}}(\bar{X}_{nj} - \bar{\bar{X}}_n) - \bar{\bar{Y}}. \quad (4A)$$

Noting that $\bar{Y}_j = a_{\bar{Y}\bar{x}_1, \ldots n} + \beta_{\bar{Y}\bar{x}_1 \cdot \bar{x}_2, \ldots \bar{x}_n}\bar{X}_{1j} + \ldots + \beta_{\bar{Y}\bar{x}_n \cdot \bar{x}_1, \ldots \bar{x}_{n-1}}\bar{X}_{nj} + e_{\bar{Y}\bar{x}_1, \ldots n}$,

and substituting (4A) into the covariance equation for n covariates, we obtain:

$$\begin{aligned} Y_{ij} = \bar{\bar{Y}} &+ \{(a_{\bar{Y}\bar{x}_1, \ldots n} + \beta_{\bar{Y}\bar{x}_1 \cdot \bar{x}_2, \ldots \bar{x}_n}\bar{X}_{1j} + \ldots + \beta_{\bar{Y}\bar{x}_n \cdot \bar{x}_1, \ldots \bar{x}_{n-1}}\bar{X}_{nj} + e_{\bar{Y}\bar{x}_1, \ldots n}) \\ &- \beta_{yx_1 \cdot x_2, \ldots x_n}(\bar{X}_{1j} - \bar{\bar{X}}_1) \\ &- \ldots - \beta_{yx_n \cdot x_1, \ldots x_{n-1}}(\bar{X}_{nj} - \bar{\bar{X}}_n) - \bar{\bar{Y}}\} \\ &+ \beta_{yx_1 \cdot x_2, \ldots x_n}(X_{1ij} - \bar{\bar{X}}_1) \\ &+ \ldots + \beta_{yx_n \cdot x_1, \ldots x_{n-1}}(X_{nij} - \bar{\bar{X}}_n) + e_{ij} \\ = a_{\bar{Y}\bar{x}_1, \ldots n} &+ \beta_{yx_1 \cdot x_2, \ldots x_n}X_{1ij} \\ &+ \ldots + \beta_{yx_n \cdot x_1, \ldots x_{n-1}}X_{nij} \\ &+ (\beta_{\bar{Y}\bar{x}_1 \cdot \bar{x}_2, \ldots \bar{x}_n} - \beta_{yx_1 \cdot x_2, \ldots x_n})\bar{X}_{1j} \\ &+ \ldots + (\beta_{\bar{Y}\bar{x}_n \cdot \bar{x}_1, \ldots \bar{x}_{n-1}} - \beta_{yx_n \cdot x_1, \ldots x_{n-1}})\bar{X}_{nj} + e. \end{aligned} \quad (5A)$$

REFERENCES

Alexander, Karl and Bruce K. Eckland
1975 "Contextual effects in the high school attainment process." American Sociological Review 40:402–16.

Alker, Hayward R., Jr.
1969 "A typology of ecological fallacies." Pp. 69–86 in M. Dogan and S. Rokkan (eds.), Quantitative Ecological Analysis in the Social Sciences. Cambridge, Ma.: MIT Press.

Alwin, Duane F.
1976 "Assessing school effects: some identities." Sociology of Education 49:294–303.

Barton, Allen H.
1970 "Allen Barton comments on Hauser's 'context and consex.' " American Journal of Sociology 76:514–7.

Blalock, Hubert M., Jr.
1964 Causal Inferences in Nonexperimental Research. Chapel Hill: University of North Carolina Press.

Blau, Peter M.
1957 "Formal organization: dimensions of analysis." American Journal of Sociology 63:58–69.
1960 "Structural effects." American Sociological Review 25:178–93.

Boudon, Raymond
1963 "Proprietes individuelles et proprietes collectives: un probleme d'analyse ecologique." Revue Française de Sociologie 4:275–99.

Burke, Peter J. and Karl F. Schuessler
1974 "Alternative approaches to analysis-of-variance tables." Pp. 145–88 in Herbert L. Costner (ed.), Sociological Methodology, 1973–74. San Francisco: Jossey-Bass.

Burnham, Walter Dean
1965 "The changing shape of the American political universe." American Political Science Review 59:7–28.
1971 "Communication." American Political Science Review 65:1149–52.

Burstein, Leigh
1974 "Issues concerning inferences from grouped observations." Paper presented at the annual meeting of the American Educational Research Association, Chicago.
1975a "Data aggregation in educational research: applications." Technical Report No. 1, March. Vasquez Associates. Also presented at AERA, Washington, D.C.
1975b The Use of Data from Groups for Inference about Individuals in Educational Research. Ph.D. dissertation, Stanford University.
1976 "Assessing differences between grouped and individual-level regression coefficients." Technical Report No. 10, April. Vasquez Associates. Also presented at AERA, San Francisco.

Chiricos, Theodore G. and Gordon P. Waldo
1970 "Punishment and crime: an examination of some empirical evidence." Social Problems 18:200–17.

Cowart, Andrew T.
1974 "A cautionary note on aggregate indicators of split ticket voting." Political Methodology 1:109–30.

Cronbach, Lee J.
1976 Research on Classrooms and Schools: Formulation of Questions, Design, and Analysis. Occasional paper of the Stanford Evaluation Consortium.

Davis, James A.
1966 "The campus as a frog pond." American Journal of Sociology 72:17–31.

Davis, James A., Joe L. Spaeth and Carolyn Huson
1961 "A technique for analyzing the effects of group composition." American Sociological Review 26:215–25.

Duncan, O. D., Raymond P. Cuzzort and Beverly Duncan
1961 Statistical Geography: Problems in Analyzing Areal Data. Glencoe: Free Press.

Duncan, O. D. and Beverly Davis
1953 "An alternative to ecological correlation." American Sociological Review 18:665–6.

Durkheim, Emile
[1897] Suicide. Glencoe: Free Press.
1951

Ehrlich, Isaac
1973 "Participation in illegitimate activities: a theoretical and empirical investigation." Journal of Political Economy 81:521–65.

1975 "The deterrent effect of capital punishment: a question of life and death." American Economic Review 65:314–25.

Farkas, George
1974 "Specification, residuals, and contextual effects." Sociological Methods and Research 2:333–63.

Feige, Edgar L. and Harold W. Watts
1972 "An investigation of the consequences of partial aggregation of micro-economic data." Econometrica 40:343–60.

Firebaugh, Glenn
1977a "Assessing group effects: a comparison of two methods." Unpublished manuscript, Department of Sociology and Anthropology, Vanderbilt University, Nashville.
1977b "Groups as contexts and frog ponds: some neglected considerations." Unpublished manuscript, Department of Sociology and Anthropology, Vanderbilt University, Nashville.

Goodman, Leo A.
1953 "Ecological regressions and behavior of individuals." American Sociological Review 18:663–4.
1959 "Some alternatives to ecological correlation." American Journal of Sociology 64:610–25.

Hammond, John L.
1973 "Two sources of error in ecological correlations." American Sociological Review 38:764–77.

Hannan, Michael T.
1971a Aggregation and Disaggregation in Sociology. Lexington: Heath-Lexington.
1971b "Problems of aggregation." Pp. 473–508 in Hubert M. Blalock, Jr. (ed.), Causal Models in the Social Sciences. Chicago: Aldine-Atherton.

Hannan, Michael T. and Leigh Burstein
1974 "Estimation from grouped observations." American Sociological Review 39:374–92.

Hannan, Michael T. and A. A. Young
1976 "Small sample results on estimation from grouped observations." Technical Report No. 24, October. Vasquez Associates.

Hanushek, Eric A., John E. Jackson and John F. Kain
1974 "Model specification, use of aggregate data, and the ecological correlation fallacy." Political Methodology 1:89–107.

Hauser, Robert M.
1970 "Context and consex: a cautionary tale." American Journal of Sociology 75:645–64.
1974 "Contextual analysis revisited." Sociological Methods and Research 2:365–75.

Hauser, Robert M., William H. Sewell and Duane F. Alwin
1976 "High school effects on achievement," Chap. 11 in William H. Sewell, Robert M. Hauser and David L. Featherman (eds.) Schooling and Achievement in American Society. New York: Academic Press.

Irwin, Laura and Allan J. Lichtman
1976 "Across the great divide: inferring individual-level behavior from aggregate data." Political Methodology 3:411–39.

Iversen, Gudmund R.
1973 "Recovering individual data in the presence of group and individual effects." American Journal of Sociology 79:420–34.

Johnston, J.
1972 Econometric Methods. 2nd ed. New York: McGraw-Hill.

Logan, Charles
1972 "General deterrent effects of punishment." Social Forces 51:64–73.

Merton, Robert K. and Alice S. Kitt
1950 "Contributions to the theory of reference group behavior." Pp. 40–105 in Robert K. Merton and Paul F. Lazarsfeld (eds.), Studies in the Scope and Methodology of "The American Soldier". Glencoe: Free Press.

Przeworski, Adam
1974 "Contextual models of political behavior." Political Methodology 1:27–61.

Robinson, W. S.
1950 "Ecological correlations and the behavior of individuals." American Sociological Review 15:351–7.

Schoenberger, Robert A. and David R. Segal
1971 "The ecology of dissent: the southern Wallace vote in 1968." Midwest Journal of Political Science 15:583–6.

Selvin, Hanan C.
1958 "Durkheim's 'Suicide' and problems of empirical research." American Journal of Sociology 63:607–19.

Selvin, Hanan C. and Warren O. Hagstrom
1963 "The empirical classification of formal groups." American Sociological Review 28:399–411.

Shively, W. Phillips
1969 " 'Ecological' inference: the use of aggregate data to study individuals." American Political Science Review 63:1183–96.

Slatin, Gerald T.
1969 "Ecological analysis of delinquency: aggregation effects." American Sociological Review 34:894–907.

Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star and R. M. Williams, Jr.
1949 The American Soldier: Adjustment During Army Life. Princeton: Princeton University Press.

Tittle, Charles R.
1969 "Crime rates and legal sanctions." Social Problems 16:409–23.

U.S. Bureau of the Census
1930 Population, Vol. 2: General Report. Washington, D.C.: U.S. Government Printing Office.

Werts, Charles E. and Robert L. Linn
1971 "Considerations when making inferences within the analysis of covariance model." Educational and Psychological Measurement 31:407–16.

Organizational Research Methods
Volume 11 Number 4
October 2008 815-852

10.1177/1094428106296642
http://orm.sagepub.com
hosted at
http://online.sagepub.com

Answers to 20 Questions About Interrater Reliability and Interrater Agreement

James M. LeBreton
Purdue University
Jenell L. Senter
Wayne State University

The use of interrater reliability (IRR) and interrater agreement (IRA) indices has increased dramatically during the past 20 years. This popularity is, at least in part, because of the increased role of multilevel modeling techniques (e.g., hierarchical linear modeling and multilevel structural equation modeling) in organizational research. IRR and IRA indices are often used to justify aggregating lower-level data used in composition models. The purpose of the current article is to expose researchers to the various issues surrounding the use of IRR and IRA indices often used in conjunction with multilevel models. To achieve this goal, the authors adopt a question-and-answer format and provide a tutorial in the appendices illustrating how these indices may be computed using the SPSS software.

Keywords: *interrater agreement; interrater reliability; aggregation; multilevel modeling*

As the use of multilevel modeling techniques has increased in the organizational sciences, the uses (and the potential for misuses) of interrater reliability (IRR) and interrater agreement (IRA) indices (often used in conjunction with multilevel modeling) have also increased. The current article seeks to provide answers to common questions pertaining to the use and application of IRR and IRA indices. Our hope is that this discussion will serve as a guide for researchers new to these indices and will help expand research possibilities to those already using these indices in their work.

Our article has three main objectives. First, we synthesize and integrate various definitional issues concerning the concepts of IRR and IRA and the indices most commonly used to assess these concepts. In doing so, we both recapitulate previous work and offer our own extensions and interpretations of this work. Second, we recognize that a number of provocative questions exist about the concepts of IRR and IRA and the primary indices used to assess these concepts. This is especially true of researchers being exposed to multilevel modeling for the first time. Thus, we also provide answers to some of the more common questions associated with using these indices when testing multilevel models. Some of these questions have been previously addressed, whereas some have not. The purpose of

Authors' Note: We would like to thank Paul Bliese, Rob Ployhart, and three anonymous reviewers for their constructive comments and feedback on earlier versions of this article. An earlier version of this article was presented at the 66th annual meeting of the Academy of Management in Atlanta, Georgia. Correspondence concerning this article should be addressed to James M. LeBreton, Department of Psychological Sciences, Purdue University, 703 Third St., West Lafayette, IN 47907-2081; e-mail: lebreton@psych.purdue.edu.

the article is to draw together, in a single resource, answers to a number of common questions pertaining to the use of IRR and IRA indices. Finally, we demonstrate the principles discussed in our answers via empirical tutorials contained in an appendix. The purpose of the last objective is to provide new researchers with concrete examples that will enable them to integrate their conceptual grasp of IRR and IRA with the technical skills necessary to answer their research questions (i.e., guidance using SPSS software). All of the data analyzed in the current article are presented in the appendix and are also available from either of the authors.

Definitional Questions About IRR and IRA

What is meant by IRR and IRA, and how are these concepts similar to and different from one another? How are IRR and IRA related to discussions of multilevel modeling? Such questions are often asked by researchers, both faculty and students, who are undertaking their first multilevel project. How one goes about answering these questions has a profound impact on (a) the approach one takes when estimating IRR and IRA, (b) the conclusions one will draw about IRR and IRA, and (c) the appropriateness of conducting a multilevel analysis. Thus, we address these definitional questions below. Throughout our article, we use the following notation:

X = an observed score, typically measured on an interval scale of measurement,
S_X^2 = the observed variance on X,
J = the number of items ranging from $j = 1$ to J,
K = the number of raters or judges ranging from $k = 1$ to K, and
N = the number of targets ranging from $i = 1$ to N.

Question 1: What is meant by IRR and IRA, and how are these concepts similar to and different from one another?

IRR refers to the relative consistency in ratings provided by multiple judges of multiple targets (Bliese, 2000; Kozlowski & Hattrup, 1992; LeBreton, Burgess, Kaiser, Atchley, & James, 2003). Estimates of IRR are used to address whether judges rank order targets in a manner that is relatively consistent with other judges. The concern here is not with the equivalence of scores but rather with the equivalence of relative rankings. In contrast, IRA refers to the absolute consensus in scores furnished by multiple judges for one or more targets (Bliese, 2000; James, Demaree, & Wolf, 1993; Kozlowski & Hattrup, 1992; LeBreton et al., 2003). Estimates of IRA are used to address whether scores furnished by judges are interchangeable or equivalent in terms of their absolute value.

The concepts of IRR and IRA both address questions concerning whether or not ratings furnished by one judge are "similar" to ratings furnished by one or more other judges (LeBreton et al., 2003). These concepts simply differ in how they go about defining interrater similarity. Agreement emphasizes the interchangeability or the absolute consensus between judges and is typically indexed via some estimate of within-group rating dispersion. Reliability emphasizes the relative consistency or the rank order similarity between judges and is typically indexed via some form of a correlation coefficient. Both IRR and

IRA are perfectly reasonable approaches to estimating rater similarity; however, they are designed to answer different research questions. Consequently, researchers need to make sure their estimates match their research questions.

Question 2: How are IRR and IRA related to discussions of multilevel modeling?

The basic idea underlying multilevel modeling is that there are variables measured at different levels of analysis (e.g., individuals, work groups, work divisions, different organizations) that affect dependent variables, typically measured at the lowest level of analysis (e.g., individuals). In some instances, the higher-level variables are actually measured at a higher level of analysis (e.g., organizational net profits). However, in other instances, higher-level variables are composites of lower-level variables (e.g., aggregated individual-level measures of affect used to measure group affective tone; George, 1990).

Depending on the theoretical nature of the aggregated construct, it may (or may not) be necessary to demonstrate that the data collected at a lower level of analysis (e.g., individual-level climate perceptions) are similar enough to one another prior to aggregating those data as an indicator of a higher-level construct (e.g., shared climate perceptions within work teams). For example, Kozlowski and Klein (2000) discussed two approaches to bottom-up processing (where individual- or lower-level data are combined to reflect a higher-level variable): composition and compilation approaches. Chan (1998) and Bliese (2000) reviewed various composition and compilation models and concluded that IRA and IRR are important when using composition models but less so for compilation models.

Compilation processes rest on the assumption that there are apparent differences between aggregated and nonaggregated data. Therefore, it is not necessary that individual- or lower-level data demonstrate consensus prior to aggregation. For example, additive models rely on a simple linear combination of lower-level data and do not require the demonstration of within-group agreement (Chan, 1998). In contrast, composition processes are often based on the assumption that individual- or lower-level data are essentially equivalent with the higher-level construct. Therefore, to justify aggregating lower-level data to approximate a higher-level construct, it is necessary to demonstrate that the lower-level data are in agreement with one another (e.g., individuals within a work group have highly similar or interchangeable levels of affect that are different from individuals' affect levels in another work group, and, thus, each work group has a unique affective tone). Because such composition models focus on the interchangeability (i.e., equivalence) of lower-level data, estimates of IRA are often used to index the extent of agreement, or lack thereof, among lower-level observations. The equivalence of lower-level data may be demonstrated via estimates of IRA or IRR + IRA. When only a single target is assessed, the empirical support needed to justify aggregation may be acquired via IRA indices such as r_{WG} (e.g., direct consensus models and referent-shift consensus models; Chan, 1998). When multiple targets are assessed, the empirical support needed to justify aggregation may be acquired via IRA indices such as r_{WG} and via IRR + IRA indices such as intraclass correlation coefficients (ICCs). In sum, when lower-level data are aggregated to form a higher-level variable, estimates of IRA or IRR + IRA are often invoked to aid in justifying this aggregation.

Question 3: Okay, so how do I figure out which form of interrater similarity is relevant to my research question?

The form of interrater similarity used to justify aggregation in multilevel modeling should depend mainly on one's research question and the type of data that one has collected. Estimates of IRA tend to be more versatile because they can be used with one or more targets, whereas estimates of IRR or IRR + IRA necessitate having multiple targets (e.g., organizations). However, it should be mentioned that because our discussion pertains to multilevel modeling and the need to provide sufficient justification for aggregation, estimates of both IRA and IRR + IRA are typically used. This is because justification of aggregating lower-level data is predicated on the consensus (i.e., interchangeability) among judges furnishing scores on these lower-level data, and estimates of IRR only measure consistency. Consequently, pure measures of IRR are rarely used in multilevel modeling because justification of aggregation is typically not predicated on the relative consistency of judges' ratings irrespective of their absolute value. The remainder of our article addresses questions primarily associated with estimating IRA or IRR + IRA.

Question 4: What are the most commonly used techniques for estimating IRA, IRR, and IRR + IRA?

Measures of IRA

r_{WG} *indices.* Table 1 summarizes the most commonly used indices of IRA, IRR, and IRR + IRA. Arguably, the most popular estimates of IRA have been James, Demaree, and Wolf's (1984, 1993) single-item r_{WG} and multi-item $r_{WG(J)}$ indices. The articles introducing these indices have been cited more than 700 times in fields ranging from strategic management to nursing. When multiple judges rate a single target on a single variable using an interval scale of measurement, IRA may be assessed using the r_{WG} index, which defines agreement in terms of the proportional reduction in error variance,

$$r_{WG} = 1 - \frac{S_X^2}{\sigma_E^2}, \tag{1}$$

where S_X^2 is the observed variance on the variable X (e.g., leader trust and support) taken over K different judges or raters and σ_E^2 is the variance expected when there is a complete lack of agreement among the judges. This is the variance obtained from a theoretical null distribution representing a complete lack of agreement among judges. As discussed under Questions 9 and 10, determining the shape of this distribution is one of the factors that most complicates the use of r_{WG}. Basically, it is the variance one would expect if all of the judges responded randomly when evaluating the target. Thus, it is both a theoretical (i.e., it is not empirically determined) and conditional (i.e., assumes random responding) distribution.

The use of r_{WG} is predicated on the assumption that each target has a single true score on the construct being assessed (e.g., leader trust and support). Consequently, any variance in judges' ratings is assumed to be error variance. Thus, it is possible to index

Table 1
Indices Used to Estimate Interrater Agreement (IRA), Interrater Reliability (IRR), and IRR + IRA

Form of Similarity	Index	Primary References
IRA	r_{WG}, $r_{WG(J)}$	James, Demaree, and Wolf (1984) James, Demaree, and Wolf (1993)
	r^*_{WG}, $r^*_{WG(J)}$	Lindell, Brandt, and Whitney (1999) Lindell and Brandt (1999) Lindell (2001)
	r_{WGp}, $r_{WGp(J)}$	LeBreton, James, and Lindell (2005) Current article (Question 7)
	SD_X, SE_M	Schmidt and Hunter (1989)
	AD_M, $AD_{M(J)}$, AD_{Md}, $AD_{Md(J)}$	Burke, Finkelstein, and Dusig (1999) Burke and Dunlap (2002)
	a_{WG}, $a_{WG(J)}$	Brown and Hauenstein (2005)
IRR	Pearson correlation	Kozlowski and Hattrup (1992) Schmidt, Viswesvaran, and Ones (2000)
IRR + IRA	*ICC(1)*, *ICC(K)*, *ICC(A,1)*, *ICC(A,K)*	McGraw and Wong (1996)

agreement among judges by comparing the observed variance to the variance expected when judges respond randomly. Basically, when all judges are in perfect agreement, they assign the same rating to the target, the observed variance among judges is 0, and $r_{WG} = 1.0$. In contrast, when judges are in total lack of agreement, the observed variance will asymptotically approach the error variance obtained from the theoretical null distribution as the number of judges increases. This leads r_{WG} to approach 0.0.

Such lack of agreement has typically assumed to be generated by a uniform (i.e., equal probability or rectangular) distribution (LeBreton et al., 2003; Schriesheim et al., 2001); however, James et al. (1984) encouraged researchers to also model other distributions, such as those that would be caused by response biases (e.g., leniency bias, central tendency bias). Issues pertaining to choosing null distributions will be discussed in greater detail under Question 10. Returning to Equation 1, S^2_X/σ^2_E represents the proportion of observed variance that is error variance caused by random responding. Consequently, r_{WG} may be interpreted as the proportional reduction in error variance.

This index has been extended to situations where a single target is rated by multiple raters on $j = 1$ to J essentially parallel items. The multi-item $r_{WG(J)}$ index is estimated by

$$r_{WG(J)} = \frac{J\left(1 - \frac{\overline{S}^2_{X_j}}{\sigma^2_E}\right)}{J\left(1 - \frac{\overline{S}^2_{X_j}}{\sigma^2_E}\right) + \left(\frac{\overline{S}^2_{X_j}}{\sigma^2_E}\right)}, \tag{2}$$

where $\overline{S}^2_{X_j}$ is the mean of the observed variances for J essentially parallel items and σ^2_E has the same meaning as above. Within the context of multilevel modeling, the r_{WG} and $r_{WG(J)}$ indices have been used by researchers to justify aggregating lower-level data

(e.g., individual affect) to represent a higher-level construct (e.g., group affective tone; George, 1990).

Standard deviation. Schmidt and Hunter (1989) critiqued the r_{WG} and $r_{WG(J)}$ indices, largely based on semantic confusion arising from earlier writers' labels of the r_{WG} indices as reliability coefficients (James et al., 1984) versus agreement coefficients (James et al., 1993; Kozlowski & Hattrup, 1992). Their primary concern with r_{WG} was that it was not conceptually anchored in classical reliability theory. Although this was an accurate statement, it is not necessarily a limitation of the r_{WG} indices because they are not reliability coefficients. In any event, Schmidt and Hunter recommended that when researchers seek to assess agreement among judges on a single target, researchers should estimate the standard deviation of ratings and the standard error of the mean rating,

$$SD_X = \sqrt{\sum_{k=1}^{K} \frac{\left(X_k - \overline{X}\right)^2}{K-1}} \tag{3}$$

$$SE_M = \frac{SD_X}{\sqrt{K}}, \tag{4}$$

where $k = 1$ to K judges, X_k is the kth judge's rating on X, and $\overline{X}$ is the mean rating on X taken over the K judges. These authors advocated using SD_X to index agreement and using the SE_M to construct 95% confidence intervals around the mean rating to assess the amount of error in the judges' mean rating. Kozlowski and Hattrup (1992) rejected this approach to estimating agreement because the SE_M is heavily dependent on the number of judges and because the Schmidt and Hunter approach failed to account for the level of agreement that could occur by chance.

We concur with other researchers that the sensitivity of the SE_M to sample size limits is usefulness as a measure of rating consensus (Lindell & Brandt, 2000; Schneider, Salvaggio, & Subirats, 2002). We also concur with these researchers that the SD_X is most appropriately conceptualized as a measure of interrater *dispersion* or disagreement (see also Roberson, Sturman, & Simons, in press). Consequently, this index is ideally suited for testing dispersion composition models (Chan, 1998; Schneider et al., 2002) but is not necessarily an optimal index of agreement.

*Average deviation (*AD*) indices.* The *AD* index has been proposed by Burke, Finkelstein, and Dusig (1999) as another measure of IRA. This measure, like r_{WG}, was developed for use with multiple judges rating a single target on a variable using an interval scale of measurement. These authors described this index as a "pragmatic" index of agreement because it estimates agreement in the metric of the original scale of the item. We concur. The *AD* index may be estimated around the mean (AD_M) or median (AD_{Md}) for a group of judges rating a single target on a single item:

$$AD_{M(j)} = \frac{\sum_{k=1}^{K} \left|X_{jk} - \overline{X_j}\right|}{K} \tag{5}$$

$$AD_{Md(j)} = \frac{\sum_{k=1}^{K} \left| X_{jk} - Md_j \right|}{K}, \tag{6}$$

where $k = 1$ to K judges, X_{jk} is the kth judge's rating on the jth item, and $\overline{X}_j$ (Md_j) is the item mean (median) taken over judges. Burke et al. noted that the use of AD for medians may be a more robust test. Similar to $r_{WG(J)}$, AD can be calculated for J essentially parallel items rated by K raters as follows:

$$AD_{M(J)} = \frac{\sum_{j=1}^{J} AD_{M(j)}}{J} \tag{7}$$

$$AD_{Md(J)} = \frac{\sum_{j=1}^{J} AD_{Md(j)}}{J}, \tag{8}$$

where all terms are as defined above and $j = 1$ to J essentially parallel items.

a_{WG} *indices.* Finally, the most recently suggested estimate of IRA, a_{WG}, was derived by Brown and Hauenstein (2005) to address limitations they suggested with the family of r_{WG} indices (especially when the uniform null distribution was used to estimate r_{WG}). First, they argued that the r_{WG} indices are scale dependent in that the lower bound of any r_{WG} index will be dependent on the number of scale anchors. For example, the authors contended that results using r_{WG} will be different depending on whether a researcher used a 5-, 7-, or 9-point Likert-type scale. Second, they suggested that the sample size (i.e., number of judges) influences the values of r_{WG}, which, consequently, influences the interpretability of the results. This sample size dependency results from the use of the ratio of observed sample variance to the population variance for the null distribution. Third, they noted that researchers inaccurately assume that the null distribution is valid, an issue discussed at great length within the current article. With consideration mainly to the issue of the null distribution often being invalid, Brown and Hauenstein developed the a_{WG} index.

This index borrows from the logic of J. Cohen's (1988) kappa, which estimates agreement by computing a ratio of the percentage of cases agreeing minus a null agreement standard to 1 minus the null agreement standard. Brown and Hauenstein (2005) extended Cohen's kappa to the single target situation. As evident in Equation 9, the logic underlying the a_{WG} index is more complicated than that of either the r_{WG} or AD indices. In short, a_{WG} may be computed for multiple judges rating a single target using an interval scale of measurement as

$$a_{WG} = 1 - \frac{2 * S_X^2}{\left[(H + L) * \overline{X} - (\overline{X}^2) - (H * L)\right] * [K/(K-1)]}, \tag{9}$$

where $\overline{X}$ is the observed mean rating taking over judges, H is the maximum possible value of the scale, L is the minimum possible value of the scale, K is the number of judges, and S_X^2 is the observed variance on X. As with other agreement indices, 1.0 indicates perfect

agreement among judges. Brown and Hauenstein (2005) noted that the main difference between r_{WG} and a_{WG} is that the former will stay constant regardless of the judges' mean rating, but the latter will vary dependent on the mean.

Similar to the $r_{WG(J)}$ and $AD_{M(J)}$ indices, a multi-item version of a_{WG} exists when J essentially parallel items are rated by K judges:

$$a_{WG(J)} = \frac{\sum_{j=1}^{J} a_{WG(j)}}{J}. \tag{10}$$

Measures of IRR

The most popular measure of IRR has been the Pearson product-moment correlation calculated by correlating ratings between judges over multiple targets (Schmidt, Viswesvaran, & Ones, 2000). The use of Pearson product-moment correlations as measures of IRR has been the source of recent debate. As noted earlier, pure measures of IRR are not commonly used in multilevel research, thus we will not go into great detail discussing these indices. Readers interested in learning more about the use of correlations to index IRR are directed to articles by Viswesvaran, Ones, and Schmidt (1996), Murphy and DeShon (2000a, 2000b), Schmidt et al. (2000), LeBreton et al. (2003), and, most recently, Viswesvaran, Schmidt, and Ones (2005).

Measures of IRA + IRR

Intraclass correlations—individual raters. Although most researchers acknowledge that ICCs furnish information about IRR (Bliese, 2000; James, 1982), few researchers have acknowledged that many of the ICCs used in multilevel modeling actually furnish information about IRR + IRA (LeBreton et al., 2003). Specifically, the one-way random effects ICCs and two-way random effects or mixed effects ICCs measuring "absolute consensus" are technically a function of both absolute rater consensus (i.e., IRA) and relative rater consistency (i.e., IRR; LeBreton et al., 2003; McGraw & Wong, 1996). In general, ICCs may be interpreted as the proportion of observed variance in ratings that is due to systematic between-target differences compared to the total variance in ratings. Within the context of multilevel modeling, the ICC based on the one-way random effects ANOVA is the most common estimate of IRR + IRA. In this case, the targets (e.g., organizations, departments, teams, supervisors) are treated as the random effect. This ICC is estimated when one is interested in understanding the IRR + IRA among multiple targets (e.g., organizations) rated by a different set of judges (e.g., different employees in each organization) on an interval measurement scale (e.g., Likert-type scale). This index has been differently labeled by different researchers. In the current article, we adopt the notation of McGraw and Wong (1996),

$$ICC(1) = \frac{MS_R - MS_W}{MS_R + (K-1)MS_W}, \tag{11}$$

where MS_R is the mean squares for rows (i.e., targets) and MS_W is the mean square within calculated from a one-way random effects ANOVA and K refers to the number of observations (e.g., ratings or judges) per target. Because this index simultaneously measures IRA and IRR, high values may only be obtained when there is both absolute consensus and relative consistency in judges' ratings. In contrast, low values may be obtained when there is low consensus, low consistency, or both (LeBreton et al., 2003). *ICC(1)* values may be interpreted as the level of consensus + consistency one would expect if a judge was randomly selected from the population of judges and his or her scores were compared to the mean score (i.e., estimated true score) obtained from the sample of judges (Bliese, 2000; James, 1982). *ICC(1)* values may also be interpreted as an effect size estimate revealing the extent to which judges' ratings were affected by the target (e.g., the extent that employee ratings of their organization's climate are affected by their membership in their organization; Bliese, 2000; Hofmann, Griffin, & Gavin, 2000).

In some instances, researchers may have multilevel data where each of the targets is rated by the same set of judges. For example, 100 job applicants may have completed a cognitive ability test, personality inventory, biodata survey, and a structured panel interview. If all applicants were assessed by the same panel of interviewers, then one might want to justify aggregating interviewer scores and then use these aggregate scores as a level 2 predictor of job performance (along with the level 1 predictors of cognitive ability, personality, and life history information). Aggregating interviewer ratings could be justified by estimating IRA using r_{WG} (i.e., calculate one r_{WG} for each of the 100 job applicants) or by estimating IRR + IRA using a two-way ICC. If the researcher were interested in generalizing to other judges, then judges would be treated as a random effects variable, and he or she would calculate the ICC using the two-way random effects ANOVA (where both the target and judge effects are random effects). If the researcher were not interested in generalizing to other judges, then judges would be treated as a fixed effects variable, and he or she would calculate the ICC using a two-way mixed effects ANOVA (where the target effect is a random effect and the judge effect is a fixed effect). Both ICCs are estimated as

$$ICC(A,1) = \frac{MS_R - MS_E}{MS_R + (K-1)MS_E + \frac{K}{N}(MS_C - MS_E)}, \tag{12}$$

where MS_R is the mean square for rows (i.e., targets), MS_C is the mean square for columns (i.e., judges), MS_E is the mean square error all obtained from a two-way ANOVA, and K refers to the number of observations (e.g., ratings or judges) for each of the N targets. The procedure for interpreting *ICC(A,1)* values is the same as for *ICC(1)* values (i.e., reliability of individual judges' ratings or an estimate of effect size).

Intraclass correlations—group mean rating. The above ICCs have been used to estimate the reliability of a single judge or rater; however, in multilevel modeling, researchers are often more interested in understanding the extent to which the mean rating assigned by a group of judges is reliable. In such cases, an ICC may be calculated that estimates the stability (i.e., reliability) of mean ratings furnished from K judges. When each target is rated

by a different set of judges on an interval scale of measurement, the average score ICC may be estimated using a one-way random effects ANOVA (where the target effect is the random effect),

$$ICC(K) = \frac{MS_R - MS_W}{MS_R}, \tag{13}$$

where K refers to the number of judges, MS_R is the mean squares for rows (i.e., targets), and MS_W is the mean square within calculated from a one-way random effects ANOVA. This index has also been labeled the *ICC(2)* (Bartko, 1976; Bliese, 2000; James, 1982) and the *ICC(1,K)* (Shrout & Fleiss, 1979); however, we retain the labels used by McGraw and Wong (1996).

When each target is rated by the same set of judges, the average score ICC may be estimated using a random (or mixed effects) two-way ANOVA,

$$ICC(A,K) = \frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{N}}, \tag{14}$$

where N is the number of targets, MS_R is the mean square for rows (i.e., targets), MS_C is the mean square for columns (i.e., judges), and MS_E is the mean square error obtained from a two-way ANOVA.

Basically, the values obtained for the *ICC(K)* and *ICC(A,K)* asymptotically approach the values one would obtain by placing the *ICC(1)* and the *ICC(A,1)* into the Spearman-Brown prophesy formula (the correction factor being K, the number of raters; Bliese, 1998, 2000). Like their single-judge counterparts, these average rating ICCs simultaneously assess IRR + IRA and may be interpreted as the IRR + IRA of a group's *mean rating*. Thus, if a new sample of targets were evaluated by a new set of K raters, then the level of IRR + IRA between the two sets of means would be approximately equal to *ICC(K)* or *ICC(A,K)* (James, 1982; LeBreton et al., 2003).

Methodological Questions Concerning the Estimation of IRA

Question 5: How do I know which type of IRA index I should use?

The particular index a researcher uses to estimate IRA is largely a matter of personal preference. These estimates of agreement tend to yield highly convergent results, which is not surprising given the similarity among their computational equations. Specifically, most of the indices are a function of each judge's deviation from the mean (or median) rating taken over judges. Whether those are absolute deviations or squared deviations varies across indices (see previous equations), but because they are all a function of rater deviations, they tend to be highly correlated with one another.

For example, Burke et al. (1999) showed that estimates of agreement calculated using the r_{WG} and *AD* indices tended to be highly correlated (often in the .90s). Brown and Hauenstein (2005) showed that, in many instances, r_{WG} and a_{WG} yielded highly similar estimates of agreement. Most recently, Roberson et al. (in press) conducted a very large

and comprehensive Monte Carlo simulation that compared various measures of IRA or dispersion. They found highly convergent results among the indices. Specifically, the average (absolute value) of the correlations among SD_X, AD, r_{WG}, and a_{WG} was .95, with a range of .93 to .98.

In sum, because all of the measures of IRA described in the current article tend to yield highly consistent conclusions, we do not see a reason to advocate one particular index over another. Instead, we actually encourage researchers to use multiple indices to aid in interpreting their data. r_{WG} and a_{WG} have the advantage of indexing agreement on a scale ranging from 0 to 1.0. AD has the advantage of indexing agreement in the metric of the original scales. It also does not require the specification of different null response distributions (although it can incorporate multiple nulls). Interpretation of multiple indices may help researchers better understand the consensus (or lack thereof) in their data. The SD_X measure is really not used as a measure of agreement but more typically as a measure of dispersion (Lindell & Brandt, 2000; Roberson et al., in press), and thus we encourage its use with dispersion composition models (Chan, 1998).

Question 6: There are many different variants of the r_{WG} index (e.g., r_{WG}, $r_{WG(J)}$, r^*_{WG}, $r^*_{WG(J)}$, r_{WGp}, $r_{WGp(J)}$); how do I know which form is correct for my multilevel analysis?

The r_{WG} and $r_{WG(J)}$ indices were initially introduced by James et al. (1984). As discussed in our answer to Question 4, these indices were designed to measure IRA by comparing the observed variance in ratings furnished by multiple judges of a single target to the variance one would expect when the judges responded randomly. If judges were in perfect agreement (i.e., all judges "see the same thing"), the observed variance would be equal to 0 (i.e., $S^2_X = 0$), and r_{WG} would be equal to 1.0, denoting perfect agreement (i.e., $r_{WG} = 1 - \frac{S^2_X}{\sigma^2_E} = 1 - \frac{0}{\sigma^2_E} = 1$). In contrast, if judges' ratings were basically due to random responding (i.e., judges lacked any level of agreement), the observed variance would approximate the variance based on the theoretical null error distribution (i.e., $S^2_X = \sigma^2_E$), and r_{WG} would be equal to 0, denoting perfect lack of agreement ($r_{WG} = 1 - \frac{S^2_X}{\sigma^2_E} = 1 - \frac{\sigma^2_E}{\sigma^2_E} = 0$). The same applies for $r_{WG(J)}$.

Thus, these indices normally assume values ranging from 0 (*perfect lack of agreement*) to 1 (*perfect agreement*). However, when computing r_{WG} and $r_{WG(J)}$, it is possible for researchers to obtain out-of-range values (i.e., values less than 0 or greater than 1). Such values occur when the observed variance exceeds the theoretical null variance (i.e., $S^2_X > \sigma^2_E$ or $\bar{S}^2_X > \sigma^2_E$). James et al. (1984) assumed that these out-of-range values were because of sampling error and that they should simply be reset to 0 to indicate a complete lack of agreement. The specific issues that give rise to out-of-range values are discussed in greater detail under Question 7.

Instead of resetting out-of-range values to 0, Lindell and his colleagues (Lindell & Brandt, 1999; Lindell, Brandt, & Whitney, 1999) recommended that researchers calculate agreement using alternative indices of agreement, denoted r^*_{WG} and $r^*_{WG(J)}$. These indices are similar to the original indices proposed by James et al. (1984). In fact, r^*_{WG} is calculated using the same equation as r_{WG} but is allowed to assume negative values rather than

have these values reset to 0. Consequently, r_{WG} typically is equal to r^*_{WG} because most values of r_{WG} are positive or 0. In contrast, the equation for $r^*_{WG(J)}$ results in greater divergences with $r_{WG(J)}$:

$$r*_{WG(J)} = 1 - \frac{\overline{S}^2_{X_j}}{\sigma^2_E}, \tag{15}$$

where $\overline{S}^2_{X_j}$ is the mean of the observed variances for J essentially parallel items and σ^2_E is the expected variance if all judges responded randomly. The motivation to create $r^*_{WG(J)}$ stemmed from a desire to create an estimate of IRA that could be used when maximum disagreement existed among the judges (i.e., when the distribution of judges' ratings is bimodal, with half of the ratings occurring in the lowest rating category—a 1 on a 5-point scale—and half occurring in the highest rating category number—a 5 on a 5-point scale). Such a bimodal distribution suggests systematic disagreement, which causes r_{WG} and $r_{WG(J)}$ values to become negative or exceed unity (i.e., observed variance because of systematic "disagreement" is greater than the theoretically expected variance because of "lack of agreement").

Another way of conceptualizing this bimodal phenomenon is to consider that a target could have multiple true scores. For example, the theory of leader-member exchange posits that, because of limited time and resources, a leader's subordinates tend to cluster into an "in-group" (typified by high levels of trust, interaction, support, loyalty, and rewards) and an "out-group" (typified by low levels of trust, interaction, support, loyalty, and rewards; Dienesch & Liden, 1986). In such an instance, it is possible that the leader has different true scores on a measure of leader trust and support, conditional on whether he or she is being rated by in-group or out-group members. When researchers believed that systematic disagreement was present among judges, $r^*_{WG(J)}$ offered a way to capture this disagreement.

As discussed below under Question 7, LeBreton, James, and Lindell (2005) recently noted that the structural equation that is the basis for r_{WG} and $r_{WG(j)}$ (and r^*_{WG} and $r^*_{WG(J)}$) does not permit a target to have multiple true scores. Consequently, the solution offered by Lindell et al. (1999) may not be as applicable to conditions of disagreement as initially thought. LeBreton et al. introduced a new index, r_{WGp}, which permits a target to have multiple true scores, one for each subgroup of raters (e.g., in-group vs. out-group); however, subgroups must be identified a priori, or else researchers run the risk of capitalizing on chance. This index was denoted r_{WGp} to emphasize that it is based on a *pooled* within-groups variance (i.e., the average of the variances obtained for each subgroup) and is estimated as

$$r_{WGp} = 1 - \frac{S^2_{X.\tau}}{\sigma^2_E}, \tag{16}$$

where $S^2_{X.\tau}$ refers to the pooled variance within groups (i.e., the variance after removing the "treatment effect," denoted τ, that distinguishes the true scores between the subgroups), and σ^2_E is the expected variance if all judges randomly responded. The pooled within-groups variance may be obtained either by computing the weighted averages of the individual variances or from the within-groups mean square from an analysis of variance (LeBreton et al., 2005).

LeBreton et al. (2005) concluded that "computing r_{WGp} in this manner eliminates the need for a distinction between r_{WG} and r^*_{WG} and, more significantly, eliminates the problem of inadmissible $r_{WG(J)}$ values reported by Lindell and Brandt (1997)" (p. 135). We concur. In short, if one does not have a priori reasons to believe that a target would have multiple true scores corresponding to multiple subgroups of raters, one should use the traditional r_{WG} or $r_{WG(J)}$ indices. However, if one has an a priori reason to believe that multiple true scores and subgroups exist, one is encouraged to estimate agreement using r_{WGp}.

Question 7: I thought that the estimates obtained using r_{WG} and $r_{WG(J)}$ were supposed to range between 0 and 1, but sometimes I observe negative values or values that exceed unity. What should I do when I obtain these out-of-range values?

In some instances, researchers may observe out-of-range r_{WG} and $r_{WG(J)}$ and values (cf. James et al., 1984; LeBreton et al., 2005; Lindell & Brandt, 1997; Lindell et al., 1999). This occurs because the observed variance exceeds the expected variance for a random response null distribution. In such instances, James et al. (1984) recommended resetting out of range values to 0. However, this recommendation is predicated on the assumption that the out-of-range values are attributed to sampling error. That is, they assumed that small negative values (e.g., $-.20 < r_{WG} < .00$; LeBreton et al., 2005) were obtained simply because of sampling error and that, with a larger sample of judges, the values would become proper (i.e., conform to the 0 to 1 range). However, there are at least two other explanations for why one may observe out-of-range values (using either a uniform null distribution or a null distribution containing a response bias).

First, out-of-range values could be obtained when the random response null distribution has been incorrectly specified. For example, if we used the variance associated with a slightly skewed null distribution when the true underlying distribution was uniform, we have a downwardly biased estimate of error variance, which could result in negative estimates of agreement. Second, LeBreton et al. (2005) noted that when a target has multiple true scores (e.g., in-group vs. out-group ratings of leader trust and support), the observed variance in judges' ratings could become multimodal, and such a distribution could easily engender out-of-range estimates of r_{WG} and $r_{WG(J)}$. In such instance, the alternative r_{WGp} index may be more appropriate (for more information, see response to Question 6 above) or its multi-item extension, which we present as

$$r_{WGp_{(J)}} = \frac{J\left(1 - \frac{\overline{S}^2_{X.\tau_{(J)}}}{\sigma^2_E}\right)}{J\left(1 - \frac{\overline{S}^2_{X.\tau_{(J)}}}{\sigma^2_E}\right) + \left(\frac{\overline{S}^2_{X.\tau_{(J)}}}{\sigma^2_E}\right)}, \tag{17}$$

where $\overline{S}^2_{X.\tau_{(J)}}$ is calculated as the average of the pooled within-groups variances across the J essentially parallel items and σ^2_E is the expected variance when judges responded randomly. We hasten to note that, similar to the single-item counterpart, the multi-item $r_{WGp(J)}$ index requires homogeneity of error variances. If this assumption cannot be met, researchers are encouraged to compute separate $r_{WG(J)}$ indices for each group.

Because the values of r_{WG} and $r_{WG(J)}$ range from 0 (*complete lack of agreement*) to 1 (*complete agreement*), negative values or values exceeding unity lack psychometric meaning. We concur with James et al. (1984) that if one believes the negative values were solely because of sampling error, they should be reset to 0—indicating a complete lack of agreement. However, we would encourage researchers to also consider the possibility that the negative values were obtained because of a misspecified null distribution or because the set of judges actually contains two subsets of judges, each assigning the target a different true score. If the problem is likely because of a misspecified null distribution, then the researcher should consider modeling alternative, *theoretically defensible* null distributions and use these to calculate agreement (see Questions 9 and 10 below). If the problem is likely because of having multiple subgroups of judges (each assigning a different true score to the target), then the researcher should consider estimating agreement using the r_{WGp} index (LeBreton et al., 2005). However, we echo the concerns raised by these authors that these subgroups should be identified a priori (e.g., based on gender, race, job types, work team, education level, leader member exchange ratings, etc.) to minimize capitalizing on chance in defining the groups.

Question 8: How many judges for items do I need when I calculate r_{WG} or $r_{WG(J)}$?

The number of judges is one factor that has been found to affect the magnitude of the family of r_{WG} indices. James et al. (1984) and Lindell et al. (1999) suggested when the number of judges (i.e., the sample size) is small, r_{WG} values are attenuated. This is especially true when the agreement between these judges is not high. Lindell et al. (1999) and Kozlowski and Hattrup (1992) suggested that 10 or more judges should be used to prevent attenuation. A. Cohen, Doveh, and Eick (2001) found that both the number of judges and the number of items should be considered when estimating $r_{WG(J)}$. For example, in one instance, Cohen et al. set the population value of $r_{WG(J)} = .75$ for $J = 6$. Using a simulation study, they found that with $K = 10$ judges, only 63% of the simulated samples resulted in $r_{WG(J)}$ values $> .70$. However, when they increased the number of judges to $K = 100$, the number of $r_{WG(J)}$ values $> .70$ increased to 84%. If instead of increasing judges, the number of items were increased to $J = 10$, then the number of $r_{WG(J)}$ values $> .70$ would increase to 81%.

Basically, adding either more judges or more items tends to increase the magnitude of $r_{WG(J)}$ values (at least when the true level of agreement is nonzero). LeBreton et al. (2005) recently showed why adding items increases estimates of IRA. Using an independent derivation of the agreement among judges on a linear composite of their judgments, they showed that the Spearman-Brown prophecy formula is not restricted to reliability indices but is also applicable to agreement indices. These authors showed that it is mathematically acceptable to input r_{WG} into the Spearman-Brown prophecy formula to estimate $r_{WG(J)}$, with the correction factor being J, the number of essentially parallel items. In short, adding items enhances the internal consistency (i.e., reliability) of the composite, thus producing a reduction in error variance. As a result, values of $r_{WG(J)}$ will tend to be larger than the estimates of agreement based on $r^*_{WG(J)}$ because the latter index does not adjust for the attenuation in the observed level of agreement that is due to the unreliability in the ratings of the individual items (LeBreton et al., 2005; Lindell, 2001). This means that adding items,

everything else being equal, will result in higher values of $r_{WG(J)}$. Similarly, adding judges, everything else being equal, will also result in higher values of $r_{WG(J)}$. This is because given a constant pattern of ratings, the estimate of observed variance (S_X^2 or $\overline{S}_X^2$) will decrease as the sample size increases (Brown & Hauenstein, 2005).

One critical, yet often unrecognized, issue concerns the pragmatics of adding truly parallel judges to estimate IRA (e.g., $r_{WG(J)}$, $AD_{M(J)}$). Judges are said to be parallel when their target true scores are identical and their error scores are identically distributed with constant variance. Consequently, in organizational research, the likelihood of having a large number of truly parallel judges may be relatively slim because, at some point, the finite pool of parallel judges will be exhausted, and, thus, many of the additional judges may tend to be less qualified (e.g., less knowledgeable, less accurate, etc.). Stated alternatively, the IRA indices reviewed in this article assume that judges are parallel; however, it is possible that although some of the judges provide truly parallel measures, others may only provide tau-equivalent or congeneric measures (Lord & Novick, 1968). A consequence of this lack of truly parallel judges is that estimates of agreement will likely be attenuated.

This fact, coupled with the findings of A. Cohen et al. (2001), suggests there is probably a trade-off between the number of judges and the number items needed to establish adequate levels of IRA. Fewer judges can be compensated for by greater numbers of items, and vice versa. In sum, in many instances, it appears that 10 judges will be sufficient for estimating r_{WG} or the $r_{WG(J)}$; however, we strongly encourage researchers to consider the number of items and the number of judges in their samples when making inferences about agreement.

Question 9: Everyone seems to use the uniform null distribution when estimating agreement using the r_{WG} indices, and it gives me the highest estimate of agreement, so why would I want to use another distribution?

To calculate the r_{WG} indices, an estimate of the expected variance when there is a total lack of agreement is needed. This estimate is based on a null distribution that represents a total lack of agreement. *Choosing the null distribution is the single greatest factor complicating the use of* r_{WG}*-based indices*. This distribution is a conditional and theoretical distribution. Basically, the researcher asks, "If raters responded randomly, then what would be the form of the distribution of scores?"

To date, the vast majority of researchers have relied on the uniform, or rectangular, null distribution (LeBreton et al., 2003; Schriesheim et al., 2001). For example, on a 5-point scale, each response option has an equal chance of being selected by a judge (i.e., 20% chance). Therefore, the distribution will be flat and rectangular. Thus, the uniform distribution is only applicable in situations where the scales are discrete and a complete lack of agreement would be hypothesized to be evenly distributed across response options (i.e., mathematically random). However, James et al. (1984) noted that in many instances, random responding will not correspond to a uniform distribution because judges' ratings could be affected by various response biases (e.g., leniency bias). In such instances, alternative distributions should be used to estimate the variance expected when judges randomly respond.

According to James et al. (1984), researchers should ask themselves, "If there is no true variance and agreement is zero, what distribution would best fit our response bias and

some measurement error?" There are several distributions commonly used for r_{WG}: triangular (corresponding to a central tendency bias), skewed (corresponding to leniency and severity biases), and uniform (corresponding to the lack of biases). James et al. specifically recommended estimating the r_{WG} indices using multiple, theoretically defensible null distributions. However, with a few notable exceptions, researchers largely ignored these recommendations and relied almost exclusively on the uniform null (Schriesheim et al., 2001).

We see two potential reasons for this preference for the uniform null. First, because the uniform null yields the largest estimate of error variance, it also yields the largest values of r_{WG}. Thus, there is a strong *disincentive* for using any of the alternative distributions because estimates of IRA will always be lower using such distributions. Second, the variance of a uniform null distribution is easily obtained:

$$\sigma^2_{EU} = \frac{(A^2 - 1)}{12}, \tag{18}$$

where A is the number of response options. Thus, for a 7-point scale, the expected null variance is $(7^2 - 1) \div 12 = 4$. Although estimating the variance of a uniform null distribution is easy and straightforward, no simple equation exists for estimating the variance of the alternative distributions. Thus, a second reason that researchers may have avoided using alternative null distributions was the lack of easily accessible estimates of error variances associated with these distributions.

Nevertheless, the uniform null distribution is only appropriate when one may assume that none of the judges' ratings were affected by biases. Given the extant literature demonstrating the pervasiveness of such cognitive and affective biases, this assumption will rarely (if ever) be fully met (cf. Baltes & Parker, 2000; Borman, 1991; Cardy & Dobbins, 1994; Cooper, 1981; Feldman, 1981; Fisicaro, 1988; Fisicaro & Lance, 1990; Fisicaro & Vance, 1994; Funder, 1987; Harris, 1994; Ilgen, Barnes-Farrell, & McKellin, 1993; Murphy & Balzer, 1989; Murphy & Cleveland, 1995; Sulsky & Balzer, 1988; Varma, DeNisi, & Peters, 1996). Consequently, reliance on the uniform null distribution will often yield inflated estimates of agreement (Brown & Hauenstein, 2005; James et al., 1984; Kozlowski & Hattrup, 1992; LeBreton et al., 2003). *Given the ubiquity of response biases in organizational research, we call for a moratorium on the unconditional (i.e., unjustified) use of any null distribution, especially the uniform null distribution*. Instead, we challenge each researcher to justify the use of a particular null distribution, uniform or otherwise, in his or her research study.

Question 10: Okay, you've convinced me that I should use multiple distributions, but how do I go about determining which distributions to use, and how do I estimate the variances for these other distributions?

As noted above, deciding which alternative null distributions to estimate r_{WG} should be based on theory. For example, LeBreton et al. (2003) examined the levels of agreement in ratings on executives furnished by various sources (e.g., self-ratings, peer ratings, subordinate ratings, supervisor ratings). These authors sought to examine the levels of agreement between and within rating sources and estimated agreement using the r_{WG} index. In addition to the

uniform or rectangular null distribution, they used references to the rating bias literature to justify null distributions based on leniency biases (i.e., slightly skewed and moderately skewed null distributions) and a central tendency bias (i.e., a quasi-normal distribution). For another example using climate data, see Kozlowski and Hults (1987).

After identifying theoretically defensible null distributions based on various response biases, it is necessary to calculate the expected variance associated with those distributions. To date, when researchers have used alternative null distributions to estimate σ_E^2, they have relied heavily on the values furnished by James et al. (1984) for 5-point scales having leniency or severity biases (i.e., skewed distributions) or a central tendency bias (i.e., a triangular distribution). To facilitate researchers incorporating multiple null distributions into their research, we extended the values presented by James et al. for 5-point scales and created a table of variance estimates for various null distributions (see Table 2). This table is not meant to be exhaustive but rather to give researchers guidance concerning variance estimates for alternative null distributions. This table furnishes the expected variances based on certain response distributions. The pattern of responses giving rise to these distributions was based on our judgment but is consistent with previous discussions of null distributions by James et al. (1984). For example, we estimated the expected null variance for a heavily skewed distribution (e.g., strong leniency bias) for judges using a 6-point scale based on 0% of judges endorsing a 1 or 2, 5% endorsing a 3, 10% endorsing a 4, 40% endorsing a 5, and 45% endorsing a 6. This yielded an expected null variance of .69.

Question 11: Okay, I've calculated r_{WG} using multiple null distributions. How do I know which r_{WG} to use when justifying aggregation, and how should I go about reporting all of the r_{WG}s that I calculated?

Once a researcher has identified multiple, theoretically defensible null distributions, he or she must then calculate separate r_{WG} estimates (for each target) using each null distribution. Even with relatively modest sample sizes, this type of analysis can result in a large number of r_{WG} indices. For example, assume that a researcher has data from 10 groups of 5 workers who were asked to rate three dimensions of climate. This researcher has identified three null distributions for use with r_{WG} and is interested in understanding whether or not she can justify aggregating individual climate perceptions to the group level. Such an analysis would result in a total of 10 (number of groups) $\times$ 3 (number of dimensions) $\times$ 3 (number of null distributions), or 90, r_{WG} estimates. How should this researcher summarize and report his or her results? One recommended solution is to report separate mean r_{WG} estimates for each climate dimension using each null distribution (A. Cohen et al., 2001; LeBreton et al., 2003). Thus, the researcher would only need to have a table reporting 9 mean r_{WG} values, not 90 individual r_{WG} values. Furthermore, he or she could furnish additional descriptive information about the distribution of observed r_{WG} values for each null distribution (e.g., mean, standard deviation, range) and the percentage of r_{WG} values that exceed his or her a priori cutoff used to demonstrate sufficient IRA exists for aggregating data. He or she may also consider a graphical display of the frequency distribution of r_{WG} scores (A. Cohen et al., 2001).

Some researchers might argue that individual r_{WG} values should be reported instead of descriptive statistics summarizing the distributions of r_{WG} values. Although we agree that

Table 2
Expected Error Variances σ_E^2 Based on the Proportion of Individuals Endorsing Each Response Option

	Proportion Endorsing Each Value (5-Point to 11-Point Scale)											
Response Option	1	2	3	4	5	6	7	8	9	10	11	σ_E^2
Distributions: 5-point scale												
Slight skew	.05	.15	.20	.35	.25							1.34
Moderate skew	.00	.10	.15	.40	.35							0.90
Heavy skew	.00	.00	.10	.40	.50							0.44
Triangular	.11	.22	.34	.22	.11							1.32
Normal	.07	.24	.38	.24	.07							1.04
Uniform	.20	.20	.20	.20	.20							2.00
Distributions: 6-point scale												
Slight skew	.05	.05	.17	.20	.33	.20						1.85
Moderate skew	.00	.05	.10	.15	.40	.30						1.26
Heavy skew	.00	.00	.05	.10	.40	.45						0.69
Triangular	.05	.15	.30	.30	.15	.05						1.45
Normal	.05	.10	.35	.35	.10	.05						1.25
Uniform	.17	.17	.17	.17	.17	.17						2.92
Distributions: 7-point scale												
Slight skew	.05	.05	.10	.15	.15	.30	.20					2.90
Moderate skew	.00	.05	.10	.10	.15	.35	.25					2.14
Heavy skew	.00	.00	.05	.10	.15	.30	.40					1.39
Triangular	.05	.10	.20	.30	.20	.10	.05					2.10
Normal	.02	.08	.20	.40	.20	.08	.02					1.40
Uniform	.14	.14	.14	.14	.14	.14	.14					4.00
Distributions: 8-point scale												
Slight skew	.03	.03	.07	.12	.12	.18	.25	.20				3.47
Moderate skew	.00	.03	.08	.11	.15	.19	.25	.19				2.79
Heavy skew	.00	.00	.05	.10	.10	.15	.25	.35				2.35
Triangular	.03	.10	.16	.21	.21	.16	.10	.03				2.81
Normal	.01	.05	.16	.28	.28	.16	.05	.01				1.73
Uniform	.13	.13	.13	.13	.13	.13	.13	.13				5.25
Distributions: 9-point scale												
Slight skew	.05	.05	.08	.08	.10	.10	.16	.22	.16			5.66
Moderate skew	.03	.03	.05	.07	.08	.10	.14	.29	.21			4.73
Heavy skew	.00	.00	.05	.05	.10	.10	.15	.25	.30			3.16
Triangular	.03	.05	.10	.17	.30	.17	.10	.05	.03			3.00
Normal	.00	.03	.08	.20	.38	.20	.08	.03	.00			1.58
Uniform	.11	.11	.11	.11	.11	.11	.11	.11	.11			6.67
Distributions: 10-point scale												
Slight skew	.03	.03	.05	.08	.10	.10	.12	.12	.20	.17		6.30
Moderate skew	.00	.03	.03	.08	.08	.10	.10	.13	.25	.20		5.09
Heavy skew	.00	.00	.00	.05	.08	.10	.10	.12	.25	.30		3.46
Triangular	.00	.05	.08	.12	.25	.25	.12	.08	.05	.00		2.89
Normal	.00	.00	.05	.15	.30	.30	.15	.05	.00	.00		1.45
Uniform	.10	.10	.10	.10	.10	.10	.10	.10	.10	.10		8.25

(continued)

Table 2 (continued)

Response Option	Proportion Endorsing Each Value (5-Point to 11-Point Scale)											σ_E^2
	1	2	3	4	5	6	7	8	9	10	11	
Distributions: 11-point scale												
Slight skew	.02	.03	.03	.08	.08	.09	.10	.11	.12	.20	.14	7.31
Moderate skew	.00	.03	.03	.05	.07	.09	.09	.12	.12	.22	.18	6.32
Heavy skew	.00	.00	.00	.03	.05	.07	.08	.12	.15	.20	.30	4.02
Triangular	.00	.03	.07	.10	.15	.30	.15	.10	.07	.03	.00	3.32
Normal	.00	.00	.02	.08	.20	.40	.20	.08	.02	.00	.00	1.40
Uniform	.09	.09	.09	.09	.09	.09	.09	.09	.09	.09	.09	10.00

Note: Proportions used to estimate variance associated with a quasinormal distribution were obtained using the procedures described in Smith (1970).

in some instances the individual r_{WG} values may be informative, in many instances the number of r_{WG} values being computed makes providing individual values impractical. Returning to the LeBreton et al. (2003) example, these authors examined IRA within and between sources of multisource performance ratings. A total of 3,851 target managers were evaluated on 16 different dimensions of performance using variance estimates from three different null response distributions. Thus, simply examining the IRA between self-ratings and boss ratings involved the calculation of $3{,}851 \times 16 \times 3 = 184{,}848$ estimates of r_{WG}. Because these authors reported all possible within- and between-source comparisons, the number of r_{WG} estimates easily exceeded 1,000,000. Consequently, reporting individual r_{WG} estimates was not feasible; instead, the researchers elected to provide descriptive information about the distributions of r_{WG} values calculated using different null response error distributions.[1]

Methodological Questions Concerning the Estimation of IRR + IRA

Question 12: If the average score ICC is always higher than the individual score ICC, then shouldn't I always use the average score ICC?

James (1982) noted that because the *ICC(K)* is basically the Spearman-Brown applied to the *ICC(1)*, it is possible to take small values of *ICC(1)* and obtain sizeable values of *ICC(K)*. For example, if the *ICC(1)* value for a particular rating scale was .20, then using 22 raters would yield an *ICC(K)* of approximately .85, and using 35 raters would yield an *ICC(K)* of approximately .90. Higher scores do not imply that researchers should always use the *ICC(K)* or *ICC(A,K)* to index agreement. The decision to compute and use ICCs based on individual ratings or average score ratings should be based on theoretical reasons. *ICC(1)* is appropriate if a researcher is interested in drawing inferences concerning IRR + IRA of *individual ratings*. This is rarely the case in organizational research because we typically sample multiple judges' ratings (e.g., employee climate ratings) for each target (e.g., organizations). Instead, within the context of multilevel modeling, the *ICC(1)* is typically used to provide an estimate of effect size (Bliese, 2000; Bryk & Raudenbush, 1992;

Hofmann et al., 2000) indicating the extent to which individual ratings (e.g., climate ratings) are attributable to group membership (e.g., organizations).

In contrast, the *ICC(K)* is appropriate when a researcher is interested in drawing inferences concerning the *reliability of mean ratings* (Bliese, 2000; James, 1982; LeBreton et al., 2003). Because multilevel composition models involve justifying aggregation based on estimates of IRA and using these mean estimates as a higher level variable (e.g., a level-two predictor in hierarchical linear modeling), the *ICC(K)* or the *ICC(A,K)* are often used to justify aggregating such data. Thus, *ICC(1)* informs a researcher as to whether judges' ratings are affected by group membership, whereas the *ICC(K)* tells him or her how reliably the mean rating (taken over judges) distinguishes between groups (Bliese, 2000; Hofmann, 2002). Consequently, because these indices answer different research questions, one's question should drive the use of *ICC(1)* (e.g., Does group membership affect judges' ratings?) or *ICC(K)* (e.g., Do judges' mean ratings reliably distinguish among the groups/targets? Is there sufficient IRR + IRA to justify aggregating my data?).

Question 13: How many judges do I need when calculating *ICC(K)* or *ICC(A,K)*?

As noted earlier, the *ICC(K)* and *ICC(A,K)* are designed to assess the stability of judges' mean ratings. These estimates furnish information about the IRR + IRA present in the mean ratings from a set of judges. As noted in our response to Question 17, researchers should set an a priori cut point indicative of minimally acceptable IRR + IRA (e.g., ICC $> .80$ demonstrates minimally acceptable IRR + IRA). Once this criterion has been set, researchers may conduct a small pilot study to estimate the ICCs. Based on these pilot studies, researchers could increase or decrease the number of judges used in the final study. This is conceptually analogous to conducting a "G Study" prior to a "D Study" using generalizability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). Alternatively, researchers could look at the research literature to help estimate the number of judges needed to obtain a particular ICC value.

For example, take a hypothetical scenario where a group of researchers have spent several years collecting data examining an aggregate measure of climate. These researchers collected data from more than 100 organizations and sampled roughly 5 employees per organization (they could have just as easily sampled 20 or 30 employees). After completing their data collection, they calculated $ICC(1) = .12$, but because they only had 5 raters per organization, their $ICC(K) = .41$. Emergent group-level effects are unlikely to appear with such a low *ICC(K)* value (Bliese, 1998, 2000). Better planning would have enabled this research team to collect the necessary data to test for multilevel effects. For a second example, a bank would like to assess the relationship between customer satisfaction and branch performance. Researchers have collected preliminary data suggesting customer satisfaction ratings have an $ICC(1) = .10$. Using this number, they estimate how many customers need to be sampled at each bank branch to produce reliable mean differences across branches—they are shooting for an $ICC(K) = .85$. They estimate that 50 customers per bank branch will suffice. This number is much fewer than the 100 customers originally proposed. Thus, the researchers were able to save thousands of dollars by better planning their multilevel project.

Methodological Questions Concerning the Interpretation of IRR, IRA, and IRR + IRA

Question 14: What values of IRA are necessary to justify aggregation in a multilevel analysis, and do these values vary as a function of null distributions?

r_{WG} *and* a_{WG} *indices.* Values of .70 have been used as the traditional cut point denoting high versus low IRA using both the r_{WG} indices (Lance, Butts, & Michels, 2006; LeBreton et al., 2003) and the a_{WG} indices (Brown & Hauenstein, 2005). Concerning the r_{WG} indices, this value was first reported by George (1990), who obtained it from a personal communication with Lawrence R. James, one of the creators of r_{WG}. LeBreton et al. (2003) recently offered an expanded explanation for the .70 cutoff. In short, r_{WG} is interpreted as the proportional reduction in error variance. Higher scores indicate greater reduction in error variance and, thus, higher levels of agreement. A value of .70 suggests that there has been a 70% reduction in error variance. Consequently, just 30% of the observed variance among judges' ratings should be credited to random responding (i.e., error variance).

Although the .70 cut point has been a useful heuristic, we advance that researchers should think more globally about the necessity of high versus low within-group agreement based on their particular research question and composition model. Clearly, some composition models do not require any level of agreement (e.g., pure compilation models—Bliese, 2000; or additive models—Chan, 1998) whereas other models require establishing some minimal level of agreement (e.g., partial isomorphism composition models—Bliese, 2000; or direct consensus models—Chan, 1998). We believe that the .70 cut point artificially dichotomizes agreement in a manner that is inconsistent with James et al.'s (1984) original intention, and it may not be useful for justifying aggregation in multilevel models.

Consequently, in Table 3, we present a more-inclusive set of guidelines for interpreting agreement that may be helpful to organizational researchers. These guidelines are simply heuristics to guide researchers when evaluating their estimates of IRA. We should mention that any heuristic, ours included, is arbitrary. Heuristics such as $p < .05$, power $> .80$, or specific values for small, medium, and large effect sizes (J. Cohen, 1988) are just arbitrary "lines in the sand." So too is the traditional .70 cutoff value used in the past and the more-inclusive set of standards we are offering. Nevertheless, we believe that this more-inclusive set of standards may be beneficial for organizational researchers because in many instances an absolute standard for IRA $> .70$ may be too high, but in other instances this value may be too low. Artificially dichotomizing decisions concerning agreement is problematic, especially when researchers fail to consider the type of composition model they are testing, the quality of measures being used to test that model, and the significance of the decisions being made as a result of aggregation.

Specifically, our set of standards was based on the logic articulated by Nunnally and Bernstein (1994). Although these authors discussed standards for interpreting reliability coefficients, we nevertheless concur with their conclusion that the quality of one's measures (i.e., judges' ratings) should be commensurate with the intensity of the decision being made based on those measures. For important decisions involving specific individuals

Table 3
Revised Standards for Interpreting Interrater Agreement (IRA) Estimates

Level of IRA	Substantive Interpretation
.00 to .30	Lack of agreement
.31 to .50	Weak agreement
.51 to .70	Moderate agreement
.71 to .90	Strong agreement
.91 to 1.00	Very strong agreement

(e.g., who gets hired, fired, or promoted), we believe that it is necessary to demonstrate very strong agreement (> .90). In contrast, other research questions may only necessitate the establishment of moderate to little agreement.

We also recommend that researchers consider the psychometric quality and validity evidence for the measures they wish to aggregate and use this information to guide the a priori determination of the minimal level of agreement needed to justify aggregation. Specifically, a value of .70 may be acceptable for newly developed measures (e.g., a new measure assessing a climate for sexual harassment), but it may not be high enough for well-established measures that have been subjected to greater psychometric tests and have greater validity evidence available (e.g., Organizational Climate Questionnaire; James & Jones, 1974). In the latter cases, higher values may be necessary to demonstrate that judges are truly "seeing the same thing." Remember, an $r_{WG} = .70$ suggests that 30% of the variance in ratings is still error variance.

We also believe the values used to justify aggregation (e.g., $r_{WG} > .80$) should not vary based on the particular null distribution being used to estimate r_{WG}. We mention this because r_{WG} values estimated using distributions containing various response biases (e.g., leniency bias) will be lower in magnitude than those estimated using the uniform null distribution (see Questions 9, 10 and Table 2). Thus, there is a strong *disincentive* for researchers to estimate r_{WG} using distributions other than the uniform distribution. We acknowledge this disincentive but challenge researchers to use sound professional judgment when choosing which null distributions to use to estimate r_{WG} (see also Question 9) and challenge reviewers to hold authors accountable for the decisions they make involving null distributions. That being said, the value used to justify aggregation ultimately should be based on a researcher's consideration of (a) the quality of the measures, (b) the seriousness of the consequences resulting from the use of aggregate scores, and (c) the particular composition model to be tested.

AD *indices.* Although the *AD* indices appear to be a useful and intuitively appealing metric for assessing IRA, the determination of the range of scores that would represent adequate agreement is yet to be finalized. Burke and Dunlap (2002) presented some initial guidelines for establishing these ranges for *AD* estimates based on a uniform response distribution. Specifically, they suggested high agreement was obtained when the *AD* values for 5-, 7-, 9-, and 11-point scales were less than 0.8, 1.2, 1.5, and 1.8, respectively. These critical values provide a useful set of heuristics for researchers interested in using the *AD* indices in their work. However, additional research is needed to delineate the standards for

establishing appropriate levels of agreement when judges' ratings may have been influenced by various response biases.

Question 15: In a related question, are empirical procedures available for drawing inferences concerning estimates of IRA?

In addition to the practically significant heuristics described above, researchers might also be interested in inferring the statistical significance of IRA values. Because the sampling distributions of most estimates of IRA remain unknown, researchers have begun using data simulations to help draw statistical conclusions about agreement (A. Cohen et al., 2001; Dunlap, Burke, & Smith-Crowe, 2003). For example, A. Cohen et al. simulated the sampling distribution for a situation where r_{WG} was computed using a uniform null distribution. Specifically, they simulated a situation where a single item was rated by a group of 10 judges using a 5-point scale. They found that the r_{WG} value at the 95th percentile of the sampling distribution was .217. Thus, one could compare one's observed r_{WG} values to this critical value to determine if the level of observed agreement was statistically greater than that expected when judges responded randomly.

Although such statistical tests permit one to infer whether the level of IRA is greater than that expected because of random responding, they do not necessarily permit the inference that judges are homogenous enough to warrant data aggregation (A. Cohen et al., 2001). Consequently, when trying to determine if data should be aggregated, we encourage researchers to combine statistical significance tests (e.g., Is the observed level of agreement greater than zero?) with practical significance tests (e.g., If agreement is greater than zero, is it greater than some practically meaningful heuristic?). Researchers interested in learning more about calculating statistical significance tests are referred to A. Cohen et al. (2001) and Dunlap et al. (2003).

Question 16: My observed values of IRA are "good" for some groups but "bad" for others. What should I do? Do I drop groups with low levels of agreement, or do I keep them in my analysis?

It is possible to have high agreement and low agreement, both within one data set. However, James et al. (1984) do not recommend focusing on point estimates of agreement but instead on calculating multiple estimates of r_{WG} based on multiple, theoretically defensible null distributions. In this case, one would make conclusions about the extent of agreement, within a data set, using a range of values. We concur with James et al. (1984) and strongly encourage researchers to model multiple, theoretically defensible null distributions and interpret the range of r_{WG} values.

Along the same lines, we suggest that researchers examine the pattern of results. For example, if 85% of the r_{WG} estimates are below the cut point a researcher set to justify aggregation (e.g., $r_{WG} > .75$), then he or she is probably not justified in aggregating his or her data, even if some of the values are much higher than the cut point. Similarly, if multiple null distributions are modeled, then a researcher is able to develop quasi–confidence intervals for agreement based on the smallest (e.g., heavily skewed distribution) and largest (e.g., uniform distribution) estimates of error variance (James et al., 1984). If only 5% of the estimates using multiple nulls contain values less than that cut point, then he or

she is probably justified in aggregating his or her data. Ultimately, researchers should make judgments based on the magnitude and pattern of r_{WG} values.

Researchers should also consider several other options when both high and low levels of agreement are demonstrated within a single data set. First, it is possible to drop those groups, divisions, or organizations that lack appropriate agreement to justify aggregation. However, we caution that this low agreement may be systematic over certain targets and important to examine, perhaps as a dispersion composition model (Chan, 1998). Also, losing potentially valuable data is never recommended. For example, if a researcher were examining organizational-level data, removing organizations with low agreement could result in deleting thousands of individual-level cases from a meaningful percentage of his or her organizations (e.g., 6 out of 30 organizations), which could be problematic. Another possible alternative is to aggregate all groups, even those that lack proper levels of r_{WG}. This could be done as long as some of the groups or organizations do have sufficiently high r_{WG} values. One problem with this approach is that researchers might be diluting their results by mixing groups that agree with those that lack agreement. Last, one might decide to aggregate all groups or organizations but include a dummy variable as a possible moderator. This dummy variable would label groups as either having sufficient or insufficient agreement. These are tough decisions, and ultimately theory and sound judgment should guide decisions regarding aggregation. Estimates of IRA simply serve to support or refute one's theory (e.g., direct consensus models vs. dispersion models; Chan, 1998).

Question 17: What values of IRR + IRA are necessary to justify aggregation in a multilevel analysis?

As noted earlier, *ICC(1)* can be interpreted as the IRR + IRA of individual ratings; however, this is rarely the case in organizational research because we typically sample multiple judges' ratings (e.g., employee climate ratings) for each target (e.g., organizations). Instead, the *ICC(1)* is typically interpreted as a measure of effect size (Bliese, 2000; Bryk & Raudenbush, 1992), revealing the extent to which individual ratings are attributable to group membership. Thus, when interpreting values for *ICC(1)*, we encourage researchers to adopt traditional conventions used when interpreting effect sizes (i.e., percentage of variance explained). Specifically, a value of .01 might be considered a "small" effect, a value of .10 might be considered a "medium" effect, and a value of .25 might be considered a "large" effect (see Murphy & Myors, 1998, p. 47). For example, an *ICC(1)* = .05 represents a small to medium effect, suggesting that group membership (e.g., employing organization) influenced judges' ratings (e.g., employees responses to a climate survey). Thus, values as small as .05 may provide prima facie evidence of a group effect. Such a finding would warrant additional investigations concerning the viability of aggregating scores within groups (e.g., estimating within-group agreement via r_{WG}).

In contrast, because *ICC(K)* and *ICC(A,K)* are often interpreted as measures of IRR + IRA, researchers may be tempted to apply traditional reliability cutoffs (e.g., Nunnally & Bernstein, 1994). However, researchers should be reminded that high values of these indices are only obtained when there is both high IRR and high IRA. Furthermore, low ICCs may be obtained when there is low IRA, low IRR, or both. When considering the

value of an ICC, it is important to understand what is driving that low value (i.e., is it because of low consistency, low consensus, or both?).

Reliability coefficients are defined as the proportion of true score variance to total score variance (Gulliksen, 1950; Lord & Novick, 1968; Nunnally, 1978). A value of .70 suggests that 70% of the variance in judges' ratings is systematic, or true score variance, whereas 30% of the variance is random measurement error variance. As described in Question 14, the minimum acceptable level of reliability for psychological measures in the early stages of development is .70 (Nunnally, 1978). Higher levels may be required of measures in the later stages of development, such as those used in advanced field research and practice. Nunnally and Bernstein (1994) noted,

> A reliability of .80 may not be nearly high enough in making decisions about individuals . . . if important decisions are being made with respect to specific test scores, a reliability of .90 is the bare minimum, and a reliability of .95 should be considered the desirable standard. (p. 265)

Lance et al. (2006) noted that many researchers have blindly relied on the .70 value without adequate consideration being given to the appropriateness of this value. Most aggregate ratings used in multilevel modeling are not used to make important decisions about specific individuals but instead about groups or organizations. So in most instances, *ICC(K)* and *ICC(A,K)* values $> .90$ are likely unnecessary. However, depending on the quality of the measures being used in the multilevel analysis, researchers will probably want to choose values between .70 and .85 to justify aggregation.

In addition to being interpreted as measures of IRR, *ICC(K)* and *ICC(A,K)* are also interpreted as measures of IRA. As such, researchers may be tempted to use traditional agreement cutoffs when interpreting ICCs. As noted earlier, a value of .70 has frequently been used as the demarcation of high versus low IRA indexed via r_{WG} (Lance et al., 2006; LeBreton et al., 2003). Although r_{WG}s and ICCs are conceptually distinct indices, both are interpreted on a scale ranging from 0 to 1.0, both inform about the proportion of observed variance that is measurement error variance, both incorporate between-judge variance into estimates of error variance, and both are reasonable indices of IRA (A. Cohen et al., 2001; James et al., 1984, 1993; Kozlowski & Hattrup, 1992; LeBreton et al., 2003; McGraw & Wong, 1996).

Although reliability and agreement are conceptually distinct concepts, they may nevertheless be estimated using a single ICC. The commonly used cut point of .70 is by no means a definitive rule; it is simply a heuristic that is frequently (and inappropriately) invoked in research (Lance et al., 2006). For researchers wanting to justify aggregation, we suggest (as above) that they set an a priori cut point for the value of ICC they feel is acceptable given their particular research question and the quality of their measures. For example, a researcher using well-established measures seeking to aggregate climate ratings to draw inferences about group-level climate → individual-level job satisfaction linkages may set a cut point of .75 or .80. In contrast, a researcher aggregating panel interviewer ratings to make inferences about aggregate-level interview performance → individual-level job performance linkages may need to set a much higher cut point (e.g., .90 or .95) because of the important consequences such inferences have for individuals.

Question 18: What is causing my estimates of IRR and IRA to diverge from one another (e.g., high agreement but low reliability)?

LeBreton et al. (2003) discussed how high levels of IRR do not guarantee high levels of IRA and vice versa. This is easily illustrated using a scenario where we have two raters evaluate three targets using a 5-point scale. These two sets of ratings may be relatively consistent with one another (Rater 1 = 1, 2, and 3; Rater 2 = 3, 4, and 5), yet they clearly lack absolute consensus. Empirically, we obtain high correlations estimating IRR ($r = 1.0$). However, the lack of rating consensus yields extremely low estimates of IRA (r_{WG}s computed using a uniform null distribution were 0.00, 0.00, and 0.00) and an extremely low estimate of IRR + IRA (*ICC(1)* = 0.00). In this instance, the divergence between estimates of IRR and IRA was because of differences in the absolute magnitude of ratings.

Because IRA assesses the interchangeability or absolute consensus in ratings, one might assume that it would always be more difficult to obtain high levels of IRA compared to IRR, which only assesses the relative consistency in ratings. However, LeBreton et al. (2003) demonstrated how it is possible to have high levels of IRA yet low levels of IRR and IRR + IRA. They showed that when between-target variance becomes substantially restricted, correlation-based estimates of IRR and IRR + IRA are attenuated. In such instances, researchers relying solely on ICCs to justify aggregation could make very erroneous decisions.

For example, in a new scenario, we ask two new raters to evaluate 10 targets using a 7-point rating scale (Rater 1 = 1, 2, 1, 1, 2, 6, 7, 6, 7, 7; Rater 2 = 2, 2, 1, 1, 2, 7, 7, 6, 6, 7). Using these data, we obtain high estimates of IRR ($r = .97$), IRA (average $r_{WG} = .95$; estimated using a uniform null distribution), and IRR + IRA (*ICC(1)* = .97). However, when we separately estimate IRR and IRA for the first 5 targets and then for the second 5 targets, we obtain very different results. Within each group, our estimates of IRR and IRR + IRA are extremely low ($r = .17$, *ICC(1)* = .27). We remind the reader that the relative rank orders within each set have not changed. The only thing that has changed is that we have restricted our variance. Mathematically, when we estimated IRR and IRR + IRA using all 10 targets, we had a 7-point scale. Our results indicated that the judges did a very good job of consistently distinguishing those targets that deserve 1s or 2s from those that deserve 6s or 7s. When we run separate analyses, our 7-point scale is mathematically transformed into a 2-point scale (i.e., we know that ratings were made on a 7-pont scale, but only 2 of the scale points were actually used). When we run separate analyses, we see that our judges did a very poor job of distinguishing 1s from 2s and 6s from 7s.

In short, it is possible for strong levels of IRA to be masked by subtle inconsistencies in rank orders. This is especially pronounced when the between-target variance is restricted (e.g., all targets are rated high or low). LeBreton et al. (2003) concluded that the "examination of both IRR and IRA statistics represents a form of psychometric checks and balances concerning [interrater similarity]" (p. 121). We concur and encourage researchers to compute both types of indices when their data permits. By calculating both sets of indices, researchers will be better able to understand if their data lack reliability, agreement, neither, or both.

Question 19: I am planning to use the r_{WG} index to justify aggregating my data prior to conducting my multilevel analysis. Can I still use r_{WG} if I have missing data?

Another way to phrase this question is to ask, "What are the implications of missing data for the computation IRA indices and how should I deal with such missing data?" This is an area with relatively little empirical work to guide our answer. How one goes about defining *missing* will shape the conclusion one draws concerning how missing data affect estimates of IRA. Newman and Sin (in press) noted that data may be missing completely at random (MCAR), where "the probability that data on a given variable unobserved is independent of the value of that variable, as well as the values of all other variables" (pp. 3-4). At the other extreme, a researcher might have data missing not-at-random (MNAR), where "the probability of missing data on a variable [such as job satisfaction] depends on the value of [job satisfaction]" (p. 4). For example, workers who are unhappy with their jobs (i.e., low scores) may be less likely to return surveys asking about their job (e.g., job satisfaction surveys). This would result in a within-group variance restriction and, one would suspect, an upwardly biased estimate of within-group agreement. In their simulation study, Newman and Sin confirmed this hypothesis and concluded that when data were MNAR, $r_{WG(J)}$ substantially overestimates IRA. Interestingly, they also found that when data were MCAR, $r_{WG(J)}$ tended to slightly underestimate IRA. Based on these findings, we urge caution when analyzing data with missing observations. Newman and Sin provided formulas for estimating the impact of missing data and gave specific recommendations for how to lessen the influence of missing data on estimates of IRA.

Thus, if a researcher has missing data, we recommend he or she examine his or her data and try to determine if data are missing randomly or systematically. If the pattern appears random, then he or she is probably safe using $r_{WG(J)}$. However, if the pattern appears systematic (or it is impossible to determine), then he or she should probably not proceed using $r_{WG(J)}$ because the estimates are likely substantially (upwardly) biased. Furthermore, researchers are referred to Newman and Sin's study for equations that can be used to correct for missing data.

Question 20: Is it appropriate to treat the values of IRA indices as actual variables in a multilevel model?

Yes. This is precisely the idea behind multilevel dispersion models (Chan, 1998). With these types of models, the psychological meaning of the higher-level variable is the dispersion or variance in the lower-level variable. For example, Schneider et al. (2002) demonstrated that the strength of climate perceptions (i.e., within-group agreement, or lack thereof) moderated the relationship between employee perceptions of customer service climate and customer ratings of satisfaction (see also González-Romá, Peiró, & Tordera, 2002). However, if one is testing a dispersion model, then conceptually it makes more sense to use a measure of interrater dispersion (e.g., SD_X) rather than a measure of IRA (e.g., r_{WG}; see Lindell & Brandt, 2000; Roberson et al., in press; Schneider et al., 2002).

Conclusion

The prevalence of multilevel modeling techniques in the organizational sciences continues to increase. This only makes sense given that organizations, by definition, contain multiple layers or levels ranging from the individual worker to the multinational

conglomerate. Multilevel researchers often rely on estimates of IRR and IRA to justify the aggregation of a lower-level variable (e.g., individual climate perceptions) into a higher-level variable (e.g., shared psychological climate). Thus, as multilevel modeling has increased in popularity, so too has the use of IRA indices and IRR indices. The purpose of the current article was to ask and answer some of the questions faced by researchers using these indices in their research. The list of questions was by no means definitive or exhaustive; other questions remain, and still more will be uncovered. We hope the current article at least addressed some of the more common questions about IRA and IRR, especially within the context of multilevel modeling.

Appendix A
Tutorial on Estimating Interrater Reliability (IRR) and Interrater Agreement (IRA)

All of the above-referenced indices are easily calculated using modern statistical software (e.g., SPSS, SAS). Although each of these indices may be calculated in SPSS using point and click, we decided to use the Syntax option in SPSS so that individuals could copy and paste our code into SPSS. To use this option, one must simply open SPSS and then open a new syntax window (File → New → Syntax). The following analyses were based on the data presented in Appendix B1.

Restructuring the Data

Prior to calculating our estimates of IRR, IRA, and IRR + IRA, we need to restructure our data. This is accomplished using the following syntax. Simply copy and paste this syntax into SPSS and click run:

```
SORT CASES BY Target.
CASESTOVARS
/ID = Target
/GROUPBY = VARIABLE.
EXECUTE.
```

Running the above syntax on the data in Appendix B1 yields the data in Appendix B2. By examining these data, you will see that each target was not necessarily rated by the same number of judges. Thus, there are some missing values in our restructured data set. The following syntax recodes these missing data to 999 and labels them as missing values in SPSS, yielding the data in Appendix B3.

```
RECODE ITEM1.1 to ITEM2.5 (MISSING = 999).
MISSING VALUES ITEM1.1 to ITEM2.5 (999).
EXECUTE.
```

Estimate r_{WG}

We will assume the data listed under the variables Item1.1 to 1.5 are ratings on a climate item measuring leader trust and support furnished by employees working in four

(continued)

Appendix A (continued)

different work teams (targets). Thus, Item1.1 is a generic label referring to the first rating furnished for each work team, Item1.2 is a generic label referring to the second rating furnished for each work team, and so on. Alternatively stated, each score in the column was furnished by a different person assigned to one of the four teams. Thus, we have a total of 18 unique "participants" or "judges" spread out over four different teams. We hope to justify aggregating climate perceptions within each work team and decide to use r_{WG} to justify aggregation. To estimate r_{WG}, we must estimate the observed variance within each team. This is accomplished using the following SPSS syntax:

```
COMPUTE obs_var1 = var(item1.1,item1.2,item1.3,item1.4,item1.5).
EXECUTE.
```

The above code computes a new variable, obs_var1, that is the variance observed within each target (work team) across each set of raters (team members) on the first climate item measuring leader trust and support. All new variables computed as part of this tutorial are presented in Appendix B4.

Next, we compare the observed variances to the variance we would expect when judges respond randomly. We believe that one form of random response might involve a rectangular or equal probability distribution (assuming a 5-point response scale, $\sigma^2_E = 2.0$). We also believe that employees might also have a tendency to "go easy" on their work teams, thus inflating their climate perceptions because of a slight leniency bias ($\sigma^2_E = 1.34$). We estimate r_{WG} using both possible null response distributions:

```
COMPUTE rwg1_un = 1-(obs_var1/2).
COMPUTE rwg1_ss = 1-(obs_var1/1.34).
EXECUTE.
```

To facilitate interpretation, we will only examine the r_{WG} values estimated using the uniform distribution. For all teams, we see that $r_{WG} > .80$, suggesting strong agreement. Thus, based on the r_{WG} estimates, we are probably justified in aggregating these data to the team level.

Estimate AD_M

We also estimated the AD about the mean rating for each team:

```
COMPUTE MEAN1 = mean(item1.1,item1.2,item1.3,item1.4,item1.5).
COMPUTE AD1 = mean(abs(item1.1-mean1),abs(item1.2-mean1),abs(item1.3-mean1),
  abs(item1.4-mean1), abs(item1.5-mean1)).
EXECUTE.
```

Burke and Dunlap (2002) suggested a critical value of .80 or less for establishing agreement when using a 5-point scale. Comparing our observed results to this critical value reveals that, in all instances, our teams demonstrated high levels of within-group agreement. Thus, the results obtained using AD_M confirmed those obtained using r_{WG}.

(continued)

Appendix A (continued)

Estimate *ICC(1)* and *ICC(K)*

We continue to treat Item1.1 to Item 1.5 as climate scores furnished by different employees working as members of different work teams and estimate the *ICC(1)* and *ICC(K)*:

```
RELIABILITY
  /VARIABLES = Item1.1 Item1.2 Item1.3 Item1.4 Item1.5
  /SCALE(ALPHA) = ALL/MODEL = ALPHA
  /ICC = MODEL(ONEWAY) CIN = 95 TESTVAL = 0.
EXECUTE.
```

Unlike the code for r_{WG}, this code does not involve computing new variables. Instead, it calculates the ICC values and prints them in an output window. Examining this window, we see that the *ICC(1)* = .60. This is a large effect size, suggesting that climate ratings were heavily influenced by team membership. The *ICC(K)* = .88 reveals high levels if IRR + IRA and suggests that the mean ratings (taken over judges) reliably distinguish the four teams. Thus, given the pattern of r_{WG} and AD_M values just obtained, along with the sizeable *ICC(K)*, we feel comfortable aggregating our data to the team level.

Estimate $r_{WG(J)}$

We will again use the data in Appendix B3 and assume that the data are measuring one dimension of climate (e.g., leader trust and support); however, now we will assume that each dimension was assessed using two items. Item1.1 to Item1.5 corresponds to the ratings on the first item (e.g., ''My supervisor is willing to listen to my problems'') and Item2.1 to Item2.5 corresponds to the ratings on the second item (e.g., ''My supervisor can be trusted''). To summarize, we have scores on two climate measures furnished by individuals nested in four different teams. Team 1 and Team 2 each have 5 members, whereas Team 3 and Team 4 each have 4 members. To estimate $r_{WG(J)}$, we must first estimate the variance on each item across the raters within each team. The variance for the first item has already been computed. Below, we estimate the variance of the second item and then calculate the average variance over the two items:

```
COMPUTE obs_var2 = var(item2.1,item2.2,item2.3,item2.4,item2.5).
COMPUTE avg_var = mean(obs_var1,obs_var2).
EXECUTE.
```

Using this new variable avg_var, we estimate $r_{WG(J)}$ using both uniform and slightly skewed null distributions:

```
COMPUTE rwgj_un = (2*(1-avg_var/2))/((2*(1-avg_var/2)) + avg_var/2).
COMPUTE rwgj_ss = (2*(1-avg_var/1.34))/((2*(1-avg_var/1.34)) + avg_var/1.34).
EXECUTE.
```

(continued)

Appendix A (continued)

Again, we limit our comments to the uniform distribution. As we can see, the $r_{WG(J)}$ values are higher when we estimated agreement based on two items compared to when we estimated agreement based on a single item. Specifically, all $r_{WG(J)}$ values exceed .85. This suggests that strong agreement exists among the judges within each team, and so we are safe to aggregate climate data to the team level.

Estimate $AD_{M(J)}$

The results obtained using $r_{WG(J)}$ are confirmed using $AD_{M(J)}$:

```
COMPUTE MEAN2 = mean(item2.1,item2.2,item2.3,item2.4,item2.5).
COMPUTE AD2 = mean(abs(item2.1-mean2),abs(item2.2-mean2),abs(item2.3-mean2),
  abs(item2.4-mean2), abs(item2.5-mean2)).
COMPUTE ADJ = mean(AD1,AD2).
EXECUTE.
```

Above, we estimated the mean for the second item and the AD for each team member's rating about the mean rating for his or her team. Finally, we calculated the average of our two AD indices. All of our obtained $AD_{M(J)}$ values were well below the critical value of .80 suggested by Burke and Dunlap (2002) for use with 5-point scales.

To recap, we found strong levels of within-group agreement using the single-item r_{WG} and AD_M indices. This agreement was confirmed using the multi-item extensions $r_{WG(J)}$ and $AD_{M(J)}$. The large *ICC(1)* suggested that group membership exerted a large influence on team member ratings, and the *ICC(K)* suggested that the mean score could reliably distinguish between teams. Taken together, these results indicate that we are probably justified in aggregating our individual climate data to the team level.

We should mention that these data were specifically contrived to yield clean and easily interpreted results. However, it is often the case that results obtained using estimates of IRA and IRR + IRA will not be so easily interpretable. For example, assume we have climate data on 10 teams. Teams 1 to 7 have r_{WG} values of .90 or higher and AD_M values of .40 or lower. However, Teams 8 to 10 have r_{WG} values less that .60 and AD_M values of .95 or higher. Furthermore, the *ICC(K)* = .68. Assuming that we had set a priori cutoffs for agreement of $AD_M < .80$ and $r_{WG} > .80$, we are now at a tough decision point. Should none of the data be aggregated to the team level? Should only those teams that exceed the minimum cutoffs be aggregated? Or should we risk it and aggregate all teams because the vast majority (70%) had adequate agreement? This is a tough question to answer. If we only aggregate the teams with adequate IRA, then we are effectively deleting 30% of our sample. If we aggregate all of the data and argue that the low agreement for Teams 8 to 10 was because of sampling error, then we could be wrong and might dilute any significant findings. If we do not aggregate any of the data, then we are unable to test our multilevel hypothesis. Clearly, tough questions do not have easy answers. In this situation, we might run both sets of aggregate analyses (i.e., aggregate all teams and only those with adequate IRA) and report any discrepancies in the tests of our multilevel hypotheses.

(continued)

Appendix A (continued)

Estimate r_{WGp}

We now turn to the data presented in Appendix C. We have slightly restructured how we present the data. We now have three different leaders, each of whom is evaluated by six subordinate employees. Focusing only on Item 1 and assuming a 5-point scale, we could estimate r_{WG} using a uniform null distribution for each set of raters as

```
SORT CASES BY Leader.
SPLIT FILE
SEPARATE BY Leader.
DESCRIPTIVES
  VARIABLES = Item1
  /STATISTICS = MEAN STDDEV VARIANCE MIN MAX.
```

The above syntax basically asks SPSS to sort the data by leader and then splits the file into three data sets (one for each leader). Next, we ask SPSS to run basic descriptive statistics on the first climate item measuring leader trust and support. We obtain the following within-leader variances: 2.667, 2.967, and 2.967. If we then manually estimate r_{WG} using Equation 1 and an expected null variance of 2.0 (uniform random response), we obtain estimates of -0.33, -0.48, and -0.48. Following the recommendations of James, Demaree, and Wolf (1984, 1993), we could reset these values to zero and conclude lack of agreement.

However, let us now assume that prior to obtaining subordinate climate ratings of leader trust and support, we first collected data on a measure of leader-member exchange and used these data to determine which subordinates were in each leader's in-group and which were in his or her out-group. These data are presented under the variable status, with 1 = in-group and 2 = out-group. We could use this a priori distinction to group individuals and estimate r_{WGp}. Provided that the homogeneity of variance assumption was not violated (LeBreton, James, & Lindell, 2005), we could proceed to estimate the pooled (i.e., weighted average) within-group variance. As noted by LeBreton et al. (2005), the easiest way to obtain this estimate is from the error mean square from an ANOVA. Because we want to come up with a unique estimate for each leader, we again sort and split the file by leader. Next, we calculate a one-way ANOVA using in-group or out-group status as our independent variable and scores on Item 1 as our dependent variable:

```
SORT CASES BY Leader.
SPLIT FILE
  SEPARATE BY Leader.
UNIANOVA
item1 BY status
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /EMMEANS = TABLES(status)
  /PRINT = HOMOGENEITY
  /CRITERIA = ALPHA(.05)
  /DESIGN = status.
```

(continued)

Appendix A (continued)

Referencing our output, we see that our three sets of analyses (i.e., split by leader) resulted in relatively small mean square error (i.e., average variance within the in-group and the out-group). For our three leaders, we obtained mean square error estimates of .667, .333, and .333, respectively. We also see that none of the analyses violated the homogeneity of variance assumption (i.e., the variance on Item 1 was statistically identical for in-group and out-group members). Using the uniform null distribution to estimate our error variance ($\sigma_E^2 = 2.0$), we manually estimated r_{WGp} using Equation 8 and obtained values of .67, .83, and .83. These results indicate that, within each subgroup of raters, there is reasonable agreement concerning climate perceptions involving leader trust and support. Examining the means for each of the three groups reveals that in-group members (status = 1) have higher ratings compared to out-group members (status = 2). Within the context of multilevel modeling, we are probably not justified in aggregating all six ratings within each leader, but we are justified in aggregating the ratings into an in-group and out-group set of climate perceptions.

Estimate $r_{WGp(J)}$

The above logic is easily extended to $J = 3$ parallel items. Essentially, we replicate the above analyses for Items 2 and 3 by running separate ANOVAs for each leader, confirming homogeneity of variance, and getting the pooled within-groups variance estimates from the mean square error for each leader analysis. Next, we calculate the average of the pooled within-groups variances. This is basically the average of the average within-group item variances. From our example data, we found average mean square error estimates of .611, .333, and .333, respectively. Using these values in conjunction with Equation 9, we estimate separate $r_{WGp(J)}$ values, one for each leader. Because we had three essentially parallel items, our estimates of $r_{WGp(J)}$ are slightly higher than those for r_{WGp}: .87, .94, and .94, respectively. Examining the means for the in-groups and out-groups, our new results confirm what we found when we used only a single item. In-groups had higher mean ratings for each leader across all three items compared to the out-groups. Although the in-groups and out-groups disagree with one another concerning the perceptions of leader trust and support, within each subgroup there is substantial agreement. Consequently, we are probably justified to aggregate data to the subgroup level (i.e., calculate aggregate scores for the in-group and out-group for each leader), but we are not justified in aggregating all data to the leader level (i.e., we should not aggregate perceptions of leader trust and support across all six raters for each leader because there is substantial lack of agreement between in-groups and out-groups).

Appendix B1
Sample Multilevel Data Set 1

The following data are arranged in a common multilevel format. In this example, there are four targets (e.g., team leaders) rated on two items. Targets 1 and 2 were each rated by five judges (e.g., team members), whereas Targets 3 and 4 were each rated by four judges.

(continued)

Appendix B (continued)

Target	Item1	Item2
1	4	4
1	5	5
1	4	4
1	5	5
1	4	4
2	4	5
2	4	4
2	3	5
2	3	4
2	3	3
3	3	3
3	3	4
3	3	4
3	4	3
4	4	3
4	5	2
4	5	4
4	4	3

Appendix B2
Sample Multilevel Data Set 1—Restructured

Below are the data from Appendix B1 after being rearranged to conform to an SPSS-friendly, multilevel format. As above, we have four targets (e.g., team leaders) rated on two items. Targets 1 and 2 were each rated by five judges (e.g., team members), whereas Targets 3 and 4 were rated by four judges.

Target	Item1.1	Item1.2	Item1.3	Item1.4	Item1.5	Item2.1	Item2.2	Item2.3	Item2.4	Item2.5
1	4	5	4	5	4	4	5	4	5	4
2	4	4	3	3	3	5	4	5	4	3
3	3	3	3	4	.	3	4	4	3	.
4	4	5	5	4	.	3	2	4	3	.

Appendix B3
Sample Multilevel Data Set 1—Restructured and Missing Data Recoded

Below are the data from Appendix B2, following the missing data recodes. As above, we have four targets (e.g., team leaders) rated on two items. Targets 1 and 2 were each rated by five judges (e.g., team members), whereas Targets 3 and 4 were rated by four judges. Missing values are now coded 999.

Target	Item1.1	Item1.2	Item1.3	Item1.4	Item1.5	Item2.1	Item2.2	Item2.3	Item2.4	Item2.5
1	4	5	4	5	4	4	5	4	5	4
2	4	4	3	3	3	5	4	5	4	3
3	3	3	3	4	999	3	4	4	3	999
4	4	5	5	4	999	3	2	4	3	999

Appendix B4
Sample Multilevel Data Set 1—New Variables

New variables are computed by applying the syntax in Appendix A to the data in Appendix B3.

Target	obs_var1	rwg1_un	rwg1_ss	Mean1	AD1	obs_var2	avg_var	rwgj_un	rwgj_ss	Mean2	AD2	ADJ
1.00	0.30	0.85	0.78	4.40	0.48	0.30	0.30	0.92	0.87	4.40	0.48	0.48
2.00	0.30	0.85	0.78	3.40	0.48	0.70	0.50	0.86	0.77	4.20	0.64	0.56
3.00	0.25	0.88	0.81	3.25	0.38	0.33	0.29	0.92	0.88	3.50	0.50	0.44
4.00	0.33	0.83	0.75	4.50	0.50	0.67	0.50	0.86	0.77	3.00	0.50	0.50

Appendix C
Sample Multilevel Data Set 2

Leader	Status	Item1	Item2	Item3
1	1	5	5	3
1	1	5	4	5
1	1	4	5	5
1	2	3	2	2
1	2	1	2	1
1	2	2	1	1
2	1	4	4	5
2	1	4	5	4
2	1	5	5	5
2	2	1	2	2
2	2	1	2	1
2	2	2	1	1
3	1	5	4	4
3	1	5	4	4
3	1	4	5	5
3	2	2	2	1
3	2	2	1	1
3	2	1	1	2

Note

1. Alternatively, if one is less interested in examining the judges' agreement for each individual target and instead seeks a global or overall estimate of agreement across targets, he or she might consider using LeBreton et al. (2005)'s r_{WGp} index. As noted under Questions 6 and 7, r_{WGp} furnishes a single estimate of agreement based on a pooled within-groups estimate of variance. For example, instead of reporting separate r_{WG}s for each of 100 work teams, a researcher may instead decide to report a single, global estimate of (pooled) within-groups agreement. Even in situations where examining the judges' agreement for each target is deemed important, researchers may still find it useful to also calculate r_{WGp} because it provides a single overall estimate of the agreement.

References

Baltes, B. B., & Parker, C. P. (2000). Reducing the effects of performance expectations on behavioral ratings. *Organizational Behavior and Human Decision Processes*, *82*, 237-267.

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

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

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

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., vol. 2, pp. 271-326). Palo Alto, CA: Consulting Psychologists Press.

Brown, R. D., & Hauenstein, *N.M.A.* (2005). Interrater agreement reconsidered: An alternative to the r_{WG} indices. *Organizational Research Methods*, *8*(2), 165-184.

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods*, *5*(2), 159-172.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods*, *2*(1), 49-68.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Application and data analysis methods*. Newbury Park, CA: Sage.

Cardy, R. L., & Dobbins, G. H. (1994). *Performance appraisal: Alternative perspectives*. Cincinnati, OH: South-Western.

Chan, D. (1998). Functional relations among constructs in the same context domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*, 234-246.

Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the $r_{WG(j)}$ index of interrater agreement. *Psychological Methods*, *6*, 297-310.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin*, *90*, 218-244.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: John Wiley.

Dienesch, R. M., & Liden, R. C. (1986). Leader-member exchange model of leadership: A critique and further development. *Academy of Management Review*, *11*, 618-634.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for rwg and average deviation interrater agreement indices. *Journal of Applied Psychology, 88*, 356-362.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology*, *66*, 127-148.

Fisicaro, S. A. (1988). A reexamination of the relation between halo error and accuracy. *Journal of Applied Psychology*, *73*, 239-244.

Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement*, *14*, 419-429.

Fisicaro, S. A., & Vance, R. J. (1994). Comments on the measurement of halo. *Educational and Psychological Measurement*, *54*, 366-371.

Funder, D. C. (1987). Errors and mistakes: Evaluating the accuracy of social judgement. *Psychological Bulletin*, *101*, 75-90.

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

González-Romá, V., Peiró, J. M., & Tordera, N. (2002). An examination of the antecedents and moderator influences of climate strength. *Journal of Applied Psychology*, *87*, 465-473.

Gulliksen, H. (1950). *Theory of mental tests*. New York: John Wiley.

Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management*, *20*, 737-756.

Hofmann, D. A. (2002). Issues in multilevel research: Theory development, measurement, and analysis. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 247-274). New York: Blackwell.

Hofmann, D. A., Griffin, M. A., & Gavin, M. B. (2000). The application of hierarchical linear modeling to organizational research. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 467-511). San Francisco: Jossey-Bass.

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes*, *54*, 321-368.

James, L. R. (1982). Aggregation in estimates of perceptual agreement. *Journal of Applied Psychology*, *67*, 219-229.

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., Demaree, R. G., & Wolf, G. (1993). r_{WG}: An assessment of within-group interrater agreement. *Journal of Applied Psychology*, *78*, 306-309.

James, L. R., & Jones, A. P. (1974). Organizational climate: A review of theory and research. *Psychological Bulletin*, *81*(12), 1096-1112.

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. (1987). 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. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349-381). San Francisco: Jossey-Bass.

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods*, *9*(2), 202-220.

LeBreton, J. M., Burgess, J.R.D., Kaiser, R. B., Atchley, E.K.P., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods*, *6*(1), 80-128.

LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding r_{WG}, r^*_{WG}, $r_{WG(J)}$, and $r^*_{WG(J)}$. *Organizational Research Methods*, *8*(1), 128-139.

Lindell, M. K. (2001). Assessing and testing interrater agreement in multi-item rating scales. *Applied Psychological Measurement, 25*, 89-99.

Lindell, M. K., & Brandt, C. J. (1997). Measuring interrater agreement for ratings of a single target. *Applied Psychological Measurement*, *21*, 271-278.

Lindell, M. K., & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of the CVI, T, and indexes. *Journal of Applied Psychology*, *84*, 640-647.

Lindell, M. K., & Brandt, C. J. (2000). Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes. *Journal of Applied Psychology*, *85*, 331-348.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of agreement for multi-item ratings of a single target. *Applied Psychological Measurement*, *23*, 127-135.

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

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30-46.

Murphy, K. R., & Balzer, W. K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology*, *74*, 619-624.

Murphy, K. R., & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal based perspectives*. Thousand Oaks, CA: Sage.

Murphy, K. R., & DeShon, R. (2000a). Interrater correlations do not estimate the reliability of job performance ratings. *Personnel Psychology*, *53*, 873-900.

Murphy, K. R., & DeShon, R. (2000b). Progress in psychometrics: Can industrial and organizational psychology catch up? *Personnel Psychology*, *53*, 913-924.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Lawrence Erlbaum.

Newman, D. A., & Sin, H. P. (in press). How do missing data bias estimates of within-group agreement? Sensitivity of SD_{WG}, CV_{WG}, $rW_{G(J)}$, $rW_{G(J)}^*$, and ICC to systematic nonresponse. *Organizational Research Methods*.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Roberson, Q. M., Sturman, M. C., & Simons, T. L. (in press). Does the measure of dispersion matter in multilevel research? A comparison of the relative performance of dispersion indices. *Organizational Research Methods*.

Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology*, *74*, 368-370.

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*, 901-912.

Schneider, B., Salvaggio, A. N., Subirats, M. (2002). Climate strength: A new direction for climate research. *Journal of Applied Psychology*, *87*, 220-229.

Schriesheim, C. A., Donovan, J. A., Zhou, X., LeBreton, J. M., Whanger, J. C., & James, L. R. (2001, August). *Use and misuse of the* r_{WG} *coefficient of within-group agreement: Review and suggestions for future research. Paper presented the annual meeting of the Academy of Management, Washington, DC.*

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420-428.

Smith, G. M. (1970). *A simplified guide to statistics for psychology and education* (4th ed.). New York: Holt, Rinehart & Winston.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology*, *73*, 497-506.

Varma, A., DeNisi, A. S., & Peters, L. H. (1996). Interpersonal affect and performance appraisal: A field study. *Personnel Psychology*, *49*, 341-360.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, *81*, 557-574.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influences. *Journal of Applied Psychology*, *90*(1), 108-131.

James M. LeBreton is an associate professor of psychological sciences at Purdue University. He earned a PhD in industrial-organizational psychology from the University of Tennessee. He also earned his BS in psychology and MS in industrial-organizational psychology from the Department of Psychology at Illinois State University. He conducts research and consults in the areas of personality measurement and quantitative methods.

Jenell L. Senter is an ABD doctoral student in industrial/organizational psychology at Wayne State University in Detroit, Michigan, where she completed her MA. She received her BS in psychology from the College of Charleston, South Carolina. Her research concerns the heterogeneity of the part-time workforce, maladaptive workplace behavior, and applied statistical methods.

Journal of Business and Psychology (2023) 38:239–258
https://doi.org/10.1007/s10869-022-09853-9

ORIGINAL PAPER

Data Aggregation in Multilevel Research: Best Practice Recommendations and Tools for Moving Forward

James M. LeBreton[1] · **Amanda N. Moeller**[1] · **Jenell L. S. Wittmer**[2]

Accepted: 16 October 2022 / Published online: 24 November 2022

Abstract

The multilevel paradigm is omnipresent in the organizational sciences, with scholars recognizing data are almost always nested – either hierarchically (e.g., individuals within teams) or temporally (e.g., repeated observations within individuals). The multilevel paradigm is moored in the assumption that relationships between constructs often reside across different levels, often requiring data from a lower-level (e.g., employee-level justice perceptions) to be aggregated to a higher-level (e.g., team-level justice climate). Given the increased scrutiny in the social sciences around issues of clarity, transparency, and reproducibility, this paper first introduces a set of data aggregation principles that are then used to guide a brief literature review. We found that reporting practices related to data aggregation are quite variable with little standardization as to what information and statistics are included by authors. We conclude our paper with a Data Aggregation Checklist and a new R package, WGA (Within-Group Agreement & Aggregation), intended to improve the clarity and transparency of future multilevel studies.

Keywords Multilevel modeling · Multilevel analysis · Interrater agreement · r_{WG} · Data aggregation

Over the last several decades, organizational scholars have embraced the multilevel paradigm as one of the primary organizing frameworks for their scholarship. It is important to recognize that phrases such as "multilevel research" and "multilevel modeling" are generic phrases that are often used when referring to three distinct but interconnected facets of the multilevel paradigm: multilevel theory, multilevel measurement & design, and multilevel analysis (Humphrey & LeBreton, 2019). *Multilevel theory* refers to theories spanning multiple construct levels, which may be hierarchically and/or temporally nested (cf. Chen et al., 2005; Cronin & Vancouver, 2019; Dansereau et al., 1999; George & Jones, 2000; Gully & Phillips, 2019; House et al., 1995; Klein et al., 1994; Klein et al., 1999; Kozlowski & Klein, 2000; Mathieu & Luciano, 2019; Mitchell & James, 2001; Morgeson & Hofmann, 1999). These theories focus on explicating the relationships that exist between constructs residing across different levels including individual, dyadic, group/team, departmental, divisional, organizational, industry, and even geographic areas. Multilevel theory allows researchers to generate hypotheses regarding the relationships between constructs, both within and across levels. To wit, a central tenet of the multilevel paradigm is that constructs at one level of analysis (e.g., climate at the team-level) may impact other constructs at similar or lower levels (e.g., efficacy at the team level; job satisfaction at the individual level).

One of the distinct qualities of multilevel theory is that higher-level constructs (e.g., team justice climate) may have origins in lower-level units (e.g., individuals' perceptions of workplace justice). The higher-level construct is said to emerge (i.e., cohere into a new structure) through the interactions of the lower-level units (e.g., team-member interactions with one another and/or their team leader; team-member exposure to critical events and experiences). The higher-level construct (e.g., team climate) may function (i.e., exert a causal impact on other variables in the multilevel-nomological network) in a manner that is distinct from the lower-level construct from which it originates (e.g., justice climate may exert a unique influence on individual-level job-satisfaction, even after controlling for individual-level perceptions of justice; cf. Firebaugh, 1978; Gully & Phillips,

✉ James M. LeBreton
james.lebreton@psu.edu

[1] Department of Psychology, Pennsylvania State University, University Park, PA 16802, USA

[2] Department of Management, University of Toledo, Toledo, USA

2019; Kozlowski & Klein, 2000; Kozlowski et al., 2013; Mathieu & Luciano, 2019; Morgeson & Hofmann, 1999; Ostroff, 1993).

Multilevel measurement (and design) refers to the methodological aspects of multilevel research. These aspects include topics such as sampling of observations at different levels of analysis and the phrasing of questions on surveys (Zhou et al., 2019), the scaling (centering) of measures (Hofmann & Gavin, 1998; Kreft & de Leeuw, 1998), estimating the statistical power of tests of multilevel hypotheses (Mathieu et al., 2012; Scherbaum & Ferreter, 2009; Scherbaum & Pesner, 2019), addressing issues of missing data (Grund et al., 2018, 2019), designing studies to capture the actual *processes* of emergence (Kozlowski et al., 2013; Mathieu & Luciano, 2019), and estimating effect sizes to describe the impact of nesting lower-level units in higher-level units, as well as the variance explained in lower-level outcomes using both higher-level and lower-level predictors (cf. Aguinis & Culpepper, 2015; Hofmann, 1997; LaHuis et al., 2014; LaHuis et al., 2019).

One of the most important elements of multilevel research involves the measurement of (and accumulation of validity evidence for inferences related to) focal constructs (Chen et al., 2005; Tay et al., 2014; Jebb et al., 2019). Complicating the measurement and validation of focal constructs is the fact that some constructs may need to be measured at a lower level (e.g., individual-level perceptions of work environment) before being aggregated to a higher level (e.g., means representing team-level climate; cf. Bliese, 2000; Bliese et al., 2019; Chan, 1998; James, 1982; James et al., 1984; Kozlowski & Hattrup, 1992; Kozlowski & Klein, 2000; Kozlowski et al., 2013; Krasikova & LeBreton, 2019; Mathieu & Luciano, 2019; Wittmer & LeBreton, 2021).

Finally, *multilevel analysis* refers to the statistical and inferential aspects of multilevel research. These aspects include broad analytic frameworks such as random coefficient regression analyses (Bliese et al., 2018; Hofmann, 1997; Hox, 2010; Raudenbush & Bryk, 2002; Shiverdecker & LeBreton, 2019; Snijders & Bosker, 2012), multilevel structural equations analyses (Heck & Thomas, 2015; Mehta & Neale, 2005; Preacher et al., 2010; Vandenberg & Richardson, 2019), growth models (Bliese & Ployhart, 2002; Byrk & Raudenbush, 1987), ecological momentary assessment (Beal & Weiss, 2003; Shiffman, 2014), multilevel social network analyses (Borgatti et al., 2013; Borgatti, Mehra, Brass & Labianca, 2009; Brass & Borgatti, 2019), and approaches for analyzing dyadic data (Atkins, 2005; Kenny et al., 2006; Knight & Humphrey, 2019; Krasikova & LeBreton, 2012).

Although multilevel theory, measurement, and analysis are often discussed as though they are separable and distinct aspects of multilevel research, they are nevertheless inextricably intertwined (Rousseau, 1985). Multilevel measurement is sterile without a deep understanding of the theoretical underpinnings of the focal constructs — Where do these constructs originate? What is their structure and function? Who is best suited to provide information about each of the focal constructs? What is the appropriate referent for evaluating focal constructs? If lower-level data (e.g., individual perceptions) are to be aggregated to a higher level for analysis (e.g., team-climate), does one's multilevel theory require substantial agreement among the lower-level units prior to aggregating the data to the higher level? Likewise, the inferences drawn from a multilevel analysis may be severely misguided without a deep understanding of how theory and measurement shape the data analytic plan — How does the scaling of variables impact the accuracy of inferences drawn from multilevel analyses? What type of analysis is appropriate for hypotheses examining reciprocity effects between members of distinguishable dyads? How should I scale my data to provide unbiased tests of a cross-level moderation hypothesis? Finally, the accuracy of one's inferences from a multilevel analysis is conditional on the proper alignment of theory, method, and analysis. Stated alternatively, failure to align one's statistical analyses with one's theory and the proper measurement of focal constructs may result in drawing improper inferences and conclusions (cf. Bliese & Hanges, 2004; Firebaugh, 1978; Greenland, 2002; James & Williams, 2000; Lopilato & Vandenberg, 2014; Mossholder & Bedeian 1983; Ostroff, 1993; Yammarino & Gooty, 2019; Zhang et al., 2009).

Much has been written about the critical role of multilevel measurement in the alignment of theory, measurement, analysis, and inference (cf. Bliese, 2000; Chan, 1998; James et al., 1984, 1993; Klein & Kozlowski, 2000; LeBreton & Senter, 2008; Ostroff, 1993). However, ambiguity in how decisions related to multilevel measurement are reported (or not reported) in contemporary multilevel studies draws into question the reproducibility of previous findings, thus undermining the confidence researchers may have in inferences and conclusions obtained in prior multilevel studies.

Toward Improved Clarity and Reproducibility of Multilevel Research

Researchers in the social sciences are navigating what some have referred to as a reproducibility or replicability crisis. This is an issue regarding the credibility surrounding scholarly knowledge due to the lack of ability of scholars to replicate the results of published research findings. For example, researchers for *The Reproducibility Project: Psychology* found that only 37% of published, statistically significant findings could be replicated. The other 63% of findings resulted in statistically non-significant results when researchers sought to replicate them (Open Science Collaboration,

2015). To address concerns related to reproducibility and replicability, researchers have begun adopting open science practices designed to provide greater levels of clarity and transparency regarding their data, methods, and analyses (Banks et al., 2019). Indeed, two of the defining characteristics of influential and robust scholarship include transparency and the ability to replicate prior findings (Grand et al., 2018). To reproduce published findings or replicate those findings using new data, the primary sources must provide clear and transparent descriptions of their methods and analyses. The purpose of the current paper is to contribute to this conversation about transparency and clarity, specifically, within the context of multilevel research.

Consider multilevel researchers who compute statistics such as r_{WG} (James et al., 1984) to help justify aggregating lower-level data to a higher-level. These researchers must make decisions concerning the estimation of the statistic (e.g., which version of r_{WG} to use; which null distribution or distributions to use; what value of r_{WG} will be considered "sufficient" for justifying data aggregation; how out-of-range values will be treated; how groups lacking sufficient agreement will be treated). The replicability of multilevel research hinges on researchers clearly reporting these decisions and the rationale they used to make them.

The purpose of the current paper is to review data aggregation reporting practices and to offer a set of recommendations that may further improve the clarity of reporting practices. Our paper is structured as follows. First, we discuss important issues related to data aggregation within the context of multilevel research. From this discussion, we extract a few key recommendations that we believe, if followed, will result in greater transparency and reproducibility of multilevel research. Next, we summarize the results of a brief review of multilevel studies and discuss the extent to which the reporting practices in these studies comport with our recommendations. Specifically, we summarize the type of information and statistics included and omitted from published multilevel studies involving data aggregation. We then provide a Data Aggregation Checklist derived from our recommendations. Finally, we introduce a new R package designed to facilitate the estimation and reporting of data aggregation statistics. We illustrate the utility of this package with a brief tutorial using publicly available data.

Moving forward, we will use an example of individuals (lower-level units) nested within teams (higher-level units). The focal construct is a higher-level construct, justice climate. This construct originates in the perceptions of individuals and is shaped over time by their interactions with one another, exposure to common environment, etc. This construct is based on a pooled constrained model of emergence (Kozlowski & Klein, 2000) and, thus, a consensus composition measurement model. Stated alternatively, we expect similar levels of justice perceptions within teams. Therefore, within-unit agreement statistics will be used to help justify data aggregation. The following notation is used throughout the remainder of this article:

X = an observed score, typically measured on an interval scale of measurement,

S_X^2 = the observed variance on X,

J = number items ranging from $j = 1$ to J,

K = number of lower-level units (e.g., team-members, raters) ranging from $k = 1$ to K,

N = number of higher-level units (e.g., teams, organizations) ranging from $i = 1$ to N,

M = the mean score on X for the lower-level units,

Md = the median score on X for the lower-level units.

Multilevel Theory: Origin, Structure, Function, and Measurement of Collective Constructs

In some instances, data aggregation is not necessary because the focal constructs are easily measured at the higher level. In these instances, the focal constructs have *global unit properties* which "originate and are manifest at the [higher] unit level" (p. 29; Kozlowski & Klein, 2000). These types of constructs are often driven by the structure or function of the higher-level unit (e.g., class size, class subject matter, experience of the teacher, age of the firm).

Alternatively, a researcher may be interested in a higher-level construct that is a function of lower-level variables. These constructs are said to "emerge" from the interactions of the lower-level units with one another and their unique environments. Kozlowski and Klein (2000) suggested that emergent constructs may be conceptualized as falling on a continuum ranging from isomorphic composition to discontinuous compilation. Composition models presume that the lower-level and higher-level constructs are the same (or very similar). In contrast, compilation models are used when the construct measured at a lower-level is believed to be functionally different from the aggregated, higher-order construct. Kozlowski and Klein's framework of emergence and multilevel measurement models extend and integrate work by Chan, (1998), and Bliese, (2000). For our purposes, we can simply distinguish between consensus composition models and compilation models.

Consensus Composition Models In the current paper, we focus on a subset of composition models that Chan, (1998) referred to as *consensus models*. Consensus models are appropriate when the existence of the higher-level construct (e.g., justice climate) is said to be conditional on the lower-level units (e.g., individuals) demonstrating sufficient agreement in their scores (e.g., shared individual-level perceptions of justice may be aggregated to form team-level justice

climate). Stated alternatively, the higher-level construct originates at the lower-level (e.g., individual-level perceptions of one's work environment) but is believed to emerge and function at the higher-level (e.g., team-level justice climate). In this instance, the higher-level construct is comprised of *shared unit properties* which "describe the characteristics that are common to – that is, shared by – the members of a [lower-level] unit" (p. 30; Kozlowski & Klein, 2000). Variations on this concept of consensus measurement models include Kozlowski & Klein's (2000) *convergent emergence model* and their *pooled constrained emergence model*, as well as Bliese's, (2000) *fuzzy composition model*. A critical aspect of using a consensus measurement model is the statistical demonstration of appropriate within-unit agreement *prior* to aggregating scores to the unit-level (Chan, 1998).

Compilation Models Compilation models are appropriate when the higher-level construct is a function of measurements taken on lower-level units, but there is no assumption that the lower-level and higher-level constructs are isomorphic with one another. For example, Chan's, (1998) *additive models* are those in which the higher-level construct is a sum or an average of lower-level ratings. These models are relatively straightforward because unlike the previously described models, neither the within-group agreement nor the within-group variance across lower-level units are theoretically important when aggregating data from the lower-level to the higher-level. Additive models are similar to Kozlowski & Klein's (2000) *pooled unconstrained* model of emergence and are most relevant when the higher-order construct is a composite of inputs from the lower-level units, but there is no assumption that the lower-level units are homogeneous in their scores. As an example, Kozlowski & Klein (2000) offered team performance — it is possible that in many teams, not all team members would contribute equally to the final team product.

Minimum/maximum models of emergence represent a notable shift in the direction of compilation measurement models (Kozlowski & Klein, 2000). Such models assign to the entire higher-level unit a score that is either the minimum or the maximum of the lower-level scores. For example, one disagreeable team member may be sufficient to engender team conflict. To test such a hypothesis, personality data from the individual level could be aggregated to the team level and correlated with team-level conflict. In this case, the lowest level of trait agreeableness within each team would be assigned to the team.

Chan's (1998) *dispersion model* is aligned with Kozlowski & Klein's (2000) *variance model* of emergence. In these models, the meaning of the higher-level construct is derived from the degree of heterogeneity, or variance, among scores taken at the lower level. In other words, the focal construct is operationalized as the degree of within-group variance or dispersion of scores. For example, the strength of team climate may be an important predictor of team-level outcomes, over and above the level of team climate. In this instance, the variability or standard deviation within teams could be computed and included in models that also contained climate level (i.e., team means; Colquitt et al., 2002; Schneider et al., 2002).

The *patterned emergence model* is anchored at the compilation end of the composition-compilation spectrum. This model "incorporates the assumption that emergence may manifest itself as different forms, and it views nonuniform patterns of dispersion as meaningful substantive phenomena" (Kozlowski & Klein, 2000; p. 73). Examples of patterned emergence could include compatible mental models, group diversity, and/or aspects of social networks (e.g., network centrality; network density).

Finally, Chan's (1998) *process model* differs from the models discussed up until this point in that it is focused on measuring processes that unfold over time rather than measuring static constructs at a single point in time. For example, a researcher might seek to measure how the climate of an organization emerges or how conflict among a team increases or decreases as time goes on. To maximize clarity and reproducibility of research findings, our first principle echoes the sentiments of Kozlowski and Klein (2000; p. 28):

Recommendation #1 Researchers should be explicit about the theory of their focal constructs. When higher-level constructs are based on the aggregation of data collected at lower levels, researchers should include a discussion of (a) where the constructs originate (i.e., level of origin), (b) the proposed process by which the lower-level data cohere into a new higher-level construct (i.e., process of emergence; the structure of the construct), and (c) the function of the higher-level construct (i.e., how the higher-level construct influences other variables).

Recommendation #2 When higher-level constructs are being estimated using the aggregation of data collected at lower levels, researchers should explicitly identify the multilevel measurement model (e.g., pooled constrained emergence model) used to connect their multilevel theory to their aggregation of data (i.e., what model was used to translate lower-level data into higher-level scores).

Multilevel Measurement: ICC(1)

One of the most commonly reported statistics in multilevel research is the *ICC(1)*, which refers to the intraclass correlation coefficient computed using the variance components obtained from a one-way random effects ANOVA (Bliese, 2000). It is simply an ANOVA where the clustering variable

(e.g., work teams) is used to partition the variance in scores obtained on lower-level units (e.g., team members' perceptions of justice and fairness) into between-groups and within-groups components. Given that multilevel designs are rarely balanced (i.e., higher-level units typically have different numbers of lower-level units nested within them), researchers are encouraged to estimate the variance components for the *ICC(1)* using a simple, "null" random coefficient regression model (see Hofmann, Griffin & Gavin, 2000; Raudenbush & Bryk, 2002; Shiverdecker & LeBreton, 2019). For a two-level model, the variance components from the null ANOVA model can be used to estimate *ICC(1)*:

$$ICC(1) = \frac{t_{00}}{t_{00} + \sigma^2} \tag{1}$$

where, t_{00} denotes the between-groups variance and σ^2 denotes the within-groups variance. Within the context of multilevel research, the ICC(1) value is interpreted as an effect size – the proportion of variability in the lower-level scores (e.g., individuals' perceptions of justice) that may be attributed to the nesting of those lower-level units (e.g., individuals) in the higher-level units (e.g., teams).

Returning to our justice climate example, if a researcher wishes to argue that team members' individual-level perceptions of justice are shaped by their environments (i.e., nesting of team members within different teams where each team has a different team leader), then it would be important to demonstrate that some of the variance in team member ratings varied across teams. Obtaining an *ICC(1)* = 0.10 indicates that 10% of the variance in individual-level perceptions may be attributed to the nesting of individuals within teams (i.e., between-groups variance); and thus, 90% of the variance in individual-level justice perceptions resides within teams. In sum, the *ICC(1)* is a critical statistic for demonstrating the non-independence of lower-level units due to their nesting within higher-level units. As LeBreton & Senter (2008) noted, even relatively small/modest *ICC(1)* values (e.g., 0.05) may underlie important emergent phenomena. Consequently, when lower-level data are going to be aggregated to a higher-level, the estimation and reporting of ICC(1) values provides important information about the extent to which lower-level scores differ across higher-level units.

Recommendation #3 Researchers should include estimates of *ICC(1)* for all lower-level variables that are being aggregated to higher-levels.

Multilevel Measurement: ICC(2)

In addition, when the multilevel measurement model involves aggregating scores by computing unit-level means (e.g., justice climate computed as the mean of the team members' scores), it is useful to have information about how reliably these means discriminate between groups. This information is furnished by the *ICC(2)*, which is sometimes denoted the *ICC(k)* or *ICC(1,k)*. This statistic is also computed the variance components from the one-way random effects ANOVA:

$$ICC(2) = \frac{t_{00}}{t_{00} + \sigma^2/K} \tag{2}$$

As Bliese (2000) noted, the *ICC(2)* is roughly equivalent to applying the Spearman-Brown prophecy equation to the *ICC(1)*:

$$ICC(2) = \frac{K(ICC(1))}{1 + (K-1)ICC(1)} \tag{3}$$

where, K denotes the number of lower-level units (e.g., individuals) nested in a particular higher-level unit (e.g., team). Bliese (1998) found that Eq. (3) yields estimates of *ICC(2)* that asymptotically approach those of Eq. (2) as the number of lower-level units nested in higher-level units increases (e.g., as team size increases). The *ICC(2)* may be interpreted as the stability or the reliability of group means. Thus, it provides researchers with a sense of how effective group mean scores are at distinguishing between the different groups. As Bliese et al. (2018) noted, "Substantial *ICC(2)* values are not necessary for identifying emergent group-level effects (but they help)." (p. 8). We also see value in estimating and reporting *ICC(2)*, as consistent reporting of these statistics will also improve the clarity and transparency of research:

Recommendation #4 Estimates of *ICC(2)* should be reported when unit-level means are computed to serve as aggregate-level variables.

When the number of individuals nested within each team is identical (e.g., each team has exactly 5 team members) then Eq. (2) yields identical estimates of *ICC(2)* for each group. Bliese (2000) noted that in this instance, "the *ICC(2)* is equivalent to the overall sample-mean reliability estimate $\hat{\lambda}$, discussed by Bryk & Raudenbush (1992, p. 63)" (p. 356). However, in many instances, k will differ across higher-level units (e.g., different team sizes). In such instances, an examination of Eq. (2) reveals that larger teams will have larger estimates of *ICC(2)* and smaller teams will have smaller estimates. This finding makes both mathematical and intuitive sense. All things being equal, group means computed using larger group sizes will be more consistent/reliable compared to group means computing using smaller group sizes.

When group sizes are unbalanced, researchers should clarify the estimation and interpretation of *ICC(2)* values. To illustrate, assume we collected data from 70 teams with

sizes ranging from 3 to 13 team members. We could report *ICC(2)* values in a number of different ways. For example, we might compute multiple estimates of *ICC(2)* using the minimum ($k = 3$) and maximum ($k = 13$) sample sizes. These estimates would convey how the range of group sample sizes impacted the precision with which we were able to estimate group means. Alternatively, we might use Eq. (2) to compute a single estimate of *ICC(2)*, where we set k equal to some specific value (e.g., minimum, maximum, mean, median, or mode of the group sample sizes). This would convey different information to our audience, conditional on the specific group size used in Eq. (2). For example, a researcher might use the minimum group size to estimate *ICC(2)*, and note that this is a very conservative estimate of mean reliability because most means were computed using larger sample sizes. Finally, we could estimate 70 different *ICC(2)* values, one for each group. We could then report the mean or median of these *ICC(2)* values. In order to improve the clarity and transparency of multilevel research:

Recommendation #5 When data are *not* perfectly balanced, researchers should clarify how they computed *ICC(2)* and how the estimate(s) should be interpreted. Specifically, researchers should explain how they selected a number to represent group sizes (k) (e.g., mean, median, minimum, or maximum group size). At a minimum, we recommend using Eq. (2) to estimate *ICC(2)* and setting k equal to the median group size.

Multilevel Measurement: Within-Group Agreement Using r_{WG} & $r_{WG(J)}$

Compilation models do not require evidence of within-group consensus (agreement) prior to aggregating data to the group-level. Thus, for researchers aggregating data using compilation models, an estimate of *ICC(1)* might be all that is needed when aggregating data. If an additive model is being used, then *ICC(2)* might also be computed and reported. In contrast, researchers using consensus composition models to aggregate data must provide additional evidence of within-group agreement *prior* to aggregating data. Unfortunately, *ICC(1)* and *ICC(2)* do not provide sufficient evidence of within-group agreement (Krasikova & LeBreton, 2019; LeBreton, Burgess, Kaiser, Atchley & James, 2003; LeBreton & Senter, 2008; Wittmer & LeBreton, 2021). Instead, consensus models require that researchers compute estimates of within-unit agreement to justify data aggregation (Chan, 1998; Kozlowski & Klein, 2000). Evidence of within-unit agreement is often based on statistics such as r_{WG} (James et al., 1984, 1993), a_{WG} (agreement within-groups; Brown & Hauenstein, 2005), *AD* (average deviation; Burke et al., 1999), or *SD* (standard deviation; Schmidt & Hunter, 1989). Prior research has demonstrated that these different measures of within-unit agreement are often highly correlated with one another (Brown & Hauenstein, 2005; Burke et al., 1999; Roberson et al., 2007). This convergence across statistics is not necessarily surprising, as each statistic is based (in part) on the deviations of scores from lower-level units (e.g., individuals) from the unit-level (e.g., team) mean or median (LeBreton & Senter, 2008). Given the convergent conclusions reached using these different statistics, we decided to focus our attention on the most common estimates of within-group agreement – r_{WG} and $r_{WG(J)}$.

r_{WG}

The most commonly used estimates of within-group agreement are James et al.'s (1984, 1993) single-item r_{WG} and multi-item $r_{WG(J)}$. If multiple judges (e.g., team members) evaluate a single target (e.g., team environment) on a single item, then agreement may be assessed using r_{WG}. When multiple judges evaluate a single target using J parallel items, then within-unit agreement may be assessed using $r_{WG(J)}$.

The r_{WG} statistic defines agreement in terms of the proportional reduction in error variance. The assumption of r_{WG} is that each higher-level unit (e.g., team) has a single "true score" on the focal construct (e.g., justice). Thus, any observed variance within units (e.g., individual-level perceptions of justice in the workplace) may be attributed to random error variance.[1]

Agreement is estimated using r_{WG} by comparing the observed variance within each group to the variance that would be expected if judges' scores were completely due to random error (James et al., 1984):

$$r_{WG} = 1 - \frac{S_X^2}{\sigma_E^2} \tag{4}$$

where, σ_E^2 is the estimate of the error variance for the null distribution of scores that would be expected when the lower-level units scores were due solely to random responding (i.e., judges or employees responded completely at random to this single item). One definition of random error is a uniform, rectangular, or equal probability distribution. However, James et al. (1984) noted that in some cases, the uniform distribution may not be appropriate. Specifically, James et al. noted that response biases in the form of "systematic errors or bias" (p. 89) may impact how researchers opt to define random responding. For example, if a *leniency bias* is present among the lower-level units, then, even

[1] Newman and Sin (2020) introduced alternative estimates of within-group agreement that allow for each group to have a specific true score and for each individual within the group to also have a separate true score. The reader is encouraged to review Newman and Sin for more information about these new statistics.

judges responding randomly would tend to assign scores from the higher end of the distribution. James et al. (1984) recommended that researchers use the extant literature to guide their selection of null distributions.

If multiple possible null distributions are identified, researchers should compute multiple versions of r_{WG} using different error variance estimates corresponding to the different null distributions. LeBreton & Senter (2008) reiterated the importance of evaluating multiple null distributions. They also provided researchers with point-estimates for the random response error variance that might be expected for different distributions (e.g., slightly skewed, triangular) using a range of different Likert-type scales (e.g., 5-point, 7-point). Meyer et al. (2014) noted that despite recommendations to use multiple null distributions, most researchers have relied on the uniform or rectangular null. These authors suggested that this reliance may be due, in part, to a lack of guidance regarding best practices for identifying and selecting alternative distributions. To address this lack of guidance, the Meyer et al. (2014) paper introduced a framework to help researchers identify, a priori, what biases are most likely to be impacting their data.

Guidelines for Selecting Null Distributions

The framework Meyer et al. (2014) offered emphasizes how target-irrelevant, nonrandom forces can engender certain response biases based on features of the environment. And, in turn, these response biases should impact how we define random responding, and thus how we select null distributions for estimating r_{WG}. Target-irrelevant, nonrandom forces are factors that systematically influence ratings, independent of the target's true standing on the construct. They can be thought of as situational cues that encourage individuals to adopt certain patterns of responding. It is important to note that these forces act upon individuals. However, when individuals are nested within a common group, it is likely that they will be exposed to similar external forces. The Meyer et al. framework lays out specific examples of target-irrelevant nonrandom forces that can be categorized into a "5Ws and an H" framework (i.e., who, what, when, where, why, and how).

Briefly, researchers must consider *who* is providing the ratings and how characteristics of raters might impact the distributions of responses. For example, individuals with above-average levels of agreeableness are more likely to provide lenient ratings (Bernardin, Cooke & Villanova, 2000). In such an instance, researchers might opt to use a skewed null distribution because of the presence of a leniency bias in the ratings.

Next, researchers must consider *what* is being rated. Meyer et al., (2014) pointed to evidence wherein individuals are more likely to select more socially desirable response options when self-reporting on affective variables, like anxiety. Here, researchers might opt to select a skewed or a slightly skewed null response distribution when calculating r_{WG}.

Researchers must also consider *when* ratings are collected and other time-sensitive information. Meyer et al. noted that anchoring effects (i.e., tendencies to respond at extreme ends of a rating scale) are more likely when raters are under time pressure (Edland & Svenson, 1993; Pennington & Roese, 2003). In such an instance, researchers might again opt to use a skewed null distribution when calculating r_{WG}.

Next, researchers might consider whether response distributions are affected by *where* rater responses are collected. This could involve whether ratings are collected in a lab environment or in a naturalistic setting, or even the broader cultural context in which ratings are provided. For example, individuals belonging to collectivistic cultures are more likely to adopt a "modesty" bias (i.e., a tendency to diminish one's own performance of individual characteristics) compared to raters belonging to individualistic cultures. Error distributions could certainly be affected by the context in which ratings are provided and should be considered when selecting a null response distribution.

Meyer et al., (2014) suggested that the most important question researchers should consider when selecting a null distribution is *why* ratings are being collected. They pointed to cases in which raters were more likely to respond with a leniency bias when informed their ratings will be used to make important administrative decisions, whereas raters tend to respond with less of a leniency bias when told their ratings are being collected for research or developmental purposes only (Cleveland, Murphy & Williams, 1989; Murphy Jako & Anhalt, 1993). Calculating r_{WG} using data collected for administrative/consequential purposes might benefit from using a moderately skewed distribution, whereas data collected for less consequential purposes might use a slightly skewed or a uniform distribution.

Finally, researchers ought to consider *how* ratings are collected. For example, socially desirable responses are more likely when ratings are not anonymous, whereas ratings provided anonymously are less likely to group toward the more favorable ends of a response scale (London, Smither & Adsit, 1997). Researchers aggregating non-anonymous ratings might consider using a skewed null distribution. The 5 Ws and an H framework proposed by Meyer et al., 2014 provides a useful starting point for considering and identifying target-irrelevant nonrandom forces that may affect error distributions under a range of conditions.

Once potential response biases have been identified and random response distributions have been specified, researchers will then need to obtain and report the specific point-estimates they used for σ_E^2 when computing estimates of r_{WG}. LeBreton & Senter (2008) and Krasikova & LeBreton (2019)

provided point-estimates for different Likert-type response scales (5-point, 6-point, 7-point, etc.) across different distributions (uniform, slightly skewed, moderately skewed, etc.). Alternatively, researchers may locally compute estimates of σ_E^2 corresponding to distributions not included LeBreton & Senter (2008) or Krasikova & LeBreton (2012).

r_{WG} is estimated as the proportional reduction in error by comparing the observed variance (i.e., S_X^2) to the variance that would be expected if judges scores were due to solely to random measurement error (i.e., random responding; σ_E^2) and subtracting this ratio from 1. If all judges provide the same rating of the target (perfect consensus or agreement), the observed variance among judges is zero, and $r_{WG} = 1.0$. If the judges have a total lack of agreement, the observed variance will asymptotically approach the error variance obtained from the null distribution (as the number of judges increases), leading r_{WG} to approach 0.0. Thus, r_{WG} is interpreted as the proportional reduction in error variance. It is important to note that unless values are bounded (due to being outside the 0 to 1 range), the different values of r_{WG} will be directly proportional to each other; it is only the absolute values that will change based upon changing the null distribution.

$r_{WG(J)}$

The r_{WG} family of indices has been extended to situations where a single target is rated by multiple raters on J items, $r_{WG(J)}$ (James et al., 1984, 1993):

$$r_{WG(J)} = \frac{J\left[1 - \frac{\overline{S_{X_j}^2}}{\sigma_E^2}\right]}{J\left[1 - \frac{\overline{S_{X_j}^2}}{\sigma_E^2}\right] + \left[\frac{\overline{S_{X_j}^2}}{\sigma_E^2}\right]} \quad (5)$$

where, $r_{WG(J)}$ furnishes an estimate of agreement for judges mean scores on J essentially parallel items and $\overline{S_{X_j}^2}$ refers to the mean of the observed item variances for the J items. Once the appropriate estimates of σ_E^2 have been obtained, $r_{WG(J)}$ is computed using Eq. (5) and is again interpreted as the proportion reduction in error variance. To maximize clarity and reproducibility:

Recommendation #6 Researchers using r_{WG} or $r_{WG(J)}$ should a) clearly identify the null distributions used to obtain the estimates of σ_E^2, b) include explanations for why those distributions were judged to be most appropriate, including a discussion of response biases (if relevant), and c) explain where/how the point-estimates for the error variances (i.e., σ_E^2) were obtained (pulled from prior studies, based on tabled values, computed locally for the current study, etc.).

Multilevel Measurement: Additional Estimates of Within-Group Agreement

Despite all estimates generally "pointing in the same direction," LeBreton & Senter (2008) suggested some researchers may wish to consider reporting multiple measures of within-group agreement. Specifically, the authors noted that it could be helpful to include one estimate of agreement scaled on the 0 to 1 metric (e.g., r_{WG}/$r_{WG(J)}$, a_{WG}/$a_{WG(J)}$) and one scaled in the metric of the original items (e.g., AD_M, SD). Using multiple measures is not required for data aggregation, but it may help readers have a better understanding of the pattern and magnitude of within-group agreement. Below we briefly summarize alternatives to r_{WG} and $r_{WG(J)}$.

a_{WG} and $a_{WG(J)}$

Brown & Hausenstein (2005) introduced a new measure of within-unit agreement denoted a_{WG}. They noted that the r_{WG} indices are scale dependent, in that the lower bound of any r_{WG} index will be conditional on the number of scale anchors. They also noted that r_{WG} is directly influenced by the number of judges, potentially complicating interpretations. To address these concerns, they introduced a new estimate of within-group agreement:

$$a_{WG} = 1 - \frac{2 * S_X^2}{S_{mpv|M}^2} \quad (6)$$

where $S_{mpv|M}^2$ refers to the maximum possible observed variance in X, given the observed sample mean for X, which may be estimated and inserted the previous equation yielding:

$$a_{WG} = 1 - \frac{2 * S_X^2}{[H + L) * M - M^2 - H * L)] * [k/(k-1)]} \quad (7)$$

This statistic assumes a range of values from -1 (perfect disagreement) to 0 (perfect lack of agreement) to + 1 perfect agreement. They noted that "dissensus is more heavily penalized in the estimate of agreement if it occurs at the extremes of the rating scale." They also offer a multi-item variant:

$$a_{WG(J)} = \frac{\Sigma a_{WG(j)}}{J} \quad (8)$$

Average Deviation

Burke et al. (1999) proposed the average deviation (AD) index to estimate within-unit agreement. This measure, like r_{WG}, was developed for situations where multiple judges (i.e., lower-level units) are providing scores on a single target (i.e., a higher-level unit). One of the advantages of the

AD metric is that it estimates agreement in the metric of the original scale of the item. This provides an intuitive and practical metric for evaluating within-unit consensus. The *AD* index for a single item evaluated by multiple judges may be estimated using deviations from either the judges' means (i.e., AD_M) or medians (i.e., AD_{Md}):

$$AD_{M(j)} = \frac{\sum_{k=1}^{K}|X_k - M|}{K} \tag{9}$$

or

$$AD_{Md(j)} = \frac{\sum_{k=1}^{K}|X_k - Md|}{K} \tag{10}$$

These statistics also have multi-item analogs that are estimated similarly to $a_{WG(J)}$:

$$AD_{M(J)} = \frac{\sum_{j=1}^{J}AD_{M(j)}}{J} \tag{11}$$

and

$$AD_{Md(J)} = \frac{\sum_{j=1}^{J}AD_{Md(j)}}{J} \tag{12}$$

Compared to the previous estimates of within-unit agreement, the *AD* statistics inform researchers about the (average) *lack of agreement* within-units. Higher scores indicate greater lack of agreement, whereas scores of 0 indicate perfect agreement.

Standard Deviation

Researchers may also estimate within-unit agreement by calculating a simple within-unit standard deviation (SD_X) and the standard error of the mean (Schmidt & Hunter, 1989). SD_X is an intuitively appealing statistic to use when the form of emergence is moored to a compilation dispersion measurement model (LeBreton & Senter, 2008). For example, researchers have tested how climate strength may be related to important outcomes and have estimated climate strength using the within-unit SD_X (Colquitt et al., 2002; Schneider et al., 2002).

Interpreting & Reporting Results: Criteria for Aggregation

Several authors have advanced different criteria or thresholds that should be met prior to aggregating lower-level data to higher levels. For example, LeBreton & Senter (2008) suggested that criteria used to make decisions about data aggregation should be considered within the broader context of the project. They suggested that the levels of within-unit agreement required for data aggregation may vary as a function of the purpose of the project (e.g., basic research study on climate vs. administrative decision about hiring a new CEO using ratings from structured panel interviews). Similarly, expectations for within-group agreement may vary as a function of the quality of measures used in a study (e.g., lower levels of agreement might be acceptable for new measures in early stages of development versus higher levels of agreement that might be expected using gold standard measures). Finally, the types of measurement models used to aggregate data and the specific use of the aggregated data will also influence the criteria or thresholds used to justify data aggregation.

Once researchers have considered the context, measures, and purpose of aggregation, they should articulate the specific statistical criteria used to guide data aggregation decisions. For example, Table 3 in LeBreton and Senter introduced a taxonomy of practical effect sizes for use with the r_{WG}/a_{WG} families of statistics. Additionally, criteria may include the use of other cutoffs for agreement or cutoffs moored to statistical significance testing (cf. Bliese & Halverson, 2002; Burke et al., 2017; Cohen et al., 2001; Dunlap et al., 2003; LeBreton & Senter, 2008; Smith-Crowe et al., 2014; Smith-Crowe et al., 2012; Woehr et al., 2015). Most importantly, researchers must clearly articulate the criteria used to guide decisions about data aggregation. These criteria should be anchored to appropriate theories of emergence and corresponding multilevel level measurement models. Guidance on specific criteria for different indices of agreement are available in previously cited works.

To illustrate, consider two research teams, both studying similar constructs, that invoked different theories of emergence requiring different measurement models and thus, different criteria (or lack thereof) to justify data aggregation. Schneider et al. (2002) sought to examine how the level and strength of customer service climate predicted customer satisfaction. In this study, they aggregated all data to the level of bank branches. To compute climate level, they relied on a direct consensus measurement model. They noted, "The direct consensus model is the one most frequently discussed in research on organizational climate because shared perceptual agreement at the individual level of analysis has been seen as functionally isomorphic to the construct at the organizational level" (p. 221). Thus, to justify data aggregation using a consensus model, their aggregation criteria consisted of $r_{WG(J)} > 0.70$. To compute climate strength, they used a dispersion/variance compilation model and simply estimated the standard deviation within each bank branch.

Similarly, Colquitt et al. (2002) sought to examine how the level and strength of justice climate at the team level was related to team-level performance and team-level absenteeism. To compute climate strength, these authors estimated the within-team standard deviation. Contrasting the study

Table 1 Examples of constructs, measures, and circumstances that warrant different levels of agreement

Level of agreement	Interpretation	Illustrative examples	Explanation
.00 to .30	Lack of agreement	N/A	If calculating interrater agreement, there are few (if any) instances where you would reasonably expect and accept a lack of agreement between raters
.31 to .50	Weak agreement	Climate strength	Ideally, we would expect a group to have some degree of a climate. However, climate does not have to be strong. Weak climates exist when variability exists in the way that group members perceive the climate, so strong agreement may not be necessary
.51 to .70	Moderate agreement	Group cohesion captured using a relatively new measure	We might expect some degree of agreement for a construct like group cohesion, but if group cohesion is assessed using a newly designed measure that has not been subjected to substantial psychometric evaluation, we might not expect strong agreement
.71 to .90	Strong agreement	Group cohesion captured using a well-established, validated measure	Again, we might expect some degree of agreement for a construct like group cohesion. If group cohesion is measured using a well-established and validated measure, we might expect stronger agreement
.91 to 1.00	Very strong agreement	Panel interview ratings for critical decisions (e.g., decisions about hiring, promotion, firing, tenure)	Agreement between raters on constructs used to make important decisions should ideally be very strong

by Schneider et al., Colquitt and his colleagues *did not* use a consensus model of aggregation to compute climate level. Instead, these authors used a simple additive (pooled unconstrained) measurement model. This model did not require within-team agreement prior to aggregating scores to the team-level. Instead, these authors noted "A consensus model was not appropriate for this study because within-team variance in justice perceptions was substantively meaningful. Indeed, within-team variance was necessary for testing the climate strength hypotheses, as a lack of variance would indicate a restriction of range in that independent variable" (p.99).

Thus, two sets of climate researchers developed similar hypotheses about how climate level and climate strength would relate to important outcomes. However, each set of researchers invoked different theoretical explanations for how climate would emerge in their specific contexts. These different explanations for emergence necessitated different multilevel measurement models with different data aggregation criteria. One could debate which approach is more appropriate given the research question — such a debate is beyond the scope of the current paper. However, what is not debatable is the *transparency and clarity that both sets of authors provided in stating their theory of the collective constructs* (Recommendation #1), *the measurement model for their collective constructs* (Recommendation #2), *and the criteria used to justify data aggregation* (Recommendation #7). To enhance the clarity and transparency of future multilevel research:

Recommendation #7 When setting criteria to guide data aggregation decisions, researchers should (a) consider the quality of their measures, the form of measurement model, and the importance of agreement within the context of their research question, and (b) explicitly state the criteria used to guide aggregation decisions (e.g., minimum cut-off values based on statistical or practical significance). See Table 1 for examples of research questions and constructs that might warrant different levels of agreement.

Interpreting & Reporting Results: Treatment of Anomalies and Inconsistencies Anomalies

Given the recommendation to compute multiple agreement indices, it is possible that researchers may obtain anomalous estimates of within-group agreement. Specifically, it is possible that estimates of $r_{WG}/r_{WG(J)}$ could fall outside the range of normal values (i.e., less than 0 or greater than 1). In such instances, researchers should be transparent about obtaining such values and how they handled such values. LeBreton and colleagues (LeBreton & Senter, 2008; LeBreton et al., 2005) have discussed potential causes of out-of-bound estimates and provided suggestions for how to handle out-of-bound estimates.

Inconsistencies

In addition, given the large number of within-group agreement estimates, it is possible that evidence supporting data aggregation could be inconsistent across groups. For example, consider a researcher who sets a cut-score for data aggregation as $r_{WG(J)} > 0.50$. This researcher finds that 65 groups had $r_{WG(J)}$ values exceeding this threshold, but 5 groups had values falling below this threshold. How is the researcher to proceed? Following the recommendations laid out by LeBreton & Senter (2008), when mixed patterns of agreement are found in the data, researchers should (a) test their hypotheses using only the groups that met criteria for aggregation, (b) rerun the analyses using all groups (i.e., those meeting and those not meeting aggregation thresholds), and (c) report any substantive differences in findings. LeBreton and Senter also suggested that researchers could create a dummy variable to distinguish groups with and without sufficient agreement and use this variable as an ad hoc moderator when testing hypotheses. Irrespective of which solution is adopted, it is important that researchers clearly explain how they treated groups that lacked sufficient agreement to justify data aggregation. To maximize transparency and reproducibility:

Recommendation #8 Researchers obtaining separate estimates of agreement for each higher-level unit should clarify (a) whether any estimates of agreement fell out-of-bounds, and if so, how those estimates were handled, and (b) whether any higher-level units lacked sufficient agreement based upon predetermined standards of agreement and how those units were handled.

Reporting Results: Patterns of Within-Unit Agreement

It is important to remember that within-unit agreement statistics are typically estimated separately for each unit. In addition, when using $r_{WG}/r_{WG(J)}$, the number of estimates will further increase if multiple null response distributions are used. Take for example a researcher who has collected data on justice perceptions from individuals nested in 70 teams. If this researcher wishes to estimate $r_{WG(J)}$ using a uniform (rectangular null distribution) and a slightly skewed null distribution, then this researcher will need to compute 140 estimates of $r_{WG(J)}$. This researcher might also decide to include estimates of agreement computed using $AD_{M(J)}$. Thus, a total of 210 estimates of within-unit agreement will be computed for this study. When many estimates of agreement are being computed, it is important that researchers provide a description of agreement that conveys the general pattern of within-unit agreement (LeBreton & Senter, 2008). Specifically, to enhance the transparency and reproducibility of research:

Recommendation #9 Researchers obtaining separate estimates of agreement for each of their high-level units (using r_{WG}, a_{WG}, AD_{Mn}, etc.) should provide information about the overall patterns of agreement across their data. This may include an online supplemental file containing (a) descriptive statistics for the estimates of agreement (mean, SD, min, max), (b) a histogram to aid in visualizing the distribution of estimates, and/or (c) a description of the proportion of units that had estimates of agreement above relevant data aggregation thresholds (e.g., 80% of teams had $r_{WG(J)}$ values greater than .80 and 97% had values greater than .70).

Reporting Results: Simplified Summaries Using $r_{WGp}/r_{WGp(J)}$

If researchers justify data aggregation using $r_{WG}/r_{WG(J)}$, $a_{WG}/a_{WG(J)}$, and/or $AD_M/AD_{M(J)}$, they will be computing a large number of statistics (i.e., one estimate for each group). If these statistics are used, researchers should provide information about the general patterns of agreement (Recommendation #9) and include information about statistical anomalies or inconsistencies (Recommendation #8). One way to simplify the reporting of agreement and reduce the likelihood of spurious anomalies and/or inconsistencies is to estimate within-unit agreement using $r_{WGp}/r_{WGp(J)}$.

r_{WGp} & $r_{WGp(J)}$

The theoretical range of the r_{WG} and $r_{WG(J)}$ spans from 0, which indicates no reduction over random responding (i.e., complete lack of agreement) to 1, which indicates the total elimination of random responding (i.e., perfect agreement). However, as Lindell et al. (1999) noted, it is possible to obtain out-of-range values for these statistics (i.e., less than 0 or greater than 1). When these out-of-range values are small in magnitude, researchers typically assume these values were engendered by sampling error (i.e., small number of judges within units) and simply reset these values to 0 (LeBreton et al., 2005). However, variations on r_{WG} and $r_{WG(J)}$ have been introduced, in part, to help address issues of out-of-range values (Lindell & Brandt, 1997; Lindell et al., 1999).

More recently, LeBreton et al. (2005) noted that one reason for obtaining out-of-range values could be the *incorrect* use of r_{WG} or $r_{WG(J)}$ in situations where the target of measurement (i.e., higher-level units) could have multiple "true scores" (e.g., a team leader may have multiple true scores on a measure of trust, conditional on whether the data are obtained from members of their in-group or out-group; an organization may not have a single justice climate, but

rather multiple justice climates, conditional on the nesting of individuals within different teams). To address this situation, LeBreton et al. (2005) introduced an additional variant of r_{WG} computed using the pooled within-groups variance rather than the traditional observed group variance:

$$r_{WG_p} = 1 - \frac{S^2_{X.\tau}}{\sigma^2_E} \tag{13}$$

where $S^2_{X.\tau}$ denotes the residual variance in judges scores after removing the "treatment effect" associated with being nested within a particular sub-group (e.g., the effect of being the in-group versus the out-group; the effect of being nested within a particular team). A multi-item version was suggested by LeBreton & Senter (2008):

$$r_{WGp_{(J)}} = \frac{J\left[1 - \frac{\overline{S^2_{X.\tau_j}}}{\sigma^2_E}\right]}{J\left[1 - \frac{\overline{S^2_{X.\tau_j}}}{\sigma^2_E}\right] + \left[\frac{\overline{S^2_{X.\tau_j}}}{\sigma^2_E}\right]} \tag{14}$$

whereas the traditional r_{WG} and $r_{WG(J)}$ statistics provide separate estimates of agreement for each higher-level unit (e.g., team), using a pooled within-groups variance results in a single, global measure of within-unit agreement. LeBreton et al. (2005) and LeBreton & Senter (2008) noted in order to compute r_{WGp} and $r_{WGp_{(J)}}$, it is critical that researchers identify the different groups a priori.

Although the examples provided by these authors emphasized sub-groups within existing groups (i.e., in-group vs. out-group), these statistics are equally applicable to nearly all data aggregation situations in the organizational sciences. For example, let us assume we have 70 work teams, each containing 3 to 13 team members. Each person is asked to rate their perceptions of justice on a 7-point Likert type scale within a measure containing $J = 5$ items. We decide to estimate agreement assuming a uniform error distribution and a slightly skewed distribution. If we approached this scenario from a traditional perspective, we would compute 140 estimates of $r_{WG(J)}$ (2 estimates for each of the 70 groups). We would then be tasked with organizing and interpreting these 140 values to determine whether sufficient agreement existed to warrant data aggregation.

Following Recommendation #9, we might report descriptive statistics for agreement estimates, including a graphical representation (e.g., histogram) of the distribution of agreement estimates, and summarizing the proportion of estimates that exceeded some a priori threshold or cutoff. An alternative or complementary approach would be to estimate $r_{WGp_{(J)}}$, which would return only two global estimates of agreement, one for each of the proposed error distributions. We would then compare these global estimates to our criteria for data aggregation (e.g., $r_{WGp(J)} > 0.50$). To be clear, we are suggesting researchers treat each of the 50 groups as different subgroups within the organization, with each subgroup having a different true score on the focal construct (i.e., justice climate).

To maximize clarity and transparency:

Recommendation #10 Researchers are encouraged to include estimates of $r_{WGp}/r_{WGp_{(J)}}$ as indicators of overall/omnibus/global within-unit agreement.

Review of Data Aggregation in Multilevel Research

To get a sense of the clarity with which data aggregation decisions, especially those based on consensus composition measurement models, are described in the organizational sciences, we conducted a circumscribed literature review. Specifically, we identified empirical articles published in the five-year period between 2017 and 2021 in six prominent organizational journals (*Journal of Applied Psychology*, *Academy of Management Journal*, *Journal of Management*, *Personnel Psychology*, *Journal of Business and Psychology*, and *Journal of Organizational Behavior*). We conducted our search using Google Scholar and included papers that used the keyword r_{WG} in their paper as well as at least one of the following keywords: *multilevel, ICC, ICC1, ICC2, data aggregat*, ad, compilation model, consensus model,* or *multilevel model.* This search returned 99 articles. Of these 99, 8 were non-empirical review papers and were excluded from the analysis, resulting in a sample of 91 papers. Table 2 provides a summary of our recommendations and Table 3 a summary of our findings.

Multilevel Theory: Domain of Composite Constructs

Of the 91 papers we reviewed, three major content areas emerged. Most papers were focused on issues related to leadership ($n = 24$), work teams ($n = 43$), or organizational culture and climate ($n = 10$). Aggregated constructs within each paper were generally relevant to these three content domains, with leadership papers including aggregating constructs such as *leader moral humility, ethical leadership, and abusive supervision* (for example) where data were aggregated from the lower-level unit (followers, subordinates) to the higher-level unit (leaders, supervisors, CEOs). Papers focused on work teams and team functioning tended to aggregate constructs such as *team cohesion, teamwork behavior,* or *team conflict,* where aggregation was from the individual-level to the shared team-level. Papers about organizational culture and climate emphasized constructs

Table 2 Data Aggregation Checklist: best practice recommendations for data aggregation in multilevel research

1. Multilevel Theory – The Theory of Collective Constructs: For each higher-level construct that is an aggregate of lower-level data, researchers should include a description of:

a. The Level of Origin: Explain where the construct is believed to originate

b. The Process of Emergence: Explain the process by which lower-level data cohere into a higher-level construct

c. The Function of Collective Constructs: Explain how the higher-level construct is believed to impact and shape other variables in the local nomological network

2. Multilevel Theory – Measurement Models of Collective Constructs: For each higher-level construct that is based on the aggregation of lower-level data, researchers should describe the specific multilevel (pooled constrained emergence; minimum/maximum, etc.) measurement model used to translate lower-level data into higher-level data

3. Multilevel Measurement – ICC(1): For each higher-level variable researchers should estimate and interpret ICC(1) values

4. Multilevel Measurement – ICC(2): For each higher-level variable that is computed as the mean estimated from lower-level scores, researchers should estimate and interpret ICC(2) values

5. Multilevel Measurement – ICC(2) for Unbalanced Designs: When data are not perfectly balanced, researchers should clarify how they computed *ICC(2)* and how the estimate(s) should be interpreted. Specifically, researchers should explain how they selected a number to represent group sizes (*k*) (e.g., mean, median, minimum, or maximum group size). At a minimum, we recommend estimating ICC(2) setting *k* equal to the median group size

6. Multilevel Measurement – $r_{WG}/r_{WG(J)}$. Researchers using r_{WG} and/or $r_{WG(J)}$ to justify data aggregation should a) clearly identify the null error distributions used to obtain estimates of σ_E^2, b) include an explanation for why those distribuitons were judged to be appropriate, and c) clarify where/how the point-estimates of σ_E^2 were obtained

7. Interpreting & Reporting Results – Criteria for Data Aggregation. Researchers should include an explicit description of the criteria (or lack thereof) used to guide data aggregation decisions (cutoffs for agreement; significance testing, etc.). Researchers are encouraged to consider the quality of their measures (newly developed vs. gold standard), the form of their multilevel measurement model (consensus versus compilation), and the context/purpose of data aggregation when finalizing data aggregation criteria (climate study vs. structured panel interviews)

8. Interpreting & Reporting Results –Anomalies & Inconsistencies. Researches should clarify a) whether any estimates of agreement fell out-of-bounds (e.g., $r_{WG}/r_{WG(J)}$ values < 0 or > 1) and b) whether any higher-level units lacked sufficient agreement to justify aggregation (i.e., did any higher-level units have estimates of agreement that fell short of the criteria set out in #7)

9. Interpreting & Reporting Results – Patterns of Local Agreement. When separate estimates of agreement are computed for each higher-level unit, researchers should provide a summary of results that conveys the overall pattern of agreement across the higher-level units. At a minimum, researchers should include basic descriptive statistics (e.g., mean and SD of $r_{WG}/r_{WG(J)}$). Researchers are also encouraged to include a visual summary (e.g., histograms) and a summary of how many higher-level units had agreement values falling above different data aggregation thresholds. This information may be included in a footnote or supplement

10. Interpreting & Reporting Results – Global Estimates of Agreement Using $r_{WG(p)}$. As an addition or alternative to Recommendation #9, researchers are encouraged to provide an estimate of the overall or global within-unit agreement using $r_{WGp}/r_{WGp(J)}$

such as *emotional culture, distributive justice climate,* and *safety climate*, where aggregation was from the individual-level to the group- or organizational-level. The remaining 14 papers focused on a range of constructs including *machiavellianism* and *narcissism* (aggregated from the within person-level to the between person-level and *performance* (again, aggregated from the within person-level to the between person-level). In the section that follows, we review these papers vis-à-vis the principles for data aggregation identified in the first part of our paper.

Recommendations #1 and #2: Theory of the Constructs and Measurement Model

Researchers tended to include descriptions of the theoretical underpinnings of the constructs and described how those constructs originated at a lower-level and were aggregated to a higher level. Given that the search criteria included "r_{WG}", one could infer that all of the papers included at least one aggregated construct that was based on a consensus measurement model. However, only 26 out of the 91 papers (approximately 29% of papers) explicitly identified the type of multilevel measurement model used to aggregate data. Of these 26, all specified the use of a composition model.

We examined these findings within the context of the focal constructs studied in these papers. As a reminder, our four broad construct domains included *work teams* (43 papers), *leadership* (24 papers), *organizational culture and climate* (10 papers), and *miscellaneous* (14 papers). We found that papers focused on teams or culture/climate were more likely to include an explicit description of their measurement models, with 33% of the teams papers and 50% of the culture/climate papers including reference to a specific measurement model compared to only 17% of leadership papers and 8% of miscellaneous papers. Although we can infer from the papers that they included at least one aggregated construct based on a consensus model, it is unclear which specific model they used, and it is also unclear the extent to which levels of agreement were considered important when aggregating constructs to the higher level. In sum,

Table 3 Literature review findings of organizational research reporting data aggregation statistics

Evaluation criteria	% reporting	Exemplar cases of reporting from the literature review
Explicitly Reported the Multilevel Measurement Model Used	27%	"The aggregation test results... showed that a significant proportion of the variance in task role enactment was accounted for by team membership and team-level means for task role enactment were reliable. Thus, we used a direct consensus model (Chan, 1998) to operationalize team task role enactment, such that aggregated team member task role behaviors reflect team task role enactment." (Li et al., 2022)
Reported ICC(1) and ICC(2) Values	90%	"Leader moral humility, ethical leadership, and leader general humility were aggregated to the team level after computing within-group interrater agreement (r_{WG}; James et al., 1993) and intraclass correlation coefficient (ICC) values (James, 1982; Schneider, White & Paul, 1998). Following methodologists' guidelines (e.g.,Bliese, 2000; Kozlowski & Klein, 2000), data aggregation is considered appropriate when ICC1 is nonzero and when ICC2 is higher than .70. Leader moral humility had an average r_{WG} value of .87 with ICC [1, 2] values of 0.46 and 0.79... Thus, we proceeded to aggregate our data for these variables." (Owens et al., 2019)
Reported Additional Within-Group Agreement Statistics	2%	"First, interrater reliability and agreement analysis was conducted with ratings of LMX, because we used group member ratings for this construct. These should be composed at the group-level of analysis to capture the common perception of LMX within groups, using a direct consensus composition model (Chan, 1998). Thus, we estimated intraclass correlation ICC(1), average deviation (AD), and r_{WG} (M. J. Burke & Dunlap, 2002; LeBreton & Senter, 2008)." (Vasquez et al., 2021)
Defined the Null Distribution	25%	"Likewise, tests of within-group agreement revealed that workgroup members generally shared corresponding perceptions of workgroup safety climates. We calculated two $r_{WG(J)}$ estimates through moderately skewed (median $r_{WG[J]} = .97$) and uniform null (median $r_{WG[J]} = .99$) response distributions since these realistically demonstrate the lower and upper bounds of within-group agreement levels across these samples' workgroups. Uniform response distributions assume that individuals within a workgroup are equally likely to endorse each response option, whereas a moderately skewed response distribution assumes negative skew, or a greater likelihood of responding favorably to items (LeBreton & Senter, 2008)." (Beus et al., 2019)
Used More than One Null Distribution to calculate r_{WG}	12%	"Following the recommendation of LeBreton & Senter (2008), we examined the value of r_{WG} with different null distributions, namely, slight skew, triangular, and rectangular distributions. The median r_{WG} (James et al., 1993) values of psychological safety were as follows: .86 for slight skew distribution, .84 for triangular distribution, and .91 for rectangular distribution.2 They were statistically significant ($p \leq .05$) compared to the corresponding critical r_{WG} values for the 5- and 10-item scales (e.g., Smith-Crowe et al., 2014). Furthermore, the intraclass correlation coefficient-1 (ICC1) was .29 and the ICC2 was .64. The values of these indices supported the aggregation (James, 1982)." (Deng et al., 2019)
Reported the Range of r_{WG} Value(s)	3%	See Table 1 in Nielsen et al., (2021)
Reported the Mean r_{WG} Value(s)	74%	
Reported the Median r_{WG} Value(s)	33%	
Reported the SD of r_{WG} Value(s)	3%	"The $r_{WG(J)}$ scores for abusive supervision ($M = .83$, $SD = .29$) and empowering leadership ($M = .86$, $SD = .27$) demonstrated a sufficient degree of agreement among team members to aggregate the follower-rated constructs of abusive and empowering leadership in order to evaluate the indirect relationship of team helping behavior to each outcome variable as mediated by team positive affective tone and moderated at the second stage by team performance." (Smallfield et al., 2020)
Reported the r_{WG} Cutoff Value Used to Justify Aggregation	23%	"We used LeBreton & Senter's (2008) guidance to judge; that is, r_{WG} from 0.51 to 1.00 indicates moderate to very strong within-group agreement, and ICC(1) at 0.01 indicates a small effect size of group membership, 0.10 a medium one, and 0.25 a large one. In 2004, 87.1% of the groups had $r_{WG} \geq 0.51$, and among all the groups, ICC (1) = 0.13, $F(12, 182) = 3.31$, $p < .01$." (Li, Shemla & Wegge, 2021)
Discussed the Pattern of r_{WG} Values in Relation to the Cutoff	10%	"The workgroup conflict $r_{WG(J)}$ group average was .67 (median = .79) with 76% of groups at or above .60. (i.e., .60 is the suggested cutoff,James, 1982). These results support the appropriateness of aggregating individual conflict scores to obtain a group-level measure." (Booth et al., 2018)

the fact that less than one-third of these papers explain to the reader the specific multilevel model used is certainly problematic and limits readers' and reviewers' sense of how important within-unit agreement was for data aggregation.

Recommendation #3 & #4: Estimating ICC(1) & ICC(2)

Out of the 91 papers we examined, 89 (approximately 98%) reported at least one ICC value, with 80 papers reporting both *ICC(1)* and *ICC(2)*, 6 reporting only *ICC(1)*, and 3 reporting only *ICC(2)*. Finally, 86 papers reported both ICC(1) and $r_{WG}/r_{WG(J)}$. When examining these findings in the context of the content domain of the papers, we found that approximately 93% of teams papers reported both *ICC(1)* and *ICC(2)* compared to 80% of organizational culture and climate papers, 79% of leadership papers, and 77% of miscellaneous papers. With nearly all papers reporting at least one ICC value, and the majority of papers reporting both *ICC(1)* and *ICC(2)* values, scholars are already helping to ensure the clarity and transparency of research.

Recommendation #5: Estimating ICC(2) for Unbalanced Designs

As a reminder, 83 out of the 91 papers calculated and reported *ICC(2)*. Most papers reported having an unbalanced design (i.e., their group sizes differed), but only four explicitly addressed that the sizes of the groups in their sample differed when describing how they calculated *ICC(2)*. Each of these four papers described calculating *ICC(2)* using the mean size of their groups. This presents a serious limitation to the reproducibility of findings in papers using data aggregation statistics for multilevel analysis. Rarely are group sizes exactly the same, so it is important for scholars to note whether mean or median group sizes were used to calculate *ICC(2)*, whether *ICC(2)* values were computed for each group and then averaged, or if other decisions were made when calculating *ICC(2)* for groups of varying sizes.

Recommendation #6: Estimating $r_{WG}/r_{WG(J)}$

Approximately 25% of papers (23 out of 91) reported the null distribution(s) used to calculate r_{WG} or $r_{WG(J)}$. Specifically, 12 articles reported using just one null distribution, and 11 reported using more than one null distribution. The most common null distributions were the uniform null (18 out of 23 papers), followed by the slightly skewed null (9 papers), heavily skewed null (2 papers) and the triangular null (1 paper).

We also examined reporting patterns separately for the different construct domains. Overall, articles where the aggregated variables reflected leadership or team-related constructs were more likely to include information about null distributions, with 30% of leadership-related papers and 33% of team-related papers including information about the null. In contrast, only 20% of the climate/culture-related papers reported their null, and none of the miscellaneous papers reported information about the null distribution. The use of the uniform null versus the slightly skewed null, heavily skewed null, and triangular null distribution was not specific to the construct domain of the papers, with a relatively even distribution of null distribution types used across domains.

Overall, we found that 75% of the articles included in our review failed to explain how r_{WG} or $r_{WG(J)}$ were computed. This pattern of non-reporting was nearly identical to the results of a larger review conducted by Meyer et al. (2014). Those authors found that information about null distributions was omitted from 75.9% of the estimates included in their review. These authors also found that when null distributions were reported, 69.8% of the estimates were computed using a uniform or rectangular null distribution. Overall, the omission of this information is extremely problematic, as it results in ambiguous conclusions about how statistics were estimated and, thus, whether the correct statistics were ultimately included in the paper or not. Looking ahead, we see the clear articulation of the null distributions used to calculate r_{WG} as a key factor for scholars being able to reproduce and replicate multilevel research.

Recommendation #7: Criteria for Justifying Data Aggregation

Only 21 of the 91 papers (23%) identified a specific $r_{WG}/r_{WG(J)}$ cutoff value used to make decisions about the appropriateness of data aggregation. Most of these (16 out of 21) used a cutoff value of 0.70, citing LeBreton & Senter (2008) or James et al. (1984). One challenge we encountered when coding these papers was that, oftentimes, only one to two sentences were provided to discuss the calculation of r_{WG} statistics. In many cases, the discussion of the calculation of the r_{WG} statistic was included in the same sentence as the discussion of the cutoff value used with only a single citation provided at the end of the sentence. This was problematic in that we were often unsure whether authors were citing certain papers as a reference to r_{WG}, to the r_{WG} cutoff statistic, or to other information such as the calculation of ICC values. For example, we found it interesting that 4 out of 21 papers cited James (1982) when discussing the r_{WG} cutoff statistic used because this paper was published two years before the r_{WG} statistic was introduced by James et al. (1984). This issue highlights the need for greater clarity in reporting data aggregation in multilevel research.

Recommendation # 8: Treatment of Inconsistencies and Anomalies

Nine of the papers discussed the general pattern of responses with respect to a cutoff, for example by explaining the proportion of $r_{WG}/r_{WG(J)}$ values that fell above or below the cutoff value. Of the 21 papers that reported a cutoff value, 2 were explicit about (a) the number or percentage of groups that fell above or below the selected cutoff and (b) how groups that did not meet the cutoff value were treated (i.e., whether they were kept or excluded from analyses). In addition, six papers discussed the number or percentage of groups with $r_{WG}/r_{WG(J)}$ values below cutoffs, but those papers failed to explain how those groups were handled (i.e., were they retained or excluded from analyses).

With respect to out-of-range values, only nine papers explicitly noted whether there were estimates that fell outside the range of acceptable values with six reporting out-of-range values and three reporting no out-of-range values. Of the six papers with out-of-range values, only two described how these out-of-range values were treated. One paper opted to exclude groups with out-of-range values, and the other opted to retain those groups.

In sum, only nine of 91 articles mentioned the possibility of out-of-range values, six indicated problematic estimates of agreement were detected, and two discussed how they handled the groups with these estimates. One major issue with the exclusion of this information is that the reproducibility of research may be limited if researchers attempting to replicate findings make different decisions about the treatment of out-of-range values and/or the treatment of groups with such estimates. This is another key area in which the transparency of multilevel research could be enhanced through the inclusion of additional descriptive information about the calculation of agreement statistics.

Recommendations #9: Patterns of Agreement

Of the 91 papers that estimated r_{WG} or $r_{WG(J)}$, only 59 included information about the mean values, 21 included information about median values, and eight included information about both the mean and median values. In addition, only five papers included information about the range of values. None of the papers included a histogram depicting the distribution of r_{WG} values.

Recommendation #10: Global Estimates of Within-Unit Agreement Using r_{WGp}

None of the 91 papers in our literature review actually calculated $r_{WGp.}$ This is unfortunate because r_{WGp} provides a concise summary of within-unit agreement for data containing more than one higher-level unit — which, by definition, represents all data used in multilevel research. LeBreton et al. (2005) introduced r_{WGp} within the broader context of how to handle out-of-range r_{WG} values. They noted that one reason that a researcher could obtain out-of-range values was due to misspecification of multilevel data structure. They illustrated this issue using an example from the leader-member exchange literature. They noted that groups of employees nested within different leaders might contain multiple subgroups of employees — some employees might be members of the "in-group" and some members of the "out-group." In this instance, the target of assessment has multiple "true scores," conditional on group membership. To address this issue, LeBreton et al. recommended estimating agreement using the pooled within-groups estimate of variance rather than the observed variance. Basically, r_{WGp} residualized the observed variance for any between-group mean differences.

Although the in-group/out-group example was an effective illustration for some scholars, LeBreton et al. recognized that r_{WGp} had much broader applications. Specifically, r_{WGp} may be applied to *any situation where group membership can be defined* a priori. Beyond simple groups and teams, LeBreton et al. suggested a priori defined groups might include "functional departments, job classifications, race, gender, educational level, marital status, geographic divisions, and hierarchical level" (p. 137). Whereas r_{WG} provides a local estimate of agreement for each specific unit (e.g., team) using the local estimate of the within-group variance, r_{WGp} provides a global estimate of agreement across all higher-level units using a global estimate of within-group variance (i.e., pooled within-groups variance).

State of the Science: Transparency in Data Aggregation

In general, we found that articles included discussions of the theory of collective constructs (Recommendation #1) and included estimates of ICC(1) and ICC(2) (Recommendations # 3 & #4). However, several areas exist for improved reporting practices (e.g., Recommendations #2, #5, #6, #8, #9, #10). We now turn to several tools that we believe may help to further improve the clarity and transparency of data aggregation in multilevel research.

Tools for Enhanced Clarity, Transparency, & Reproducibility

We propose that the recommendations in Table 2 could be used as a Data Aggregation Checklist. We are optimistic that this checklist, when used in conjunction with other important resources (cf. Kozlowski & Klein, 2000; Meyer et al., 2014; Woehr et al., 2015), should lead to greater clarity and transparency in multilevel research. This checklist is offered

as a general template to help guide multilevel researchers (and those tasked with reviewing multilevel studies). Using these recommendations, we developed a data aggregation package (WGA: Within-Group Agreement & Aggregation) available in R. We hope this package also contributes to greater clarity and transparency in multilevel research. We have included a brief tutorial on using WGA as an online supplemental file.

Conclusion

The proliferation of the use of multilevel modeling in organizational research necessitates greater structure and consistency in reporting relevant statistics related to the aggregation of data across levels. Multilevel researchers frequently rely on estimates of $r_{WG,}$ *ICC(1),* and *ICC(2)* to justify decisions to aggregate data from a lower-level to a higher-level. The goal of this paper was to discuss general principles to guide data aggregation decisions and to provide recommendations for those principles that may be adopted to improve the clarity, transparency, and reproducibility of future multilevel research. The principles and resulting Data Aggregation Checklist, while not exhaustive, do provide a basic framework for improving the consistency in how multilevel scholars address data aggregation issues.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Aguinis, H., & Culpepper, S. A. (2015). An expanded decision-making procedure for examining cross-level interaction effects with multilevel modeling. *Organizational Research Methods, 18*, 155–176.

Atkins, D. C. (2005). Using multilevel models to analyze couple and family treatment data: Basic and advanced issues. *Journal of Family Psychology, 19*, 98–110.

Banks, G. C., Field, J. G., Oswald, F. L., O'Boyle, E. H., Landis, R. S., Rupp, D. E., & Rogelberg, S. G. (2019). Answers to 18 questions about open sciences practices. *Journal of Business and Psychology, 34*, 257–270.

Beal, D. J., & Weiss, H. M. (2003). Methods of ecological momentary assessment in organizational research. *Organizational Research Methods, 6*(4), 440–464.

Bernardin, H. J., Cooke, D. K., & Villanova, P. (2000). Conscientiousness and agreeableness as predictors of rating leniency. *Journal of Applied Psychology, 85*(2), 232–236.

Beus, J. M., Payne, S. C., Arthur, W., Jr., & Muñoz, G. J. (2019). The development and validation of a cross-industry safety climate measure: Resolving conceptual and operational issues. *Journal of Management, 45*(5), 1987–2013.

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

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel Theory, Research, and Methods in Organizations* (pp. 349–381). Jossey-Bass.

Bliese, P. D., & Halverson, R. R. (2002). Using random group resampling in multilevel research: An example of the buffering effects of leadership climate. *The Leadership Quarterly, 13*(1), 53–68.

Bliese, P. D., & Hanges, J. (2004). Being both too liberal and too conservative: The perils of treating group data as though they were independent. *Organizational Research Methods, 7*(4), 400–417.

Bliese, P. D., Maltarich, M. A., & Hendricks, J. L. (2018). Back to basics with mixed effects models: Nine take-away points. *Journal of Business and Psychology, 33*, 1–23.

Bliese, P. D., Maltarich, M. A., Hendricks, J. L., Hofmann, D. A., & Adler, A. B. (2019). Improving the measurement of group-level constructs by optimizing between-group differentiation. *Journal of Applied Psychology, 104*(2), 293–302.

Bliese, P. D., & Ployhart, R. E. (2002). Growth modeling using random coefficient models: Model building, testing, and illustration. *Organizational Research Methods, 5*(4), 362–387.

Booth, J. E., Park, T. Y., Zhu, L. L., Beauregard, T. A., Gu, F., & Emery, C. (2018). Prosocial response to client-instigated victimization: The roles of forgiveness and workgroup conflict. *Journal of Applied Psychology, 103*(5), 513.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing social networks*. UK: Sage Publications.

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the social sciences. *Science, 323*(5916), 892–895.

Brass, D. J., & Borgatti, S. P. (2019). Multilevel thoughts on social networks. In *The Handbook of Multilevel Theory, Measurement, and Analysis.* (pp. 187–200). American Psychological Association

Brown, R. D., & Hauenstein, N. M. A. (2005). Interrater agreement reconsidered: An alternative to the rWG indices. *Organizational Research Methods, 8*(2), 165–184.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Sage Publications, Inc.

Burke, M. I., Landis, R. S., & Burke, M. J. (2017). Estimating group-level relationships: General recommendations and considerations for the use of intraclass correlation coefficients. *Journal of Business and Psychology, 32*, 611–626.

Burke, M. J., & Dunlap, W. P. (2002). Estimating interrater agreement with the average deviation index: A user's guide. *Organizational Research Methods, 5*, 159–172.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2*(1), 49–68.

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*(2), 234–246.

Chen, G., Bliese, P. D., & Mathieu, J. E. (2005). Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology. *Organizational Research Methods, 8*(4), 375–409.

Cleveland, J. N., Murphy, K. R., & Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*(1), 130–135.

Cohen, A., Doveh, E., & Eick, U. (2001). Statistical properties of the $r_{WG(J)}$ index of agreement. *Psychological Methods, 6*(3), 297.

Colquitt, J. A., Noe, R. A., & Jackson, C. L. (2002). Justice in teams: Antecedents and consequences of procedural justice climate. *Personnel Psychology, 55*(1), 83–109.

Cronin, M. A., & Vancouver, J. B. (2019). The only constant is change: Expanding theory by incorporating dynamic properties into one's models. In S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook of Multilevel Theory, Measurement, and Analysis* (pp. 89–114). American Psychological Association.

Dansereau, F., Yammarino, F. J., & Kohles, J. C. (1999). Multiple levels of analysis from a longitudinal perspective: Implications for theory building. *Academy of Management Review, 24*(2), 346–357.

Deng, H., Leung, K., Lam, C. K., & Huang, X. (2019). Slacking off in comfort: A dual-pathway model for psychological safety climate. *Journal of Management, 45*(3), 1114–1144.

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for r_{WG} and average deviation interrater agreement indexes. *Journal of Applied Psychology, 88*(2), 256–362.

Edland, A., & Svenson, O. (1993). Judgment and decision making under time pressure. In *Time pressure and stress in human judgment and decision making* (pp. 27–40). Springer.

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

George, J. M., & Jones, G. R. (2000). The role of time in theory and theory building. *Journal of Management, 26*(4), 657–684.

Grand, J. A., Rogelberg, S. G., Allen, T. D., Landis, R. S., Reynolds, D. H., Scott, J. C., Tonidandel, S., & Truxillo, D. M. (2018). A Systems-based approach to fostering robust science in industrial and organizational psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 11*(1), 4–42.

Greenland, S. (2002). A review of multilevel theory for ecologic analysis. *Statistics in Medicine, 21*, 389–395.

Grund, S., Lüdtke, O., & Robitzsch, A. (2018). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods, 21*(1), 111–149.

Grund, S., Lüdtke, O., & Robitzsch, A. (2019). Missing data in multilevel research. In *The handbook of multilevel theory, measurement, and analysis*. (pp. 365–386). American Psychological Association

Gully, S. M., & Phillips, J. M. (2019). On finding your level. In S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook of Multilevel Theory, Measurement, and Analysis* (pp. 11–38). American Psychological Association.

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques: MLM and SEM Approaches Using Mplus*. Routledge.

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management, 23*, 723–744.

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

Hofmann, D. A., Griffin, M. A., & Gavin, M. B. (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: Foundations, extensions, and new directions* (pp. 467–511). Jossey-Bass.

House, R., Rousseau, D. M., & Thomas-Hunt, M. (1995). The meso paradigm: A framework for the integration of micro and macro organizational behavior. *Research in Organizational Behavior, 17*, 71–114.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Routledge.

Humphrey, S. E., & LeBreton, J. M. (2019). *The Handbook of Multilevel Theory, Measurement, and Analysis*. American Psychological Association.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67*(2), 219–229.

James, L. R., & Williams, L. J. (2000). The cross-level operator in regression, ANCOVA, and contextual analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 382–424). Jossey-Bass.

James, L. R., Demaree, R. G., & Wolf, G. (1984). Estimating within-group interrater reliability with and without response bias. *Journal of Applied Psychology, 69*(1), 85–98.

James, L. R., Demaree, R. G., & Wolf, G. (1993). r_{WG}: An assessment of within-group interrater agreement. *Journal of Applied Psychology, 78*(2), 306–309.

Jebb, A. T., Tay, L., Ng, V., Woo, S. E. (2019). Construct validation in multilevel studies. S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook of Multilevel Theory, Measurement, and Analysis* (pp. 253–278). Washington, D.C.: American Psychological Association

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic Data Analysis*. Guilford.

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.

Klein, K. J., & Kozlowski, S. W. (2000). *Multilevel theory, research, and methods in organizations*. Jossey-Bass.

Klein, K. J., Tosi, H., & Cannella, A. A., Jr. (1999). Multilevel theory building: Benefits, barriers, and new developments. *Academy of Management Review, 24*(2), 243–248.

Knight, A. P., & Humphrey, S. E. (2019). Dyadic data analysis. In S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook of Multilevel Theory, Measurement, and Analysis* (pp. 423–448). American Psychological Association.

Kozlowski, S. W. J., & Hattrup, K. (1992). A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology, 77*(2), 161–167.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multi-level Theory, Research and Methods in Organizations* (pp. 3–90). Jossey-Bass.

Kozlowski, S. W. J., Chao, G. T., Grand, J. A., Braun, M. T., & Kuljanin, G. (2013). Advancing multilevel research design: Capturing the dynamics of emergence. *Organizational Research Methods, 16*(4), 581–615.

Krasikova, D., & LeBreton, J. M. (2012). Just the two of us: Misalignment of theory and methods in examining dyadic phenomena. *Journal of Applied Psychology, 97*(4), 739–757.

Krasikova, D. V., & LeBreton, J. M. (2019). Multilevel measurement: Agreement, reliability, and non-independence. S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook of Multilevel Theory, Measurement, and Analysis* (pp. 279–304). Washington, D.C.: American Psychological Association

Kreft, I., & de Leeuw, J. (1998). *Introducing Multilevel Modeling*. Sage Publications.

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods, 17*, 433–451.

LaHuis, D. M., Blackmore, C. E., & Bryant-Lees, K. B. (2019). Explained variance measures for multilevel models. In S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook of Multilevel Theory, Measurement, and Analysis* (pp. 353–364). American Psychological Association.

LeBreton, J. M., & Senter, J. L. (2008). Answers to twenty questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*(4), 815–852.

LeBreton, J. M., James, L. R., & Lindell, M. K. (2005). Recent issues regarding r_{WG}, r^*_{WG}, $r_{WG(J),}$ $r^*_{WG(J)}$. *Organizational Research Methods, 8*(1), 128–138.

LeBreton, J. M., Burgess, J. R., Kaiser, R. B., Atchley, E. K., & James, L. R. (2003). The restriction of variance hypothesis and interrater reliability and agreement: Are ratings from multiple sources really dissimilar? *Organizational Research Methods, 6*(1), 80–128.

Li, Y., Koopmann, J., Lanaj, K., & Hollenbeck, J. R. (2022). An integration-and-learning perspective on gender diversity in self-managing teams: The roles of learning goal orientation and shared leadership. *Journal of Applied Psychology, 107*(9), 1628–1639.

Li, J., Shemla, M., & Wegge, J. (2021). The preventative benefit of group diversification on group performance decline: An investigation with latent growth models. *Journal of Organizational Behavior, 42*(3), 332–348.

Lindell, M. K., & Brandt, C. J. (1997). Measuring interrater agreement for ratings of a single target. *Applied Psychological Measurement, 21*(3), 271–278.

Lindell, M. K., Brandt, C. J., & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement, 23*(2), 127–135.

London, M., Smither, J. W., & Adsit, D. J. (1997). Accountability: The Achilles' heel of multisource feedback. *Group & Organization Management, 22*(2), 162–184.

LoPilato, A. C., & Vandenberg, R. J. (2014). The not-so-direct cross-level direct effect. In *More Statistical and Methodological Myths and Urban Legends,* (pp. 302–320). Routledge

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level effects in multilevel modeling. *Journal of Applied Psychology, 97*(5), 951–966.

Mathieu, J. E., & Luciano, M. M. (2019). Multilevel emergence in work collectives. In S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook of Multilevel Theory, Measurement, and Analysis* (pp. 163–186). American Psychological Association.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*(3), 259.

Meyer, R. D., Mumford, T. V., Burrus, C. J., Campion, M. A., & James, L. R. (2014). Selecting null distributions when calculating r_{WG}: A tutorial and review. *Organizational Research Methods, 17*(3), 324–345.

Mitchell, T. R., & James, L. R. (2001). Building better theory: Time and the specification of when things happen. *Academy of Management Review, 26*(4), 530–547.

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

Mossholder, K. W., & Bedeian, A. G. (1983). Cross-level inference in organizational research: Perspectives on interpretation and application. *Academy of Management Review, 8*(4), 547–558.

Murphy, K. R., Jako, R. A., & Anhalt, R. L. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78*(2), 218–225.

Newman, D. A., & Sin, H. P. (2020). Within-group agreement (r_{WG}): Two theoretical parameters and their estimators. *Organizational Research Methods, 23*(1), 30–64.

Nielsen, K., Tafvelin, S., von Thiele Schwarz, U., & Hasson, H. (2021). In the eye of the beholder: How self-other agreements influence leadership training outcomes as perceived by leaders and their followers. *Journal of Business and Psychology*, 1–18.

Open Science Collaboration. (2015). Psychology. Estimating the Reproducibility of Psychological Science. *Science, 349*(6251), aac4716.

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

Owens, B. P., Yam, K. C., Bednar, J. S., Mao, J., & Hart, D. W. (2019). The impact of leader moral humility on follower moral self-efficacy and behavior. *Journal of Applied Psychology, 104*(1), 146.

Pennington, G. L., & Roese, N. J. (2003). Regulatory focus and temporal distance. *Journal of Experimental Social Psychology, 39*(6), 563–576.

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

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

Roberson, Q. M., Sturman, M. C., & Simons, T. L. (2007). Does the measure of dispersion matter in multilevel research? A comparison of the relative performance of dispersion indexes. *Organizational Research Methods, 10*, 564–588.

Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. In L. L. Cummings & B. M. Staw (Eds.), *Research in Organizational Behavior* (Vol. 7, pp. 1–37)

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

Scherbaum, C. A., & Pesner, E. (2019). Power analysis for multilevel research. In *The handbook of multilevel theory, measurement, and analysis*. (pp. 329–352). American Psychological Association

Schmidt, F. L., & Hunter, J. E. (1989). Interrater reliability coefficients cannot be computed when only one stimulus is rated. *Journal of Applied Psychology, 74*(2), 368.

Schneider, B., Salvaggio, A. N., & Subirats, M. (2002). Climate strength: A new direction for climate research. *Journal of Applied Psychology, 87*(2), 220.

Schneider, B., White, S. S., & Paul, M. C. (1998). Linking service climate and customer perceptions of service quality: Tests of a causal model. *Journal of Applied Psychology, 83*(2), 150.

Shiffman, S. (2014). Conceptualizing analyses of ecological momentary assessment data. *Nicotine & Tobacco Research, 16*, S76–S87.

Shiverdecker, L. K., & LeBreton, J. M. (2019). A primer on multilevel (random coefficient) regression modeling. In S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook of Multilevel Theory, Measurement, and Analysis* (pp. 389–422). American Psychological Association.

Smallfield, J., Hoobler, J. M., & Kluemper, D. H. (2020). How team helping influences abusive and empowering leadership: The roles of team affective tone and performance. *Journal of Organizational Behavior, 41*(8), 757–781.

Smith-Crowe, K., Burke, M. J., Cohen, A., & Doveh, E. (2014). Statistical significance criteria for the r_{WG} and average deviation interrater agreement indices. *Journal of Applied Psychology, 99*(2), 239–261.

Smith-Crowe, K., Burke, M. J., Kouchaki, M., & Signal, S. M. (2012). Assessing interrater agreement via the average deviation index given a variety of theoretical and methodological problems. *Organizational Research Methods, 16*(1), 127–151.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Tay, L., Woo, S. E., & Vermunt, J. K. (2014). A conceptual and methodological framework for psychometric isomorphism: Validation of multilevel construct measures. *Organizational Research Methods, 17*(1), 77–106.

Vandenberg, R. J., & Richardson, H. A. (2019). A primer on multilevel structural modeling: User-friendly guidelines. In *The handbook of multilevel theory, measurement, and analysis.* (pp. 449–472). American Psychological Association

Vasquez, C. A., Madrid, H. P., & Niven, K. (2021). Leader interpersonal emotion regulation motives, group leader–member exchange, and leader effectiveness in work groups. *Journal of Organizational Behavior, 42*(9), 1168–1185.

Wittmer, J., & LeBreton, J. (2021). Interrater agreement and interrater reliability: Implications for multilevel research. *Oxford Research Encyclopedia of Business and Management.* Retrieved 28 Dec. 2021, from https://oxfordre.com/business/view/10.1093/acrefore/9780190224851.001.0001/acrefore-9780190224851-e-222

Woehr, D. J., Loignon, A. C., Schmidt, P. B., Loughry, M. L., & Ohland, M. W. (2015). Justifying aggregation with consensus-based constructs: A review and examination of cutoff values for common aggregation indices. *Organizational Research Methods, 18*(4), 704–737.

Yammarino, F. J., & Gooty, J. (2019). Cross-level models. In S. E. Humphrey & J. M. LeBreton (Eds.), *The handbook of multilevel theory, measurement, and analysis* (pp. 563–585). American Psychological Association.

Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods, 12*(4), 695–719.

Zhou, L., Song, Y., Alterman, V., Liu, Y., & Wang, M. (2019). Introduction to data collection in multilevel research. S. E. Humphrey & J. M. LeBreton (Eds.), *Handbook of multilevel theory, measurement, and analysis* (pp. 225–252). Washington, D.C.: American Psychological Association.

Multilevel Modeling in R (2.7)

A Brief Introduction to R, the multilevel package and the nlme package

Paul Bliese (pdbliese@gmail.com)

March 10, 2022

Table of Contents

1 Introduction

This is an introduction to how R can be used to perform multilevel analyses typical to organizational researchers. Multilevel analyses are applied to data that have some form of a nested structure. For instance, individuals may be nested within workgroups, or repeated measures may be nested within individuals, or firms may provide several years of data in what is referred to as panel data. Nested structures are often accompanied by some form of non-independence. In work settings, individuals in the same workgroup typically display some similarity with respect to performance or they provide similar responses to questions about aspects of the work environment. Likewise, in repeated measures data, individuals or firms usually display a high degree of similarity in responses over time. Non-independence may be considered either a nuisance variable or something to be substantively understood but working with nested data requires tools to deal with non-independence.

The term "multilevel analysis" is used to describe a set of analyses also referred to as random coefficient models, random effects, and mixed-effects models (see Bryk & Raudenbush, 1992; Clark & Linzer, 2014; Kreft & De leeuw, 1998; Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). Mixed-effects models (the term primarily used in this document) are not without limitations (e.g., Clark & Linzer, 2014), but are generally well-suited for dealing with non-independence (Bliese, Schepker, Essman & Ployhart, 2020). Prior to the widespread use of mixed-effects models, analysts used a variety of techniques to analyze data with nested structures and many of these techniques such as the econometric fixed-effect model are still widely used. In organizational research, mixed-effects models are often augmented by tools designed to quantify within-group agreement and group-mean reliability and the `multilevel` package contains many functions designed around testing within-group agreement and reliability.

This document is designed to cover a broad range of tools and approaches for analyzing multilevel data. Having worked for over two decades with both R and with multilevel data from numerous contexts, I routinely leverage different approaches and different packages depending upon the specific circumstances. Therefore, my goal in writing this document is to show how R can cover a wide range of inter-related topics related to multilevel analyses including:

- Data aggregation and merging for multilevel analyses
- Within-group agreement and reliability
- Contextual and basic econometric fixed-effect OLS models
- Covariance theorem decomposition of correlations
- Random Group Resampling
- Mixed Effects Models for nested group data
- Variants of Mixed Effects Models for Repeated Measures Data

Some of the basic analyses can conducted using R's base packages, but many of the analyses use functions in the `multilevel` package. As a broad overview, the `multilevel` package provides (a) functions for estimating within-group agreement and reliability indices, (b) functions for manipulating multilevel and longitudinal (panel) data, (c) simulations for

estimating power and generating multilevel data, and (d) miscellaneous functions for estimating reliability and performing simple calculations and data transformations. The `multilevel` package also contains several datasets to illustrate concepts.

The other library that is frequently used is the non-linear and linear mixed-effects (`nlme`) model package, (Pinheiro & Bates, 2000). The `nlme` package provides functions to estimate a variety of models for both data nested in groups and for repeated measures data collected over time (growth models). Functions in the `nlme` package have remarkable flexibility and can estimate a variety of alternative statistical models. In some cases, the `lme4` package developed by Doug Bates after the `nlme` package provides additional flexibility, so some functions from the `lme4` package are also detailed. I tend to use `lme4` when dealing with dichotomous dependent variables, or when data are partially or fully crossed, or when I want to generate an interaction plot (many more recent plotting packages were designed to work with `lme4` rather than `nlme`).

2 Reading data from files

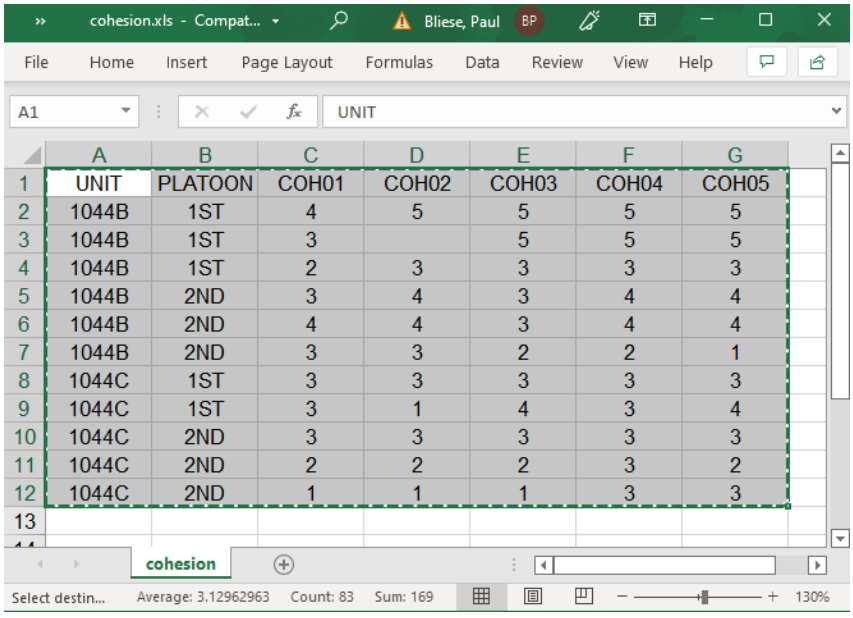
Before detailing multilevel analyses, I provide a short section on reading in data. There are numerous options for reading in data, so this section is in no way exhaustive. I provide what has been a simple and reliable way to import external files into dataframes.

In almost all cases working with research partners either in industry or academia, I have found that EXCEL files are a common platform particularly since EXCEL can read comma-delimited (csv) files. One additional advantage to EXCEL is that it is easy to quickly scan the data file for potential problems. I tend to avoid bringing in columns containing large amounts of text, and I often add an additional row under the header row with new R-friendly names (some research partners provide column headers the length of a small novel).

2.1.1 Reading data directly from EXCEL (Windows and MAC)

2.1.1.1 Windows

Consider the following data and notice how it has been highlighted and copied into the Window's clipboard (Ctrl-C):



	A	B	C	D	E	F	G
1	UNIT	PLATOON	COH01	COH02	COH03	COH04	COH05
2	1044B	1ST	4	5	5	5	5
3	1044B	1ST	3		5	5	5
4	1044B	1ST	2	3	3	3	3
5	1044B	2ND	3	4	3	4	4
6	1044B	2ND	4	4	3	4	4
7	1044B	2ND	3	3	2	2	1
8	1044C	1ST	3	3	3	3	3
9	1044C	1ST	3	1	4	3	4
10	1044C	2ND	3	3	3	3	3
11	1044C	2ND	2	2	2	3	2
12	1044C	2ND	1	1	1	3	3
13							

Once the file is in the Windows "clipboard", the following command reads the data into R:

```
> cohesion<-read.table(file="clipboard",sep="\t",header=T)
```

An even simpler variation is to use:

```
> cohesion<-read.delim(file="clipboard")
```

The `read.delim` function is variant of `read.table` that assumes the data are tab-delimited with a header. I have found that this simple approach covers about 95% of all my data entry needs to include importing either csv or EXCEL files with tens of thousands of observations.

2.1.1.2 MAC

If using a MAC, the basic ideas are the same, but the clipboard is accessed differently using pipe.

```
> cohesion<-read.delim(pipe("pbpaste"))
```

2.1.2 Reading external csv files with `file.choose` (Windows and MAC)

In cases where datasets are too large to read into EXCEL using the `file.choose()` function with `read.csv` or other `read.table` functions helps having to specify the path as in:

```
>cohesion<-read.csv(file.choose())
```

Using `file.choose()` opens the graphic user interface (gui) so one can select the file using a mouse or other device. This option is particularly useful when data are stored in complex network file structures.

2.1.3 Writing R files to EXCEL (Windows and MAC)

2.1.3.1 Windows

Because the "`clipboard`" option also works with `write.table` it is also a useful way to export the results of data analyses to EXCEL or other programs. For instance, if we create a correlation matrix from the cohesion data set, we can export this correlation table directly to EXCEL.

```
> CORMAT<-cor(cohesion[,3:7],use="pairwise.complete.obs")
> CORMAT
          COH01     COH02     COH03     COH04     COH05
COH01 1.0000000 0.7329843 0.6730782 0.4788431 0.4485426
COH02 0.7329843 1.0000000 0.5414305 0.6608190 0.3955316
COH03 0.6730782 0.5414305 1.0000000 0.7491526 0.7901837
COH04 0.4788431 0.6608190 0.7491526 1.0000000 0.9036961
COH05 0.4485426 0.3955316 0.7901837 0.9036961 1.0000000

> write.table(CORMAT,file="clipboard",sep="\t",col.names=NA)
```

Going to EXCEL and issuing the Windows "paste" command (or Ctrl-V) will insert the matrix into the EXCEL worksheet. Note the somewhat counter-intuitive use of `col.names=NA` in this example. This command does *not* mean omit the column names (achieved using `col.names=F`); instead the command puts an extra blank in the first row of the column names to line up the column names with the correct columns. Alternatively, one can use the option `row.names=F` to omit the row numbers.

Written objects may be too large for the default memory limit of the Window's clipboard. For instance, writing the full `bh1996` dataset from the `multilevel` package into the clipboard results in the following error (truncated):

```
> library(multilevel)
> data(b1996)  #Bring data from the library to the workspace
> write.table(bh1996,file="clipboard",sep="\t",col.names=NA)
Warning message:
In write.table(x, file, nrow(x),... as.integer(quote),  :
  clipboard buffer is full and output lost
```

To increase the size of the clipboard to 1.5MG (or any other arbitrary size), the `"clipboard"` option can be modified as follows: `"clipboard-1500"`. The options surrounding the use of the `clipboard` are specific to various operating systems and may change with different versions of R so it will be worth periodically referring to the help files.

2.1.3.2 MAC

If using a MAC, the "clipboard" option does not work, so the command line would be:

```
> write.table(bh1996, file=pipe("pbcopy"),sep="\t",col.names=NA)
```

Unlike Windows, the pipe option does not appear to need to be resized to accommodate large files.

2.1.4 The `foreign` package and SPSS files

The `foreign` package contains functions to import SPSS, SAS, Stata and minitab files. Help files are available for different formats. Below is a command to bring in an SPSS file as a dataframe and numbers (e.g., 4) instead of the number's value label (e.g., "agree").

```
> library(foreign)
> help(read.spss)     #look at the documentation on read.spss
> cohesion<-read.spss(file.choose(),use.value.labels=F, to.data.frame=T)
> cohesion
    UNIT PLATOON COH01 COH02 COH03 COH04 COH05
1  1044B     1ST     4     5     5     5     5
2  1044B     1ST     3    NA     5     5     5
3  1044B     1ST     2     3     3     3     3
4  1044B     2ND     3     4     3     4     4
5  1044B     2ND     4     4     3     4     4
6  1044B     2ND     3     3     2     2     1
7  1044C     1ST     3     3     3     3     3
8  1044C     1ST     3     1     4     3     4
9  1044C     2ND     3     3     3     3     3
10 1044C     2ND     2     2     2     3     2
11 1044C     2ND     1     1     1     3     3
```

2.1.5 Checking your dataframes with `str` and `summary`

With small data sets it is easy to verify that the data has been read in correctly. Often, however, one will be working with large data sets that are difficult to visual verify. Consequently, functions such as `str` (structure) and `summary` provide easy ways to examine dataframes.

```
> str(cohesion)
`data.frame':   11 obs. of  7 variables:
 $ UNIT   : Factor w/ 2 levels "1044B","1044C": 1 1 1 1 1 1 2 2 2 2 ...
 $ PLATOON: Factor w/ 2 levels "1ST","2ND": 1 1 1 2 2 2 1 1 2 2 ...
 $ COH01  : int  4 3 2 3 4 3 3 3 3 2 ...
 $ COH02  : int  5 NA 3 4 4 3 3 1 3 2 ...
 $ COH03  : int  5 5 3 3 3 2 3 4 3 2 ...
 $ COH04  : int  5 5 3 4 4 2 3 3 3 3 ...
 $ COH05  : int  5 5 3 4 4 1 3 4 3 2 ...

> summary(cohesion)
    UNIT   PLATOON     COH01           COH02          COH03
 1044B:6   1ST:5   Min.   :1.000   Min.   :1.00   Min.   :1.000
 1044C:5   2ND:6   1st Qu.:2.500   1st Qu.:2.25   1st Qu.:2.500
                   Median :3.000   Median :3.00   Median :3.000
                   Mean   :2.818   Mean   :2.90   Mean   :3.091
                   3rd Qu.:3.000   3rd Qu.:3.75   3rd Qu.:3.500
                   Max.   :4.000   Max.   :5.00   Max.   :5.000
```

```
                                    NA's   :1.00
     COH04           COH05
 Min.   :2.000   Min.   :1.000
 1st Qu.:3.000   1st Qu.:3.000
 Median :3.000   Median :3.000
 Mean   :3.455   Mean   :3.364
 3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :5.000   Max.   :5.000
```

2.1.6 Loading data from packages

One of the useful attributes of R is that the data used in the examples are almost always available to the user. These data are associated with specific packages. For instance, the multilevel package uses a variety of data files to illustrate specific functions. To gain access to these data, one uses the `data` command:

```
>data(package="multilevel")
```

This command lists the data sets associated with the multilevel package, and the command

```
>data(bh1996, package="multilevel")
```

copies the `bh1996` data set to the workspace making it possible to work with the `bh1996` dataframe. If a package has been attached by the `library` function its datasets are automatically included in the search, so that if

```
>library(multilevel)
```

has been run, then

```
>data(bh1996)
```

copies the data from the package to the workspace without specifying the package.

2.2 A Brief Review of Matrix Brackets

One of the unique aspects of R is the use of matrix brackets to access various parts of a dataframe. While the bracket notation may initially appear cumbersome, mastering the use of matrix brackets provides considerable control.

The overall notation is [rows, columns]. So accessing rows 1,5,and 8 and columns 3 and 4 of the `cohesion` dataframe would be done like so:

```
> cohesion[c(1,5,8),3:4]
  COH01 COH02
1     4     5
5     4     4
8     3     1
```

Alternatively, we can specify the column names (this helps avoid picking the wrong columns).

```
> cohesion[c(1,5,8),c("COH01","COH02")]
  COH01 COH02
1     4     5
5     4     4
8     3     1
```

It is often useful to pick specific rows that meet some criteria. So, for example, we might want to pick rows that are from the 1ST PLATOON

```
> cohesion[cohesion$PLATOON=="1ST",]
   UNIT PLATOON COH01 COH02 COH03 COH04 COH05
1 1044B     1ST     4     5     5     5     5
2 1044B     1ST     3    NA     5     5     5
3 1044B     1ST     2     3     3     3     3
7 1044C     1ST     3     3     3     3     3
8 1044C     1ST     3     1     4     3     4
```

Upon inspection, we might want to further refine our choice and exclude missing values. We do this by adding another condition using `AND` operator `"&"` along with the `NOT` operator `"!".`

```
> cohesion[cohesion$PLATOON=="1ST"&!is.na(cohesion$COH02),]
   UNIT PLATOON COH01 COH02 COH03 COH04 COH05
1 1044B     1ST     4     5     5     5     5
3 1044B     1ST     2     3     3     3     3
7 1044C     1ST     3     3     3     3     3
8 1044C     1ST     3     1     4     3     4
```

These simple examples should provide an idea of how to subset large datasets when conducting analyses.

3 Multilevel Analyses

The remainder of this document illustrates how R can be used in multilevel modeling beginning with several R functions particularly useful for preparing data for subsequent analyses

3.1 Multilevel data manipulation functions

3.1.1 The `merge` Function

One of the key data manipulation tasks that must be accomplished prior to estimating several of the multilevel models (specifically contextual models and mixed-effects models) is that group-level variables must be "assigned down" to the individual. To make a dataframe containing both individual and group-level variables, one typically begins with two separate dataframes. One dataframe contains individual-level data, and the other dataframe contains group-level data. Combining these two dataframes using a group identifying variable common to both produces a single dataframe containing both individual and group data. In R, combining dataframes is accomplished using the `merge` function.

For instance, consider the `cohesion` data introduced when showing how to read data from external files. The `cohesion` data is included as a multilevel data set, so we can use the `data` function to bring it from the multilevel package to the working environment

```
>data(cohesion)
>cohesion
    UNIT PLATOON COH01 COH02 COH03 COH04 COH05
1  1044B     1ST     4     5     5     5     5
2  1044B     1ST     3    NA     5     5     5
3  1044B     1ST     2     3     3     3     3
4  1044B     2ND     3     4     3     4     4
```

```
5  1044B     2ND     4     4     3     4     4
6  1044B     2ND     3     3     2     2     1
7  1044C     1ST     3     3     3     3     3
8  1044C     1ST     3     1     4     3     4
9  1044C     2ND     3     3     3     3     3
10 1044C     2ND     2     2     2     3     2
11 1044C     2ND     1     1     1     3     3
```

Now assume that we have another dataframe with platoon sizes. We can create this dataframe as follows:

```
> group.size<-data.frame(UNIT=c("1044B","1044B","1044C","1044C"),
PLATOON=c("1ST","2ND","1ST","2ND"),PSIZE=c(3,3,2,3))
> group.size  #look at the group.size dataframe
   UNIT PLATOON PSIZE
1 1044B     1ST     3
2 1044B     2ND     3
3 1044C     1ST     2
4 1044C     2ND     3
```

To create a single file (`new.cohesion`) that contains both individual and platoon information, use the `merge` command.

```
> new.cohesion<-merge(cohesion,group.size,by=c("UNIT","PLATOON"))
> new.cohesion
     UNIT PLATOON COH01 COH02 COH03 COH04 COH05 PSIZE
1  1044B     1ST     4     5     5     5     5     3
2  1044B     1ST     3    NA     5     5     5     3
3  1044B     1ST     2     3     3     3     3     3
4  1044B     2ND     3     4     3     4     4     3
5  1044B     2ND     4     4     3     4     4     3
6  1044B     2ND     3     3     2     2     1     3
7  1044C     1ST     3     3     3     3     3     2
8  1044C     1ST     3     1     4     3     4     2
9  1044C     2ND     3     3     3     3     3     3
10 1044C     2ND     2     2     2     3     2     3
11 1044C     2ND     1     1     1     3     3     3
```

Notice that every individual now has a value for `PSIZE` – a value that reflects the number of individuals in the platoon.

In situations where there is a single unique group identifier, the `by` option can be simplified to include just one variable. For instance, if the group-level data had reflected values for each UNIT instead of PLATOON nested in unit, the by option would simply read `by="UNIT"`. In the case of PLATOON, however, there are numerous platoons with the same name (1ST, 2ND), so unique platoons need to be identified within the nesting of the larger UNIT.

3.1.2 The `aggregate` function

In many cases in multilevel analyses, we create a group-level variable by mean aggregating individual responses. Consequently, the `aggregate` function is often used in combination with the `merge` function. In our cohesion example, we can assign platoon means for the variables COH01 and COH02 back to the individuals using `aggregate` and `merge`.

The first step is to create a dataframe with group means using the `aggregate` function. The `aggregate` function has three key arguments: the first is matrix of variables to convert to group-level variables. Second is the grouping variable(s) as a list, and third is the function (`mean`, `var`, `length`, etc.) executed on the variables. To calculate the means of COH01 and COH02 (columns 3 and 4 of the cohesion dataframe) issue the command:

```
> TEMP<-aggregate(cohesion[,3:4],list(cohesion$UNIT,cohesion$PLATOON),mean)
> TEMP
  Group.1 Group.2    COH01    COH02
1   1044B     1ST 3.000000       NA
2   1044C     1ST 3.000000 2.000000
3   1044B     2ND 3.333333 3.666667
4   1044C     2ND 2.000000 2.000000
```

Notice that COH02 has an "NA" value for the mean. The NA value occurs because there was a missing value in the individual-level file. If we decide to base the group mean on the non-missing individual values from group members we can add the parameter `na.rm=T`, to designate that NA values should be removed prior to calculating the group mean.

```
> TEMP<-aggregate(cohesion[,3:4],list(cohesion$UNIT,cohesion$PLATOON),
  mean,na.rm=T)
> TEMP
  Group.1 Group.2    COH01    COH02
1   1044B     1ST 3.000000 4.000000
2   1044C     1ST 3.000000 2.000000
3   1044B     2ND 3.333333 3.666667
4   1044C     2ND 2.000000 2.000000
```

To merge the `TEMP` dataframe with the `new.cohesion` dataframe, we need to align the merge columns from both dataframes and control how the merge handles variables with the same names using the `suffixes= c("",".G")` option which leaves the variable name unchanged in the first dataframe but adds a .G suffix on the second dataframe.

```
> final.cohesion<-merge(new.cohesion,TEMP,by.x=c("UNIT","PLATOON"),
+ by.y=c("Group.1","Group.2"),suffixes=c("",".G"))
> final.cohesion
    UNIT PLATOON COH01 COH02 COH03 COH04 COH05 PSIZE  COH01.G  COH02.G
1  1044B     1ST     4     5     5     5     5     3 3.000000 4.000000
2  1044B     1ST     3    NA     5     5     5     3 3.000000 4.000000
3  1044B     1ST     2     3     3     3     3     3 3.000000 4.000000
4  1044B     2ND     3     4     3     4     4     3 3.333333 3.666667
5  1044B     2ND     4     4     3     4     4     3 3.333333 3.666667
6  1044B     2ND     3     3     2     2     1     3 3.333333 3.666667
7  1044C     1ST     3     3     3     3     3     2 3.000000 2.000000
8  1044C     1ST     3     1     4     3     4     2 3.000000 2.000000
9  1044C     2ND     3     3     3     3     3     3 2.000000 2.000000
10 1044C     2ND     2     2     2     3     2     3 2.000000 2.000000
11 1044C     2ND     1     1     1     3     3     3 2.000000 2.000000
```

The `aggregate` and `merge` functions provide tools necessary to manipulate data and prepare it for subsequent multilevel analyses. Again, note that this illustration uses a relatively complex situation where there are two levels of nesting (Platoon within Unit). In cases where

there is only one grouping variable (for example, UNIT) the commands for aggregate and merge contain the name of a single grouping variable. For instance,

```
>TEMP<-aggregate(cohesion[,3:4],list(cohesion$UNIT),mean,na.rm=T)
```

3.2 Within-Group Agreement and Reliability

The data used in this section are taken from Bliese, Halverson & Rothberg (2000) using the bhr2000 data set from the multilevel package.

```
> data(bhr2000)#imports the data into the working environment
> names(bhr2000)
 [1] "GRP"   "AF06"  "AF07"  "AP12"  "AP17"  "AP33"  "AP34"
 "AS14"  "AS15" "AS16"  "AS17"  "AS28"  "HRS"   "RELIG"
> nrow(bhr2000)
[1] 5400
```

The names function identifies 14 variables. The first one, GRP, is the group identifier. The variables in columns 2 through 12 are individual responses on 11 items that make up a leadership scale. HRS represents individuals' reports of work hours, and RELIG represents individuals' reports of the degree to which religion is a useful coping mechanism. The nrow command indicates that there are 5400 observations. To find out how many groups there are we can use the length command in conjunction with the unique command

```
> length(unique(bhr2000$GRP))
[1] 99
```

There are several functions in the multilevel library that are useful for calculating and interpreting agreement indices. These functions are rwg, rwg.j, rwg.sim, rwg.j.sim, rwg.j.lindell, awg, ad.m, ad.m.sim and rgr.agree. The rwg function calculates the James, Demaree & Wolf (1984) r_{wg} for single item measures; the rwg.j function calculates the James et al. (1984) $r_{wg(j)}$ for multi-item scales. The rwg.j.lindell function calculates $r^*_{wg(j)}$ (Lindell, & Brandt, 1997; 1999). The awg function calculates the a_{wg} agreement index proposed by Brown and Hauenstein (2005). The ad.m function calculates average deviation (AD) values for the mean or median (Burke, Finkelstein & Dusig, 1999).

A series of functions with "sim" in the name (rwg.sim, rwg.j.sim and ad.m.sim) can be used to simulate agreement values from a random null distributions to test for statistical significance agreement. The simulation functions are based on work by Dunlap, Burke and Smith-Crowe (2003); Cohen, Doveh and Eich (2001) and Cohen, Doveh and Nuham-Shani (2009). Finally, the rgr.agree function performs a Random Group Resampling (RGR) agreement test (see Bliese, et al., 2000).

In addition to the agreement measures, there are two multilevel reliability measures, ICC1 and ICC2 than can be used on ANOVA models. As Bliese (2000) and others (e.g., Kozlowski & Hattrup, 1992; Tinsley & Weiss, 1975) have noted, reliability measures such as the ICC(1) and ICC(2) are fundamentally different from agreement measures; nonetheless, they often provide complementary information to agreement measures, so this section illustrates the use of each of these functions using the dataframe bhr2000.

3.2.1 Agreement: r_{wg}, $r_{wg(j)}$, and $r^*_{wg(j)}$

Both the `rwg` and `rwg.j` functions are based upon the formulations described in James et al. (1984). he functions require three pieces of information. The first piece is the variable of interest (`x`), the second is the grouping variable (`grpid`), and third is the expected random variance (`ranvar`). The default estimate of `ranvar` is 2, which is the expected random variance based upon the rectangular distribution for a 5-point item (i.e., σ_{EU}^2) calculated using the formula ranvar=(A^2-1)/12 where A represents the number of response options associated with the scale anchors. See `help(rwg)`, James et al., (1984), or Bliese et al., (2000) for details on selecting appropriate `ranvar` values.

Below is an example using the `rwg` function to calculate agreement for the "coping using religion" item:

```
> RWG.RELIG<-rwg(bhr2000$RELIG,bhr2000$GRP,ranvar=2)
> RWG.RELIG[1:10,]  #examine first 10 rows of data
   grpid        rwg gsize
1      1 0.11046172    59
2      2 0.26363636    45
3      3 0.21818983    83
4      4 0.31923077    26
5      5 0.22064137    82
6      6 0.41875000    16
7      7 0.05882353    18
8      8 0.38333333    21
9      9 0.14838710    31
10    10 0.13865546    35
```

The function returns a dataframe with three columns. The first column contains the group names (`grpid`), the second column contains the 99 r_{wg} values – one for each group. The third column contains the group size. To calculate the mean r_{wg} value use the `summary` command:

```
> summary(RWG.RELIG)
     grpid             rwg              gsize
 1      : 1       Min.   :0.0000   Min.   :  8.00
 10     : 1       1st Qu.:0.1046   1st Qu.: 29.50
 11     : 1       Median :0.1899   Median : 45.00
 12     : 1       Mean   :0.1864   Mean   : 54.55
 13     : 1       3rd Qu.:0.2630   3rd Qu.: 72.50
 14     : 1       Max.   :0.4328   Max.   :188.00
 (Other):93
```

The `summary` command informs us that the average r_{wg} value is .186 and the range is from 0 to 0.433. By convention, values at or above 0.70 are considered good agreement, so there appears to be low agreement among individuals within the same work groups with respect to coping using religion. The `summary` command also provides information about the group sizes.

To calculate r_{wg} for work hours, the expected random variance (EV) needs to be changed from its default value of 2. Work hours was asked using an 11-point item, so EV based on the rectangular distribution (σ_{EU}^2) is 10.00 ($\sigma_{EU}^2=(11^2-1)/12$) – see the `rwg` help file for details).

```
> RWG.HRS<-rwg(bhr2000$HRS,bhr2000$GRP,ranvar=10.00)
> mean(RWG.HRS[,2])
```

```
[1] 0.7353417
```

There is apparently much higher agreement about work hours within groups than there was about using religion as a coping mechanism in this sample. By convention, this mean value would indicate agreement because r_{wg} (and $r_{wg(j)}$) values above .70 are considered to provide evidence of agreement.

The use of the `rwg.j` function is nearly identical to the use of the `rwg` function except that the first argument to `rwg.j` is a matrix instead of a vector. In the matrix, each column represents one item in the multi-item scale, and each row represents an individual response. For instance, columns 2-12 in `bhr2000` represent 11 items comprising a leadership scale. The items were assessed using 5-point response options (Strongly Disagree to Strongly Agree), so the expected random variance is (5^2-1)/12 or 2.

```
> RWGJ.LEAD<-rwg.j(bhr2000[,2:12],bhr2000$GRP,ranvar=2)
> summary(RWGJ.LEAD)
     grpid            rwg.j            gsize
 1       : 1       Min.   :0.7859   Min.   :  8.00
 10      : 1       1st Qu.:0.8708   1st Qu.: 29.50
 11      : 1       Median :0.8925   Median : 45.00
 12      : 1       Mean   :0.8876   Mean   : 54.55
 13      : 1       3rd Qu.:0.9088   3rd Qu.: 72.50
 14      : 1       Max.   :0.9440   Max.   :188.00
 (Other):93
```

Note that Lindell and colleagues (Lindell & Brandt, 1997, 1999; 2000; Lindell, Brandt & Whitney, 1999) have raised concerns about the mathematical underpinnings of the $r_{wg(j)}$ formula. Specifically, they note that this formula is based upon the Spearman-Brown reliability estimator. Generalizability theory provides a basis to believe that reliability should increase as the number of measurements increase, so the Spearman-Brown formula is defensible for measures of reliability. In contrast, there may be no theoretical grounds to believe that generalizability theory applies to measures of agreement. That is, there may be no reason to believe that agreement should increase as the number of measurements on a scale increase (but also see LeBreton, James & Lindell, 2005).

To address this potential concern with the $r_{wg(j)}$, Lindell and colleagues have proposed the $r^*_{wg(j)}$. The $r^*_{wg(j)}$ is calculated by substituting the average variance of the items in the scale into the numerator of r_{wg} formula in lieu of using the $r_{wg(j)}$ formula (r_{wg} = 1- Observed Group Variance/Expected Random Variance). Note that Lindell and colleagues also recommend against truncating the Observed Group Variance value so that it matches the Expected Random Variance value in cases where the observed variance is larger than the expected variance. Their modification results $r^*_{wg(j)}$ values being able to take on negative values. We can use the function `rwg.j.lindell` to estimate the $r^*_{wg(j)}$ values for leadership.

```
> RWGJ.LEAD.LIN<-rwg.j.lindell(bhr2000[,2:12],
bhr2000$GRP,ranvar=2)
> summary(RWGJ.LEAD.LIN)
     grpid          rwg.lindell         gsize
 1       : 1       Min.   :0.2502   Min.   :  8.00
 10      : 1       1st Qu.:0.3799   1st Qu.: 29.50
 11      : 1       Median :0.4300   Median : 45.00
 12      : 1       Mean   :0.4289   Mean   : 54.55
```

```
 13      : 1        3rd Qu.:0.4753   3rd Qu.: 72.50
 14      : 1        Max.   :0.6049   Max.   :188.00
 (Other):93
```

The average $r^*_{wg(j)}$ value of .43 is considerably lower than the average $r_{wg(j)}$ value of .89 listed earlier.

3.2.2 The a_{wg} Index

Brown and Hauenstein (2005) argue that the r_{wg} family of agreement indices have three major limitations: (1) the magnitude of the measures are dependent on sample size, (2) the scale used to assess the construct influences the magnitude of the measure, and (3) the use of the uniform null distribution is an invalid comparison upon which to base an estimate of agreement. To overcome these limitations, Brown and Hauenstein proposed the a_{wg} index as a multi-rater agreement measure analogous to Cohen's kappa. The a_{wg} index is calculated using the `awg` function.

The `awg` function has three arguments: `x`, `grpid`, and `range`. The `x` argument represents the item or scale upon which to calculate a_{wg} values. The `awg` function determines whether `x` is a vector (single item) or multiple item matrix (representing the items in a scale), and performs the a_{wg} calculation appropriate for the type of input. The second function, `grpid`, is a vector containing the group ids associated with the `x` argument. The third argument, `range`, represents the upper and lower limits of the response options. The `range` defaults to `c(1,5)` which represents a 5-point scale from (for instance) strongly disagree (1) to strongly agree (5).

The code below illustrates the use of the `awg` function for the multi-item leadership scale.

```
> AWG.LEAD<-awg(bhr2000[,2:12],bhr2000$GRP)
> summary(AWG.LEAD)
     grpid          a.wg            nitems      nraters        avg.grp.var
 1       : 1   Min.   :0.2223   Min.   :11   Min.   :  8.00   Min.   :0.2787
 10      : 1   1st Qu.:0.3654   1st Qu.:11   1st Qu.: 29.50   1st Qu.:0.4348
 11      : 1   Median :0.4193   Median :11   Median : 45.00   Median :0.5166
 12      : 1   Mean   :0.4125   Mean   :11   Mean   : 54.55   Mean   :0.5157
 13      : 1   3rd Qu.:0.4635   3rd Qu.:11   3rd Qu.: 72.50   3rd Qu.:0.5692
 14      : 1   Max.   :0.5781   Max.   :11   Max.   :188.00   Max.   :0.9144
 (Other):93
```

Notice that ratings of the a_{wg} tend to more similar in magnitude to the $r^*_{wg(j)}$ which likely reflects the facts that (a) large variances can result in negative ratings reflecting disagreement, and (b) the denominator for the measure is fundamentally based upon the idea of maximum possible variance (similarly to the $r^*_{wg(j)}$) rather than a uniform distribution.

One final note is that Brown and Hauenstein (2005) contend that the class of r_{wg} agreement indices should not be estimated in cases where group size (or number of raters) is less than the number of response options (scale anchors) associated with the items (A). In this example, A is 5 representing the scale anchors from strongly disagree to strongly agree. In contrast, however, Brown and Hauenstein (2005) state that it is appropriate to estimate a_{wg} on the number of anchors minus 1. The reason why a_{wg} can be estimated on smaller groups is that a_{wg} (unlike r_{wg}) uses a sample-based variance estimate in the denominator whereas r_{wg} uses a population-based variance estimate (recall that the formula for the rectangular variance distribution is ranvar=(A^2-1)/12 which represents a population-based value (σ_{EU}^2)). In the example there is no issue with group size given that the smallest group has eight members.

3.2.3 Significance testing using `rwg.sim` and `rwg.j.sim`

As noted in section 3.2.1, r_{wg} and $r_{wg(j)}$ values at or above .70 are conventionally considered providing evidence of within-group agreement. A series of studies by Charnes and Schriesheim (1995); Cohen, Doveh and Eick (2001); Dunlap, Burke, and Smith-Crowe (2003) and Cohen, Doveh and Nahum-Shani (2009) lay the groundwork for establishing tests of statistical significance for r_{wg} and $r_{wg(j)}$. The basic idea behind these simulations is to draw observations from a known null distribution (generally a uniform or rectangular null), and repeatedly estimate r_{wg} or $r_{wg(j)}$. Because the observations are drawn from a uniform random null, r_{wg} or $r_{wg(j)}$ estimates in the simulation should be zero. Occasionally, however, the r_{wg} or $r_{wg(j)}$ values will be larger than zero reflecting sampling variability associated with the specific attributes of the simulation. Repeatedly drawing random numbers and estimating r_{wg} and $r_{wg(j)}$ provides a way to calculate expected null values and confidence intervals.

The simulations conducted by Cohen et al., (2001) varied several factors, but the two factors found to be most important for the expected null values of the $r_{wg(j)}$ were (a) group size and (b) the number of items. Indeed, Cohen et al., (2001) found that the expected null $r_{wg(j)}$ values in the simulations differed considerably as group size and the number of items varied. These findings imply that the conventional value of .70 may be a reasonable cut-off value for significance with some configurations of group sizes and items but may not be reasonable for others. Thus, Cohen et al., (2001) recommended researchers simulate parameters based on the specific characteristics of the researchers' samples when determining whether $r_{wg(j)}$ values are significant.

In 2003, Dunlap and colleagues estimated 95% confidence intervals for the single item r_{wg} using the idea of simulating null distributions. Their work showed that the 95% confidence interval for the single item measure varied as a function of (a) group size and (b) the number of response options. In the case of 5 response options (e.g., strongly disagree, disagree, neither, agree, strongly agree), the 95% confidence interval estimate varied from 1.00 with a group of 3 to 0.12 for a group of 150. That is, one would need an r_{wg} estimate of 1.00 with groups of size three to be 95% certain the groups agreed more than chance levels, but with groups of size 150 any value equal to or greater than 0.12 would represent significant agreement.

The function `rwg.sim` provides a way to replicate the results presented by Dunlap and colleagues. For instance, to estimate the 95% confidence interval for a group of size 10 on an item with 5 response options one would provide the following parameters to the `rwg.sim` function keeping in mind that the results from a separate run will not match these results exactly because no random seed was set:

```
> RWG.OUT<-rwg.sim(gsize=10, nresp=5, nrep=10000)
> summary(RWG.OUT)
$rwg
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.1221  0.2167  0.8667

$gsize
[1] 10
$nresp
[1] 5
$nitems
[1] 1
$rwg.95
```

```
[1] 0.5277778
```

The results in the preceding example are based on 10,000 simulation runs. In contrast, Dunlap et al., (2003) used 100,000 simulation runs. Nonetheless, both Table 2 from Dunlap et al., (2003) and the example above suggest that 0.53 is the 95% confidence interval estimate for a group of size 10 with five response options.

Because the estimation of r_{wg} in the simulations produces a limited number of possible responses, the typical methods for establishing confidence intervals (e.g., the generic function `quantile`) cannot be used. Thus, the multilevel package provides a `quantile` method for the objects of class `agree.sim` created using `rwg.sim`. To obtain 90%, 95% and 99% confidence interval estimates (or any other values) one would issue the following command:

```
> quantile(RWG.OUT,c(.90,.95,.99))
  quantile.values confint.estimate
1            0.90        0.4222222
2            0.95        0.5277778
3            0.99        0.6666667
```

Cohen et al. (2009) expanded upon the work of Dunlap et al., (2003) and the early work by Cohen et al. (2001) by demonstrating how confidence interval estimation could be applied to multiple item scales in the case of $r_{wg(j)}$ values. The function `rwg.j.sim` is based upon the work of Cohen et al., (2009) and simulates $r_{wg(j)}$ values from a uniform null distribution for user supplied values of (a) group size, (b) number of items in the scale, and (c) number of response options on the items. Users also provide the number of simulation runs (repetitions) upon which to base the estimates. In most cases, the number of simulation runs will be 10,000 or more although the examples illustrated here will be limited to 1,000.

The final optional argument to `rwg.j.sim` is `itemcors`. If this argument is omitted, the simulated items used to comprise the scale are assumed to be independent (non-correlated). If the argument is provided, the items comprising the scale are simulated to reflect a given correlational structure. Cohen et al., (2001) showed that results based on independent (non-correlated) items were similar to results based on correlated items; nonetheless, the model with correlated items is more realistic and thereby preferable (see Cohen et al., 2009).

For an example of using `rwg.j.sim` with non-correlated items, consider a case where a researcher was estimating the expected value and confidence intervals of $r_{wg(j)}$ on a sample where group size was 15 using a 7-item scale with 5 response options for the items (A=5). The call to `rwg.j.sim` would be:

```
> RWG.J.OUT<-rwg.j.sim(gsize=15,nitems=7,nresp=5,nrep=1000)

> summary(RWG.J.OUT)
$rwg.j
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000000 0.000000 0.009447 0.161800 0.333900 0.713700
$gsize
[1] 15
$nresp
[1] 5
$nitems
[1] 7
$rwg.j.95
```

```
[1] 0.5559117
```

In this example, the upper expected 95% confidence interval is 0.56. This is lower than 0.70, and suggests that in this case the value of 0.70 might be too stringent. Based on this simulation, one might justifiably conclude that a value of 0.56 is evidence of significant agreement ($p<.05$). Using the simulation, one can show that as group size increases and the number of items increase, the criteria for what constitutes significant agreement decreases.

To illustrate how significance testing of $r_{wg(j)}$ might be used in a realistic setting, we will examine whether group members agreed about three questions specific to mission importance in the `lq2002` data set. These data were also analyzed in Cohen et al., 2009. We begin by estimating the mean $r_{wg(j)}$ for the 49 groups in the sample and obtaining a value of .58. This value is below the .70 conventional criteria and suggests a lack of agreement.

```
> RWG.J<-rwg.j(lq2002[,c("TSIG01","TSIG02","TSIG03")],
  lq2002$COMPID,ranvar=2)
> summary(RWG.J)
     grpid        rwg.j            gsize
 10     : 1   Min.   :0.0000   Min.   :10.00
 13     : 1   1st Qu.:0.5099   1st Qu.:18.00
 14     : 1   Median :0.6066   Median :30.00
 15     : 1   Mean   :0.5847   Mean   :41.67
 16     : 1   3rd Qu.:0.7091   3rd Qu.:68.00
 17     : 1   Max.   :0.8195   Max.   :99.00
 (Other):43
```

To determine whether the value of .58 is significant, we use the `rwg.j.sim` function using item correlations and average group size (41.67 rounded to 42). In this case, notice the simulation suggests that a value of .35 is significant providing evidence of significant agreement.

```
> RWG.J.OUT<-rwg.j.sim(gsize=42,nitems=3,nresp=5,
   itemcors=cor(lq2002[,c("TSIG01","TSIG02","TSIG03")]),
   nrep=1000)
> summary(RWG.J.OUT)
$rwg.j
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.000000 0.000000 0.007224 0.088520 0.162500 0.548600
$gsize
[1] 42
$nresp
[1] 5
$nitems
[1] 3
$rwg.j.95
[1] 0.346875
```

3.2.4 Average Deviation (AD) Agreement using `ad.m`

Burke, Finkelstein and Dusig (1999) proposed using average deviation (AD) indices as measures of within-group agreement. Cohen et al., (2009) note that AD indices are also referred to as Mean or Median Average Deviation. AD indices are calculated by first computing the absolute deviation of each observation from the mean or median. Second, the absolute deviations

are averaged to produce a single AD estimate for each group. The formula for AD calculation on a single item is:

$AD = \Sigma|x_{ij} - X_j|/N$

where x_{ij} represents an individual observation (i) in group j; X_j represents the group mean or median, and N represents the group size. When AD is calculated on a scale, the AD formula above is estimated for each item on the scale, and each item's AD value is averaged to compute the scale AD score.

AD values are considered practically significant when the values are less than A/6 where A represents the number of response options on the item. For instance, A is 5 when items are asked on a Strongly Disagree, Disagree, Neither, Agree and Strongly Agree format so any value less than .83 (5/6) would be considered practically significant.

The function `ad.m` is used to compute the average deviation of the mean or median. The function requires the two arguments, `x` and `grpid`. The `x` argument represents the item or scale upon which to estimate the AD value. The `ad.m` function determines whether `x` is a vector (single item) or multiple item matrix (multiple items representing a scale), and performs the AD calculation appropriate for the nature of the input variable. The second function, `grpid`, is a vector containing the group ids of the `x` argument. The third argument is optional. The default value is to compute the Average Deviation of the mean. The other option is to change the `type` argument to "`median`" and compute the Average Deviation of the median.

For instance, recall that columns 2-12 in `bhr2000` represent 11 items comprising a leadership scale. The items were assessed using 5-point response options (Strongly Disagree to Strongly Agree), so the practical significance of the AD estimate is 5/6 or 0.83. The AD estimates for the first five groups and the mean of the overall sample are provided below:

```
> data(bhr2000)
> AD.VAL <- ad.m(bhr2000[, 2:12], bhr2000$GRP)
> AD.VAL[1:5,]
  grpid      AD.M gsize
1     1 0.8481366    59
2     2 0.8261279    45
3     3 0.8809829    83
4     4 0.8227542    26
5     5 0.8341355    82
> mean(AD.VAL[,2:3])
      AD.M      gsize
 0.8690723 54.5454545
```

Two of the estimates are less than 0.833 suggesting these two groups (2 and 4) agree about ratings of leadership. The overall AD estimate is 0.87, which is also higher than 0.83 and suggests a general lack of agreement.

The AD value estimated using the median instead of the mean, in contrast, suggests practically significant agreement for the sample as a whole.

```
> AD.VAL <- ad.m(bhr2000[, 2:12], bhr2000$GRP,type="median")
> mean(AD.VAL[,2:3])
      AD.M      gsize
 0.8297882 54.5454545
```

To use the ad.m function for single item variables such as the work hours (HRS) variable in the bhr2000 data simply include the HRS vector instead of a matrix as the first argument. Recall that work hours is asked on an 11-point response format scale so practical significance is 11/6 or 1.83. The average observed AD value of 1.25 suggests within-group agreement about work hours across the sample as a whole.

```
> AD.VAL.HRS <- ad.m(bhr2000$HRS, bhr2000$GRP)
> mean(AD.VAL.HRS[,2:3])
     AD.M     gsize
 1.249275 54.545455
```

3.2.5 Significance testing ad.m.sim

The function ad.m.sim is used to simulate AD values and test for significance of various combinations of group size, number of response options and number of items in multiple-item scales. The ad.m.sim function is similar to the rwg.sim and rwg.j.sim functions used to test the significance of r_{wg} and $r_{wg(j)}$; however, unlike the functions for the two forms of the r_{wg}, the ad.m.sim function works with both single items and multiple-item scales.

The ad.m.sim function is based upon the work of Cohen et al. (2009) and of Dunlap et al., (2003). The function simulates AD values from a uniform null distribution for user supplied values of (a) group size, (b) number of items in the scale, and (c) number of response options on the items. Based on Cohen et al. (2009), the final optional parameter can include a correlation matrix when simulating multiple-item scales. The user also provides the number of simulation runs (repetitions) upon which to base the estimates. Again in practice, the number of simulation runs will typically be 10,000 or more although the examples illustrated here will be limited to 1,000.

To illustrate the ad.m.sim function, consider the 11 leadership items in the bhr2000 dataframe. Recall the AD value based on the mean suggested that groups failed to agree about leadership. In contrast, the AD value based on the median suggested that groups agreed. To determine whether the overall AD value based on the mean is statistically significant, one can simulate data matching the characteristics of the bhr2000 sample:

```
> AD.SIM<-ad.m.sim(gsize=55,nresp=5,
itemcors=cor(bhr2000[,2:12]),type="mean",nrep=1000)
> summary(AD.SIM)
$ad.m
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.087   1.182   1.208   1.209   1.236   1.340

$gsize
[1] 55

$nresp
[1] 5

$nitems
[1] 11

$ad.m.05
[1] 1.138212
```

```
$pract.sig
[1] 0.8333333
```

The simulation suggests that any AD mean value less than or equal to 1.14 is statistically significant. Thus, while the AD value for the leadership items (0.87) may not meet the criteria for practical significance, it does for statistical significance. As with the r_{wg} simulation functions, the `ad.m.sim` function has a specifically associated `quantile` function to identify different cut-off points. The example below illustrates how to identify values corresponding to the .90 (.10), .95 (.05) and .99 (.01) significance levels. That is, to be 99% certain that a value was significant, it would need to be smaller than or equal to 1.114.

```
> quantile(AD.SIM,c(.10,.05,.01))
  quantile.values confint.estimate
1            0.10         1.155763
2            0.05         1.138212
3            0.01         1.114170
```

3.2.6 Agreement: Random Group Resampling

The final agreement related function in the multilevel library is `rgr.agree`. In some ways this function is similar to the `rwg.j.sim` function in that it uses repeated simulations of data to draw inferences about agreement. The difference is that the `rgr.agree` function uses the actual group data, while the `rwg.j.sim` function simulates from an expected distribution (the uniform null).

The `rgr.agree` function (a) uses Random Group Resampling to create pseudo groups and calculate pseudo group variances, (b) estimates actual group variances, and (c) performs tests of significance to determine whether actual group and pseudo group variances differ. To use `rgr.agree`, one must provide three variables. The first is a vector representing the variable upon which one wishes to estimate agreement. The second is group membership (`grpid`). The third parameter is the number of pseudo groups to generate.

The third parameter requires a little explanation, because in many cases the number of pseudo groups returned in the output will not exactly match the third parameter. For instance, in our example, we will request 1000 pseudo groups, but the output will return only 990. This is because the `rgr.agree` algorithm creates pseudo groups that are identical in size characteristics to the actual groups. In so doing, the algorithm creates sets of pseudo groups in "chunks." The size of each chunk is based upon the number of actual groups. So, if there are 99 actual groups, then the total number of pseudo groups must be evenly divisible by 99. Nine-hundred-and-ninety is evenly divisible by 99, while 1000 is not. Rather than require the user to determine what is evenly divisible by the number of groups, `rgr.agree` will do this automatically. Below is an example of using `rgr.agree` on the work hours variable.

```
> RGR.HRS<-rgr.agree(bhr2000$HRS,bhr2000$GRP,1000)
```

The first step is to create an RGR Agreement object named `RGR.HRS`. The object contains several components. In most cases, however, users will be interested in the estimated z-value indicating whether the within-group variances from the actual groups are smaller than the variances from the pseudo groups. A useful way to get this information is to use the `summary`

command. When `summary` is applied to the RGR agreement object it provides standard deviations, variance estimates, an estimate of the z-value, and upper and lower confidence intervals.

```
> summary(RGR.HRS)
$"Summary Statistics for Random and Real Groups"
  N.RanGrps Av.RanGrp.Var SD.Rangrp.Var Av.RealGrp.Var  Z-value
1       990      3.322772      0.762333       2.646583 -8.82554

$"Lower Confidence Intervals (one-tailed)"
    0.5%       1%     2.5%       5%      10%
1.648162 1.795134 1.974839 2.168830 2.407337

$"Upper Confidence Intervals (one-Tailed)"
     90%      95%    97.5%      99%    99.5%
4.251676 4.545078 4.832813 5.642410 5.845143
```

The first section of the summary provides key statistics for contrasting within-group variances from real group with within-group variances from random groups. The second and third sections provide lower and upper confidence intervals. Keep in mind that if one replicates this example one is likely to get slightly different results because no random seed was set. While the exact numbers may differ, the conclusions drawn should be the same.

The first section of the summary shows that the average within-group variance for the random groups was 3.32 with a Standard Deviation of 0.76. In contrast, the average within-group variance for the real groups was considerably smaller at 2.65. The estimated z-value suggests that, overall, the within-group variances in ratings of work hours from real groups were significantly smaller than the within-group variances from the random groups. These results suggest that group members agree about work hours. Recall that a z-value greater than or less than 1.96 signifies significance at $p<.05$, two-tailed.

The upper and lower confidence interval information allows one to estimate whether specific groups do or do not display agreement. For instance, only 5% of the pseudo groups had a variance less than 2.17. Thus, if we observed a real group with a variance smaller than 2.17, we could be 95% confident this group variance was smaller than the variances from the pseudo groups. Likewise, if we want to be 90% confident we were selecting groups showing agreement, we could identify real groups with variances less than 2.41.

To see which groups meet this criterion, use the `tapply` function in conjunction with the `sort` function. The `tapply` function partitions the first variable by levels of the second variable and performs a specified function much like the `aggregate` function (see section 3.1.2). Below we partition HRS into separate Groups (GRP) and calculate the variance for each group (`var`). Using `sort` in front of this command makes the output easier to read.

```
> sort(tapply(bhr2000$HRS,bhr2000$GRP,var))
       33        43        38        19         6        39        69        17
0.8242754 1.0697636 1.1295681 1.2783251 1.3166667 1.3620690 1.4566667 1.4630282
       20        99        98        44         4        53        61        63
1.5009740 1.5087719 1.5256410 1.5848739 1.6384615 1.6503623 1.6623656 1.7341430
```

```
       66        14        76        71        21        18        59        50
1.7354302 1.7367089 1.7466200 1.7597586 1.7808500 1.7916027 1.8112599 1.8666667
       48        60        83         8        22         2        75        11
1.8753968 1.9267300 1.9436796 1.9476190 1.9679144 2.0282828 2.1533101 2.1578947
       96        23        54        47        55        26        74        57
2.1835358 2.1864802 2.2091787 2.2165242 2.2518939 2.2579365 2.2747748 2.2808858
       45        97        64        35        32        41         1        24
2.2975687 2.3386525 2.3535762 2.3563495 2.3747899 2.4096154 2.4284044 2.4391678
       82        37        81        68        42        73        34        25
2.4429679 2.4493927 2.5014570 2.5369458 2.5796371 2.6046154 2.6476418 2.6500000
       93        62        92        12        40        88         5        29
2.6602168 2.7341080 2.7746106 2.7906404 2.7916084 2.8505650 2.8672087 2.8748616
       85        70        77        51         3        13        79        87
2.8974843 2.9938483 3.0084034 3.0333333 3.0764032 3.1643892 3.1996997 3.2664569
        7        95        78        84        46        27        36        15
3.2712418 3.2804878 3.3839038 3.3886048 3.4084211 3.4309008 3.4398064 3.4425287
       89        16        58        49         9        31        90        72
3.4444444 3.4461538 3.4949020 3.5323440 3.6258065 3.6798419 3.8352838 3.9285714
       91        80        86        10        94        28        30        56
3.9565960 3.9729730 3.9753195 4.0336134 4.0984900 4.0994152 4.6476190 4.7070707
       65        52        67
4.7537594 5.2252964 5.3168148
```

If we start counting from group 33 (the group with the lowest variance of 0.82) we find 46 groups with variances smaller than 2.41. That is, we find 46 groups that have smaller than expected variance using the 90% confidence estimate.

It may also be interesting to see what a "large" variance is when defined in terms of pseudo group variances. This information is found in the third part of the summary of the `RGR.HRS` object. A variance of 4.55 is in the upper 95% of all random group variances. Given this criterion, we have five groups that meet or exceed this standard. In an applied setting, one might be very interested in examining this apparent lack of agreement in groups 30, 56, 65, 52 and 67. That is, one might be interested in determining what drives certain groups to have very large differences in how individuals perceive work hours.

Finally, for confidence intervals not given in the summary, one can use the `quantile` function with the random variances (`RGRVARS`) in the `RGR.HRS` object. For instance to get the lower .20 confidence interval:

```
> quantile(RGR.HRS$RGRVARS, c(.20))
     20%
2.695619
```

Note that `rgr.agree` only works on vectors. Consequently, to use `rgr.agree` with the leadership scale we would need to create a leadership scale score. We can do this using the

rowMeans function. We will create a leadership scale (LEAD) and put it in the bhr2000 dataframe, so the specific command we issue is:

```
>bhr2000$LEAD<-rowMeans(bhr2000[,2:12],na.rm=TRUE)
```

Now that we have created a leadership scale score, we can perform the RGR agreement analysis on the variable.

```
> summary(rgr.agree(bhr2000$LEAD,bhr2000$GRP,1000))

$"Summary Statistics for Random and Real Groups"
  N.RanGrps Av.RanGrp.Var SD.Rangrp.Var Av.RealGrp.Var  Z-value
1       990     0.6011976     0.1317229      0.5156757 -6.46002

$"Lower Confidence Intervals (one-tailed)"
     0.5%        1%      2.5%        5%       10%
0.2701002 0.3081618 0.3605966 0.3939504 0.4432335

$"Upper Confidence Intervals (one-Tailed)"
      90%       95%     97.5%       99%     99.5%
0.7727185 0.8284755 0.8969857 0.9651415 1.0331922
```

The results indicate that the variance in actual groups about leadership ratings is significantly smaller than the variance in randomly created groups (i.e., individuals agree about leadership). For interesting cases examining situations where group members do not agree see Bliese & Halverson (1998a) and Bliese and Britt (2001).

Ongoing research continues to examine the nature of RGR based agreement indices relative to ICC(1), ICC(2) and other measures of agreement such as the r_{wg} (e.g., Lüdtke & Robitzsch, 2009). This work indicates that measures of RGR agreement are strongly related to the magnitude of the ICC values.

3.2.7 Reliability: ICC(1) and ICC(2)

Reliability indices differ from agreement indices (see Bliese, 2000; LeBreton & Senter, 2008), and the multilevel package contains the ICC1 and ICC2 functions to estimate reliability. These two functions are applied to ANOVA models and are used to estimate ICC(1) and ICC(2) as described by Bartko, (1976), James (1982), and Bliese (2000).

These two functions are applied to a one-way analysis of variance model using aov. Notice the as.factor function on GRP in the command below which designates GRP (a numeric vector) as being categorical or nominal. Once specified as categorical, R creates N-1 dummy codes in the model matrix using GRP 1 as the referent. More specifically, the contrast default in as.factor is contr.treatment which uses the first factor as the referent; however, R provides numerous options for controlling dummy and effects coding – see help(contrasts) for details. In the present example, the 99 groups result in 98 dummy-coded categories (98 df).

```
> data(bhr2000)
> hrs.mod<-aov(HRS~as.factor(GRP),data=bhr2000)
> summary(hrs.mod)
                 Df  Sum Sq Mean Sq F value    Pr(>F)
```

```
as.factor(GRP)   98  3371.4    34.4  12.498 < 2.2e-16 ***
Residuals      5301 14591.4     2.8
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

The `ICC1` and `ICC2` functions are then applied to the `aov` object.

```
> ICC1(hrs.mod)
[1] 0.1741008
> ICC2(hrs.mod)
[1] 0.9199889
```

The ICC(1) value is equivalent to the ICC term referred to the mixed-effects model literature (e.g., Bryk & Raudenbush, 1992; 2002) and a value of .17 indicates that 17% of the variance in individual perceptions of work hours can be "explained" by group membership. The ICC(2) is a measure of group-mean reliability and a value of .92 indicates that groups can be reliably differentiated in terms of average work hours (see Bliese, 2000).

3.2.8 Estimate multiple ICC values: `mult.icc`

The `mult.icc` function can be used to estimate multiple ICC(1) and ICC(2) values in a given data set. Code to estimate the ICC(1) and ICC(2) values for work hours, coping with religion, and three different leadership items in the `bhr2000` data set is provided below. In the function, the first element is a subset of the dataframe with the variables of interest and the second element is the grouping variable.

```
> mult.icc(bhr2000[,c("HRS","RELIG","AF06","AF07","AP12")],bhr2000$GRP)
  Variable        ICC1      ICC2
1      HRS 0.177543969 0.9217206
2    RELIG 0.009801542 0.3506163
3     AF06 0.103492912 0.8629524
4     AF07 0.087490365 0.8394800
5     AP12 0.149052933 0.9052514
```

The results suggest that individuals use of religion as a coping mechanism had the lowest ICC(1) value (less than 1% of the variance in an individual's response can be explained by group membership). The `mult.icc` function is based upon `lme` from the `nlme` package so it returns slightly different ICC(1) and ICC(2) estimates for Work Hours (0.178 and 0.922, respectively) than estimates based on the `aov` models (0.174 and 0.920). If group sizes equal, the `lme` and `aov` approach would provide virtually identical values. In general, the preferred method with unbalanced data would be to use `lme`. One other difference (not illustrated here) is that ICC(1) values estimated in OLS can be negative, but ICC(1) values based on mixed-effects models have a lower bound of zero.

3.2.9 Comparing ICC values with a two-stage bootstrap: `boot.icc`

When examining ICC values, it can often be informative to estimate a sampling distribution to determine whether ICC values differ. For instance, the ICC(1) values for Work Hours is 0.178 (mixed-effects model), but it is not clear whether the other values which are lower significantly differ from 0.178. One way to answer the question of whether ICC values differ is to estimate a measure of variability around the point estimates. The `boot.icc` is an experimental function

that performs a two-stage bootstrap. A two-stage first samples with replacement from level-2 units (the groups) followed by sampling with replacement from individuals within the level-2 units. The function is computationally intensive, but is illustrated below both with using `lme` (the default) and `aov` (an option) as the computational algorithm underlying the ICC(1) estimate:

```
> system.time(OUT.HRS.lme<-boot.icc3(bhr2000$HRS,bhr2000$GRP,1000))
   user  system elapsed
 292.87    0.53  295.86
> quantile(OUT.HRS.lme,c(0.025,.975))
     2.5%     97.5%
0.1372000 0.2211409

> system.time(OUT.HRS.aov<-boot.icc3(bhr2000$HRS,bhr2000$GRP,1000,
  aov.est=TRUE))
   user  system elapsed
 301.93    3.35  307.89
> quantile(OUT.HRS.aov,c(0.025,.975))
     2.5%     97.5%
0.1302396 0.2160199
```

Notice that the `aov` option is slightly slower and the values are slightly smaller which is not surprising given that the `aov` estimate of the ICC(1) is smaller than the `lme` estimate. The `lme` percentile-based 95% confidence interval for the ICC(1) for work hours is [0.137, 0.221] suggesting that single point estimates of ICC(1) values outside this range would significantly differ from those associated with Work Hours. In the example using `mult.icc` everything except AP12 (I am impressed by the quality of leadership in this company) has a smaller ICC(1) value than the lower confidence interval of 0.137 for work hours. A more thorough comparison would involve estimating confidence intervals for AP12 and using both sets of confidence intervals to draw inferences (Cummings & Finch, 2005). Finally note that performing a non-parametric bootstrap of nested data is controversial because it is not clear how to best sample with replacement.

3.2.10 Visualizing an ICC(1) with `graph.ran.mean`

It is often valuable to visually examine the group-level properties of data to see the form of the group-level effects. Levin (1967) observed that high ICC(1) values can be the product of one or two highly aberrant groups rather than indicating generally shared group properties among the entire sample.

One way to examine the group-level properties of the data is to contrast the observed group means with group means that are the result of randomly assigning individuals to pseudo groups. If the actual group means and the pseudo-group means are identical, there is no evidence of group effects. If one or two groups are clearly different from the pseudo-group distribution it suggests the ICC(1) value is simply caused by a few aberrant observations. If a number of groups have higher than expected means, and a number have lower than expected means, it suggests fairly well-distributed group-level properties.

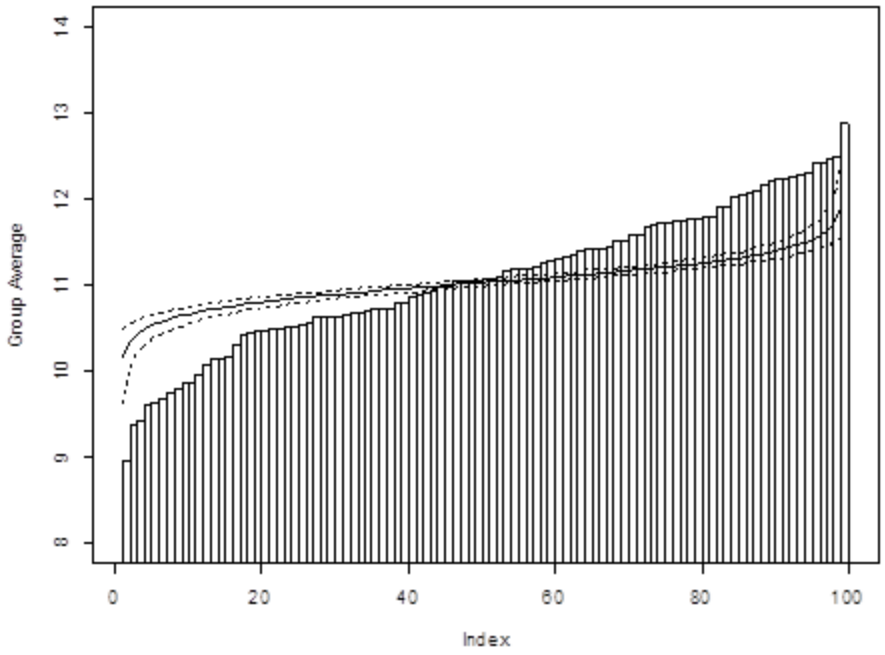
The `graph.ran.mean` function can be used to visually contrast actual group means with pseudo group means. The function requires three parameters. The first is the variable of interest. The second is the group designator, and the third is a smoothing parameter (`nreps`) determining how many sets of pseudo groups should be created to create the pseudo group curve. Low

numbers (<10) for this last parameter create a choppy line while high numbers (>25) create smooth lines. In cases where the parameter `bootci` is TRUE (see optional parameters), `nreps` should equal 1000 or more.

Three optional parameters control the y axis limits (`limits`); whether a plot is created (`graph=TRUE`) or a dataframe is returned (`graph=FALSE`); and whether bootstrap confidence intervals are estimated and plotted (`bootci=TRUE`). The default for `limits` is to use the lower 10% and upper 90% values of the raw data. The default for `graph` is to produce a plot, but returning a dataframe can be useful for exporting results for subsequent graphing in `ggplot2` or other packages. Finally, the default for `bootci` is to return a plot or a dataframe without bootstrap confidence interval estimates. In the following example, we plot the observed and pseudo group distribution of the work hours variable from the data set bhr2000.

```
> data(bhr2000)
> graph.ran.mean(bhr2000$HRS,bhr2000$GRP,nreps=1000,
limits=c(8,14),bootci=TRUE)
```

The function produces the resulting plot where the bar chart represents each groups' average rating of work hours sorted from highest to lowest, and the line represents a random distribution where 99 pseudo groups (with exact size characteristics of the actual groups) were created 1000 times and the sorted values were averaged across the 1000 iterations. The dotted lines represent the upper and lower 95% confidence interval estimates. In short, the line represents the expected distribution if there were no group-level properties associated with these data. The graph suggests fairly evenly distributed group-level properties associated with the data although two groups do stand out – one on the extreme high end and one on the extreme low end. In the end, though, the graph along with the results from the two-stage bootstrap analyses (section 3.2.11) which placed the ICC(1) estimate of 0.178 fairly close to the center of the 95% confidence interval of [0.137, 0.221] suggests that the ICC(1) values are not being driven by extreme groups (experience with other data suggests that a few extreme groups stand out in graphs and they also produce confidence intervals asymmetrical to the point estimate).



3.2.11 Simulating ICC(1) values with `sim.icc`

ICC(1) values play a key role in multilevel data; therefore, the ability to simulate ICC(1) values can be a valuable tool to help understand multilevel data and analyses. The `sim.icc` function generates data with specific ICC(1) values. Multiple vectors (items) can be generated in one of two ways: either with or without level-1 correlations. The function is used to generate a single vector (`VAR1`) below:

```
> set.seed(1535324)
> ICC.SIM<-sim.icc(gsize=10,ngrp=100,icc1=.15) #Simulate a single vector
> ICC.SIM[c(1:3,11:13),]  # Examine a few rows of simulated data
   GRP       VAR1
1    1  0.2800938
2    1 -1.4002869
3    1 -2.1422593
11   2 -1.3098119
12   2 -2.7164491
13   2 -0.3160884

> ICC1(aov(VAR1~as.factor(GRP), ICC.SIM))
[1] 0.16681
```

In the next example, four items are generated without any level-1 correlation among items. These data would represent a situation in which any observed raw correlation would be the due to the ICC(1) value. The example below uses the `waba` function discussed in section 3.4.1 to perform a variance decomposition of several raw correlations.

```
> set.seed(15324)
> ICC.SIM<-sim.icc(gsize=10,ngrp=100,icc1=.15,nitems=4)
```

```
> mult.icc(ICC.SIM[,2:5],ICC.SIM$GRP)
  Variable      ICC1      ICC2
1     VAR1 0.2035837 0.7188047
2     VAR2 0.1442111 0.6275778
3     VAR3 0.2229725 0.7415725
4     VAR4 0.1549414 0.6470794

> with(ICC.SIM,waba(VAR1,VAR2,GRP))$Cov.Theorem  #Examine CorrW
     RawCorr    EtaBx     EtaBy     CorrB    EtaWx     EtaWy       CorrW
1 0.07728039 0.530273 0.4775097 0.5939511 0.847827 0.8786265 -0.09815005

> with(ICC.SIM,waba(VAR1,VAR3,GRP))$Cov.Theorem  #Examine CorrW
    RawCorr    EtaBx     EtaBy     CorrB    EtaWx     EtaWy       CorrW
1 0.1769287 0.530273 0.5464122 0.6723887 0.847827 0.8375164 -0.02520087

> with(ICC.SIM,waba(VAR1,VAR4,GRP))$Cov.Theorem  #Examine CorrW
    RawCorr    EtaBx     EtaBy     CorrB    EtaWx     EtaWy      CorrW
1 0.1943248 0.530273 0.4874644 0.6127858 0.847827 0.8731429 0.04853107
```

Notice that the ICC(1) values for each item are variable (a function of small group sizes and a relatively small number of groups). Notice also that the `CorrW` (within-group correlation) values for three of the bivariate correlations vary around zero while `RawCorr` (the raw correlations) varies around .15 which corresponds to the simulated ICC(1) value.

As a final example, the code below incorporates a level-1 correlation of .30 among variables. Notice that the within-group correlation varies around .30 and the raw correlation increases as a function of the level-1 correlation and the ICC(1) value.

```
> set.seed(15324)
> ICC.SIM<-sim.icc(gsize=10,ngrp=100,icc1=.15,nitems=4,item.cor=.3)
> mult.icc(ICC.SIM[,2:5],ICC.SIM$GRP)
  Variable      ICC1      ICC2
1     VAR1 0.1669452 0.6671118
2     VAR2 0.1558558 0.6486689
3     VAR3 0.1381652 0.6158502
4     VAR4 0.1715351 0.6743219

> with(ICC.SIM,waba(VAR1,VAR2,GRP))$Cov.Theorem  #Examine CorrW
    RawCorr    EtaBx     EtaBy     CorrB     EtaWx     EtaWy     CorrW
1 0.3987741 0.498367 0.4883034 0.6976093 0.8669662 0.8726739 0.3026887

> with(ICC.SIM,waba(VAR1,VAR3,GRP))$Cov.Theorem  #Examine CorrW
    RawCorr    EtaBx     EtaBy     CorrB     EtaWx     EtaWy     CorrW
1 0.3746905 0.498367 0.4718088 0.7083573 0.8669662 0.8817009 0.2722794

> with(ICC.SIM,waba(VAR1,VAR4,GRP))$Cov.Theorem  #Examine CorrW
    RawCorr    EtaBx     EtaBy     CorrB     EtaWx     EtaWy     CorrW
1 0.3732463 0.498367 0.5024739 0.7104143 0.8669662 0.8645924 0.2606111
```

3.3 Regression and Contextual OLS Models

Contextual models represent a basic form a multilevel model where both the raw predictor and the group-mean of the same predictor are included in the model. For instance, regressing

Well-Being on individual work hours and group average work hours would represent a basic contextual model. A significant effect for the group-mean predictor indicates that the slope for the group-means differs from the slope for the individual-level variables and suggests a contextual effect is present (Firebaugh, 1978; Snijders & Bosker, 1999).

Prior to the introduction of multilevel mixed-effects models, OLS regression models were widely used to detect contextual effects. Firebaugh (1978) provides a good methodological discussion of these types of contextual models as does Kreft and De Leeuw (1998) and James and Williams (2000). While OLS regression has historically been used to estimate contextual regression models, the models can severely underestimate the standard error associated with the group-level effect producing tests that are too liberal. For this reason, mixed-effects models are the more appropriate way to identify contextual effects.

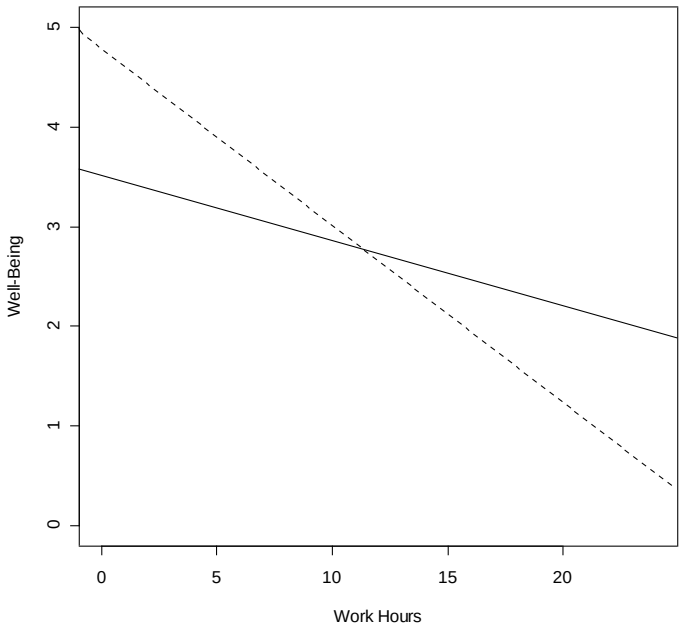
3.3.1 Contextual Effect Example

In this example, we use the `bh1996` dataframe to illustrate a contextual model involving work hours, group work hours and well-being presented in Bliese (2002). The `bh1996` dataframe has group mean variables included along with the group-mean center or demeaned variables.

```
> data(bh1996)
> names(bh1996)
 [1] "GRP"      "COHES"    "G.COHES"  "W.COHES"  "LEAD"     "G.LEAD"
 [7] "W.LEAD"   "HRS"      "G.HRS"    "W.HRS"    "WBEING"   "G.WBEING"
[13] "W.WBEING"

> tmod<-lm(WBEING~HRS+G.HRS,data=bh1996)
> round(summary(tmod)$coef,4)
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)   4.7831     0.1364  35.0680        0
HRS          -0.0465     0.0049  -9.4307        0
G.HRS        -0.1308     0.0130 -10.0596        0
```

Notice that `G.HRS` is significant with a t-value of –10.06 suggesting a significant contextual effect. Later we show that this t-value is highly inflated by a standard error that is too small. Nonetheless, it is informative to plot the form of the relationship showing that the group-mean slope (the dotted line) is considerably steeper than the individual slope (the solid line). Notice the use of `!duplicated(bh1996$GRP)` to select only the first row with a specific group's group-level data effectively reducing the sample size to 99 group means:

```
> plot(bh1996$HRS,bh1996$WBEING,xlab="Work Hours",
  ylab="Well-Being",type="n") #type = n omits the 7,382 points

> abline(lm(WBEING~HRS,data=bh1996)) # plots the individual-level slope
> abline(lm(G.WBEING~G.HRS,data=bh1996[!duplicated(bh1996$GRP),]),
  lty=2) #group-level slope
```



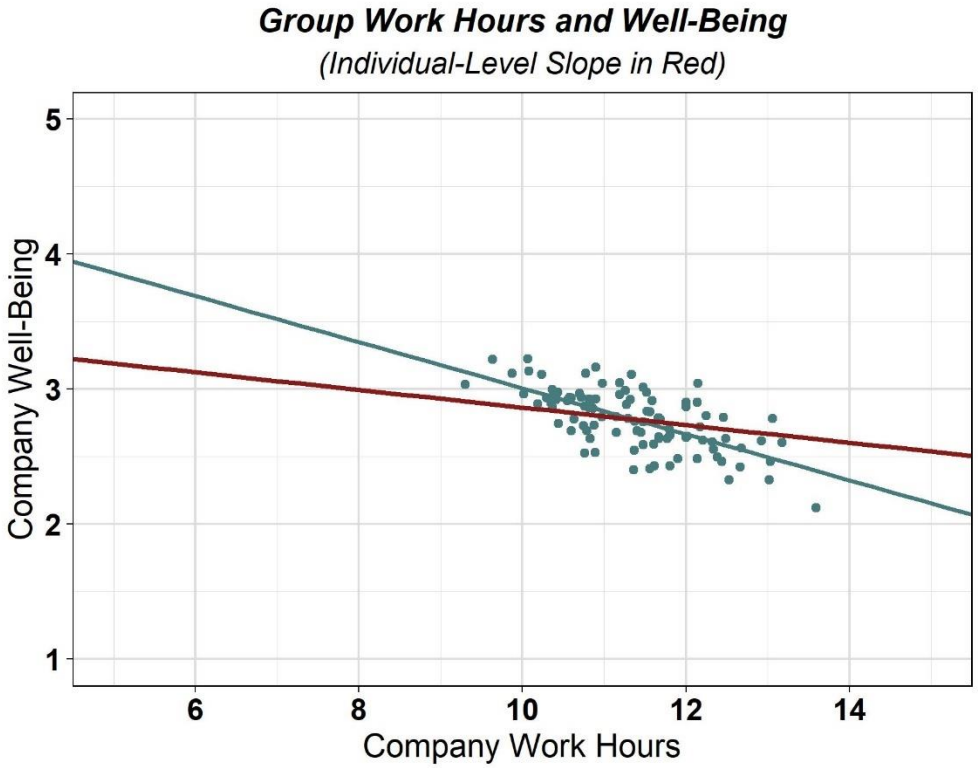
The idea that relationship strength might differ across levels is fundamental to multilevel analyses, so the basic idea of contextual regression is important. Fortunately, the problem with using OLS regression and having a standard error that is too small can be fixed in mixed-effect models (illustrated in section 4). For more details on the effects of non-independence see Bliese (2002); Bliese and Hanges (2004); Kenny and Judd, (1986) and Snijders and Bosker, (1999).

3.3.2 Contextual Effect Plot Using `ggplot2`

As an example of some of R's graphics capabilities, I reproduce the contextual effect using `ggplot2`.

```
library(ggplot2)

win.graph(height=4.75,width=6) #quartz() for MAC

data(bh1996)

bh1996.grp<-bh1996[!duplicated(bh1996$GRP),
      c("G.COHES","G.LEAD","G.HRS","G.WBEING")]

g <- ggplot(bh1996.grp, aes(x=G.HRS, y=G.WBEING))+
  labs(title = "Group Work Hours and Well-Being",
       subtitle = "(Individual-Level Slope in Red)",
       x = "Company Work Hours",
       y = "Company Well-Being")

g+coord_cartesian(xlim = c(5, 15),ylim=c(1,5))+
  geom_point(color="#477b7d")+
  geom_smooth(method="lm",fullrange=TRUE,
              se=FALSE,color="#477b7d")+
```

```
  geom_smooth(data=bh1996,aes(x=HRS,y=WBEING),
              method="lm",color="firebrick4")+
  scale_x_continuous(breaks=seq(0,24,by=2))+
  theme(
    plot.title = element_text(color="black", size=14,
                      hjust=0.5,face="bold.italic"),
    plot.subtitle = element_text(color="black", size=13,
                              hjust=0.5,face="italic"),
    axis.title.x = element_text(color="black", size=14),
    axis.title.y = element_text(color="black", size=14),
    axis.text = element_text(color="black",size=13,face="bold"),
    panel.border = element_rect(fill = NA, colour = "black",
                                size = rel(1)),
    panel.background = element_rect(fill = "transparent",
                                    colour = NA),
    panel.grid = element_line(colour = "grey87"),
    panel.grid.major = element_line(size = rel(1)),
    panel.grid.minor = element_line(size = rel(0.25)),
    axis.ticks = element_line(colour = "black",
                              size = rel(0.5))
  )
ggsave(filename = "c:\\temp\\plotgg.jpg",
       device = "jpeg")
```



3.4 Correlation Decomposition and the Covariance Theorem

OLS contextual models provide a way to determine whether regression slopes based on group means differ from regression slopes based on individual-level variables (while the OLS

contextual model for the group-mean predictor is biased by being too liberal, a null effect from the group-mean is informative). The covariance theorem provides a contextual model analog for correlations by breaking down a raw correlation into two separate components – the portion of the raw correlation attributable to within-group (individual) processes, and the portion of the correlation attributable to between-group (group-level) processes.

Robinson (1950) proposed the covariance theorem, and Dansereau and colleagues incorporated the theorem it into an analysis system they labeled WABA for Within-And-Between-Analyses (Dansereau, Alutto & Yammarino, 1984). WABA has two components: WABA I and WABA II. The first component (WABA I) uses decision tools based on eta values to inform decisions about the individual or group-level nature of the data. Eta values, however, are highly influenced by group size and unfortunately WABA I makes no group size adjustments; consequently, there is little value in using WABA I criteria unless one is working with dyads (see Bliese, 2000; Bliese & Halverson, 1998b).

Arguably a more useful way to draw inferences from eta-values is to contrast eta-values from actual groups to eta-values from pseudo groups. I illustrate this in a Random Group Resampling extension of the covariance theorem decomposition (see section 3.4.2). We begin, however, with a simple WABA analysis.

3.4.1 The `waba` and `cordif` functions

WABA II revolves around estimating the covariance theorem components, and the `waba` function in the multilevel library provides these components. The example partitions the raw correlation between work hours and well-being using the same data as used in the OLS contextual model example (section 3.3.1). The within-group correlation (`CorrW`) is group-mean centered (or demeaned) X and Y values. The group-level correlation (`CorrB`) represents the correlation between group means weighted by the size of each group.

```
> waba(bh1996$HRS,bh1996$WBEING,bh1996$GRP)
$Cov.Theorem
     RawCorr     EtaBX     EtaBY      CorrB     EtaWX     EtaWY      CorrW
1 -0.1632064 0.3787881 0.2359287 -0.7121729 0.9254834 0.9717704 -0.1107031
$n.obs
[1] 7382
$n.grps
[1] 99
```

The `waba` function returns a list with three elements. The first is the covariance theorem with all its components. The second is the number of observations, and the third is the number of groups. The latter two elements should routinely be examined because the `waba` function, by default, performs listwise deletion of missing values.

The raw correlation = (EtaBX*EtaBY*CorrB) + (EtaWX*EtaWY*CorrW) or

```
> (.379*.236*-.712)+(.925*.972*-.111)
[1] -0.1634842
```

The first set of parentheses represents the between-group component of the correlation, and the second set of parentheses represents the within-group component.

The weighted group-mean correlation of -.71 appears significantly larger than the within-group correlation of -.11. Since these two correlations are independent, we can contrast them using the `cordif` function. This function performs an r to z' transformation of the two correlations (see also the `rtoz` function) and then tests for differences between the two z' values using the formula provided in Cohen and Cohen (1983, p. 54). Four arguments are provided to `cordif`: (1) the first correlation of interest, (2) the second correlation of interest, (3) the N on which the first correlation is based, and (4) the N on which the second correlation is based. In our example, we already have the two correlations of interest (-.11 and -.71) and the relevant N values for the number of groups (99). The N for the within-group correlation is calculated as the total N minus the number of groups (see Dansereau, et al., 1984) which is 7,382 minus 99 or 7,283.

```
> cordif(rvalue1=-.1107, rvalue2=-.7122, n1=7283, n2=99)
$"z value"
[1] 7.597172
```

The z-value is larger than 1.96, so we conclude that the two correlations are significantly different for each other. This finding mirrors what we found in our contextual analysis but with an appropriate z-value.

3.4.2 Random Group Resampling of Covariance Theorem (`rgr.waba`)

As noted above, it may be interesting to see how the eta-between, eta-within, between group and within-group correlations vary as a function of the group-level properties of the data. The `rgr.waba` function provides a way to examine the group-level properties of elements of the covariance theorem. Essentially, the `rgr.waba` function allows one to answer questions such as "is the eta-between value for x larger than would be expected by chance?". The `rgr.waba` function randomly assigns individuals into pseudo groups having the exact size characteristics as the actual groups, and then calculates the covariance theorem parameters. By repeatedly assigning individuals to pseudo groups and re-estimating the covariance theorem components, one can create sampling distributions of the covariance theorem components to see if actual group results differ from pseudo group results (see Bliese & Halverson, 2002). Below I illustrate the use of `rgr.waba`. Note that this is a very computationally intensive routine, so it may take some time to complete.

```
> with(bh1996, waba(HRS,WBEING,GRP))$Cov.Theorem
     RawCorr     EtaBx     EtaBy      CorrB     EtaWx     EtaWy      CorrW
1 -0.1632064 0.3787881 0.2359287 -0.7121729 0.9254834 0.9717704 -0.1107031
> RGR.WABA<-rgr.waba(bh1996$HRS,bh1996$WBEING,bh1996$GRP,1000)
> round(summary(RGR.WABA),dig=4)
       RawCorr     EtaBx     EtaBy     CorrB     EtaWx     EtaWy     CorrW
NRep 1000.0000 1000.0000 1000.0000 1000.0000 1000.0000 1000.0000 1000.0000
Mean   -0.1632    0.1154    0.1151   -0.1614    0.9933    0.9933   -0.1632
SD      0.0000    0.0082    0.0081    0.0961    0.0010    0.0009    0.0013
```

The `summary` of the `rgr.waba` object produces a table giving the number of random repetitions, the means and the standard deviations from analysis. Notice that when there are no meaningful group differences, the between-group correlation, the raw correlation, and the within-group correlation all have the same value (with some rounding error). The raw correlation has a

standard deviation of zero because it does not change. In contrast, the between-group correlation has the highest standard deviation (.096) indicating that it varied the most across the pseudo group runs. All of covariance theorem components in the actual groups significantly vary from their counterparts in the pseudo group analysis because most actual group values are more than two standard deviations different from the pseudo group means.

To further test for significant differences, we can examine the sampling distribution of the random runs, and use the 2.5% and 97.5% sorted values to approximate 95% confidence intervals. Any values outside of this range would be considered significantly different from their pseudo group counterparts.

```
> quantile(RGR.WABA,c(.025,.975))
           EtaBx      EtaBy       CorrB     EtaWx     EtaWy      CorrW
2.5%  0.09944367 0.09916248 -0.34021577 0.9913014 0.9914049 -0.1658118
97.5% 0.13161137 0.13082964  0.03106165 0.9950432 0.9950713 -0.1607501
```

All of the covariance theorem values based on the actual groups are outside of the 95% confidence interval estimates. In other words, all the actual group results are significantly different than would be expected if individuals had been randomly assigned to groups ($p<.05$). The 99% confidence intervals draw the same conclusion at a more stringent confidence level.

```
> quantile(RGR.WABA,c(.005,.995))
           EtaBx      EtaBy      CorrB     EtaWx     EtaWy      CorrW
0.5%  0.09307571 0.09416619 -0.4065661 0.9907133 0.9908819 -0.1666120
99.5% 0.13596781 0.13473339  0.1049678 0.9956590 0.9955565 -0.1596676
```

Keep in mind that a replication is likely to differ slightly from results presented here because we did not start by setting a random seed.

3.5 Simulate Multilevel Correlations (`sim.mlcor`)

Contextual effects where relationships significantly differ across levels such as the illustration involving work hours and well-being are common. In many cases, the effects are less dramatic than having a within-group correlation of -.11 and a between-group correlation of -.71, but contextual effects exist and what drives them is relatively unexplored. One necessary, but not sufficient, condition for observing contextual effects is that both variables must have non-zero ICC(1) values (see Bliese, 1998). For this reason, researchers who are focused on modeling shared properties of constructs such as safety climate, cohesion, or team emotional cultures need to develop measures that have good ICC1 values and differentiate groups (see Bliese, Maltarich, Hendricks, Hofmann & Adler, 2019).

The `sim.mlcor` (simulate multilevel correlation) function was designed to help explore how measurement properties at different levels impact observed raw, within, and between-group correlations. We could examine, for example, how correlations would have differed if we had been able to increase the ICC(1) values or alpha values of the variables.

In the function, users provide group size, the number of groups, a between-group correlation, a within-group correlation, an ICC(1) for x, an ICC(1) for y, and alpha values for both x and y. The function returns a simulated dataset.

We can create a simulated dataset for our running example involving work hours and well-being by first obtaining the values from the actual data:

```
> data(bh1996)
> with(bh1996,waba(HRS,WBEING,GRP))
$Cov.Theorem
     RawCorr     EtaBx     EtaBy      CorrB     EtaWx     EtaWy      CorrW
1 -0.1632064 0.3787881 0.2359287 -0.7121729 0.9254834 0.9717704 -0.1107031

$n.obs
[1] 7382

$n.grps
[1] 99

> mult.icc(bh1996[,c("HRS","WBEING")],bh1996$GRP)
  Variable       ICC1      ICC2
1      HRS 0.12923699 0.9171286
2   WBEING 0.04337922 0.7717561
```

In this case, the group-level correlation of -.71 is smaller than it would have been if group means had reliabilities of 1. Instead, the ICC(2) values show that the group-mean reliability for work hours is .92 and for well-being the value is .77. We can correct the -.71 value by adjusting the incremental effect (the difference between the within-group and between-group correlation) for attenuation using ICC(2) values and adding this effect back to the within-group correlation.

```
> (-0.7121729--0.1107031)/sqrt(0.9171286*0.7717561)+-0.1107031
[1] -0.8256251
```

From this correction we can assume that if the ICC(2) values for both variables had been 1, the group-mean correlation would have been -.826. Using these data in the simulation and assuming an average group sizes of 75 (7382/99) and alpha values of 1, we obtain the following simulated dataset with results that mirror our actual data. Here I set a seed so exact results can be replicated.

```
> set.seed(578323)
> SIM.ML.COR<-sim.mlcor(gsize=75,ngrp=99,gcor=-.8256,wcor=-.1107,
+                       icc1x=0.04338,icc1y=0.12924,alphax=1,alphay=1)

> with(SIM.ML.COR,waba(X,Y,GRP))
$Cov.Theorem
     RawCorr     EtaBx     EtaBy      CorrB     EtaWx     EtaWy      CorrW
1 -0.1699012 0.2317119 0.3804799 -0.7353652 0.9727844 0.9247892 -0.1167938

$n.obs
[1] 7425

$n.grps
[1] 99

> mult.icc(SIM.ML.COR[,c("X","Y")],SIM.ML.COR$GRP)
  Variable       ICC1      ICC2
```

```
1        X 0.04142764 0.7642263
2        Y 0.13448630 0.9209720
```

To see the implications of having had a zero ICC(1) for the one of the variables, we can rerun the simulation and show that the between-group correlation no longer differs from the within or raw. This result is entirely expected because a necessary condition for contextual effects is a non-zero ICC(1) on both variables.

```
> SIM.ML.COR<-sim.mlcor(gsize=75,ngrp=99,gcor=-.8256,wcor=-.1107,
+                       icc1x=0,icc1y=0.12924,alphax=1,alphay=1)

> with(SIM.ML.COR,waba(X,Y,GRP))$Cov.Theorem
     RawCorr     EtaBx     EtaBy      CorrB     EtaWx     EtaWy      CorrW
1 -0.1304409 0.1256461 0.3688716 -0.1483209 0.9920751 0.9294803 -0.1340036

> mult.icc(SIM.ML.COR[,c("X","Y")],SIM.ML.COR$GRP)
  Variable        ICC1      ICC2
1        X 0.002640832 0.1656842
2        Y 0.125605587 0.9150646
```

To see the implications of improved the group-level measurement properties of the well-being measure to better differentiate groups, we can increase the ICC(1) for X to be .10 which produces a between-group correlation of -.76 in this particular run. The raw correlation also inherits more from the group correlation and increases to -.19.

```
> SIM.ML.COR<-sim.mlcor(gsize=75,ngrp=99,gcor=-.8256,wcor=-.1107,
+             icc1x=.10,icc1y=0.12924,alphax=1,alphay=1)

> with(SIM.ML.COR,waba(X,Y,GRP))$Cov.Theorem
     RawCorr     EtaBx     EtaBy      CorrB     EtaWx     EtaWy      CorrW
1 -0.1947374 0.3607571 0.3796512 -0.7638527 0.9326598 0.9251297 -0.1044453

> mult.icc(SIM.ML.COR[,c("X","Y")],SIM.ML.COR$GRP)
  Variable      ICC1      ICC2
1        X 0.1195602 0.9105922
2        Y 0.1338431 0.9205681
```

Finally, to illustrate one of Bliese et al.'s (2019) main points that individual reliability indices such as alpha are largely irrelevant to the magnitude of between-group correlations, we can change the alpha for both X and Y to be .70. In this case, note that the within-correlation is now -.08 and would be adjusted back to -.11 if corrected for attenuation (`-.08/sqrt(.7*.7)`)

```
> SIM.ML.COR<-sim.mlcor(gsize=75,ngrp=99,gcor=-.8256,wcor=-.1107,
+             icc1x=0.04338,icc1y=0.12924,alphax=.7,alphay=.7)
> with(SIM.ML.COR,waba(X,Y,GRP))$Cov.Theorem
     RawCorr     EtaBx     EtaBy      CorrB     EtaWx     EtaWy       CorrW
1 -0.1356601 0.2287517 0.3741004 -0.6967766 0.9734848 0.9273882 -0.08421891
> mult.icc(SIM.ML.COR[,c("X","Y")],SIM.ML.COR$GRP)
  Variable       ICC1      ICC2
1        X 0.04003354 0.7577361
2        Y 0.12957194 0.9177936
```

For detailed examinations of measurement properties, the examples presented would need to be put within a Monte Carlo function and averaged across multiple iterations, but the `sim.mlcor` function provides a way to generate multilevel correlations.

4 Mixed-Effects Models for Multilevel Data

This section illustrates the use of mixed-effects models to analyze multilevel data using the `nlme` package (Pinhiero & Bates, 2000). Most of the examples described in this section are taken from Bliese (2002) and use the Bliese and Halverson (1996) data (`bh1996`). Model notation is based on Bryk and Raudenbush's (1992) and Raudenbush and Bryk (2002).

A complete description of mixed-effects modeling is beyond the scope of this document; nonetheless, a short overview is presented to help facilitate the illustration of the methods. For more detailed discussions see Bliese, (2002); Bliese, Maltarich and Hendricks, 2018; Bryk and Raudenbush, (1992); Hofmann, (1997); Hox (2002); Kreft and De Leeuw, (1998); Pinheiro and Bates (2000); Raudenbush and Bryk (2002) and Snidjers and Bosker (1999).

One can think of mixed-effects models as ordinary regression models that have additional variance terms for handling non-independence due to group membership. The key to mixed-effects models is to understand how nesting individuals within groups can produce additional sources of variance (non-independence) in data.

The first variance term that distinguishes a mixed-effects model from a regression model is a term that reflects the degree to which groups differ in their mean values (intercepts) on the dependent variable (DV). A significant variance term (τ_{00}) indicates that groups significantly differ in terms of the DV and further suggests that it may be useful to include group-level variables as predictors. Group-level variables (or level-2 variables) differ across groups but are consistent for members within the same groups. For example, group average work hours are the same across all members of the same group and represents a level-2 variable that could potentially be used to predict group-level variance (τ_{00}) in well-being.

The second variance term that distinguishes a mixed-effects model from typical regression reflects the degree to which slopes between independent and dependent variables vary across groups (τ_{11}). Single-level regression models generally assume that the relationship between the IV and DV is constant across groups. In contrast, mixed-effects models permit testing whether the slope varies among groups. If slopes significantly vary, we can explain the variation by including a cross-level interaction using a level-2 variable such as average group work hours to explain why the slope between IV and DV in some groups is stronger than the slopes in other groups.

A third variance term is common to both mixed-effects models and regression models. This variance term, σ^2, reflects the degree to which an individual score differs from its predicted value within a specific group. σ^2 represents the within-group variance and is predicted individual-level or level-1 variables. Level-1 variables differ among members of the same group. For instance, a level-1 variable such as participant age would vary among members of the same group.

In summary, in a complete mixed-effect model analysis, one examines (1) level-1 factors related to the within-group variance σ^2; (2) group-level factors related to the between-group variation in intercepts τ_{00}; and (3) group-level factors related to within-group slope differences,

τ_{11}. The next sections re-analyze portions of the Bliese and Halverson (1996) data set to illustrate a typical sequence of steps used in multilevel modeling.

4.1 Steps in multilevel modeling

4.1.1 Step 1: Examine the ICC for the Outcome

Because multilevel modeling involves predicting variance at different levels, it is important to begin by determining the levels where significant variation exists. In the case of the two-level model (the only models considered here) we can assume there is significant variation in the within-group variance, σ^2. We do not necessarily assume there will be significant intercept variation (τ_{00}) or between-group slope variation (τ_{11}) so modeling often begins with variance decomposition of intercept variance (see Bryk & Raudenbush, 1992; Hofmann, 1997). If τ_{00} does not differ by more than chance levels, there may be little reason to use mixed-effects models as simpler OLS models will suffice (though see Bliese et al., 2018 who argue that there is virtually no downside to estimating mixed-effect models even when If τ_{00} is small or non-significant because in these cases the mixed-effect models just return the OLS estimates). Note that if slopes randomly vary (τ_{11}) even if intercepts (τ_{00}) do not, there may still be reason to estimate mixed-effects models (see Snijders & Bosker, 1999).

In Step 1, we first examine the group-level properties of the outcome variable to estimate the ICC(1) (commonly referred to simply as the ICC in mixed-effect models). Second, we determine whether the variance of the intercept (τ_{00}) is significantly larger than zero.

These two aspects of the outcome variable are examined by estimating an unconditional means or null model. An unconditional means model does not contain any predictors but includes a random intercept variance term for groups. The model estimates how much variability there is in mean Y values (i.e., how much variability there is in the intercept) relative to the total variability. In the two stage HLM notation, the model is:

$$Y_{ij} = \beta_{0j} + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$

In combined form, the model is: $Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}$. The null model states that the dependent variable is a function of a common intercept γ_{00}, and two error terms: the between-group error term, u_{0j}, and the within-group error term, r_{ij}. The model essentially states that any Y value can be described in terms of an overall mean plus some error associated with group membership and some individual error. A summary of the variance components of the null model provides two estimates of variance; τ_{00} associated with u_{0j} reflecting the variance in how much each groups' intercept varies from the overall intercept (γ_{00}), and σ^2 associated with r_{ij} reflecting how much each individual's score differs from the group mean. Bryk and Raudenbush (1992) note that the null model is directly equivalent to a one-way random effects ANOVA – an ANOVA model where one predicts the dependent variable as a function of group membership.

We estimate the unconditional means model and other mixed-effects models using the `lme` (for linear mixed effects) function in the `nlme` package (see Pinheiro & Bates, 2000). There are two formulas that must be specified in any `lme` call: a fixed effects formula and a random effects formula.

In the unconditional means model, the fixed portion of the model is γ_{00} (an intercept term) and the random component is $u_{0j}+r_{ij}$. The random portion of the model states that intercepts can vary among groups. We begin the analysis by attaching the `multilevel` package (which also loads the `nlme` package) and making the `bh1996` data set in the `multilevel` package available for analysis.

```
> library(multilevel)
> data(bh1996)
> Null.Model<-lme(WBEING~1,random=~1|GRP,data=bh1996,
  control=list(opt="optim"))
```

In the model, the fixed formula is `WBEING~1` indicating that the only predictor of well-being is an intercept term. The model assumes that in the absence of any predictors, the best estimate of any specific outcome value is the mean value on the outcome. The random formula is `random=~1|GRP` which specifies that the intercept can vary as a function of group membership. A random intercept model is the most basic random formula, and in many situations a random intercept model may be all that is required to adequately account for the nested nature of the grouped data. The option `control=list(opt="optim")` in the call to `lme` instructs the program to use R's general purpose optimization routine. Versions of `lme` after 2.2 default to `nlmimb` which has several advantages including better diagnostics when optimization fails. In practice, however, `nlmimb` tends to converge less often than the general purpose optimizater. Furthermore, the examples in this document were estimated under `"optim"`, so for consistency we will revert back to the original optimizer. In practice, users likely want to use the default `"nlmimb"` optimizer; however, if models fail to converge it may be useful to revert to `"optim"`.

Estimating ICC. The unconditional means model provides between-group and within-group variance estimates in the form of τ_{00} and σ^2, respectively. The formula for the ICC is $\tau_{00}/(\tau_{00} + \sigma^2)$ (see, Bryk & Raudenbush, 1992; Kreft & De Leeuw, 1998). Bliese (2000) notes that the ICC is equivalent to Bartko's ICC(1) formula (Bartko, 1976) and to Shrout and Fleiss's ICC(1,1) formula (Shrout & Fleiss, 1979). The `VarCorr` function provides estimates of variance for an `lme` object.

```
> VarCorr(Null.Model)
GRP = pdSymm(1)
            Variance   StdDev
(Intercept) 0.03580079 0.1892110
Residual    0.78949727 0.8885366
> 0.03580079/(0.03580079+0.78949727) #Calculate ICC
[1] 0.04337922
```

The estimate of τ_{00} (between-group or Intercept variance) is 0.036, and the estimate of σ^2 (within-group or residual variancel) is 0.789. The ICC estimate ($\tau_{00}/(\tau_{00} + \sigma^2)$) is .04.

To verify that the ICC results from the mixed-effects models are similar to those from an ANOVA model and the `ICC1` function (see section 0) we can perform an ANOVA analysis on the same data.

```
> tmod<-aov(WBEING~as.factor(GRP),data=bh1996)
> ICC1(tmod)
```

```
[1] 0.04336905
```

The ICC value from the mixed-effects model and the ICC(1) from the ANOVA model are similar although they will tend to differ if group sizes vary dramatically given that the ANOVA models assume equal group sizes.

Determining whether τ_{00} is significant. Returning to our original analysis involving well-being from the `bh1996` data set, we would likely be interested in knowing whether the intercept variance (i.e.,τ_{00}) estimate of 0.036 is significantly different from zero. In mixed-effects models, we perform this test by comparing –2 log likelihood values between (1) a model with a random intercept, and (2) a model without a random intercept.

A model without a random intercept can be estimated using the `gls` function in the `nlme` package. The –2 log likelihood values (i.e., Deviance) for an `lme` or `gls` object are obtained using the `logLik` function and multiplying the returned value by –2. If the –2 log likelihood value for the model with the random intercept is significantly smaller than the model without the random intercept (based on a Chi-square distribution), then we conclude that the model with the random intercept fits the data significantly "better" than does the model without the random intercept. In the R, model contrasts are conducted using the `anova` function.

```
> Null.gls<-gls(WBEING~1,data=bh1996,
  control=list(opt="optim"))

> logLik(Null.gls)*-2
`log Lik.' 19536.17 (df=2)

> logLik(Null.Model)*-2
`log Lik.' 19347.34 (df=3)

> 19536.17-19347.34
[1] 188.83

> anova(Null.gls, Null.Model)
           Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Null.gls       1  2 19540.17 19553.98 -9768.084
Null.Model     2  3 19353.34 19374.06 -9673.669 1 vs 2 188.8303  <.0001
```

The –2 log likelihood value for the `gls` model without the random intercept is 19536.17. The difference of 188.8 is significant on a Chi-Squared distribution with one degree of freedom (one model estimated a variance term associated with a random intercept, the other did not, and this results in the one df difference). These results indicate significant intercept variation.

In summary, we would conclude that there is significant intercept variation in terms of general well-being scores across the 99 Army companies in our sample. We also estimate that 4% of the variation in individuals' well-being score is a function of the group to which he or she belongs. Thus, a model that allows for random variation in well-being among Army companies is a better fit than a model that does not allow for this random variation.

4.1.2 Step 2: Explain Level 1 and 2 Intercept Variance

At this point, we have two sources of variation that we can attempt to explain in subsequent modeling – within-group variation (σ^2) and between-group intercept (i.e., mean) variation (τ_{00}).

In many cases, these may be the only two sources of variation we are interested in explaining so let us begin by building a model that predicts these two sources of variation.

In our running example, we assume that individual well-being is negatively related to individual reports of work hours. At the same time, however, we assume that average work hours in an Army Company are related to the average well-being of the Company over-and-above the individual-level work hours and well-being relationship. Using Hofmann and Gavin's (1998) terminology, we are testing an incremental model where the level-2 variable predicts unique variance after controlling for level-1 variables. Our model is directly equivalent to the contextual model that we estimated in section 3.3.1 but we now use mixed-effect models rather than OLS regression.

The form of the model using Bryk and Raudenbush's (1992) notation is:

$$WBEING_{ij} = \beta_{0j} + \beta_{1j}(HRS_{ij}) + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}(G.HRS_j) + u_{0j}$$
$$\beta_{1j} = \gamma_{10}$$

The first line indicates that individual well-being is a function of the groups' intercept plus a component that reflects the linear effect of individual reports of work hours plus some random error. The second line indicates that each groups' intercept (mean) is a function of some common intercept (γ_{00}) plus a component that reflects the linear effect of average group work hours plus some random between-group error. The third line states that the slope between individual work hours and well-being is fixed—it is not allowed to randomly vary across groups. Stated another way, we assume that the relationship between work hours and well-being varies by no more than chance levels among groups.

When we combine the three rows into a single equation, we get an equation that looks like a common regression equation with an extra error term (u_{0j}). This error term indicates that WBEING intercepts (i.e., means) can randomly differ across groups. The combined model is:

$$WBEING_{ij} = \gamma_{00} + \gamma_{10}(HRS_{ij}) + \gamma_{01}(G.HRS_j) + u_{0j} + r_{ij}$$

This model is specified in `lme` as:

```
> Model.1<-lme(WBEING~HRS+G.HRS,random=~1|GRP,data=bh1996,
  control=list(opt="optim"))

> summary(Model.1)
Linear mixed-effects model fit by REML
 Data: bh1996
       AIC      BIC    logLik
  19222.28 19256.81 -9606.14

Random effects:
 Formula: ~1 | GRP
        (Intercept)  Residual
StdDev:   0.1163900 0.8832353

Fixed effects: WBEING ~ HRS + G.HRS
                Value  Std.Error   DF   t-value p-value
(Intercept)  4.740829 0.21368746 7282 22.185808  <.0001
HRS         -0.046461 0.00488798 7282 -9.505056  <.0001
```

```
G.HRS       -0.126926 0.01940357   97 -6.541368  <.0001
 Correlation:
      (Intr) HRS
HRS    0.000
G.HRS -0.965 -0.252

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-3.35320562 -0.65024982  0.03760797  0.71319835  2.70917777

Number of Observations: 7382
Number of Groups: 99
```

Notice that work hours are significantly negatively related to individual well-being. Furthermore, after controlling the individual-level relationship, average work hours (G.HRS) are related to the average well-being in a group. The interpretation of this model, like the interpretation of the contextual effect model (section 3.3.1) indicates that the slope at the group-level significantly differs from the slope at the individual level. Indeed, in this example, each hour increase at the group level is associated with a -.163 (-.046+-.127) decrease in average well-being. The coefficient of -.127 reflects the degree of difference between the two slopes. Importantly, in the mixed-effect model, the t-value for G.HRS is -6.54 whereas in the OLS model the t-value was upwardly biased at -10.06.

In this basic model, we can also estimate how much of the variance was explained by these two predictors. Because individual work hours were significantly related to well-being, we expect that it will have "explained" some of the within-group variance σ^2. Similarly, since average work hours were related to the group well-being intercept we expect that it will have "explained" some of intercept variance, τ_{00}. Recall that in the null model, the variance estimate for the within-group residuals, σ^2, was 0.789; and the variance estimate for the intercept, τ_{00}, was 0.036. The `VarCorr` function on the `Model.1` object reveals that each variance component has changed slightly.

```
> VarCorr(Model.1)
GRP = pdSymm(1)
            Variance   StdDev
(Intercept) 0.01354663 0.1163900
Residual    0.78010466 0.8832353
```

Specifically, the variance estimates from the model with the two predictors are 0.780 and 0.014. That is, the variance of the within-group residuals decreased from 0.789 to 0.780 and the variance of the between-group intercepts decreased from 0.036 to 0.014. We can calculate the percent of variance explained by using the following formula:

Variance Explained = 1 – (Var with Predictor/Var without Predictor)

To follow through with our example, work hours explained 1 – (0.780/0.789) or 0.011 (1%) of the within-group variance in σ^2, and group-mean work hours explained 1 – (0.014/0.036) or 0.611 (61%) of the between-group intercept variance τ_{00}. While the logic behind variance estimates appears straightforward (at least in models without random slopes), the variance estimates should be treated with some degree of caution because they are partially dependent

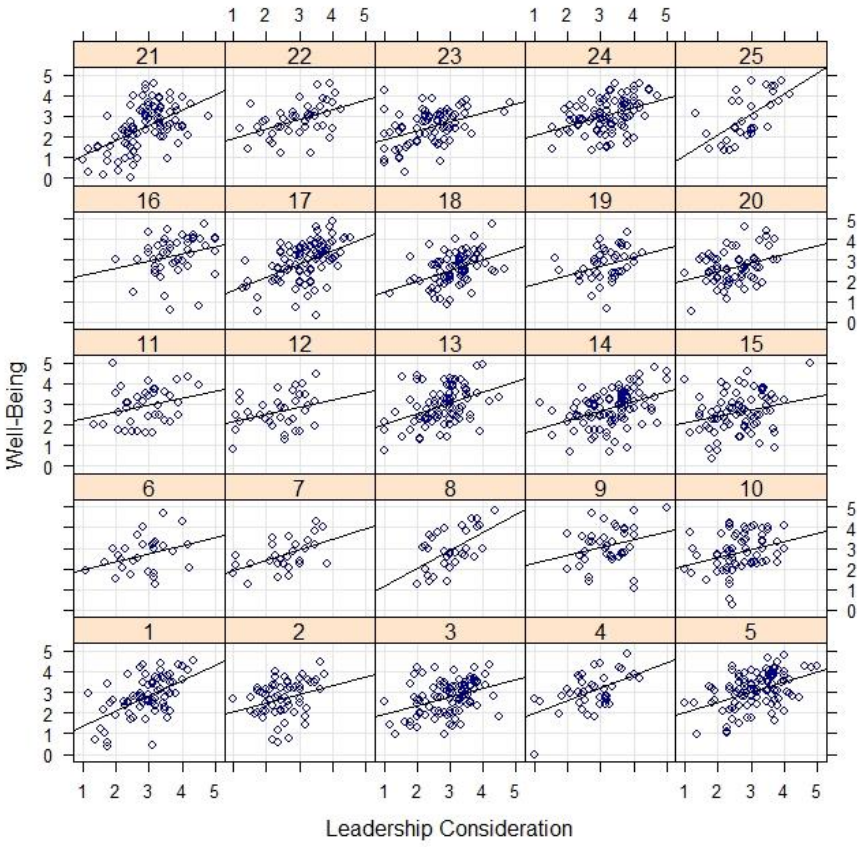
upon how one specifies the models. Interested readers are directed to Snijders and Bosker (1994; 1999) for an in-depth discussion of variance estimates.

4.1.3 Step 3: Examine and Predict Slope Variance

Let us continue our analysis by trying to explain the third source of variation, namely, variation in our slopes (τ_{11}, τ_{12}, etc.). To do this, we examine another variable from `bh1996`. This variable represents Army Company members' ratings of leadership consideration (LEAD). Generally, individual soldiers' ratings of leadership are related to well-being. In this analysis, however, we will consider the possibility that the strength of the relationship between individual ratings of leadership consideration and well-being varies among groups.

We begin by examining slope variation among the first 25 groups using `xyplot` from the `lattice` package.

```
> library(lattice)
> xyplot(WBEING~LEAD|as.factor(GRP),data=bh1996[1:1582,],
  type=c("p","g","r"),col="dark blue",col.line="black",
  xlab="Leadership Consideration",
  ylab="Well-Being")
```



From the plot of the first 25 groups in the `bh1996` data set, it seems likely that there is some slope variation. The plot, however, does not tell us whether this variation is significant. We begin our analysis of slope variability by adding leadership consideration to our model and testing whether there is significant variation in the leadership consideration and well-being slopes across groups. Our base model is:

$$
\begin{aligned}
WBEING_{ij} = \ & \beta_{0j} + \beta_{1j}(HRS_{ij}) + \beta_{2j}(LEAD_{ij}) + r_{ij} \\
& \beta_{0j} = \gamma_{00} + \gamma_{01}(G.HRS_j) + u_{0j} \\
& \beta_{1j} = \gamma_{10} \\
& \beta_{2j} = \gamma_{20}
\end{aligned}
$$

The last two lines include that neither the slope for HRS or LEAD is allowed to vary across groups. In combined form the model is:

$WBEING_{ij} = \gamma_{00} + \gamma_{10}(HRS_{ij}) + \gamma_{20}(LEAD_{ij}) + \gamma_{01}(G.HRS_j) + u_{0j} + r_{ij}$.

The model specification in `lme` is:

```
> Model.2<-lme(WBEING~HRS+LEAD+G.HRS,random=~1|GRP, data=bh1996,
+                control=list(opt="optim"))

> round(summary(Model.2)$tTable,digit=3)
             Value Std.Error   DF t-value p-value
(Intercept)  2.559     0.216 7281  11.859       0
HRS         -0.028     0.004 7281  -6.317       0
LEAD         0.496     0.013 7281  38.786       0
G.HRS       -0.079     0.019   97  -4.185       0

> VarCorr(Model.2)
GRP = pdLogChol(1)
            Variance   StdDev
(Intercept) 0.01418026 0.1190809
Residual    0.64704412 0.8043905
```

Across the sample, individuals' perceptions of leadership have a strong, positive relationship to their well-being. To determine whether the strength of this relationship differs across groups, we need to estimate a model with a random slope for LEAD. This alternative model is:

$$
\begin{aligned}
WBEING_{ij} = \ & \beta_{0j} + \beta_{1j}(HRS_{ij}) + \beta_{2j}(LEAD_{ij}) + r_{ij} \\
& \beta_{0j} = \gamma_{00} + \gamma_{01}(G.HRS_j) + u_{0j} \\
& \beta_{1j} = \gamma_{10} \\
& \beta_{2j} = \gamma_{20} + u_{2j}
\end{aligned}
$$

The last line indicates that the slope between leadership consideration and well-being for any specific group is a function of a common slope γ_{20} and a group-specific error term u_{2j}. The variance term associated with u_{2j} is τ_{12}. In this model, we have not permitted the slope between individual work hours and individual well-being to vary across groups.

In combined form the model is:

$$WBEING_{ij} = \gamma_{00} + \gamma_{10}(HRS_{ij}) + \gamma_{20}(LEAD_{ij}) + \gamma_{01}(G.HRS_j) + u_{0j} + u_{2j} * LEAD_{ij} + r_{ij}.$$

The model specification in `lme` and the relevant changes to the variance components are:

```
> Model.2a<-lme(WBEING~HRS+LEAD+G.HRS,random=~LEAD|GRP, data=bh1996,
+               control=list(opt="optim"))

> VarCorr(Model.2a)
GRP = pdLogChol(LEAD)
            Variance   StdDev    Corr
(Intercept) 0.14401197 0.3794891 (Intr)
LEAD        0.01044352 0.1021935 -0.97
Residual    0.64129330 0.8008079
```

Changing the random component to `(random=~LEAD|GRP)` produces an estimate of the slope variance, τ_{12}, (.01) and an estimate of the correlation between the intercept and slope (-.97). To test whether this model provides significantly better fit, we test the –2 log likelihood ratios between a model with and a model without a random slope for leadership consideration and well-being.

```
> anova(Model.2,Model.2a)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
Model.2      1  6 17862.68 17904.12 -8925.341
Model.2a     2  8 17838.58 17893.83 -8911.290 1 vs 2 28.10254  <.0001
```

This comparison test is known to be conservative and we could halve the p-value (LaHuis & Ferguson, 2009), but even so the difference of 28.10 is significant on two degrees of freedom. The -2 log likelihood results indicate the model with the random effect for the leadership consideration and well-being slope provides a significantly better fit than the model without this random effect implying that the strength of the slope differs across groups.

Another way to consider the differences between the two models is to examine the empirical Bayes' estimates for each group. The values for the first five groups with the random intercept model are:

```
> coef(Model.2)[1:5,]
  (Intercept)         HRS      LEAD       G.HRS
1    2.534036 -0.02827849 0.4956385 -0.07900961
2    2.694639 -0.02827849 0.4956385 -0.07900961
3    2.458733 -0.02827849 0.4956385 -0.07900961
4    2.764899 -0.02827849 0.4956385 -0.07900961
5    2.616261 -0.02827849 0.4956385 -0.07900961
```

In this specification, group 4 has the highest level of well-being, and group 3 has the lowest, but these intercept (mean) differences are the only model parameters varying across groups. The slopes match the values from the summary of the t-table presented previously. In contrast, the empirical Bayes' estimates for model with the random slope are:

```
> coef(Model.2a)[1:5,]
  (Intercept)         HRS      LEAD       G.HRS
1    2.195403 -0.02847764 0.5715939 -0.07050472
2    2.839074 -0.02847764 0.4071772 -0.07050472
3    2.398461 -0.02847764 0.4910177 -0.07050472
4    2.846874 -0.02847764 0.4247142 -0.07050472
5    2.608235 -0.02847764 0.4679652 -0.07050472
```

In this specification, the slope indicated the strength of the relationship between individuals' perceptions of leadership consideration and their well-being also varies by group. In group 1, the relationship between the two variables is stronger (.57) than in group 2 (.41).

Given significant variation in the leadership and well-being slope, we can attempt to see what group-level properties are related to this variation. We propose that when groups are under a lot of strain from work requirements, the relationship between leadership consideration and well-being will be relatively strong. In contrast, when groups are under little strain, we expect a relatively weak relationship between leadership consideration and well-being. Our proposition represents a contextual effect in an occupational stress model (see Bliese & Jex, 2002).

Our proposition represents a cross-level interaction where the slope between individuals' perceptions of leadership consideration and their ratings of well-being varies as a function of the level-2 variable of group work demands. In mixed-effects models, we test this hypothesis by examining whether a level-2 variable explains a significant amount of the level-1 slope variation among groups. In our example, we test whether average work hours in the group "explains" group-by-group variation in the relationship between individual perceptions of leadership consideration and individual reports of well-being. In Bryk and Raudenbush's (1992) notation, the model that we are testing is:

$$
\begin{aligned}
WBEING_{ij} &= \beta_{0j} + \beta_{1j}(HRS_{ij}) + \beta_{2j}(LEAD_{ij}) + r_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}(G.HRS_j) + u_{0j} \\
\beta_{1j} &= \gamma_{10} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}(G.HRS_j) + u_{2j}
\end{aligned}
$$

In combined form the model is:

$$WBEING_{ij} = \gamma_{00} + \gamma_{10}(HRS_{ij}) + \gamma_{20}(LEAD_{ij}) + \gamma_{01}(G.HRS_j) + \gamma_{21}(LEAD_{ij} * G.HRS_j) + u_{0j} + u_{2j} * LEAD_{ij} + r_{ij}.$$

In `lme`, we specify the cross-level interaction by adding an interaction term between leadership (LEAD) and average group work hours (G.HRS). Specifically, the model is:

```
> Final.Model<-lme(WBEING~HRS+LEAD+G.HRS+LEAD:G.HRS,
random=~LEAD|GRP,data=bh1996,control=list(opt="optim"))

> round(summary(Final.Model)$tTable,dig=3)
             Value Std.Error   DF t-value p-value
(Intercept)  3.654     0.726 7280   5.032   0.000
HRS         -0.029     0.004 7280  -6.391   0.000
LEAD         0.126     0.217 7280   0.578   0.564
G.HRS       -0.175     0.064   97  -2.751   0.007
```

```
LEAD:G.HRS   0.032     0.019 7280   1.703   0.089
```

The `tTable` results from the final model indicate there is a significant cross-level interaction (the last row using a liberal p-value of less than .10). This result indicates that average work hours "explained" a significant portion of the variation in τ_{12} – the vertical cohesion and well-being slope.

4.1.4 Step 3 using the lme4 Package and Interaction Plot

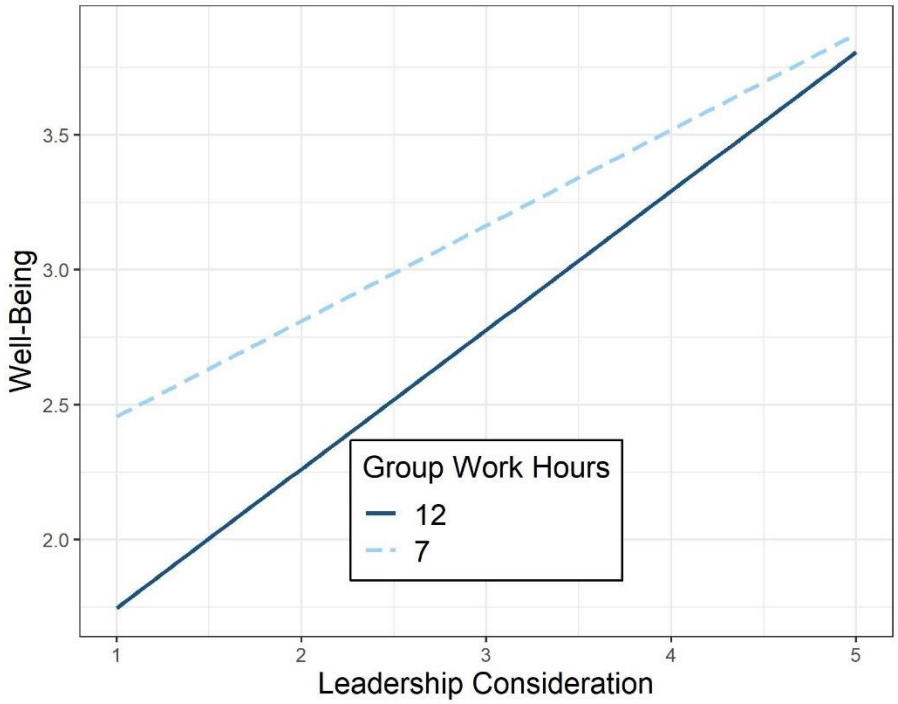
To plot the form of the interaction and make use of the graphics capabilities of `ggplot2`, we can use the lme4 package and rerun the model using `lmer`. The code also uses the lmerTest package for p-values and degrees of freedom and changes the optimizer because the default failed to converge.

```
> library(lme4)
> library(lmerTest)

> Model.2b<-lmer(WBEING~HRS+LEAD*G.HRS+(LEAD|GRP), data=bh1996,
+                control=lmerControl(optimizer = "Nelder_Mead"))

> summary(Model.2b)$coef
               Estimate  Std. Error         df   t value     Pr(>|t|)
(Intercept)  3.64325839 0.732553188   87.67621  4.973370 3.243398e-06
HRS         -0.02855876 0.004468026 7287.99657 -6.391807 1.740410e-10
LEAD         0.12894421 0.218811339   89.83115  0.589294 5.571432e-01
G.HRS       -0.17401949 0.064152902   87.45942 -2.712574 8.038535e-03
LEAD:G.HRS   0.03216543 0.019187381   89.79663  1.676384 9.714129e-02
```

With a lmer model, we can use the interactions library and the following code to plot values for group averages of 7 hours versus 12 hours of work.

```
library(interactions)
library(ggplot2)
win.graph(height=4.75,width=6) #quartz() for MAC

interact_plot(Model.2b,pred=LEAD,modx=G.HRS,
              modx.values = c(7,12),
              x.label = "Leadership Consideration",
              y.label = "Well-Being",
              legend.main="Group Work Hours")+
  theme_bw()+
  theme(legend.background=element_rect(fill="white",
                  size=0.5, linetype="solid",color ="black"),
        legend.position = c(0.5, 0.2),
        axis.title.x = element_text(color="black", size=14),
        axis.title.y = element_text(color="black", size=14),
        legend.title = element_text(color="black", size=14),
        legend.text = element_text(color="black", size=14)
  )

ggsave(filename = "c:\\temp\\plotgg.jpg",
```

```
device = "jpeg")
```

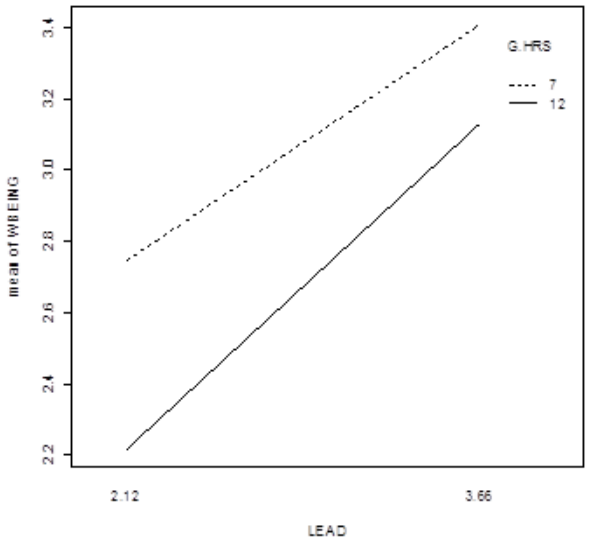


Soldiers' perceptions of leadership consideration are positively related to their well-being regardless of the group average work hours. The relationship between individuals' ratings of leadership consideration and their well-being is stronger (steeper slope) in groups with high work hours than in groups with low work hours. Another way to think about the interaction is to note that well-being really drops (in relative terms) when a soldier perceives that leadership is low in consideration and one is a member of a group with high average work hours. This pattern supports our proposition that considerate leadership is relatively more important in a high work demand context.

4.2 Plotting with `interaction.plot`

The previous example used the lme4, interactions, and ggplot2 library to make a publication quality plot. A quick alternative is to use the `interaction.plot` function illustrated below.

```
> Final.Model<-lme(WBEING~HRS+LEAD+G.HRS+LEAD:G.HRS,
  random=~LEAD|GRP,data=bh1996,control=list(opt="optim"))

> TDAT<-data.frame(HRS=c(11.2987,11.2987,11.2987,11.2987),
                   LEAD=c(2.12,2.12,3.66,3.66),
                   G.HRS=c(7, 12, 7, 12),
                   GRP=c(1,1,1,1))

> TDAT$WBEING<-predict(Final.Model,TDAT,level=1)
> with(TDAT,interaction.plot(LEAD,G.HRS,WBEING))
```



4.3 Some Notes on Centering

In multilevel modeling, centering issues is a major consideration. In our examples, we have used raw variables as predictors. In some cases, though, there may be good reasons to consider centering the level-1 variables with one of two centering options.

Level-1 variables such as leadership can be grand-mean centered or group-mean centered. Grand-mean centering is often worth considering because doing so helps reduce multicollinearity among predictors and random effect terms. In cases where interactive terms are included, grand-mean centering can be particularly helpful in reducing correlations between main-effect and interactive terms. Hofmann and Gavin (1998) and others have shown that grand-mean centered and raw variable models produce identical results for the predictors; however, grand-mean centered models may converge in situations where a model based on raw variables will not.

Grand-mean centering can be accomplished in one of two ways. The explicit way is to subtract the overall mean from the raw variable. The less obvious way is to use the `scale` function. The `scale` function is used to standardize (mean=0, sd=1) variables, but can also be used to grand-mean center if the `scale=FALSE` option is selected. Below I create grand-mean centered variables for leadership both ways.

```
> bh1996$GRAND.CENT.LEAD<-bh1996$LEAD-mean(bh1996$LEAD)
> bh1996$GRAND.CENT.LEAD<-scale(bh1996$LEAD,scale=FALSE)
```

Group-mean centering (demeaning) is another centering option with level-1 variables. In group-mean centering, each individual score is subtracted from the group mean. Review section 3.1 and the `aggregate` and `merge` functions for assigning a group-mean variable back to each individual. Once a group mean is assigned back to the individual, simply subtract the group mean from the raw score. A group-mean centered variable reflects how much an individual differs from their group average. Group-mean centering represents a different parameterization of the model than does the raw or grand-mean centered version (Hofmann & Gavin, 1998; Hox, 2002; Snijders & Bosker, 1999).

4.3.1.1 Centering and Cross-Level Interactions

There is value in using group-mean centering when testing a cross-level interaction. Bryk and Raudenbush (1992) and Hofmann and Gavin (1998) point out that group-mean centering provides the "purest" estimate of the within-group slope in these situations. That is, slope estimates based on raw variables and grand-mean centered variables can be partially influenced by between-group factors. In contrast, group-mean centered variables have had between-group effects removed. Bryk and Raudenbush (1992) show that group-level interactions can sometimes pose as cross-level interactions, so a logical strategy is to use raw or grand-mean centered variables to test for cross-level interactions but verify the final results with group-mean centered variables.

The `bh1996` dataframe has group-mean centered variables for all the predictors beginning with a "W" for "within". For comparisons, the first model uses a raw leadership variable and the second model below uses the group-mean centered leadership variable in both the fixed part of the model and in the random statement.

```
> Final.Model<-lme(WBEING~HRS+LEAD+G.HRS+LEAD:G.HRS,
+                  random=~LEAD|GRP,data=bh1996, control=list(opt="optim"))
> round(summary(Final.Model)$tTable,dig=3)
             Value Std.Error   DF t-value p-value
(Intercept)  3.654     0.726 7280   5.032   0.000
HRS         -0.029     0.004 7280  -6.391   0.000
LEAD         0.126     0.217 7280   0.578   0.564
G.HRS       -0.175     0.064   97  -2.751   0.007
LEAD:G.HRS   0.032     0.019 7280   1.703   0.089

> Final.Model.R<-lme(WBEING~HRS+W.LEAD+G.HRS+W.LEAD:G.HRS,
+              random=~W.LEAD|GRP,data=bh1996, control=list(opt="optim"))
> round(summary(Final.Model.R)$tTable,dig=3)
              Value Std.Error   DF t-value p-value
(Intercept)   4.733     0.214 7280  22.080   0.000
HRS          -0.028     0.004 7280  -6.271   0.000
W.LEAD        0.055     0.223 7280   0.249   0.804
G.HRS        -0.145     0.019   97  -7.471   0.000
W.LEAD:G.HRS  0.040     0.020 7280   2.037   0.042
```

Notice that the cross-level interaction is now significant with a t-value of 2.037 versus 1.703 in the model with raw variable. Thus, there are some minor differences between the two model specifications, but it would appear there is a significant cross-level interaction ($p<.05$) in the pure specification. For an interesting example of trying to determine whether cohesion buffering effects are cross-level or group-mean interactions see Campbell-Sills et al., (2022).

4.3.1.2 Centering and Contextual Models

Centering choice also has important implications for interpreting contextual models. When contextual models are based on raw level-1 variables, the level-2 coefficient represents the difference between the two slopes. In contrast, when the level-1 variable is group-mean centered, the level-2 coefficient captures the total effect (the level-1 slope plus any difference) and tests whether this total effect is different from zero. Below are the two models.

```
> tmod.raw<-lme(WBEING~HRS+G.HRS,random=~1|GRP,bh1996)
> round(summary(tmod.raw)$tTable, dig=3)
             Value Std.Error   DF t-value p-value
(Intercept)  4.741     0.214 7282  22.187       0
HRS         -0.046     0.005 7282  -9.505       0
G.HRS       -0.127     0.019   97  -6.542       0
>
> tmod.cent<-lme(WBEING~W.HRS+G.HRS,random=~1|GRP,bh1996)
> round(summary(tmod.cent)$tTable, dig=3)
             Value Std.Error   DF t-value p-value
(Intercept)  4.741     0.214 7282  22.187       0
W.HRS       -0.046     0.005 7282  -9.505       0
G.HRS       -0.173     0.019   97  -9.234       0
```

The first model indicates that the G.HRS slope is -0.127 stronger than the within slope of -0.046. The model represents a relative test. The second model tests whether the total between-group slope of -0.173 differs from zero. It is relatively common for researchers to make errors when interpreting these two variants of the model (see Bliese et al., 2018).

4.4 Estimating Group-Mean Reliability (ICC2) with `gmeanrel`

In mixed-effects models, it is possible to obtain an estimate of the group-mean reliability analogous to the ICC(2) (see section 3.2.7). Group mean reliability estimates are a function of the ICC and group size (see Bliese, 2000; Bryk & Raudenbush, 1992), and the `gmeanrel` function from the multilevel package calculates the ICC, the group size, and the group mean reliability for each group.

The code below illustrates the `gmeanrel` function on the `bhr2000` data set to show how the results compare to results in section 3.2.7 where the ICC(1) estimate from the ANOVA model was 0.174 and the ICC(2) estimate was 0.920.

```
> Null.Model<-lme(HRS~1,random=~1|GRP,data=bhr2000,
  control=list(opt="optim"))

> GREL.DAT<-gmeanrel(Null.Model)
> names(GREL.DAT)
[1] "ICC"     "Group"   "GrpSize" "MeanRel"

> GREL.DAT$ICC  #ICC estimate
[1] 0.177544

> GREL.DAT$MeanRel[1:20] #First 20 Reliability Estimates
 [1] 0.9272005 0.9066657 0.9471382 0.8487743 0.9465280
 [6] 0.7754791 0.7953197 0.8192754 0.8699945 0.8831157
[11] 0.8119385 0.8622636 0.9379303 0.9452644 0.9260382
[16] 0.8487743 0.9395503 0.9315061 0.8622636 0.9235985


> mean(GREL.DAT$MeanRel)
[1] 0.8955047
```

The ICC estimate is 0.178 (the same as the value produced by `mult.icc` in section 3.2.8) and slightly higher than the ANOVA based estimate of 0.174. The average group-mean

reliability from `gmeanrel` is 0.896 which is smaller (but close) to the value of 0.920 from the ANOVA model. The output also illustrates that each group receives a separate estimate of group-mean reliability. Values vary as a function of group size.

5 Growth Modeling Repeated Measures Data

Growth models are an important variation of multilevel models (see section 4). In growth models repeated observations from an individual represent the level-1 variables, and the attributes of the individual represent the level-2 variables. The fact that the level-1 variables are repeated over time poses some additional analytic issues; however, the steps used to analyze the basic growth model and the steps used to analyze a multilevel model share many key similarities.

This chapter begins by briefly reviewing some of the methodological challenges associated with growth modeling. Following this, the chapter illustrates how data must be configured to conduct growth modeling. Finally, the chapter illustrates a complete growth modeling analysis using the `nlme` package. Much of this material is adapted from Bliese and Ployhart (2002).

5.1 Methodological challenges

Since longitudinal data is collected from single entities over multiple times, it is likely that there will be a high degree of non-independence in the responses. Multiple responses from an individual will tend to be related by virtue of being provided by the same person, and this non-independence violates the statistical assumption of independence underlying many common data analytic techniques (Kenny & Judd, 1986).

Issues about non-independence are similar to those that occur when working with lower-level data nested in higher-level groups. In longitudinal designs, however, there are additional complications associated with the lower-level responses. First, it is likely that responses temporally close to each other (e.g., responses 1 and 2) will be more strongly related than responses temporally far apart (e.g., responses 1 and 4). This pattern is defined as a simplex pattern or lag 1 autocorrelation in the residuals. Second, it is likely that responses will tend to become either more variable over time or less variable over time. For instance, individuals starting jobs may initially have a low degree of variability in performance, but over time the variance in job performance may increase. In statistical terms, outcome variables in longitudinal data are likely to display heteroscedasticity. To obtain correct standard errors and to draw the correct statistical inferences, autocorrelation, and heteroscedasticity both need to be incorporated into the statistical model.

The need to test for both autocorrelation and heteroscedasticity in growth models arises because the level-1 variables (repeated measures from an individual) are ordered by time. One of the main differences between growth models and multilevel models revolves around understanding how to properly account for time in both the statistical models and in the data structures.

In R, growth modeling can be estimated using the `lme` function from the `nlme` package (Pinhiero & Bates, 2000). The `lme` function is the same function used in multilevel modeling (see section 4); however, the `nlme` package has a variety of options available for handling autocorrelation and heteroscedasticity in growth models.

Before conducting growth modeling, the data has to be set up in a way that explicitly includes time as a variable. This data transformation is referred to as changing a data set from multivariate to stacked, long, or univariate form. In the next section, we show how to create a dataframe for growth modeling.

5.2 Data Structure and the `make.univ` Function

Often data are stored in a format where each row represents observations from one individual. For instance, an individual might provide three measures of job satisfaction in a longitudinal study, and the data might be arranged in multivariate form such that column 1 is the subject number; column 2 is job satisfaction at time 1; column 3 is job satisfaction at time 2, and column 4 is job satisfaction at time 3, etc.

The `univbct` dataframe in the multilevel library allows us to illustrate a common way of storing repeated measures data. This data set contains three measures taken six-months apart on three variables – job satisfaction, commitment, and readiness. It also contains some stable individual characteristics such as respondent gender, marital status and age at the initial data collection time. These latter variables are treated as level-2 predictors in subsequent modeling.

The `univbct` dataframe is already in univariate form; however, for the purposes of illustration, we will select a subset of the entire `univbct` dataframe and transform it back into multivariate form. With this subset we will illustrate how to convert a typical multivariate dataframe back into the univariate form necessary for growth modeling.

```
> library(multilevel)
> data(univbct)
> names(univbct)
 [1] "BTN"     "COMPANY" "MARITAL" "GENDER"  "HOWLONG" "RANK"    "EDUCATE"
 [8] "AGE"     "JOBSAT1" "COMMIT1" "READY1"  "JOBSAT2" "COMMIT2" "READY2"
[15] "JOBSAT3" "COMMIT3" "READY3"  "TIME"    "JSAT"    "COMMIT"  "READY"
[22] "SUBNUM"
> nrow(univbct)
[1] 1485
> length(unique(univbct$SUBNUM))
[1] 495
```

These commands indicate there are 1485 rows in the data set representing 495 individuals so each individual provides three rows of data. To create a multivariate data set out of the `univbct` dataframe, we can select the first row for each participant in the `univbct` dataframe. In this illustration we restrict our analyses to the three job satisfaction scores and to respondent age at the initial data collection period.

```
> GROWDAT<-univbct[!duplicated(univbct$SUBNUM),c(22,8,9,12,15)]
> GROWDAT[1:3,]
  SUBNUM AGE  JOBSAT1 JOBSAT2 JOBSAT3
1      1  20 1.666667       1       3
4      2  24 3.666667       4       4
7      3  24 4.000000       4       4
```

The dataframe GROWDAT now contains data from 495 individuals. The first individual was 20 years old at the first data collection time. At time 1, the first individual's job satisfaction score was 1.67; at time 2 it was 1.0, and at time 3 it was 3.0.

Because the univbct dataframe in the multilevel package was already in univariate form, we illustrated the additional steps of converting it back to multivariate form. From a practical standpoint, though, the important issue is that the GROWDAT dataframe now represents a typical multivariate data set containing repeated measures. Specifically, the GROWDAT dataframe contains one row of information for each subject, and the repeated measures (job satisfaction) are represented by three different variables.

From a growth modeling perspective, the key problem with multivariate dataframes like GROWDAT is that they do not contain a variable that indexes time. That is, we know time is an attribute of this data because we have three different measures of job satisfaction; however, analytically we have no way of explicitly modeling time in the multivariate form of the data. Therefore, it is critical to create a new variable that explicitly indexes time which requires transforming the data to univariate or a stacked format.

The make.univ function from the multilevel package provides a simple way to perform this transformation. Two arguments are required (x and dvs), and two are optional (tname and outname). The first required argument is the dataframe in multivariate or wide format. The second required argument is a subset of the entire dataframe identifying the columns containing the repeated measures. The second required argument must be time-sorted -- column 1 must be time 1, column 2 must be time 2, and so on. The two optional arguments control the names of the two created variables: tname defaults to "TIME" and outname defaults to "MULTDV".

For instance, to convert GROWDAT into univariate form we issue the following command:

```
> UNIV.GROW<-make.univ(GROWDAT,GROWDAT[,3:5])
> UNIV.GROW[1:9,]
    SUBNUM AGE  JOBSAT1 JOBSAT2 JOBSAT3 TIME   MULTDV
1        1  20 1.666667       1       3    0 1.666667
1.1      1  20 1.666667       1       3    1 1.000000
1.2      1  20 1.666667       1       3    2 3.000000
4        2  24 3.666667       4       4    0 3.666667
4.1      2  24 3.666667       4       4    1 4.000000
4.2      2  24 3.666667       4       4    2 4.000000
7        3  24 4.000000       4       4    0 4.000000
7.1      3  24 4.000000       4       4    1 4.000000
7.2      3  24 4.000000       4       4    2 4.000000
```

Note that each individual now has three rows of data indexed by the variable "TIME". Time ranges from 0 to 2. To facilitate model interpretation, the first time is coded as 0 instead of as 1. Doing so allows one to interpret the intercept in subsequent models as the level of job satisfaction at the initial data collection time. Second, notice that the make.univ function has created a variable called "MULTDV". This variable represents the multivariate dependent variable. The variable "TIME" and the variable "MULTDV" are two of the primary variables used in growth modeling. Finally, notice that AGE, SUBNUM and the values for the three job satisfaction variables were repeated three times for each individual. By repeating the individual variables, the make.univ function has essentially created a dataframe with level-2 variables in

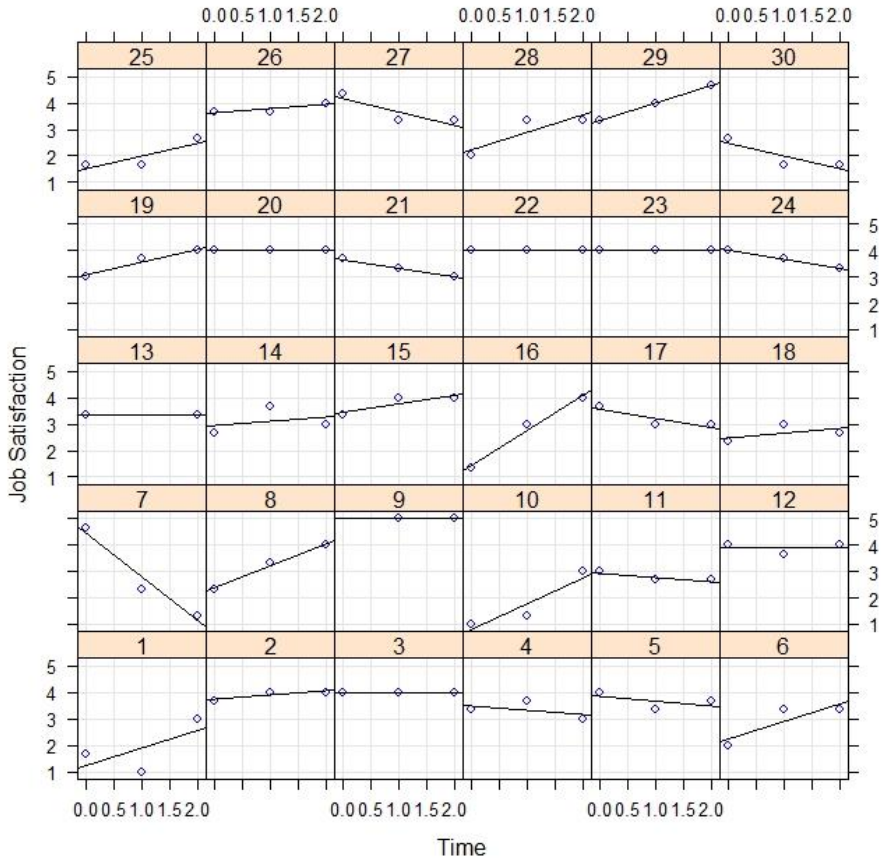
the proper format. For instance, subject age can now be used as a level-2 predictor in subsequent modeling.

In many cases, one may have only one dependent variable that needs to be converted into univariate or stacked format and therefore the `make.univ` function will suffice. If, however, it is necessary to create a univariate dataframe with multiple variables indexed by time, the `mult.make.univ` function in the `multilevel` package is available as is the `reshape` function in the base R program (see help files).

5.3 Growth Modeling Illustration

With the data in univariate form, we can begin to visually examine whether we see patterns between time and the outcome. For instance, the commands below use the `lattice` package to produce a plot of the first 30 individuals:

```
>library(lattice)
>xyplot(MULTDV~TIME|as.factor(SUBNUM),data=UNIV.GROW[1:90,],
 type=c("p","r","g"),col="blue",col.line="black",
 xlab="Time",ylab="Job Satisfaction")
```



From this plot, it appears as though there is considerable variability both in overall levels of job satisfaction and in how job satisfaction changes over time. The goal in growth modeling is to determine whether we can find consistent patterns in the relationship between time and job

satisfaction. Therefore, we are now ready to illustrate growth modeling in a step-by-step approach. In this illustration, we follow the model comparison approach outlined by Bliese and Ployhart (2002) and in also discussed in Ployhart, Holtz and Bliese (2002).

As an overview of the steps, the basic procedure is to start by examining the nature of the outcome. Much as we did in multilevel modeling, we are interested in estimating the ICC and determining whether the outcome (job satisfaction) randomly varies among individuals. Second, we are interested in examining the form of the relationship between time and the outcome. Basically, we want to know whether the outcome generally increases, decreases, or shows some other type of relationship with time. The plot of the first 30 individuals shows no clear pattern in how job satisfaction is changing over time, but the analysis might identify an overall trend among the 495 respondents. Third, we attempt to determine whether the relationship between time and the outcome is constant among individuals or whether it varies on an individual-by-individual basis. Fourth, we model in more complicated error structures such as autocorrelation, and finally we add level-2 predictors of intercept and slope variances.

5.3.1 Step 1: Examine the DV

The first step in growth modeling is to examine the properties of the dependent variable by estimating a null model and calculating the ICC.

```
> null.model<-lme(MULTDV~1,random=~1|SUBNUM,data=UNIV.GROW,
na.action=na.omit, control=list(opt="optim"))

> VarCorr(null.model)
SUBNUM = pdLogChol(1)
            Variance  StdDev
(Intercept) 0.4337729 0.6586144
Residual    0.4319055 0.6571952

> 0.4337729/(0.4337729+0.4319055)
[1] 0.5010786
```

In our example, the ICC associated with job satisfaction is .50 indicating that 50% of the variance in any individual report of job satisfaction can be explained by the properties of the individual who provided the rating. Another way to think about this is that individuals tend to remain consistent in ratings over time (a person who has high job satisfaction at one time will then to have high job satisfaction at other times). At the same time, an ICC of .50 is low enough to allow for within-person change over time. In practice, ICC values between .30 and .70 tend to be good when modeling change over time.

5.3.2 Step 2: Model Time

Step two involves modeling the fixed relationship between time and the dependent variable. In almost all cases, it is logical to begin by modeling a linear relationship and progressively add more complicated relationships such as quadratic, cubic, etc. To test whether there is a linear relationship between time and job satisfaction, we regress job satisfaction on time in a model with a random intercept.

```
> model.2<-lme(MULTDV~TIME,random=~1|SUBNUM,data=UNIV.GROW,
na.action=na.omit,control=list(opt="optim"))
> summary(model.2)$tTable
```

```
                Value  Std.Error  DF   t-value    p-value
(Intercept) 3.21886617 0.04075699 903 78.977040 0.00000000
TIME        0.05176461 0.02168024 903  2.387640 0.01716169
```

Results indicate a significant linear relationship between time and job satisfaction such that job satisfaction increases by .05 each time period. Because the first time period was coded as 0, the intercept value of 3.22 represents the average level of job satisfaction at the first time period.

More complicated time functions can be included in one of two ways – either through raising the time variable to various powers, or by converting time into power polynomials. Both techniques are illustrated.

```
> model.2b<-lme(MULTDV~TIME+I(TIME^2),random=~1|SUBNUM,
data=UNIV.GROW,na.action=na.omit,control=list(opt="optim"))

> summary(model.2b)$tTable
                  Value  Std.Error  DF    t-value   p-value
(Intercept)  3.23308157 0.04262697 902 75.8459120 0.0000000
TIME        -0.03373846 0.07816572 902 -0.4316273 0.6661154
I(TIME^2)    0.04276425 0.03756137 902  1.1385167 0.2552071

> model.2c<-lme(MULTDV~poly(TIME,2),random=~1|SUBNUM,
data=UNIV.GROW,na.action=na.omit,control=list(opt="optim"))
> summary(model.2c)$tTable
                   Value Std.Error  DF   t-value    p-value
(Intercept)    3.2704416 0.0346156 902 94.478836 0.00000000
poly(TIME, 2)1 1.5778835 0.6613714 902  2.385775 0.01724863
poly(TIME, 2)2 0.7530736 0.6614515 902  1.138517 0.25520707
```

Neither model finds evidence of a significant quadratic trend. Note that a key advantage of the power polynomials is that the linear and quadratic effects are orthogonal. Consequently, in the second model the linear effect of time is still significant even with the quadratic effect in the model so only one model needs to be estimated to identify both the linear and quadratic effects. When squaring time, it is important to run the linear model before running the model with both the linear and quadratic effect to ensure that the linear effect is identified.

5.3.3 Step 3: Model Slope Variability

A potential limitation with model 2 is that it assumes that the relationship between time and job satisfaction is constant for all individuals. Specifically, it assumes that each individual increases job satisfaction by .05 points at each time. An alternative model is one that allows slopes to vary. Given the degree of variability in the graph of the first 30 respondents, a random slope model seems like a plausible alternative. The random slope model is tested by adding the linear effect for time as a random effect. In the running example, we can `update` model.2 by adding a different random effect component and contrast model 2 and model 3.

```
> model.3<-update(model.2,random=~TIME|SUBNUM)
> anova(model.2,model.3)
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model.2     1  4 3461.234 3482.194 -1726.617
model.3     2  6 3434.132 3465.571 -1711.066 1 vs 2 31.10262  <.0001
```

The results show that a model that allows the slope between time and job satisfaction to vary across individuals fits the data better than a model that fixes the slope to be a constant value. In cases where higher-level trends were also significant, one would also be interested in determining whether allowing the slopes of the higher-level variables to randomly vary also improved model fit. For instance, one might find that a quadratic relationship varied in strength among individuals.

5.3.4 Step 4: Modeling Error Structures

The fourth step in developing the level-1 model involves assessing the error structure of the model. It is important to scrutinize the level-1 error structure because significance tests may be affected if error structures are not properly specified. The goal of step 4 is to determine whether one's model fit improves by incorporating (a) an autoregressive structure with serial correlations and (b) heterogeneity in the error structures.

Tests for autoregressive structure (autocorrelation) are conducted by including the `correlation` option in `lme`. For instance, we can update model.3 and include lag 1 autocorrelation as follows:

```
> model.4a<-update(model.3,correlation=corAR1())
> anova(model.3,model.4a)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model.3      1  6 3434.132 3465.571 -1711.066
model.4a     2  7 3429.771 3466.451 -1707.886 1 vs 2 6.360465  0.0117
```

A model that allows for autocorrelation fits the data better than does a model that assumes no autocorrelation. A summary of model 4a reveals that the autocorrelation estimate is .367 (see the Phi coefficient).

```
> summary(model.4a)
Linear mixed-effects model fit by REML
 Data: UNIV.GROW
       AIC      BIC    logLik
  3429.771 3466.451 -1707.886
.....
Correlation Structure: AR(1)
 Formula: ~1 | SUBNUM
 Parameter estimate(s):
      Phi
0.3676831
```

It is important to note that the use of `correlation=corAR1()` in the default mode assumes data is structured such that time increases for each individual. Stacked data created using `make.univ` has this structure. If data are imported or otherwise manipulated so that this order is not maintained, it will be necessary either to re-order the dataframe or to specify the structure to `corAR1()` with more detail (see help files). For example, if the rows in GROW.UNIV are randomly ordered, the estimate for AR 1 changes:

```
> UNIV.GROW2<-UNIV.GROW[order(rnorm(1485)),]
> UNIV.GROW2[1:10,]
       SUBNUM AGE  JOBSAT1  JOBSAT2  JOBSAT3 TIME   MULTDV
6           2  24 3.666667 4.000000 4.000000    0 3.666667
285.2      93  20 2.333333 3.000000 3.000000    2 3.000000
```

```
339.2     109  33 3.666667 3.000000 3.333333    2 3.333333
228        74  23 5.000000       NA 5.000000    0 5.000000
894       294  37 4.000000 4.000000 4.000000    0 4.000000
1029.1    339  20 3.000000 3.333333 3.000000    1 3.333333
1416      468  20 3.333333 3.333333 3.666667    0 3.333333
696.2     228  19 4.000000 2.666667 3.333333    2 3.333333
735.1     241  25 3.666667 3.000000 3.000000    1 3.000000
51         17  20 3.666667 3.000000 3.000000    0 3.666667

> tmod<-lme(MULTDV~TIME,random=~1|TIME,na.action=na.omit,
data=UNIV.GROW2,corAR1())

> summary(tmod)
Linear mixed-effects model fit by REML
 Data: UNIV.GROW2
       AIC      BIC    logLik
  3766.914 3793.113 -1878.457
...
Correlation Structure: AR(1)
 Formula: ~1 | TIME
 Parameter estimate(s):
       Phi
0.05763463
```

In the truncated results, notice how the estimate of the phi-coefficient changed (replications will result in different estimates of the phi-coefficient because of different structures associated with the random sorting of the data). To ensure the data is in the proper structure the `order` function can be used on any dataframe to restructure by higher-level entity and time:

```
> UNIV.GROW3<-UNIV.GROW2[order(UNIV.GROW2$SUBNUM,UNIV.GROW2$TIME),]
> UNIV.GROW3[1:10,]
    SUBNUM AGE  JOBSAT1  JOBSAT2 JOBSAT3 TIME   MULTDV
3        1  20 1.666667 1.000000       3    0 1.666667
3.1      1  20 1.666667 1.000000       3    1 1.000000
3.2      1  20 1.666667 1.000000       3    2 3.000000
6        2  24 3.666667 4.000000       4    0 3.666667
6.1      2  24 3.666667 4.000000       4    1 4.000000
6.2      2  24 3.666667 4.000000       4    2 4.000000
9        3  24 4.000000 4.000000       4    0 4.000000
9.1      3  24 4.000000 4.000000       4    1 4.000000
9.2      3  24 4.000000 4.000000       4    2 4.000000
12       4  23 3.333333 3.666667       3    0 3.333333
```

Finally, we can examine the degree to which the variance of the responses changes over time using the `varExp` option (see Pinheiro & Bates, 2000 for details).

```
> model.4b<-update(model.4a,weights=varExp(form=~TIME))
> anova(model.4a,model.4b)
         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
model.4a     1  7 3429.771 3466.451 -1707.886
model.4b     2  8 3428.390 3470.309 -1706.195 1 vs 2 3.381686  0.0659
```

The model that includes both autocorrelation and allows for decreases in variance fits the data marginally better (using a liberal p-value) than does the model that only includes autocorrelation.

In subsequent analyses, however, `model.4b` ran into convergence problems. Consequently, we elect to use `model.4a` as our final level-1 model.

With the completion of step 4, we have exhaustively examined the form of the level-1 relationship between time and job satisfaction. This analysis has revealed that (a) individuals vary in terms of their mean levels of job satisfaction, (b) there is a linear, but not quadratic, relationship between time and job satisfaction, (c) the strength of the linear relationships varies among individuals, and (d) there is significant autocorrelation in the data. At this point, we are ready to add level-2 variables to try and explain the random variation in intercepts (i.e., mean job satisfaction) and in the time-job satisfaction slope.

5.3.5 Step 5: Predicting Intercept Variation

Step 5 in growth modeling is to examine factors that can potentially explain intercept variation. In our case, we are interested in examining factors that explain why some individuals have high job satisfaction while other individuals have low job satisfaction. In this example, we explore the idea that age at the first data collection time is related to intercept variation.

To model this relationship, the individual-level characteristic, age, is used as a predictor of the job satisfaction intercept. The model that we will test is represented below using the Bryk and Raudenbush (1992) notation.

$$Y_{ij} = \pi_{0j} + \pi_{1j}(Time_{ij}) + r_{ij}$$
$$\pi_{0j} = \beta_{00} + \beta_{01}(Age_j) + u_{0j}$$
$$\pi_{1j} = \beta_{10} + u_{1j}$$

This equation states that respondent j's mean level of job satisfaction (π_{0j}) can be modeled as a function of two things. One is the mean level of job satisfaction (β_{00}) for all respondents. The second is a coefficient associated with the individual's age (β_{01}). Note that the error term for the intercept (u_{0j}) now represents the difference between an individuals' mean job satisfaction and the overall job satisfaction after accounting for age. In `lme` the model is specified as:

```
> model.5<-lme(MULTDV~TIME+AGE,random=~TIME|SUBNUM,
  correlation=corAR1(),na.action=na.omit,data=UNIV.GROW,
  control=list(opt="optim"))

> round(summary(model.5)$tTable,dig=3)
            Value Std.Error  DF t-value p-value
(Intercept) 2.347     0.146 897  16.086   0.000
TIME        0.053     0.024 897   2.205   0.028
AGE         0.034     0.005 486   6.241   0.000
```

Model 5 differs only from Model 4a in that Model 5 includes AGE (age at the baseline survey). Notice that AGE is positively related to levels of job satisfaction. Also notice that there are fewer degrees of freedom for AGE than for TIME since AGE is an individual (level-2) attribute. The AGE parameter indicates that a 23-year-old in the baseline survey would have average job satisfaction scores across the three times that were 0.034 higher than a 22-year-old in the baseline survey.

5.3.6 Step 6: Predicting Slope Variation

The final aspect of growth modeling involves attempting to determine attributes of individual respondents that are related to slope variability. In this section, we attempt to determine whether respondent age can explain some of the variation in the time-job satisfaction slope. The model that we test is presented below:

$$Y_{ij} = \pi_{0j} + \pi_{1j}(Time_{ij}) + r_{ij}$$
$$\pi_{0j} = \beta_{00} + \beta_{01}(Age_j) + u_{0j}$$
$$\pi_{1j} = \beta_{10} + \beta_{11}(Age_j) + u_{1j}$$

This model is similar to the model specified in step 5 except that we now test the assumption that the slope between time and job satisfaction for an individual (π_{1j}) is a function of an overall slope (β_{10}), individual age (β_{11}), and an error term (u_{1j}). In `lme`, the model is specified as:

```
> model.6<-lme(MULTDV~TIME*AGE,random=~TIME|SUBNUM,
  correlation=corAR1(),na.action=na.omit,data=UNIV.GROW,
  control=list(opt="optim"))
```

Note that the only difference between model 5 and model 6 is that we include an interaction term for TIME and AGE. A summary of model 6 reveals a significant interaction.

```
> round(summary(model.6)$tTable,dig=3)
             Value Std.Error  DF t-value p-value
(Intercept)  2.098     0.186 896  11.264   0.000
TIME         0.271     0.104 896   2.608   0.009
AGE          0.043     0.007 486   6.180   0.000
TIME:AGE    -0.008     0.004 896  -2.153   0.032
```

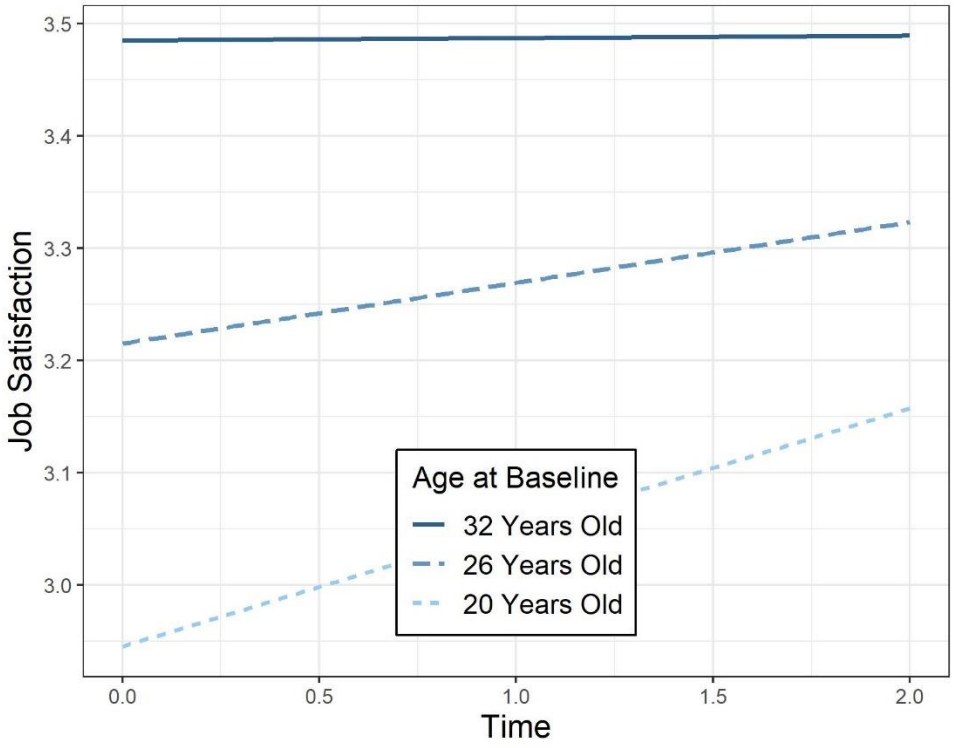
5.3.7 Plot Growth Model Using the lme4 Package and Interactions Library

To plot we first re-estimate the model in the lme4 package. The `lmer` function does not have the option to control for autocorrelation, but we can see that omitting this option does not change our substantive interpretation.

```
> library(lme4)
> library(lmerTest)

> model.6a<-lmer(MULTDV~TIME*AGE+(TIME|SUBNUM),data=UNIV.GROW)
> round(summary(model.6a)$coef,dig=3)
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)    2.078      0.186 470.301  11.176    0.000
TIME           0.273      0.104 462.965   2.630    0.009
AGE            0.044      0.007 469.523   6.276    0.000
TIME:AGE      -0.008      0.004 461.280  -2.169    0.031
```

The code below uses the `lmer` model to produce a plot using the defaults of the mean and one standard deviation above and below the mean AGE (a 32, 26 and 20 year old).

```
library(interactions)
library(ggplot2)
win.graph(height=4.75,width=6)
interact_plot(model.6a,pred=TIME,modx=AGE,
              modx.labels = c("20 Years Old","26 Years Old",
                              "32 Years Old"),
              x.label = "Time",
              y.label = "Job Satisfaction",
              legend.main="Age at Baseline")+
  theme_bw()+
  theme(legend.background=element_rect(fill="white",
           size=0.5, linetype="solid",color ="black"),
        legend.position = c(0.5, 0.2),
        axis.title.x = element_text(color="black", size=14),
        axis.title.y = element_text(color="black", size=14),
        legend.title = element_text(color="black", size=13,
                                    hjust=.5),
        legend.text = element_text(color="black", size=12)
  )
ggsave(filename = "c:\\temp\\plotgg.jpg",
       device = "jpeg")
```



Older individuals at baseline reported higher job satisfaction initially and tended to show a very slight increase over time. In contrast, younger respondents tended to report lower initial job satisfaction, but showed a more pronounced increase in job satisfaction over time.

5.4 Discontinuous Growth Models

In the previous example (section 5.3.2), two variants of time were examined (linear and quadratic). Indeed, with only three periods it is difficult to explore more than a linear and quadratic trend (through one could treat time as a categorical variable and make no assumptions about trends). In situations where numerous repeated measures are collected, however, a variety of interesting options exist for modeling time.

One particularly interesting variant is the discontinuous growth model (DGM) a model also referred to as the piecewise hierarchical linear model (Raudenbush & Bryk, 2002; Hernández-Lloreda et al., 2004) or the multiphase mixed-effects model (Cudeck & Klebe, 2002). The basic idea behind the DGM is to simultaneously use a set of two or three time-related covariates to capture a known discontinuity.

For instance, Lang and Bliese (2009) use the DGM to model the performance impact of unexpectedly changing key elements of a complex computer-based task. In the design, participants worked on the task for six trials and then on the seventh trial the task became substantially more difficult. Although there are numerous variants for modeling a discontinuity of this nature (see Bliese & Lang, 2016), the basic form can be captured by the three terms TIME, TRANS, and POST. Because these time-varying predictors represent a system of equations, TIME captures the initial linear trend; TRANS captures the immediate response to the event, and POST captures the post-transition slope change. A fourth useful variant is to include a TIME.A (for absolute) that results in expressing the TRANS and POST parameters in absolute versus relative terms.

5.4.1 Coding for DGM Simple Cases

The data set `tankdat` from Lang and Bliese (2009) was used to illustrate variants of the DGM in Bliese and Lang (2016). Below we apply a subset of the R code from Appendix B of Bliese and Lang to illustrate basic form of the DGM.

```
> data(tankdat)

> tankdat$TRANS<-ifelse(tankdat$TIME<6,0,1)
> tankdat$POST<-ifelse(tankdat$TIME>5,tankdat$TIME-6,0)
> tankdat$TIME.A<-ifelse(tankdat$TIME<5,tankdat$TIME,5)

> tankdat[1:12,c("TIME","TRANS","POST","TIME.A")]
   TIME TRANS POST TIME.A
1     0     0    0      0
2     1     0    0      1
3     2     0    0      2
4     3     0    0      3
5     4     0    0      4
6     5     0    0      5
7     6     1    0      5
8     7     1    1      5
9     8     1    2      5
10    9     1    3      5
11   10     1    4      5
12   11     1    5      5
```

TRANS represents a dummy-coded variable that is zero before the event and one after the event. POST is slightly more complex in that it begins with a zero and then begins recounting (starting with zero) after the event occurs. TIME.A begins similarly to TIME, but holds the pre-transition element (5 in this case) constant once the change has occurred.

Below the basic DGM mixed-effect model is estimated and used to illustrate the difference between TIME and TIME.A.

```
> tmod<-lme(SCORE~TIME+TRANS+POST, random=~1|ID,tankdat)
> round(summary(tmod)$tTable,dig=3)
             Value Std.Error   DF t-value p-value
(Intercept) -3.686     0.631 2021  -5.837       0
TIME         1.814     0.125 2021  14.461       0
TRANS       -4.980     0.619 2021  -8.049       0
POST        -1.220     0.177 2021  -6.880       0

> tmod.a<-lme(SCORE~TIME.A+TRANS+POST, random=~1|ID,tankdat)
> round(summary(tmod.a)$tTable,dig=3)
             Value Std.Error   DF t-value p-value
(Intercept) -3.686     0.631 2021  -5.837       0
TIME.A       1.814     0.125 2021  14.461       0
TRANS       -3.166     0.537 2021  -5.895       0
POST         0.593     0.125 2021   4.732       0
```

Notice that TIME and TIME.A have the same parameter estimate and standard errors and both indicate that the performance score increased by 1.81 each trial. In the top model (TIME), the parameter estimate for TRANS is -4.98 and the POST estimate is -1.22 (both are significant). When using TIME, both TRANS and POST represent change relative to TIME, so the decline of -4.98 assumes this time period would have increased by 1.81. Likewise, the POST slope of -1.22 indicates a slope that is 1.22 less steep than the 1.81 increase associated with TIME.

The parameters associated with TIME.A are absolute, so in the lower model the value of -3.17 represents the absolute change (relative to zero) in performance. Likewise, the now positive slope of 0.59 indicates that while the recovery slope is significantly less strong than the pre-transition slope associated with TIME, the recovery slope is still significantly positive.

The DGM model, like the growth model, can be examined in a series of steps examining person-level variability in each parameter and including predictors of this variability. Interested readers are directed to Bliese and Lang (2106) and Bliese, Kautz, and Lang (2020) for additional details. Several examples using the DGM include Kim and Ployhart, (2014); Li, Hausknecht and Dragoni (2020); Pagiavlas, et al., (2021) and Rupp et al., 2009; Stewart et al., (2017).

5.4.2 Coding for DGM Complex Cases (`dgm.code`)

In cases such as with the tank data from Lang and Bliese (2009), the coding of the time-varying parameters is simple. In many applied settings, however, the coding can be more complicated for three reasons. First the longitudinal or panel data may be unbalanced such that each higher-level entity has a different number of repeated measures. Second, the event of interest may occur at different time points for each entity. Third, entities might not have the same number of events or any events at all.

For instance, a study of the impact of employee turnover on store performance might have panel data with thousands of stores providing quarterly data for 2 years. In each quarter, turnover may or may not have occurred, so each store has a unique pattern of turnover. Attempting to code the DGM time-varying covariates on a store-by-store basis would be challenging and time consuming.

The `dgm.code` function was designed to produce a design matrix for cases where events occur on an irregular basis and/or where entities have different number of observations. Details on the using dgm.code are in the help files, but below I reproduce one example.

```
> data(tankdat)
>
> # Add a marker (1 or 0) indicating an event at random
> set.seed(343227)
> tankdat$taskchange<-rbinom(nrow(tankdat),1,prob=.1)
> tankdat[1:24,]
   ID    CONSC TIME SCORE taskchange
1   1 1.041923    0    -5          0
2   1 1.041923    1     0          0
3   1 1.041923    2    -3          0
4   1 1.041923    3    -9          0
5   1 1.041923    4    -7          0
6   1 1.041923    5    -3          0
7   1 1.041923    6    -7          0
8   1 1.041923    7    -3          0
9   1 1.041923    8   -11          0
10  1 1.041923    9    -5          0
11  1 1.041923   10    -1          1
12  1 1.041923   11    -4          0

13  2 1.426890    0     3          0
14  2 1.426890    1    17          1
15  2 1.426890    2    18          0
16  2 1.426890    3    10          0
17  2 1.426890    4    22          1
18  2 1.426890    5    14          0
19  2 1.426890    6    -3          1
20  2 1.426890    7     6          0
21  2 1.426890    8    10          0
22  2 1.426890    9    15          0
23  2 1.426890   10    14          0
24  2 1.426890   11     7          0
```

In this example, the first individual (ID=1) had a taskchange at time 10 while the second individual (ID=2) had a task change at times, 1, 4, and 6. This example illustrates several issues. First, there are clearly different patterns of events. Second, it is not clear how events to code. An additional issue is that the event may occur on the first observation in which case the TRANS and POST time-varying vectors cannot be estimated. If we attempt to create the DGM design matrix we get the following error identifying groups that start with a taskchange (truncated output):

```
> OUT<-with(tankdat,dgm.code(ID,TIME,taskchange))
```

```
[1] "The following groups start with an event"
     grp time event
97     9    0     1
169   15    0     1
193   17    0     1
241   21    0     1
337   29    0     1
373   32    0     1
385   33    0     1
Truncated…
```

To handle both the issue of multiple events and an event starting on the first occasion, the `dgm.code` function contains two control options. By setting first.obs=TRUE we can recode the first observation to zero keeping a marker for whether we made this change. By setting n.events we can limit the design matrix to code only the first few events. For instance, to code only two events and recode the first event to a zero the command would be:

```
> OUT<-with(tankdat,dgm.code(ID,TIME,taskchange,n.events=2,first.obs=TRUE))
> OUT[1:24,]
   grp time event trans1 trans2 post1 post2 time.a tot.events event.first
1    1    0     0      0      0     0     0      0          1           0
2    1    1     0      0      0     0     0      1          1           0
3    1    2     0      0      0     0     0      2          1           0
4    1    3     0      0      0     0     0      3          1           0
5    1    4     0      0      0     0     0      4          1           0
6    1    5     0      0      0     0     0      5          1           0
7    1    6     0      0      0     0     0      6          1           0
8    1    7     0      0      0     0     0      7          1           0
9    1    8     0      0      0     0     0      8          1           0
10   1    9     0      0      0     0     0      9          1           0
11   1   10     1      1      0     0     0      9          1           0
12   1   11     0      1      0     1     0      9          1           0
13   2    0     0      0      0     0     0      0          3           0
14   2    1     1      1      0     0     0      0          3           0
15   2    2     0      1      0     1     0      0          3           0
16   2    3     0      1      0     2     0      0          3           0
17   2    4     1      0      1     0     0      0          3           0
18   2    5     0      0      1     0     1      0          3           0
19   2    6     1      0      1     0     2      0          3           0
20   2    7     0      0      1     0     3      0          3           0
21   2    8     0      0      1     0     4      0          3           0
22   2    9     0      0      1     0     5      0          3           0
23   2   10     0      0      1     0     6      0          3           0
24   2   11     0      0      1     0     7      0          3           0
```

The output returns a time, time.a, trans1, trans2, post1 and post2 to model the design matrix for two events. It also records the total events for each entity (tot.events) and indicates whether the first observation was an event.

Finally, to make use of this design matrix, it would need to be merged with the original data and reordered as follows:

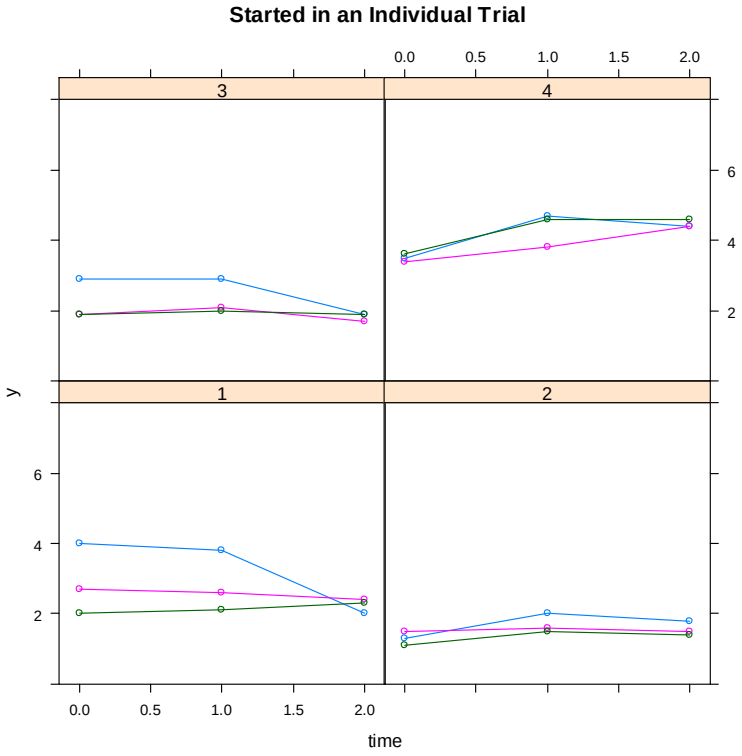
```
> tankdat.dgm<-merge(tankdat,OUT,by.x=c("ID","TIME"),by.y=c("grp","time"))
> tankdat.dgm<-tankdat.dgm[order(tankdat.dgm$ID,tankdat.dgm$TIME),]
```

5.5 Testing Emergence by Examining Error Structure

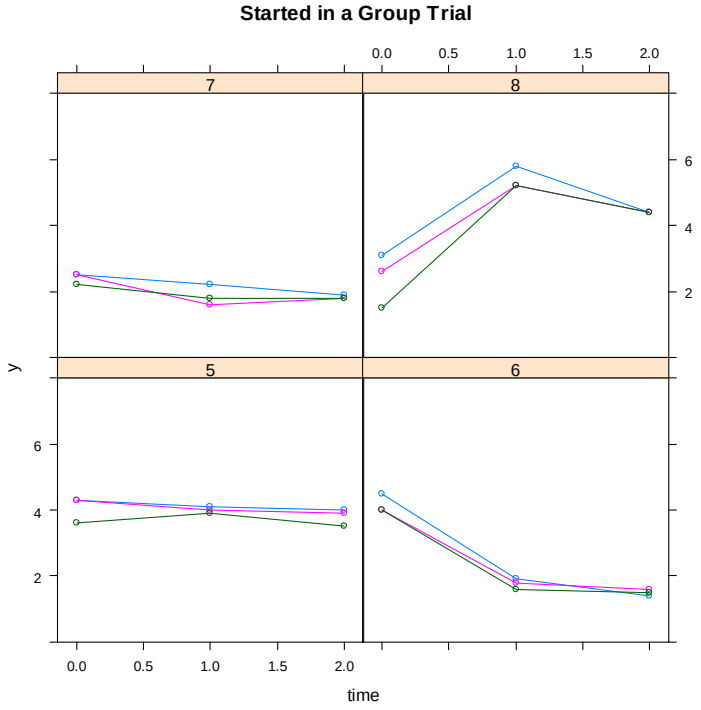
In most treatments of growth models heteroscedasticity in error structures are considered a form of model miss-specification that should be controlled (see section 5.3.4). Variants of mixed-effects models, however, have been suggested as a tool to formally test whether patterns of change in residual error variance over time have substantive meaning (Lang & Bliese, 2019; Lang et al., 2018; Lang et al., 2019).

For instance, consider the patterns displayed by participants over time in Sherif's (1935) classic experiment on group influence. In the experimental paradigm participants estimated movement of a small light (in inches) in a completely dark room. Participants either made initial estimates alone or with other group members and Sherif provided a plot of the results over three group-based trails. The data set `sherifdat` contains the values presented in Sherif's plot. The first set of figures below present the pattern for participants who began making estimates alone (and then transitioned to three trails where they made estimates with other group members). The second set of figures presents the pattern for participants who began making estimates with other group members over three trials.

```
> data(sherifdat)
> library(lattice)
> xyplot(y~time|as.factor(group),sherifdat[sherifdat$condition==1,],
  groups=person,type=c("p","l"),ylim=c(0,8),
  main="Started in an Individual Trial")
```



```
> xyplot(y~time|as.factor(group),sherifdat[sherifdat$condition==0,],
  groups=person,type=c("p","l"),ylim=c(0,8),
  main="Started in a Group Trial")
```



In both cases (either starting as an individual or starting in a group setting), the plots suggest that group members influence each other such that consensus emerges. The idea of consensus emergence appears stronger in cases where individuals started their first trial as an individual, but both conditions appear to show this effect. Lang and Bliese (2019) and Lang et al. (2018) provide details on how a three-level mixed-effect model (the census emergence model or CEM) can be estimated and how the -2log likelihood values can be contrasted to formally test whether emergence is present. Details are beyond the scope of this manual, but the basic formal test of emergence is provided below:

```
> threelevel<-lme(y ~ time,
  random = list(group=pdLogChol(~time),person=pdIdent(~1)),
  data=sherifdat,control=lmeControl(opt="optim",maxIter=3000,
  msMaxIter=3000))

> threelevelCEM<-update(threelevel,weights=varExp( form = ~ time))

> anova(threelevel,threelevelCEM)
              Model df      AIC      BIC    logLik   Test  L.Ratio p-value
threelevel        1  7 182.3422 198.0817 -84.17112
threelevelCEM     2  8 155.8097 173.7977 -69.90485 1 vs 2 28.53253  <.0001
```

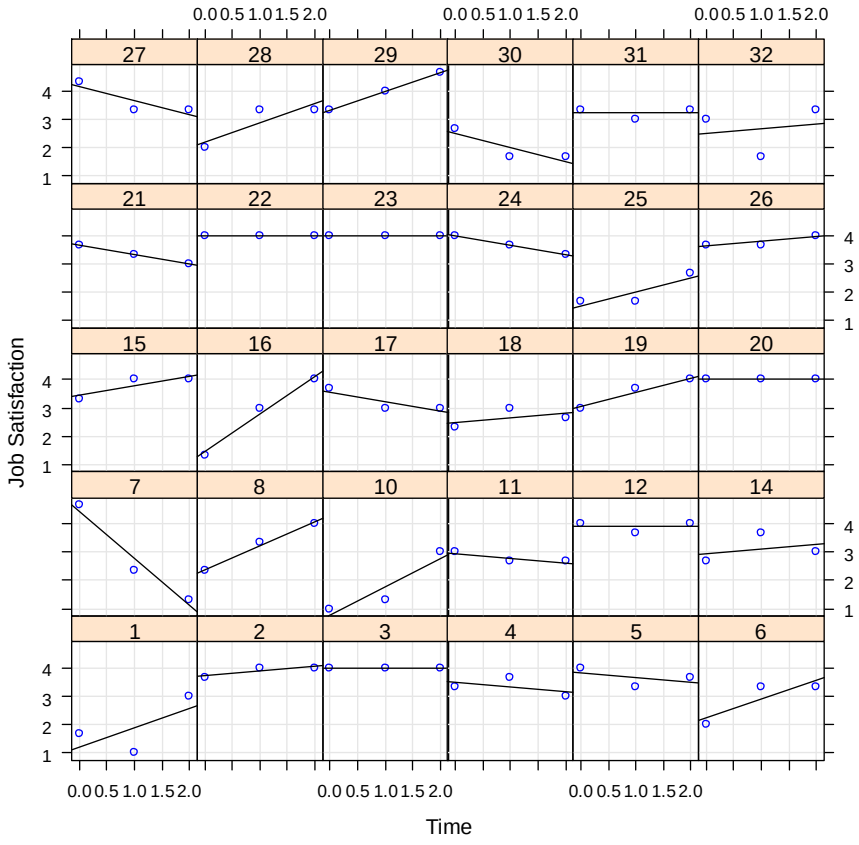
In both models, the random statement is a complex form of a three-level model that allows the slope for each group to randomly vary while fixing the time slope for individuals. A summary of the model `threelevelCEM` (not shown) provides the estimate for varExp as -1.017 indicating

an overall reduction in residual variance within groups (emergence). Including a variance term leads to a significant improvement in model fit suggesting that a significant emergence effect exists. Finally, while not demonstrated here, the models can be modified to formally test whether the emergence effect is stronger under the two conditions of starting individually or in a group.

5.6 Empirical Bayes estimates

While briefly introduced previously, one of the useful aspects of examining repeated measures in mixed-effects models is the ability to estimate predicted intercepts and slopes for individuals using (a) information about the individual along with (b) information from the rest of the sample. For instance, consider the growth modeling data presented in section 5.3. In this example, we modify the data so that only those with responses at all three times are included. We do so only to show that OLS-based estimates and empirical Bayes estimate differ even when data are complete.

```
> data(univbct)
> TEMP<-univbct[3*1:495,c(22,1:17)]  #convert to multivariate form
> TEMP<-na.exclude(TEMP[,c("SUBNUM","JOBSAT1","JOBSAT2","JOBSAT3")])
> TEMP.UNIV<-make.univ(TEMP,TEMP[,2:4],outname="JSAT")

> library(lattice)
> xyplot(JSAT~TIME|as.factor(SUBNUM),data=TEMP.UNIV[1:90,],
    type=c("p","r","g"),col="blue",col.line="black",
    xlab="Time",ylab="Job Satisfaction")
```



The figure shows large differences in intercepts and in slopes, yet each panel is estimated separately without taking into consideration any of the data from other respondents. An alternative would be to estimate a simple growth model and use data from model parameters to estimate values for each individual.

```
>tmod<-lme(JSAT~TIME,random=~TIME|SUBNUM, TEMP.UNIV,
    na.action=na.omit,control=list(opt="optim"))
```

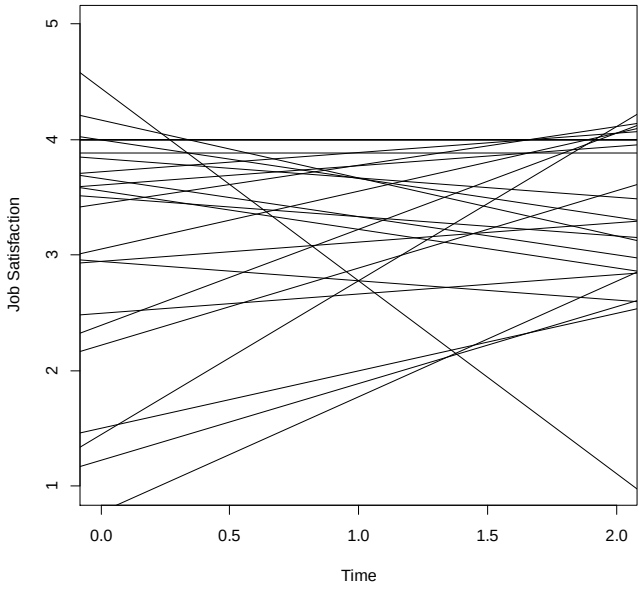
From this model, one can extract the empirical Bayes estimates for both the intercept and the slope by using the `coef` function: the first 12 values (bottom two rows) are listed.

```
> coef(tmod)[1:12,]
   (Intercept)         TIME
1     1.771548  0.358222009
2     3.701752  0.069173239
3     3.868707 -0.002492476
4     3.368637 -0.039600872
5     3.654505 -0.054411154
6     2.629151  0.313791178
7     3.537183 -0.615478500
8     2.843353  0.365710056
10    1.532927  0.496616898
11    2.892191 -0.014917079
12    3.773418  0.002444280
14    3.034727  0.103730558
```

The empirical Bayes estimates returned from `coef` correspond to what is displayed in the lattice plot. Individual 1, for instance, has a low value for satisfaction and a positive slope and individual 7 has a moderately high value and a strong negative slope.

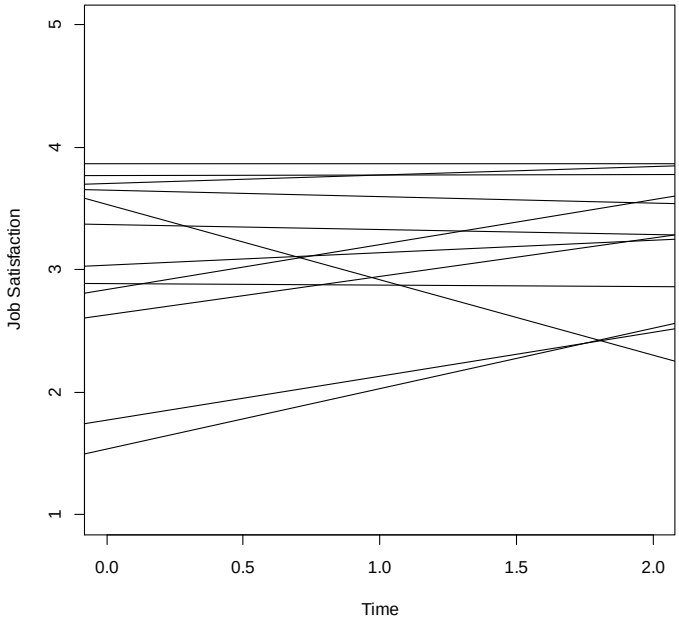
The differences can be more easily visualized by plotting all 30 individuals on a single plot. The plot represents the intercept and slope estimates from 30 separate linear regression equations.

```
>tmod3<-lmList(JSAT~TIME|SUBNUM, data=TEMP.UNIV[1:90,])

>plot(TEMP.UNIV$TIME,TEMP.UNIV$JSAT, xlab="Time",
     ylab="Job Satisfaction",type="n")

>lmplot<-function(X){
   for (i in 1:25){
   abline(X[[i]])
 }}

>lmplot(tmod3)
```



The second plot is for the same 30 individuals, but is based off of the empirical Bayes estimates.

```
>plot(TEMP.UNIV$TIME,TEMP.UNIV$JSAT, xlab="Time",
     ylab="Job Satisfaction",type="n")
>apply(coef(tmod)[1:12,],1,abline)
```



The fact that each individual's estimates are partially based on information from the rest of the sample adjusts some of the more extreme response (and explains why these are sometimes referred to as shrunken estimates). Empirical Bayes estimates may be particularly useful in situations where intercepts and slopes are used to predict other outcomes. For instance, Chen, Ployhart, Thomas, Anderson, & Bliese (2011) used empirical Bayes estimates of slope changes in job satisfaction and showed that the nature of the change (increase or decrease) was the primary predictor of turnover intentions.

It may go without saying, but one can also extract empirical Bayes estimates from non-longitudinal nested models such as those considered in section 4. In the context of non-longitudinal models, the values provides estimates of intercepts and slopes for each group adjusted for the overall intercept and slope. As a general rule, when ICC(1) values are small, the empirical Bayes estimates are more strongly adjusted to the rest of the sample (more shrinkage) than when ICC(1) values are large (see Gelman & Pardoe, 2006).

6 More on `lme4`

While the current document has focused on the `nlme` package for mixed-effects models, the `lme4` package in R provides additional flexibility in terms of specifying models. The `lme4` package is particularly valuable in dealing with (a) non-normally distributed outcomes and (b) partially crossed data structures.

6.1 Dichotomous outcomes

When the dependent variable is dichotomous or otherwise non-normally distributed, it may be useful to estimate a generalized linear mixed effects model (glmm) rather than a linear mixed effects model. Below we dichotomize WBEING and use `glmer` from the `lme4` package with a binomial link function to estimate a mixed-effects logistic regression model.

```
>library(multilevel)
>library(lme4)
>data(bh1996)
>tmod<-glmer(ifelse(WBEING>3.5,1,0)~HRS+G.HRS+(1|GRP),
       family="binomial",control=glmerControl(optimizer="bobyqa"),bh1996)

>summary(tmod)

Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: ifelse(WBEING > 3.5, 1, 0) ~ HRS + G.HRS + (1 | GRP)
   Data: bh1996
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
  7572.1   7599.7  -3782.0   7564.1     7378

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.9902 -0.5559 -0.4672 -0.3587  4.6130

Random effects:
```

```
 Groups Name        Variance Std.Dev.
 GRP    (Intercept) 0.06323  0.2515
Number of obs: 7382, groups:  GRP, 99

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.80660    0.53504   5.246 1.56e-07 ***
HRS         -0.09860    0.01465  -6.731 1.69e-11 ***
G.HRS       -0.26784    0.04923  -5.440 5.31e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      (Intr) HRS
HRS   -0.020
G.HRS -0.954 -0.272
```

The precision of the model in terms of log likelihood can be improved by including the `nAGQ` option with a value greater than 1 (100 in this case). Notice the slight change in log likelihood values and the minor changes in parameter estimates and standard errors between the model based on `nAGQ=1` (above) and `nAGQ=25` (below). In practice, one would likely want to change `nAGQ` values to (a) verify parameter estimates and standard errors and (b) verity that contrasts of -2log likelihood values contrasting models with `anova` are similar with higher `nAGQ` values. In my experience using values above 100 is rarely useful.

```
> tmod.r<-glmer(ifelse(WBEING>3.5,1,0)~HRS+G.HRS+(1|GRP),
  family="binomial", control=glmerControl(optimizer="bobyqa"),
  bh1996,nAGQ=25)

> logLik(tmod) # Original model with nAGQ=1
'log Lik.' -3782.036 (df=4)

> logLik(tmod.r) # Model with nAGQ = 25
'log Lik.' -3781.999 (df=4)

> summary(tmod.r)$coef
               Estimate Std. Error   z value     Pr(>|z|)
(Intercept)  2.80640657 0.53692297  5.226833 1.724383e-07
HRS         -0.09861117 0.01466700 -6.723335 1.776112e-11
G.HRS       -0.26782094 0.04939543 -5.421978 5.894300e-08
```

6.2 Crossed and partially crossed models

The second situation in which `lme4` is particularly valuable is in cases where data are partially or fully crossed. For instance, in a longitudinal study individuals might be nested within groups, but over time some individuals might switch from one group to another. If no participants switched groups, the data would be fully nested with repeated observations nested within individuals nested within groups (a three-level model). In `lme` the three-level nested model would be specified as `random= ~1|GRP/IND`. If individuals switch groups, though, the fully nested structure no longer holds. In `lme4` and the `lmer` function, however, the structure could be specified as `(1|GRP)+(1|IND)`. The `lmer` specification does not assume fully nested data and will provide variance estimates if the data are partially crossed.

6.3 Predicting values in `lme4`

As illustrated in the text, statistical models can be used to predict levels of an outcome variable given specific values of predictors. R has a number of `predict` functions linked to specific models (e.g., `lm, glm, lme, lmer, glmer`). The `predict` functions are generally consistent in terms of usage; however, there are minor differences when applied to specific models. Recall, for instance, that one must specify `level=0` to obtain overall sample based predictions when using `lme`.

In most cases in mixed-effects models, one will be interested in obtaining predictions for the overall sample rather than predictions for any specific unit; however, in the `lmer` and `glmer` functions associated with lme4, the predict command uses the option `re.form=NA` rather than `level=0` to indicate that predictions should be made based on the parameter estimates from the overall sample. An example is provided below:

```
> library(multilevel)
> library(lme4)
> data(bh1996)

> tmod<-lmer(WBEING~HRS*LEAD+(1|GRP),bh1996)

> TDAT<-data.frame(HRS=c(7,7,12,12),LEAD=c(2.12,2.12,3.66,3.66))
> predict(tmod,TDAT,re.form=NA)
       1        2        3        4
2.519160 2.519160 3.137911 3.137911
```

As another example, the code below illustrates the use of the `type="response"` option with models that have a dichotomous variable as the outcome. Notice that one can transform the prediction to a percent (-2.377 to 0.085 or 8.5%), but it is often easier to use `type="response"`.

```
> tmod<-glmer(ifelse(WBEING>3.5,1,0)~LEAD+(1|GRP),family="binomial",bh1996,
  control=glmerControl(optimizer="bobyqa"))

> TDAT<-data.frame(LEAD=c(2.12,3.66))

> predict(tmod,TDAT,re.form=NA)
         1          2
-2.3774501 -0.6565601

> exp(-2.3774501)/(1+exp(-2.3774501))
[1] 0.08490848

> predict(tmod,TDAT,re.form=NA,type="response")
         1          2
0.08490848 0.34151277
```

7 Miscellaneous Functions and Tips

The multilevel package has a number of other functions that have either been referenced in appendices of published papers, or are of basic utility to applied organizational researchers. This section briefly describes these functions. Complete help files are available in the `multilevel` package for each of the functions discussed.

7.1 Scale reliability: `cronbach` and `item.total`

Two functions that are can be particularly useful in estimating the reliability of multi-item scales are the `cronbach` and the `item.total` functions. Both functions take a single argument, a dataframe with multiple columns where each column represents one item in a multi-item scale.

7.2 Random Group Resampling for OLS Regression Models

The function `rgr.OLS` allows one to contrast a group-level hierarchical regression model with an identically specified model where group identifiers are randomly generated. This type of model was estimated in Bliese and Halverson (2002).

7.3 Estimating bias in nested regression models: `simbias`

Bliese and Hanges (2004) showed that a failure to model the nested properties of data in ordinary least squares regression could lead to a loss of power in terms of detecting effects. The article provided the `simbias` function to help estimate the degree of power loss in complex situations.

7.4 Detecting mediation effects: `sobel`

MacKinnon, Lockwood, Hoffman, West and Sheets (2002) showed that many of the mediation tests used in psychology tend to have low power. One test that had reasonable power was Sobel's (1982) indirect test for mediation. The `sobel` function provides a simple way to run Sobel's (1982) test for mediation. Details on the use of the `sobel` function are available in the help files.

8 References

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

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The New S Language*. New York: Chapman & Hall.

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

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and Analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel*

Theory, Research, and Methods in Organizations (pp. 349-381). San Francisco, CA: Jossey-Bass, Inc.

Bliese, P. D. (2002). Multilevel random coefficient modeling in organizational research: Examples using SAS and S-PLUS. In F. Drasgow & N. Schmitt (Eds.), *Modeling in Organizational Research: Measuring and Analyzing Behavior in Organizations* (pp. 401-445). San Francisco, CA: Jossey-Bass, Inc.

Bliese, P. D., & Britt, T. W. (2001). Social support, group consensus and stressor-strain relationships: Social context matters. *Journal of Organizational Behavior, 22*, 425-436.

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

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

Bliese, P. D., & Halverson, R. R. (1998a). Group consensus and psychological well-being: A large field study. *Journal of Applied Social Psychology, 28*, 563-580.

Bliese, P. D., & Halverson, R. R. (1998b). Group size and measures of group-level properties: An examination of eta-squared and ICC values. *Journal of Management, 24*, 157-172.

Bliese, P. D., & Halverson, R. R. (2002). Using random group resampling in multilevel research. *Leadership Quarterly, 13*, 53-68.

Bliese, P. D., & Halverson, R.R. & Rothberg, J. (2000). Using random group resampling (RGR) to estimate within-group agreement with examples using the statistical language R. *Unpublished Manuscript*.

Bliese, P. D. & Jex, S. M. (2002). Incorporating a multilevel perspective into occupational stress research: Theoretical, methodological, and practical implications. *Journal of Occupational Health Psychology, 7*, 265-276.

Bliese, P. D., & Jex S. M. (1999). Incorporating multiple levels of analysis into occupational stress research. *Work and Stress, 13*, 1-6.

Bliese, P. D., Kautz, J., & Lang, J. W. (2020). Discontinuous growth models: Illustrations, recommendations, and an R function for generating the design matrix. In Y. Griep & S. D. Hansen (Eds.), *Handbook on the Temporal Dynamics of Organizational Behavior* (pp. 319–350). Northampton, MA: Edward Elgar Publishers, Inc. DOI: https://doi.org/10.4337/9781788974387

Bliese, P. D., & Lang, J. W. B. (2016). Understanding relative and absolute change in discontinuous growth models: Coding alternatives and implications for hypothesis testing. *Organizational Research Methods, 19,* 562-592.

Bliese, P. D., Maltarich, M. A., Hendricks, J. L., Hofmann, D. A., & Adler, A. B. (2019). Improving the measurement of group-level constructs by optimizing between-group differentiation. *Journal of Applied Psychology, 104,* 293-302.

Bliese, P. D., Maltarich, M. A., & Hendricks, J. L. (2018). Back to Basics with Mixed-Effects Models: Nine Take-Away Points. *Journal of Business and Psychology, 33,* 1-23.

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

Brown, R. D., & Hauenstein, N. M. A. (2005). Interrater agreement reconsidered: An alternative to the rwg indices. *Organizational Research Methods, 8,* 165-184.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Burke, M. J., Finkelstein, L. M., & Dusig, M. S. (1999). On average deviation indices for estimating interrater agreement. *Organizational Research Methods, 2,* 49-68.

Campbell-Sills, L., Flynn, P. J., Choi, K. W., Ng, T. H., Aliaga, P. A., Broshek, C., Jain, S., Kessler, R. C., Stein, M. B., Ursano, R. J. & Bliese, P. D., (2022). Unit cohesion during deployment and post-deployment mental health: Is cohesion an individual- or unit-level buffer for combat-exposed soldiers? *Psychological Medicine, 52,* 121-131.

Chambers, J. M. & Hastie, T. J. (1992). *Statistical Models in S*. New York: Chapman & Hall.

Chen, G., Ployhart, R. E., Thomas, H. C., Anderson, N. & Bliese, P. D. (2011). The power of momentum: A new model of dynamic relationships between job satisfaction change and turnover intentions. *Academy of Management Journal, 54,* 159-181.

Clark, T. S., & Linzer, D. A. (2015). Should I use fixed or random effects? *Political Science Research and Methods, 3,* 399-408.

Cohen, J. & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, A., Doveh, E. & Eick, U. (2001). Statistical properties of the rwg(j) index of agreement. *Psychological Methods, 6,* 297-310.

Cohen, A., Doveh, E. & Nahum-Shani, I. (2009). Testing agreement for multi-item scales with the indices rwg(j) and ADM(J). *Organizational Research Methods, 12,* 148-164.

Cudeck, R., & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. *Psychological Methods, 7*, 41–63.

Cummings, G. & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 170-180.

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

Dunlap, W. P., Burke, M. J., & Smith-Crowe, K. (2003). Accurate tests of statistical significance for rwg and average deviation interrater agreement indices. *Journal of Applied Psychology, 88*, 356-362.

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

Gelman, A. & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics, 48*, 241-251.

Hernández-Lloreta, M. V., Colmenares, F., & Martínez-Arias, R. (2004). Application of piecewise hierarchical linear growth modeling to the study of continuity in behavioral development of baboons (Papio hamadryas). *Journal of Comparative Psychology, 118,* 316–324.

Hofmann, D. A. (1997). An overview of the logic and rationale of Hierarchical Linear Models. *Journal of Management, 23*, 723-744.

Hofmann, D. A. & Gavin, M. (1998). Centering decisions in hierarchical linear models: Theoretical and methodological implications for research in organizations. *Journal of Management, 24*, 623-641.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67*, 219-229.

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. & Williams, L. J. (2000). The cross-level operator in regression, ANCOVA, and contextual analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel Theory, Research, and Methods in Organizations* (pp. 382-424). San Francisco, CA: Jossey-Bass, Inc.

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

Klein, K. J. & Kozlowski, S. W. J, (2000). *Multilevel theory, research, and methods in organizations*. San Francisco, CA: Jossey-Bass, Inc.

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. Kozlowski (Eds.), *Multilevel Theory, Research, and Methods in Organizations* (pp. 512-553). San Francisco, CA: Jossey-Bass, Inc.

Kim, Y. & Ployhart, R. E., (2014). The effects of staffing and training on firm productivity and profit growth before, during, and after the great recession. *Journal of Applied Psychology, 99*, 361-389.

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.

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

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

Lang, J. W. B., & Bliese, P. D., (2019). A Temporal Perspective on Emergence: Using 3-level Mixed Effects Models to Track Consensus Emergence in Groups. In S. E. Humphrey & J. M. LeBreton (Eds.), *The Handbook for Multilevel Theory, Measurement, and Analysis*. Washington, DC: American Psychological Association.

Lang, J. W. B., & Bliese, P. D. (2009). General mental ability and two types of adaptation to unforeseen change: Applying discontinuous growth models to the task-change paradigm. *Journal of Applied Psychology, 92*, 411-428.

Lang, J. W. B., Bliese, P. D., & Adler, A. B. (2019). Opening the Black Box: A Multilevel Framework for Studying Group Processes. *Advances in Methods and Practices in Psychological Science, 2*, 271-287.

Lang, J. W. B., Bliese, P. D., & de Voogt, A. (2018). Modeling Consensus Emergence in Groups Using Longitudinal Multilevel Methods. *Personnel Psychology, 71*, 255-281.

LeBreton, J. M., James, L. R. & Lindell, M. K. (2005). Recent issues regarding r_{WG}, r^*_{WG}, $r_{WG(J)}$, and $r^*_{WG(J)}$. *Organizational Research Methods, 8*, 128-138.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational research methods, 11(4)*, 815-852.

Levin, J. R. (1967). Comment: Misinterpreting the significance of "explained variation." *American Psychologist, 22,* 675-676.

Li, H., Hausknecht, J. P., & Dragoni, L. (2020). Initial and Longer-Term Change in Unit-Level Turnover Following Leader Succession: Contingent Effects of Outgoing and Incoming Leader Characteristics. *Organization Science, 31(2),* 458-476.

Lindell, M. K. & Brandt, C. J. (1997). Measuring interrater agreement for ratings of a single target. *Applied Psychological Measurement, 21,* 271-278.

Lindell, M. K. & Brandt, C. J. (1999). Assessing interrater agreement on the job relevance of a test: A comparison of CVI, T, rWG(J), and r*WG(J) indexes. *Journal of Applied Psychology, 84,* 640-647.

Lindell, M. K. & Brandt, C. J. (2000). Climate quality and climate consensus as mediators of the relationship between organizational antecedents and outcomes. *Journal of Applied Psychology, 85,* 331-348.

Lindell, M. K., Brandt, C. J. & Whitney, D. J. (1999). A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement, 23,* 127-135.

Lüdtke, O., & Robitzsch, A. (2009). Assessing within-group agreement: A critical examination of a random-group resampling approach. *Organizational Research Methods, 12,* 461-487.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7,* 83-104.

Pagiavlas, S., Kalaignanam, K., Gill, M., & Bliese, P. D. (2021). EXPRESS: Regulating Product Recall Compliance in the Digital Age: Evidence from the "Safe Cars Save Lives" Campaign. *Journal of Marketing,* DOI: 00222429211023016.

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

Ployhart, R. E., Holtz, B. C. & Bliese, P. D. (2002). Longitudinal data analysis: Applications of random coefficient modeling to leadership research. *Leadership Quarterly, 13,* 455-486.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.

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

Rupp, T. L., Wesensten, N. J., Bliese, P. D., & Balkin, T. J. (2009). Banking sleep: Realization of benefits during subsequent sleep restriction and recovery. *Sleep, 32*, 311-321.

Sherif, M. (1935). A study of some social factors in perception: Chapter 3. *Archives of Psychology, 27*, 23- 46.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 24*, 323-355.

Snijders, T. A. B. & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods and Research, 22*, 342-363.

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

Sobel, M. E., (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological Methodology 1982* (pp. 290-312). Washington, DC: American Sociological Association.

Stewart, G. L., Astrove, S. L., Reeves, C. J. , Crawford, E. R., & Solimeo, S. L. (2017). Those with the Most Find It Hardest to Share: Exploring Leader Resistance to the Implementation of Team-based Empowerment. *Academy of Management Journal, 60*, 2266–2293.

Tinsley, H. E. A., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgements. *Journal of Counseling Psychology, 22*, 358-376.

2 The Logic of Hierarchical Linear Models

- Preliminaries
- A General Model and Simpler Submodels
- Generalizations of the Basic Hierarchical Linear Model
- Choosing the Location of X and W (*Centering*)
- Summary of Terms and Notation Introduced in This Chapter

This chapter introduces the logic of hierarchical linear models. We begin with a simple example that builds upon the reader's understanding of familiar ideas from regression and analysis of variance (ANOVA). We show how these common statistical models can be viewed as special cases of the hierarchical linear model. The chapter concludes with a summary of some definitions and notation that are used throughout the book.

Preliminaries

A Study of the SES-Achievement Relationship in One School

We begin by considering the relationship between a single student-level predictor variable (say, socioeconomic status [SES]) and one student-level outcome variable (mathematics achievement) within a single, hypothetical school. Figure 2.1 provides a scatterplot of this relationship. The scatter of points is well represented by a straight line with intercept β_0 and slope β_1. Thus, the regression equation for the data is

$$Y_i = \beta_0 + \beta_1 X_i + r_i. \qquad [2.1]$$

The intercept, β_0, is defined as the expected math achievement of a student whose SES is zero. The slope, β_1, is the expected change in math achievement associated with a unit increase in SES. The error term, r_i, represents a unique effect associated with person i. Typically, we assume that r_i is normally distributed with a mean of zero and variance σ^2, that is, $r_i \sim \mathrm{N}(0, \sigma^2)$.

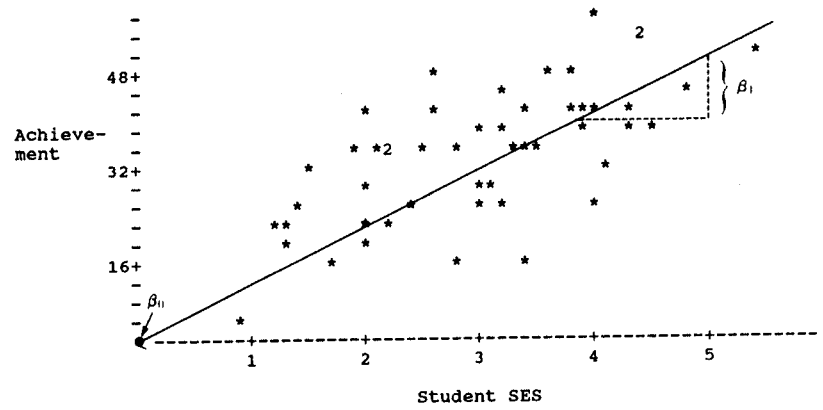


Figure 2.1. Scatterplot Showing the Relationship Between Achievement and SES in One Hypothetical School

It is often helpful to scale the independent variable, X, so that the intercept will be meaningful. For example, suppose we "center" SES by subtracting the mean SES from each score: $X_i - \overline{X}.$ where $\overline{X}.$ is the mean SES in the school. If we now plot Y_i as a function of $X_i - \overline{X}.$ (see Figure 2.2) with the

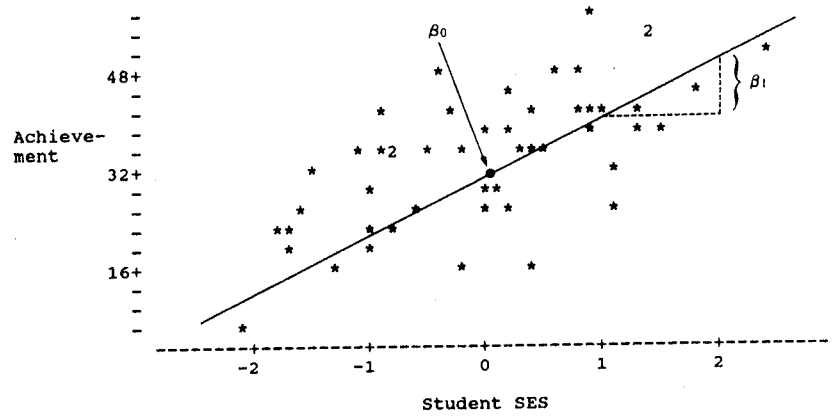


Figure 2.2. Scatterplot Showing the Relationship Between Achievement and SES (Centered) in One Hypothetical School

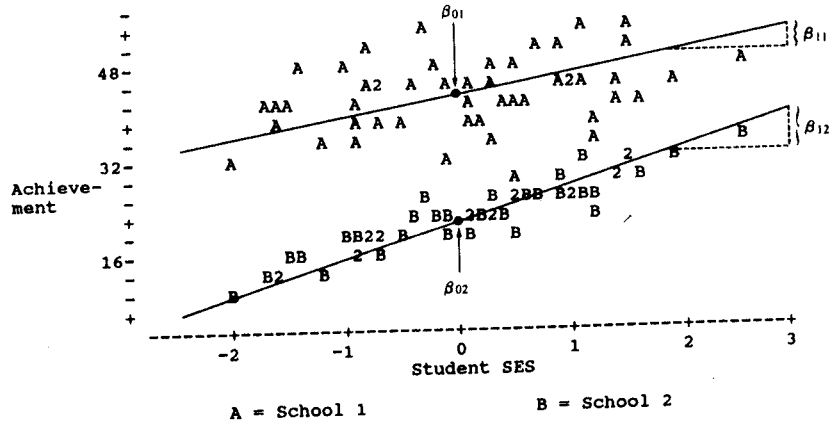


Figure 2.3. Scatterplot Showing the Relationship Between Achievement and SES Within Two Hypothetical Schools

regression line superimposed, we see that the intercept, β_0, is now the mean math achievement while the slope remains unchanged.

A Study of the SES-Achievement Relationship in Two Schools

Let us now consider separate regressions for two hypothetical schools. These are displayed in Figure 2.3. The two lines indicate that School 1 and School 2 differ in two ways. First, School 1 has a higher mean than does School 2. This difference is reflected in the two intercepts, that is, $\beta_{01} > \beta_{02}$. Second, SES is less predictive of math achievement in School 1 than in School 2, as indicated by comparing the two slopes, that is, $\beta_{11} < \beta_{12}$.

If students had been randomly assigned to the two schools, we could say that School 1 is both more "effective" and more "equitable" than School 2. The greater effectiveness is indicated by the higher mean level of achievement in School 1 (i.e., $\beta_{01} > \beta_{02}$). The greater equity is indicated by the weaker slope (i.e., $\beta_{11} < \beta_{12}$). Of course, students are not typically assigned at random to schools, so much interpretations of school effects are unwarranted without taking into account differences in student composition. Nevertheless, the assumption of random assignment clarifies the goals of the analysis and simplifies our presentation.

A Study of the SES-Achievement Relationship in J Schools

We now consider the study of the SES-math achievement relationship within an entire *population* of schools. Suppose that we now have a random

sample of J schools from a population, where J is a large number. It is no longer practical to summarize the data with a scatterplot for each school. Nevertheless, we can describe this relationship within any school j by the equation

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \overline{X}_{.j}) + r_{ij}, \qquad [2.2]$$

where for simplicity we assume that r_{ij} is normally distributed with homogeneous variance across schools, that is, $r_{ij} \sim \mathrm{N}(0, \sigma^2)$. Notice that the intercept and slope are now subscripted by j, which allows each school to have a unique intercept and slope. For each school, effectiveness and equity are described by the pair of values (β_{0j}, β_{1j}). It is often sensible and convenient to assume that the intercept and slope have a bivariate normal distribution across the population of schools. Let

$$\mathrm{E}(\beta_{0j}) = \gamma_0, \qquad \mathrm{Var}(\beta_{0j}) = \tau_{00},$$
$$\mathrm{E}(\beta_{1j}) = \gamma_1, \qquad \mathrm{Var}(\beta_{1j}) = \tau_{11},$$
$$\mathrm{Cov}(\beta_{0j}, \beta_{1j}) = \tau_{01},$$

where

- γ_0 is the average school mean for the population of schools;
- τ_{00} is the population variance among the school means;
- γ_1 is the average SES-achievement slope for the population;
- τ_{11} is the population variance among the slopes; and
- τ_{01} is the population covariance between slopes and intercepts.

A positive value of τ_{01} implies that schools with high means tend also to have positive slopes. Knowledge of these variances and of the covariance leads directly to a formula for calculating the population correlation between the means and slopes:

$$\rho(\beta_{0j}, \beta_{1j}) = \tau_{01}/(\tau_{00}\,\tau_{11})^{1/2}. \qquad [2.3]$$

In reality, we rarely know the true values of the population parameters we have introduced $(\gamma_0, \gamma_1, \tau_{11}, \tau_{00}, \tau_{01})$ nor of the true individual school means and slopes (β_{0j} and β_{1j}). Rather, all of these must be estimated from the data. Our focus in this chapter is simply to clarify the meaning of the parameters. The actual procedures used to estimate them are introduced in Chapter 3 and are discussed more extensively in Chapter 14.

Suppose we did know the true values of the means and slopes for each school. Figure 2.4 provides a scatterplot of the relationship between β_{0j} and

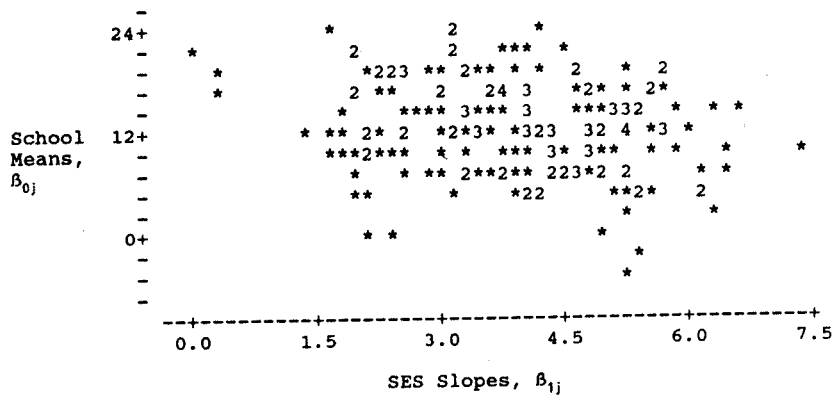


Figure 2.4. Plot of School Means (vertical axis) and SES Slopes (horizontal axis) for 200 Hypothetical Schools

β_{1j} for a hypothetical sample of schools. This plot tells us about how schools vary in terms of their means and slopes. Notice, for example, that there is more dispersion among the means (vertical axis) than the slopes (horizontal axis). Symbolically, this implies that $\tau_{00} > \tau_{11}$. Notice also that the two effects tend to be negatively correlated: Schools with high average achievement, β_{0j}, tend to have weak SES-achievement relationships, β_{1j}. Symbolically, $\tau_{01} < 0$. Schools that are effective and egalitarian—that is, with high average achievement (large values of β_{0j}) and weak SES effects (small values of β_{1j})—are found in the upper left quadrant of the scatterplot.

Having examined graphically how schools vary in terms of their intercepts and slopes, we may wish to develop a model to predict β_{0j} and β_{1j}. Specifically, we could use school characteristics (e.g., levels of funding, organizational features, policies) to predict effectiveness and equity. For instance, consider a simple indicator variable, W_j, which takes on a value of one for Catholic schools and a value of zero for public schools. Coleman, Hoffer, and Kilgore (1982) argued that W_j is positively related to effectiveness (Catholic schools have higher average achievement than do public schools) and negatively related to the slope (SES effects on math achievement are smaller in Catholic than in public schools). We represent these two hypotheses via two regression equations:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \qquad [2.4a]$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j}, \qquad [2.4b]$$

where

γ_{00} is the mean achievement for public schools;

γ_{01} is the mean achievement difference between Catholic and public schools (i.e., the Catholic school "effectiveness" advantage);

γ_{10} is the average SES-achievement slope in public schools;

γ_{11} is the mean difference in SES-achievement slopes between Catholic and public schools (i.e., the Catholic school "equity" advantage);

u_{0j} is the unique effect of school j on mean achievement holding W_j constant (or conditioning on W_j); and

u_{1j} is the unique effect of school j on the SES-achievement slope holding W_j constant (or conditioning on W_j).

We assume u_{0j} and u_{1j} are random variables with zero means, variances τ_{00} and τ_{11}, respectively, and covariance τ_{01}. Note these variance-covariance components are now *conditional* or *residual* variance-covariance components. That is, they represent the variability in β_{0j} and β_{1j} remaining after controlling for W_j.

It is not possible to estimate the parameters of these regression equations directly, because the outcomes (β_{0j}, β_{1j}) are not observed. However, the data contain information needed for this estimation. This becomes clear if we substitute Equations 2.4a and 2.4b into Equation 2.2, yielding the single prediction equation for the outcome

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{01} W_j + \gamma_{10}(X_{ij} - \overline{X}_{.j}) + \gamma_{11} W_j (X_{ij} - \overline{X}_{.j}) \\ & + u_{0j} + u_{1j}(X_{ij} - \overline{X}_{.j}) + r_{ij}. \end{aligned} \qquad [2.5]$$

Notice that Equation 2.5 is not the typical linear model assumed in standard ordinary least squares (OLS). Efficient estimation and accurate hypothesis testing based on OLS require that the random errors are independent, normally distributed, and have constant variance. In contrast, the random error in Equation 2.5 is of a more complex form, $u_{0j} + u_{1j}(X_{ij} - \overline{X}_{.j}) + r_{ij}$. Such errors are dependent within each school because the components u_{0j} and u_{1j} are common to every student within school j. The errors also have unequal variances, because $u_{0j} + u_{1j}(X_{ij} - \overline{X}_{.j})$ depend on u_{0j} and u_{1j}, which vary across schools, and on the value of $(X_{ij} - \overline{X}_{.j})$, which varies across students. Though standard regression analysis is inappropriate, such models can be estimated by iterative maximum likelihood procedures described in the

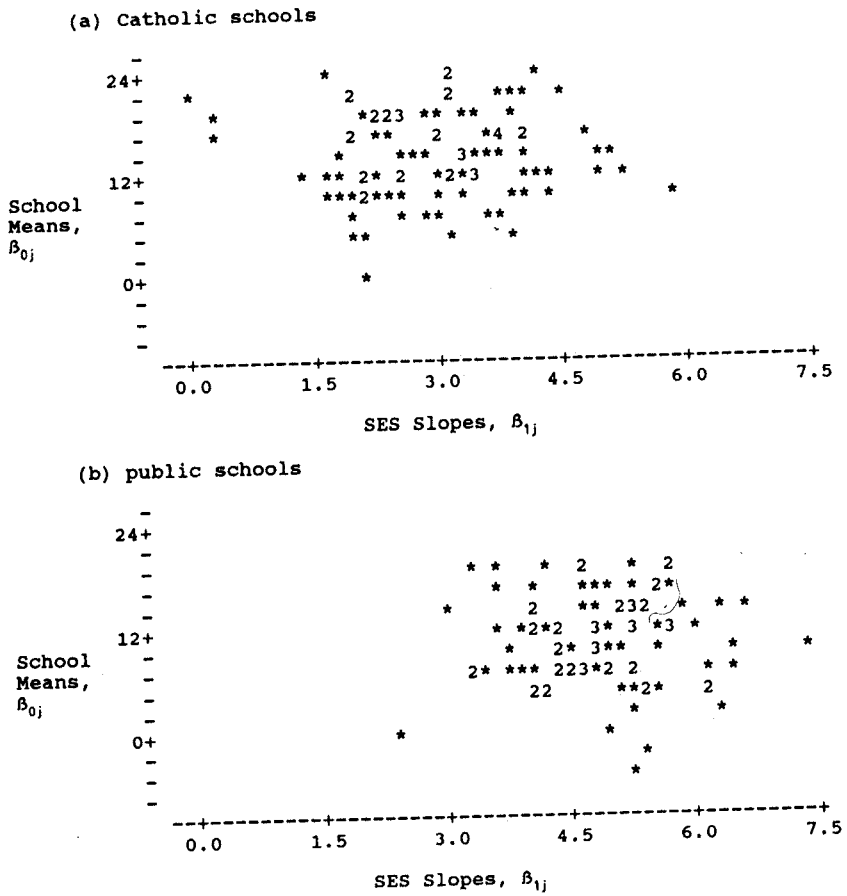


Figure 2.5. Plot of School Means (vertical axis) and SES Slopes (horizontal axis) for 100 Hypothetical Catholic Schools and 100 Hypothetical Public Schools

next chapter. We note that if u_{0j} and u_{1j} were null for every j, Equation 2.5 would be equivalent to an OLS regression model.

Figure 2.5 provides a graphical representation of the model specified in Equation 2.4. Here we see two hypothetical plots of the association between β_{0j} and β_{1j}, one for public and one for Catholic schools. The plots were constructed to reflect Coleman et al.'s (1982) contention that Catholic schools have both higher mean achievement and weaker SES effects than do the public schools.

A General Model and Simpler Submodels

We now generalize our terminology a bit so that it applies to any two-level hierarchical data structure. Equation 2.2 may be labeled the *level-1* model; Equation 2.4 is the *level-2* model, and Equation 2.5 is the *combined* model. In the school-effects application, the level-1 units are students and the level-2 units are schools. The errors r_{ij} are the level-1 random effects and the errors u_{0j} and u_{1j} are level-2 random effects. Moreover, $\text{Var}(r_{ij})$ is the level-1 variance, and $\text{Var}(u_{0j})$, $\text{Var}(u_{1j})$, and $\text{Cov}(u_{0j}, u_{1j})$ are the level-2 variance-covariance components. The β parameters in the level-1 model are level-1 coefficients and the γs are the level-2 coefficients.

Given a single level-1 predictor, X_{ij}, and a single level-2 predictor, W_j, the model given by Equations 2.2, 2.4, and 2.5 is the simplest example of a full hierarchical linear model. When certain sets of terms in this model are set equal to zero, we are left with a set of simpler models, some of which are quite familiar. It is instructive to examine these, both to demonstrate the range of applications of hierarchical linear models and to draw out the connections to more common data analysis methods. The submodels, running from the simpler to the more complex, include the one-way ANOVA model with random effects; a regression model with means-as-outcomes; a one-way analysis of covariance (ANCOVA) model with random effects; a random-coefficients regression model; a model with intercepts- and slopes-as-outcomes; and a model with nonrandomly varying slopes.

One-Way ANOVA with Random Effects

The simplest possible hierarchical linear model is equivalent to a one-way ANOVA with random effects. In this case, β_{1j} in the level-1 model is set to zero for all j, yielding

$$Y_{ij} = \beta_{0j} + r_{ij}. \qquad [2.6]$$

We assume that each level-1 error, r_{ij}, is normally distributed with a mean of zero and a constant level-1 variance, σ^2. Notice that this model predicts the outcome within each level-1 unit with just one level-2 parameter, the intercept, β_{0j}. In this case, β_{0j} is just the mean outcome for the jth unit. That is, $\beta_{0j} = \mu_{Y_j}$.

The level-2 model for the one-way ANOVA with random effects is Equation 2.4a with γ_{01} set to zero:

$$\beta_{0j} = \gamma_{00} + u_{0j}, \qquad [2.7]$$

where γ_{00} represents the grand-mean outcome in the population, and u_{0j} is the random effect associated with unit j and is assumed to have a mean of zero and variance τ_{00}.

Substituting Equation 2.7 into Equation 2.6 yields the combined model

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}, \qquad [2.8]$$

which is, indeed, the one-way ANOVA model with grand mean γ_{00}; with a group (level-2) effect, u_{0j}; and with a person (level-1) effect, r_{ij}. It is a random-effects model because the group effects are construed as random. Notice that the variance of the outcome is

$$\text{Var}(Y_{ij}) = \text{Var}(u_{0j} + r_{ij}) = \tau_{00} + \sigma^2. \qquad [2.9]$$

Estimating the one-way ANOVA model is often useful as a preliminary step in a hierarchical data analysis. It produces a point estimate and confidence interval for the grand mean, γ_{00}. More important, it provides information about the outcome variability at each of the two levels. The σ^2 parameter represents the within-group variability, and τ_{00} captures the between-group variability. We refer to the hierarchical model of Equations 2.6 and 2.7 as *fully unconditional* in that no predictors are specified at either level 1 or 2.

A useful parameter associated with the one-way random-effects ANOVA is the intraclass correlation coefficient. This coefficient is given by the formula

$$\rho = \tau_{00}/(\tau_{00} + \sigma^2) \qquad [2.10]$$

and measures the proportion of the variance in the outcome that is between the level-2 units. See Chapter 4 for an application of the one-way random-effects submodel.

Means-as-Outcomes Regression

Another common statistical problem involves the means from each of many groups as an outcome to be predicted by group characteristics. This submodel consists of Equation 2.6 as the level-1 model and, for the level-2 model,

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}, \qquad [2.11]$$

where in this simple case we have one level-2 predictor W_j. Substituting Equation 2.11 into Equation 2.6 yields the combined model:

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + u_{0j} + r_{ij}. \qquad [2.12]$$

We note that u_{0j} now has a different meaning as contrasted with that in Equation 2.7. Whereas the random variable u_{0j} had been the deviation of unit j's mean from the grand mean, it now represents the residual

$$u_{0j} = \beta_{0j} - \gamma_{00} - \gamma_{01} W_j.$$

Similarly, the variance in u_{0j}, τ_{00}, is now the residual or conditional variance in β_{0j} after controlling for W_j. The advantages of estimating Equation 2.12 rather than performing a standard regression using sample means-as-outcomes are discussed in Chapter 5.

One-Way ANCOVA with Random Effects

Referring again to the full model (Equations 2.2 and 2.4), let us constrain the level-2 coefficients γ_{01} and γ_{11} and the random effects u_{1j} (for all j) equal to 0. The resulting model would be a one-factor ANCOVA with random effects and a single level-1 predictor as a covariate. The level-1 model is Equation 2.2, but now the predictor X_{ij} is centered around the grand mean. That is,

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \overline{X}..) + r_{ij}. \qquad [2.13]$$

The level-2 model becomes

$$\beta_{0j} = \gamma_{00} + u_{0j}, \qquad [2.14a]$$

$$\beta_{1j} = \gamma_{10}. \qquad [2.14b]$$

Notice that the effect of X_{ij} is constrained to be the same fixed value for each level-2 unit as is indicated by Equation 2.14b.

The combined model becomes

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \overline{X}..) + u_{0j} + r_{ij}. \qquad [2.15]$$

The only difference between Equation 2.15 and the standard ANCOVA model (cf. Kirk, 1995, chap. 15) is that the group effect here, u_{0j}, is conceived as random rather than fixed. As in ANCOVA, γ_{10} is the pooled within-group regression coefficient of Y_{ij} on X_{ij}. Each β_{0j} is now the mean outcome for each level-2 unit adjusted for differences among these units in X_{ij}. Specifically, $\beta_{0j} = \mu_{Y_j} - \gamma_{10}(\overline{X}.j - \overline{X}..)$, where μ_{Y_j} is the mean outcome in school j. We also note that the $\text{Var}(r_{ij}) = \sigma^2$ is now a residual variance after adjusting for the level-1 covariate, X_{ij}.

An extension of the random-effects ANCOVA allows for the introduction of level-2 covariates. For example, if the coefficient γ_{01} is nonnull, the combined model becomes

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + \gamma_{10}(X_{ij} - \overline{X}..) + u_{0j} + r_{ij}. \qquad [2.16]$$

This model provides for a level-2 covariate, W_j, while also controlling for the effect of a level-1 covariate, X_{ij}, and the random effects of the level-2 units, u_{0j}. Interestingly, all of the parameters of Equation 2.16 can be estimated using the methods introduced in the next chapter. This is not the case, however, for a classical fixed-effects ANCOVA. Also, the classical ANCOVA model assumes that the covariate effect, γ_{10}, is identical for every group. This homogeneity of regression assumption is easily relaxed using the models described in the next three sections (for randomly varying and nonrandomly varying slopes). We illustrate use of the random-effects ANCOVA model in Chapter 5 in analyzing data on the effectiveness of an instructional innovation on students' writing.

Random-Coefficients Regression Model

All of the submodels discussed above are examples of *random-intercept models*. Only the level-1 intercept coefficient, β_{0j}, was viewed as random. The level-1 slope did not exist in the one-way ANOVA or the means-as-outcomes cases. In the random-effects ANCOVA model, β_{1j} was included but constrained to have a common effect for all groups.

A major class of applications of hierarchical linear models involves studies in which level-1 slopes are conceived as varying randomly over the population of level-2 units. The simplest case of this type is the random-coefficients regression model. In these models, both the level-1 intercept and one or more level-1 slopes vary randomly, but no attempt is made to predict this variation.

Specifically, the level-1 model is identical to Equation 2.2. The level-2 model is still a simplification of Equation 2.4 in that both γ_{01} and γ_{11} are constrained to be null. Hence, the level-2 model becomes

$$\beta_{0j} = \gamma_{00} + u_{0j}, \qquad [2.17a]$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \qquad [2.17b]$$

where

γ_{00} is the average intercept across the level-2 units;
γ_{10} is the average regression slope across the level-2 units;
u_{0j} is the unique increment to the intercept associated with level-2 unit j; and
u_{1j} is the unique increment to the slope associated with level-2 unit j.

We formally represent the dispersion of the level-2 random effects as a variance-covariance matrix:

$$\text{Var}\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} = \mathbf{T}, \qquad [2.18]$$

where

$\text{Var}(u_{0j}) = \tau_{00}$ = unconditional variance in the level-1 intercepts;
$\text{Var}(u_{1j}) = \tau_{11}$ = unconditional variance in the level-1 slopes; and
$\text{Cov}(u_{0j}, u_{1j}) = \tau_{01}$ = unconditional covariance between the level-1 intercepts and slopes.

Note that we refer to these as unconditional variance-covariance components because no level-2 predictors are included in either Equation 2.17a or 2.17b. Similarly, we refer to Equations 2.17a and 2.17b as an *unconditional* level-2 model.

Substitution of the expressions for β_{0j} and β_{1j} in Equations 2.17a and 2.17b into Equation 2.2 yields a combined model:

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \overline{X}_{.j}) + u_{0j} + u_{1j}(X_{ij} - \overline{X}_{.j}) + r_{ij}. \qquad [2.19]$$

This model implies that the outcome Y_{ij} is a function of the average regression equation, $\gamma_{00} + \gamma_{10}(X_{ij} - \overline{X}_{.j})$ plus a random error having three components: u_{0j}, the random effect of unit j on the mean; $u_{1j}(X_{ij} - \overline{X}_{.j})$, where u_{1j} is the random effect of unit j on the slope β_{1j}; and the level-1 error, r_{ij}.

Intercepts- and Slopes-as-Outcomes

The random-coefficients regression model allows us to estimate the variability in the regression coefficients (both intercepts and slopes) across the level-2 units. The next logical step is to model this variability. For example, in Chapter 4, we ask "What characteristics of schools (the level-2 units) help predict why some schools have higher means than others and why some schools have greater SES effects than others?"

Given one level-1 predictor, X_{ij}, and one level-2 predictor, W_j, these questions may be addressed by employing the "full model" of Equations 2.2 and 2.4. Of course, this model may be readily expanded to incorporate the effects of multiple Xs and of multiple Ws (see "Generalizations of the Basic Hierarchical Linear Model").

A Model with Nonrandomly Varying Slopes

In some cases, the analyst will prove quite successful in predicting the variability in the regression slopes, β_{1j}. For example, it might be found that the level-2 predictor W_j in Equation 2.4b does indeed predict the level-1 slope β_{1j}. In fact, the analyst might find that after controlling for W_j the residual variance of β_{1j} (i.e., the variance of the residuals, u_{1j} in Equation 2.4b) is very close to zero. The implication would be that once W_j is controlled, little or no variance in the slopes remains to be explained. For reasons of both statistical efficiency and computational stability (as discussed in Chapter 9), it would be sensible, then, to constrain the values of u_{1j} to be zero. This eliminates τ_{11}, the residual variance of the slope, and τ_{01}, the residual covariance between the slope and the intercept, as parameters to be estimated.

If the residuals u_{1j} in Equation 2.4b are indeed set to zero, the level-2 model for the slopes becomes

$$\beta_{1j} = \gamma_{10} + \gamma_{11} W_j, \qquad [2.20]$$

and this model, when combined with Equations 2.2 and 2.4a, yields the combined model

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{01} W_j + \gamma_{10}(X_{ij} - \overline{X}_{\cdot j}) \\ & + \gamma_{11} W_j (X_{ij} - \overline{X}_{\cdot j}) + u_{0j} + r_{ij}. \end{aligned} \qquad [2.21]$$

In this model, the slopes do vary from group to group, but their variation is nonrandom. Specifically, as Equation 2.20 shows, the slopes β_{1j} vary strictly as a function of W_j.

We note that Equation 2.21 can be viewed as another example of what we have called a random-intercept model, because β_{0j} is the only component that varies randomly across level-2 units. In general, hierarchical linear models may involve multiple level-1 predictors where any combination of random, nonrandomly varying, and fixed slopes can be specified.

Section Recap

We have been considering a simple hierarchical linear model with a single level-1 predictor, X_{ij}, and a single level-2 predictor, W_j. In this scenario, the level-1 model (Equation 2.2) defines two parameters, the intercept and the slope. At level 2, each of these may be predicted by W_j and each may have a random component of variation, as in Equations 2.4a and 2.4b. The resulting full model, summarized by Equation 2.5, is the most general model we have considered so far. If certain elements of the full model are constrained to be null, we are left with a submodel that may be useful either as preliminary to a full hierarchical analysis or as a more parsimonious summary than the full model.

The six submodels we have considered may be classified in several different ways. We have distinguished between random-intercept models and randomly varying slope models. The one-way random-effects ANOVA model, the means-as-outcomes model, the one-way ANCOVA model, and the model with nonrandomly varying slopes are all random-intercept models. In such models, the variance components are just the level-1 variance, σ^2, and the level-2 variance, τ_{00}. We noted that in the ANOVA and means-as-outcomes models, no level-1 slope exists. In the ANCOVA model, the level-1 slope exists but is constrained or fixed to be invariant across level-2 units. In the nonrandomly varying slope model, slopes were allowed to vary strictly as a function of a known W_j with no additional random component. In contrast, the random-coefficients model and the slopes- and intercepts-as-outcomes models allowed random variation for both the intercepts and slopes.

Another distinction is whether models include *cross-level interaction terms* such as $\gamma_{11} W_j (X_{ij} - \overline{X}_{\cdot j})$. In general, the combined model will include such cross-level interaction terms whenever we seek to predict variation in a slope. Such terms appear in two of our submodels: the intercepts- and slopes-as-outcomes model and the nonrandomly varying slope model.

Generalizations of the Basic Hierarchical Linear Model

Multiple *X*s and Multiple *W*s

Suppose now that the analyst wishes to use information about a second level-1 predictor. Let X_{1ij} denote the original X discussed above and let X_{2ij} denote the second level-1 predictor. For now, assume that there is still just a single level-2 predictor, W_j. The level-1 model, assuming group-mean centering for both X_{1ij} and X_{2ij}, becomes

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{1ij} - \overline{X}_{1\cdot j}) + \beta_{2j}(X_{2ij} - \overline{X}_{2\cdot j}) + r_{ij}. \qquad [2.22]$$

Again, we have three options for modeling β_{2j}. One option is that the effect of X_{2ij} is constrained to be invariant across level-2 units, implying

$$\beta_{2j} = \gamma_{20},$$

where γ_{20} is the common effect of X_{2ij} in every level-2 unit. We say that the effect of β_{2j} is *fixed* across level-2 units.

A second option would be to model the slope β_{2j} as a function of an average value, γ_{20}, plus a random effect associated with each level-2 unit:

$$\beta_{2j} = \gamma_{20} + u_{2j}. \qquad [2.23]$$

Here β_{2j} is *random*. Notice that Equation 2.23 specifies no predictors for β_{2j}. Suppose, however, that this slope depends on W_j. One might then formulate the slopes-as-outcomes model:

$$\beta_{2j} = \gamma_{20} + \gamma_{21}W_j + u_{2j}. \qquad [2.24]$$

According to this model, part of the variation of the slope β_{2j} can be predicted by W_j, but a random component, u_{2j}, remains unexplained. On the other hand, it may be that once the effect of W_j is taken into account, the residual variation in β_{2j}—that is, $\text{Var}(u_{2j}) = \tau_{22}$—is negligible. Then a model constraining that residual variation to be null would be sensible:

$$\beta_{2j} = \gamma_{20} + \gamma_{21}W_j. \qquad [2.25]$$

In this third case, β_{2j} is a *nonrandomly varying* slope because it varies strictly as a function of the predictor W_j.

So far we have been interested in just a single level-2 predictor, W_j. The introduction of multiple W_js is straightforward. Further, the level-2 model does not need to be identical for each equation. One set of W_js may apply for the intercept, a different set be used for β_{1j}, another set for β_{2j}, and so on. When nonparallel specification is employed, however, extra care must be exercised in the interpretation of the results (see Chapter 9).

Generalization of the Error Structures at Level 1 and Level 2

The model specified in Equations 2.2 and 2.4 assumes homogeneous errors at both level 1 and level 2. This assumption is quite acceptable for a broad class of multilevel problems. Most published applications have been based on this assumption, as are most of the examples discussed in Chapter 5 through 8.

The model can easily be extended, however, to more complex error structures at both levels. The level-1 variance might be different for each level-2 unit and denoted σ_j^2, or it might be a function of some measured level-1 characteristic. (The modeling framework for this extension appears in Chapter 5.) Similarly, at level 2, a different covariance structure might exist for distinct subsets of level-2 units. This would result in different T matrices estimated for different subsets of level-2 units.

Extensions Beyond the Basic Two-Level Hierarchical Linear Model

The core ideas introduced in this chapter in the context of two-level models extend directly to models with three or more levels. These extensions are described and illustrated in Chapter 8. A common feature of the basic hierarchical linear model, regardless of the number of levels, is that the outcome variable at level 1, Y, is continuous and assumed normally distributed, conditional on the level-1 predictors included in the model. Over the last decade, extensions beyond the basic hierarchical linear model framework have been advanced to include dichotomous level-1 outcomes, count data, and categorical outcomes. Models for missing data, latent variable effects, and more complex data designs, including crossed random effects, have also appeared. Although the estimation methods are more complex for these extensions, the basic conceptual ideas and modeling framework extend quite naturally. In general, the range of modeling possibilities is now much richer than when we authored the first edition of this book. Part III, which is new to the second edition, introduces these new developments.

Choosing the Location of X and W (*Centering*)

In all quantitative research, it is essential that the variables under study have precise meaning so that statistical results can be related to the theoretical concerns that motivate the research. In the case of hierarchical linear models, the intercept and slopes in the level-1 model become outcome variables at level 2. It is vital that the meaning of these outcome variables be clearly understood.

The meaning of the intercept in the level-1 model depends on the location of the level-1 predictor variables, the Xs. We know, for example, that in the simple model

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}, \qquad [2.26]$$

the intercept, β_{0j}, is defined as the expected outcome for a student attending school j who has a value of zero on X_{ij}. If the researcher is to make sense of models that account for variation in β_{0j}, the choice of a metric for all level-1 predictors becomes important. In particular, if an X_{ij} value of zero is not meaningful, then the researcher may want to transform X_{ij}, or "choose a location for X_{ij}" that will render β_{0j} more meaningful. In some cases, a proper choice of location will be required in order to ensure numerical stability in estimating hierarchical linear models.

Similarly, interpretations regarding the intercepts in the level-2 models (i.e., γ_{00} and γ_{10} in Equations 2.4a and 2.4b) depend on the location of the W_j variables. The numerical stability of estimation is not affected by the location for the Ws, but a suitable choice will ease interpretation of results. We describe below some common choices for the location of the Xs and Ws.

Location of the Xs

We consider four possibilities for the location of X: the natural X metric, centering around the grand mean, centering around the group mean, and other locations for X. We assume that X is measured on an interval scale. The case of dummy variables is considered separately.

The Natural X *Metric.* Although the natural X metric may be quite appropriate in some applications, in others this may lead to nonsensical results. For example, suppose X is a score on the Scholastic Aptitude Test (SAT), which ranges from 200 to 800. Then the intercept, β_{0j}, will be the expected outcome for a student in school j who had an SAT of zero. The β_{0j} parameter is meaningless in this instance because the minimum score on the test is 200. In such cases, the correlation between the intercept and slope will tend toward -1.0. As a result, the intercept is essentially determined by the slope. Schools with strong positive SAT-outcome slopes will tend to have very low intercepts. In contrast, schools where the SAT slope is negligible will tend to have much higher intercepts.

In some applications, of course, an X value of zero will in fact be meaningful. For example, if X is the dosage of an experimental drug, $X_{ij} = 0$ implies that subject i in group j had no exposure to the drug. As a result, the intercept β_{0j} is the expected outcome for such a subject. That is, $\beta_{0j} = \mathrm{E}(Y_{ij}|X_{ij} = 0)$. We wish to emphasize that it is always important to consider the meaning of $X_{ij} = 0$ because it determines the interpretation of β_{0j}.

Centering Around the Grand Mean. It is often useful to center the variable X around the grand mean, as discussed earlier (see "One-Way ANCOVA with Random Effects"). In this case, the level-1 predictors are of the form

$$(X_{ij} - \overline{X}..). \qquad [2.27]$$

Now, the intercept, β_{0j}, is the expected outcome for a subject whose value on X_{ij} is equal to the grand mean, $\overline{X}..$. This is the standard choice of location for X_{ij} in the classical ANCOVA model. As is the case in ANCOVA, grand-mean centering yields an intercept that can be interpreted as an adjusted mean for group j,

$$\beta_{0j} = \mu_{Y_j} - \beta_{1j}(\overline{X}.{}_j - \overline{X}..).$$

Similarly, the $\mathrm{Var}(\beta_{0j}) = \tau_{00}$ is the variance among the level-2 units in the adjusted means.

Centering Around the Level-2 Mean (Group-Mean Centering). Another option is to center the original predictors around their corresponding level-2 unit means:

$$(X_{ij} - \overline{X}.{}_j). \qquad [2.28]$$

In this case, the intercept β_{0j} becomes the unadjusted mean for group j. That is,

$$\beta_{0j} = \mu_{Y_j} \qquad [2.29]$$

and $\mathrm{Var}(\beta_{0j})$ is now just the variance among the level-2 unit means, μ_{Y_j}.

Other Locations for X. Specialized choices of location for X are often sensible. In some cases, the population mean for a predictor may be known and the investigator may wish to define the intercept β_{0j} as the expected outcome in group j for the "average person in the population." In this case, the level-1 predictor would be the original value of X_{ij} minus the population mean.

In applications of two-level hierarchical linear models to the study of growth, the data involve time-series observations so that the level-1 units are occasions and the level-2 units are persons. The investigator may wish to define the metric of the level-1 predictors such that the intercept is the expected outcome for person i at a specific time point of theoretical interest (e.g., entry to school). So long as the data encompass this time point, such a definition is quite appropriate. Examples of this sort are illustrated in Chapters 6 and 8.

Dummy Variables. Consider the familiar level-1 model

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}, \qquad [2.30]$$

where X_{ij} is now an indicator or dummy variable. Suppose, for example, that X_{ij} takes on a value of 1 if subject i in school j is a female and 0 if not. In this case, the intercept β_{0j} is defined as the expected outcome for a male student in group j (i.e., the predicted value for student with $X_{ij} = 0$). We note in this case that $\text{Var}(\beta_{0j}) = \tau_{00}$ will be the variance in the male outcome means across schools.

Although it may seem strange at first to center a level-1 dummy variable, this is appropriate and often quite useful. Suppose, for example, that the indicator variable for sex is centered around the grand mean, $\overline{X}..$. This centered predictor can take on two values. If the subject is female, $X_{ij} - \overline{X}..$ will equal the proportion of male students in the sample. If the subject is male, $X_{ij} - \overline{X}..$ will be equal to minus the proportion of female students. As in the case of continuous level-1 predictors centered around the respective grand means, the intercept, β_{0j}, is the adjusted mean outcome in unit j. In this case, it is adjusted for differences among units in the percentage of female students.

Alternatively, we might use group-mean centering. For females, $X_{ij} - \overline{X}_{.j}$ will take on the value equal to the proportion of male students in school j; for males, $X_{ij} - \overline{X}_{.j}$ will take on a value equal to minus the proportion of female students in school j. The fact that X_{ij} is a dummy variable does not change the interpretation given to β_{0j} when group-mean centering is employed. The intercept still represents the average outcome for unit j, μ_{Y_j}.

In sum, several locations of dichotomous predictors will produce meaningful intercepts. Again, it is incumbent on the researcher to take this location into account in interpreting results. Care is especially needed when there are multiple dummy variables. For example, in a school-effects study with indicators for whites, females, and students with preprimary education, the intercept for school j might be the expected outcome for a non-white male student with no preprimary experience. This may or may not be the intercept the investigator wants. Again we offer the general caveat—be conscious of the choice of location for each level-1 predictor because it has implications for interpretation of β_{0j}, $\text{Var}(\beta_{0j})$, and by implication, all of the covariances involving β_{0j}.

In general, sensible choices of location depend on the purposes of the research. No single rule covers all cases. It is important, however, that the researcher carefully consider choices of location in light of those purposes; and it is vital to keep the location in mind while interpreting results.

In addition, the choice of location for the level-1 predictors can, under certain circumstances, also influence the estimation of the level-2 variance-covariance components, $\mathbf{T}$, and random level-1 coefficients, β_{qj}. Complications can occur in the context of both organizational research and growth curve applications. The reader is referred to Chapters 5 and 6, respectively, for a further discussion of these technical considerations.

Location of *W*s

In general, the choice of location for the Ws is not as critical as for the level-1 predictors. Problems of numerical instability are less likely, except when cross-product terms are introduced at level 2 (e.g., a predictor set of the form W_{1j}, W_{2j}, and $W_{1j}W_{2j}$). All of the γ coefficients can be easily interpreted whatever choice of metric (or nonchoice) is made for level-2 predictors. Nevertheless, it is often convenient to center all of the level-2 predictors around their corresponding grand means, for example, $W_{1j} - \overline{W}_{1}$. .

Summary of Terms and Notation Introduced in This Chapter

A Simple Two-Level Model

Hierarchical form:

$$\begin{aligned} &\text{Level 1 (e.g., students)} && Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij}, \\ &\text{Level 2 (e.g., schools)} && \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}, \\ & && \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j}. \end{aligned}$$

Model in *combined form:*

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}X_{ij}W_j + u_{0j} + u_{1j}X_{ij} + r_{ij},$$

where we assume:

$$\text{E}(r_{ij}) = 0, \qquad \text{Var}(r_{ij}) = \sigma^2,$$

$$\text{E}\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \text{Var}\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix} = \mathbf{T},$$

$$\text{Cov}(u_{0j}, r_{ij}) = \text{Cov}(u_{1j}, r_{ij}) = 0.$$

Notation and Terminology Summary

There are $i = 1, \ldots, n_j$ level-1 units nested with $j = 1, \ldots, J$ level-2 units. We speak of student i nested within school j.

β_{0j}, β_{1j} are level-1 coefficients. These can be of three forms:

fixed level-1 coefficients (e.g., β_{1j} in the one-way random-effects ANCOVA model, Equation 2.14b)

nonrandomly varying level-1 coefficients (e.g., β_{1j} in the nonrandomly-varying-slopes model, Equation 2.20)

random level-1 coefficients (e.g., β_{0j} and β_{1j} in the random-coefficient regression model [Equations 2.17a and 2.17b] and in the intercepts- and slopes-as-outcomes model [Equations 2.4a and 2.4b])

$\gamma_{00}, \ldots, \gamma_{11}$ are level-2 coefficients and are also called fixed effects.
X_{ij} is a level-1 predictor (e.g., student social class, race, and ability).
W_j is a level-2 predictor (e.g., school size, sector, social composition).
r_{ij} is a level-1 random effect.
u_{0j}, u_{1j} are level-2 random effects.
σ^2 is the level-1 variance.
$\tau_{00}, \tau_{01}, \tau_{11}$ are level-2 variance-covariance components.

Some Definitions

Intraclass correlation coefficient (see "One-Way ANOVA with Random Effects"):

$$\rho = \tau_{00}/(\sigma^2 + \tau_{00}).$$

This coefficient measures the proportion of variance in the outcome that is between groups (i.e., the level-2 units). It is also sometimes called the *cluster effect*. It applies only to random-intercept models (i.e., $\tau_{11} = 0$).

Unconditional variance-covariance of β_{0j}, β_{1j} are the values of the level-2 variances and covariances based on the random-coefficient regression model.

Conditional or residual variance-covariance of β_{0j}, β_{1j} are the values of the level-2 variances and covariances after level-2 predictors have been added for β_{0j} and β_{1j} (see, e.g., Equations 2.4a and 2.4b).

Submodel Types

One-way random-effects ANOVA model involves no level-1 or level-2 predictors. We call this a *fully unconditional* model.

Random-intercept model has only one random level-1 coefficient, β_{0j}.

Means-as-outcomes regression model is one form of a random-intercept model.

One-way random-effects ANCOVA model is a classic ANCOVA model, except that the level-2 effects are viewed as random.

Random-coefficients regression model allows all level-1 coefficients to vary randomly. This model is *unconditional at level 2*.

Centering Definitions	**Implications for β_{0j}**
X_{ij} in the natural metric	$\beta_{0j} = \mathrm{E}(Y_{ij}\|X_{ij} = 0)$
$(X_{ij} - \overline{X}..)$ called grand-mean centering	$\beta_{0j} = \mu_{Y_j} - \beta_{1j}(X_{ij} - \overline{X}..)$ (i.e., adjusted level-2 means)
$(X_{ij} - \overline{X}._j)$ called group-mean centering	$\beta_{0j} = \mu_{Y_j}$ (i.e., level-2 means)
X_{ij} centered at some theoretically chosen location for X	$\beta_{0j} = \mathrm{E}(Y_{ij}\|X_{ij} =$ chosen centering location for $X)$

Advanced Quantitative Techniques in the Social Sciences

VOLUMES IN THE SERIES

1. **HIERARCHICAL LINEAR MODELS: Applications and Data Analysis Methods**
   Anthony S. Bryk and Stephen W. Raudenbush
2. **MULTIVARIATE ANALYSIS OF CATEGORICAL DATA: Theory**
   John P. Van de Geer
3. **MULTIVARIATE ANALYSIS OF CATEGORICAL DATA: Applications**
   John P. Van de Geer
4. **STATISTICAL MODELS FOR ORDINAL VARIABLES**
   Clifford C. Clogg and Edward S. Shihadeh
5. **FACET THEORY: Form and Content**
   Ingwer Borg and Samuel Shye
6. **LATENT CLASS AND DISCRETE LATENT TRAIT MODELS: Similarities and Differences**
   Ton Heinen
7. **REGRESSION MODELS FOR CATEGORICAL AND LIMITED DEPENDENT VARIABLES**
   J. Scott Long
8. **LOG-LINEAR MODELS FOR EVENT HISTORIES**
   Jeroen K. Vermunt
9. **MULTIVARIATE TAXOMETRIC PROCEDURES: Distinguishing Types From Continua**
   Niels G. Waller and Paul E. Meehl
10. **STRUCTURAL EQUATION MODELING: Foundations and Extensions**
    David Kaplan

Hierarchical Linear Models

Applications and Data Analysis Methods

Second Edition

Stephen W. Raudenbush
Anthony S. Bryk

Advanced Quantitative Techniques in the Social Sciences Series 1

Sage Publications
International Educational and Professional Publisher
Thousand Oaks ■ London ■ New Delhi

For information:

Sage Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

Sage Publications, Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

Sage Publications India Pvt. Ltd.
M-32 Market
Greater Kailash I
New Delhi 110 048 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

06 10 9 8 7 6

Acquiring Editor: C. Deborah Laughton
Editorial Assistant: Veronica Novak
Production Editor: Sanford Robinson
Typesetter/Designer: Technical Typesetting

Contents

PART III

MULTILEVEL ANALYSIS

Chapter 17

A PRIMER ON MULTILEVEL (RANDOM COEFFICIENT) REGRESSION MODELING

Levi K. Shiverdecker and James M. LeBreton

Hierarchical nesting is a fundamental property woven into the fabric of existence itself: from the microscopic organelles nested within our cells to the Milky Way galaxy nested within an infinitely expanding universe. Residing at a level somewhere in between, social scientists find themselves pondering questions relating to both hierarchical and temporal nesting. For example, clinical psychologists may want to investigate the efficacy of therapy for clients nested within different therapists. Similarly, organizational psychologists may want to examine the moderating role of a group-level phenomenon (e.g., team cohesion) on the relationship between person-level variables (e.g., employee-level burnout and employee-level turnover). Alternatively, developmental psychologists might wish to investigate within-person trends (e.g., growth and/or decline in cognitive abilities) across the lifespan. These hierarchically and temporally nested structures may be thought of as *multilevel* structures because they span multiple conceptual levels (e.g., repeated observations nested in persons nested in groups). The first part of this handbook focuses on issues related to multilevel constructs and multilevel theories. The second part of this handbook focuses on issues related to multilevel measurement and multilevel design. The third section of the handbook transitions to discussing various multilevel analytic tools and issues. The purpose of the current chapter is to provide the reader with a basic grounding in the classic multilevel regression (MLR) model.

To effectively tackle the analyses for research questions involving nested data structures, particular statistical analytic techniques must be employed to ensure that the subsequent results are unbiased and as consistent as possible to the true population parameters. Specifically, MLR has been, and continues to be, one of the most popular data analytic techniques used to test multilevel research questions and hypotheses. MLR is sometimes referred to by other names including hierarchical linear modeling, random coefficients regression (RCR), mixed effects modeling, mixed determinants modeling, or most commonly, multilevel modeling (MLM). We elected to use the MLR label rather than MLM in order to distinguish this approach to multilevel analysis from the other approaches presented in this handbook.

The onset of this chapter provides a brief introduction to MLR using an illustrative example that will be the focal example utilized for the subsequent sections of the chapter. The introduction to our illustrative example and the MLR analytic framework will be followed by a description of why MLR is necessary for nested data by examining how alternative methods may be inappropriate and how MLR circumvents the shortcomings of these alternatives. The chapter progresses into a step-by-step introduction to a model building/comparison

http://dx.doi.org/10.1037/0000115-018
The Handbook of Multilevel Theory, Measurement, and Analysis, S. E. Humphrey and J. M. LeBreton (Editors-in-Chief)

approach for using MLR analyses for hierarchically nested data.

Caveat. The models described in this chapter are applicable only when the dependent variable is measured at the lowest level (Level 1) in the nested structure. Independent variables may be measured at either the lower or at higher levels (Level 1, Level 2, etc.).

ILLUSTRATIVE EXAMPLE

One way to initially conceptualize MLR is to imagine taking a single-level research question and testing that research question across multiple samples. For the remainder of our chapter, we will rely on an illustrative example where we initially wish to test the hypothesis that there is a positive linear relationship between employees' levels of trait aggression (AGG; James & LeBreton, 2010, 2012) and their subsequent levels of counterproductive workplace behaviors (CWBs; e.g., harassment, lying, theft, sabotage; Bennett & Robinson, 2000). We generated a data set corresponding to data from 600 employees uniquely nested in 60 different teams. To simplify our presentation, the data are balanced with 10 employees assigned to each of the 60 teams.

We begin by regressing CWBs onto AGG using the data from the first 10 employees nested in Group 1. We repeat this analysis for each of the remaining groups, leading to 60 independent estimates of the regression coefficients (i.e., intercept and slope). Figures 17.1 through 17.4 provide a summary of the results for the first four groups. Table 17.1 contains the regression weights and R^2 estimates for each of the 60 groups and Figure 17.5 provides a visual representation of the 60 separate regressions superimposed on a single graph. Appendix 17.1 contains a copy of the R code used to generate all of the analyses and figures presented in this chapter. A brief review of Table 17.1 and Figures 17.1 through 17.5 reveals that there is substantial variability in the results across these 60 groups. For example, when trait AGG is zero (i.e., the minimum score on our survey), the predicted levels of employees' CWB (i.e., intercept coefficients) varies across the groups ranging from –0.37 to 3.15. Similarly, the strength of the relationship varies

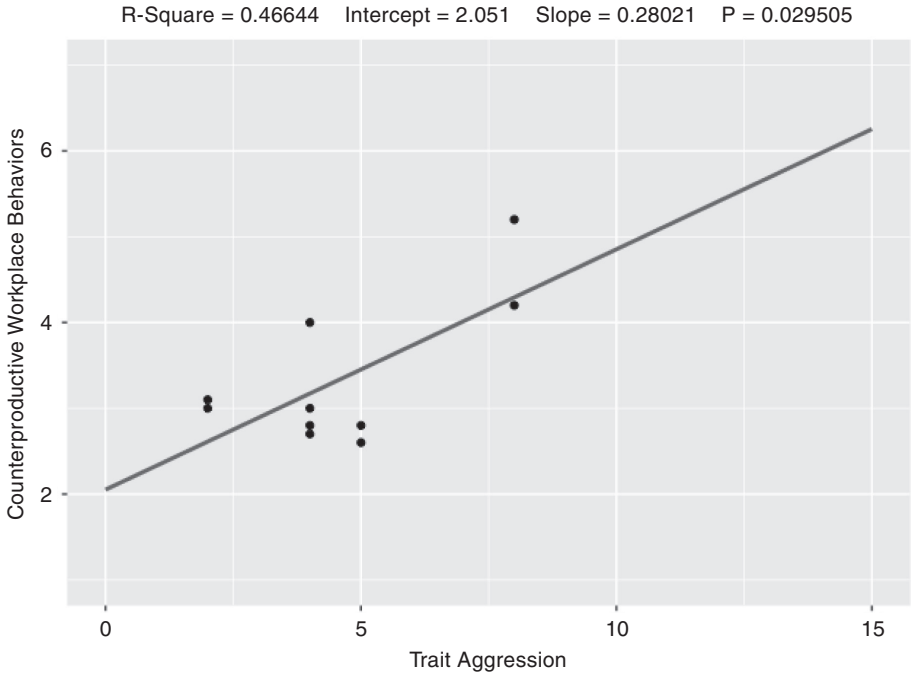


FIGURE 17.1. Simple linear regression of counterproductive workplace behaviors onto trait aggression using data for the 10 employees nested in Group 1.

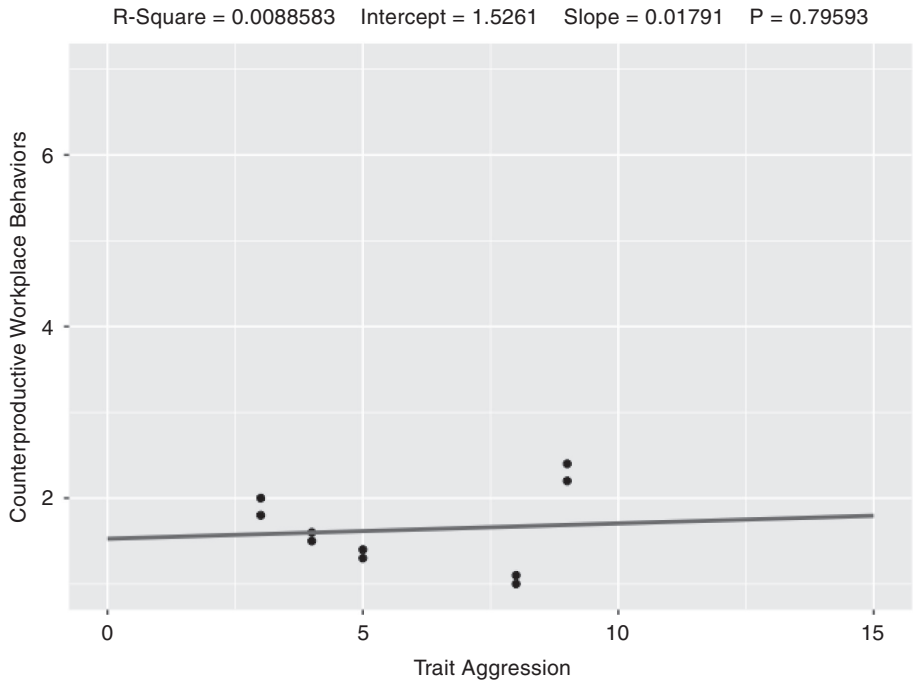


FIGURE 17.2. Simple linear regression of counterproductive workplace behaviors onto trait aggression using data for the 10 employees nested in Group 2.

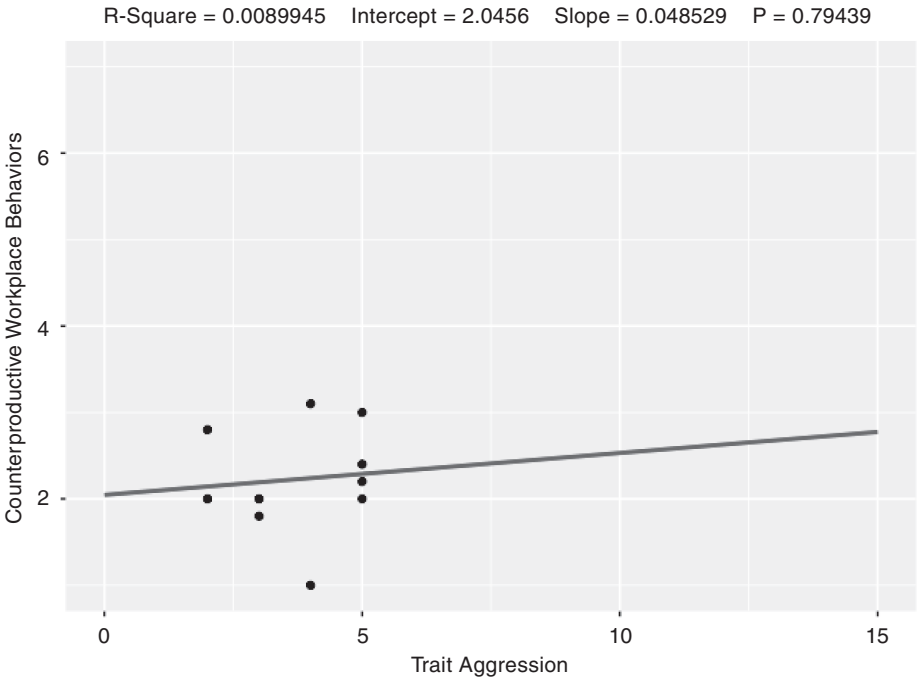


FIGURE 17.3. Simple linear regression of counterproductive workplace behaviors onto trait aggression using data for the 10 employees nested in Group 3.

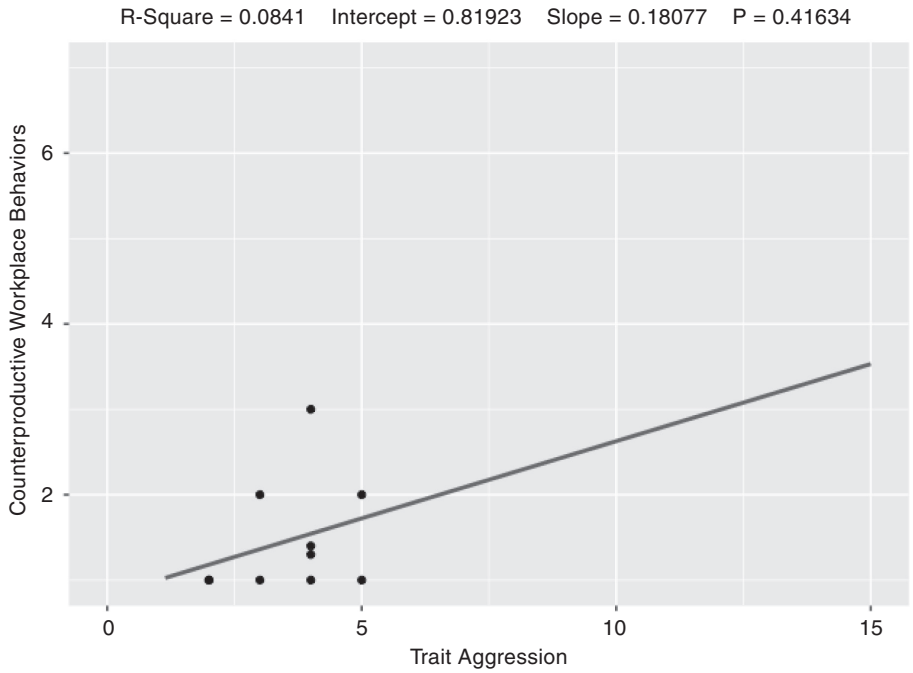


FIGURE 17.4. Simple linear regression of counterproductive workplace behaviors onto trait aggression using data for the 10 employees nested in Group 4.

TABLE 17.1

Results of Simple Linear Regression Analyses Repeated for the 60 Groups

Group	Intercept	Slope	R^2
1	2.051041667	0.280208333	0.4664359014
2	1.526119403	0.017910448	0.0088583358
3	2.045588235	0.048529412	0.0089944992
4	0.819230769	0.180769231	0.0840995184
5	1.906250000	–0.006250000	0.0009360599
6	1.441791045	0.011194030	0.0030809256
7	1.352941176	–0.011764706	0.0008650519
8	0.991428571	0.042857143	0.2008928571
9	2.685294118	–0.132352941	0.1153681812
10	1.072727273	0.403030303	0.6802415013
11	–0.065753425	0.526027397	0.6173426667
12	1.686363636	–0.022727273	0.0025826446
13	3.050000000	–0.316666667	0.2103729604
14	0.733644860	0.192990654	0.5455519520
15	0.144927536	0.585507246	0.2992724286
16	–0.370000000	0.450000000	0.3806390977
17	1.968421053	0.456578947	0.4242980577
18	1.180000000	0.075000000	0.1785714286
19	0.881250000	0.218750000	0.2734375000
20	6.250000000	–1.150000000	0.4467905405
21	0.600000000	0.200000000	0.5977011494

TABLE 17.1

Results of Simple Linear Regression Analyses Repeated for the 60 Groups (*Continued*)

Group	Intercept	Slope	R^2
22	1.471698113	0.037735849	0.0314465409
23	2.764705882	–0.376470588	0.8366013072
24	–0.600000000	0.550000000	0.8897058824
25	1.375852273	0.076988636	0.1862018782
26	2.718518519	–0.237037037	0.3108682453
27	2.023076923	–0.053846154	0.0088066139
28	1.200000000	0.200000000	0.1397849462
29	1.127777778	0.138888889	0.0964506173
30	1.267924528	0.109433962	0.1823899371
31	0.719230769	0.219230769	0.2431158336
32	1.626415094	–0.015094340	0.0058055152
33	1.157142857	0.328571429	0.1420515575
34	1.925000000	–0.025000000	0.0125000000
35	0.978378378	0.237837838	0.2198501022
36	1.448648649	–0.035135135	0.0496474736
37	0.727058824	0.270588235	0.4227941176
38	1.360000000	0.080000000	0.0421052632
39	3.100000000	–0.212500000	0.1619955157
40	1.066666667	–0.008333333	0.0104166667
41	1.672727273	–0.018181818	0.0014204545
42	0.873913043	0.082608696	0.0421926134
43	0.695652174	0.313043478	0.2860295740
44	0.991304348	0.026086957	0.0489130435
45	2.900000000	–0.550000000	0.7438524590
46	0.813235294	0.704411765	0.5582152954
47	3.145454545	–0.345454545	0.2369543814
48	0.900000000	0.250000000	0.3409090909
49	1.337096774	0.111290323	0.2666330645
50	1.561194030	0.216417910	0.2120310609
51	2.884210526	–0.273684211	0.2928308425
52	1.610000000	0.150000000	0.0757575758
53	0.561111111	0.181944444	0.7696896304
54	1.854022989	0.101149425	0.0519168820
55	0.542718447	0.190291262	0.7698057250
56	0.908108108	0.216216216	0.3931203931
57	1.849411765	0.094117647	0.1742919390
58	0.003030303	0.453030303	0.2730418476
59	1.305000000	0.075000000	0.0546116505
60	–0.088888889	0.459259259	0.8043523750

across the groups with slopes ranging from –0.55 (and an R^2 of 0.00) to 0.70 (and an R^2 of 0.84).

Now, obviously, much of the variability is driven by sampling error. After all, we are conducting regression analyses using the samples size of $N = 10$. However, we also suspect that some of the variance may be attributed to differences in the social relationships that have formed within these groups. Specifically, we hypothesize that scores on a measure of group cohesiveness (COH) might be negatively related to CWBs even after controlling for employee-level AGG. Essentially, we predict that employees working in a highly cohesive group will be less likely to engage in CWBs than those working in fragmented or uncohesive groups. In addition, we hypothesize that the variability in slopes observed in

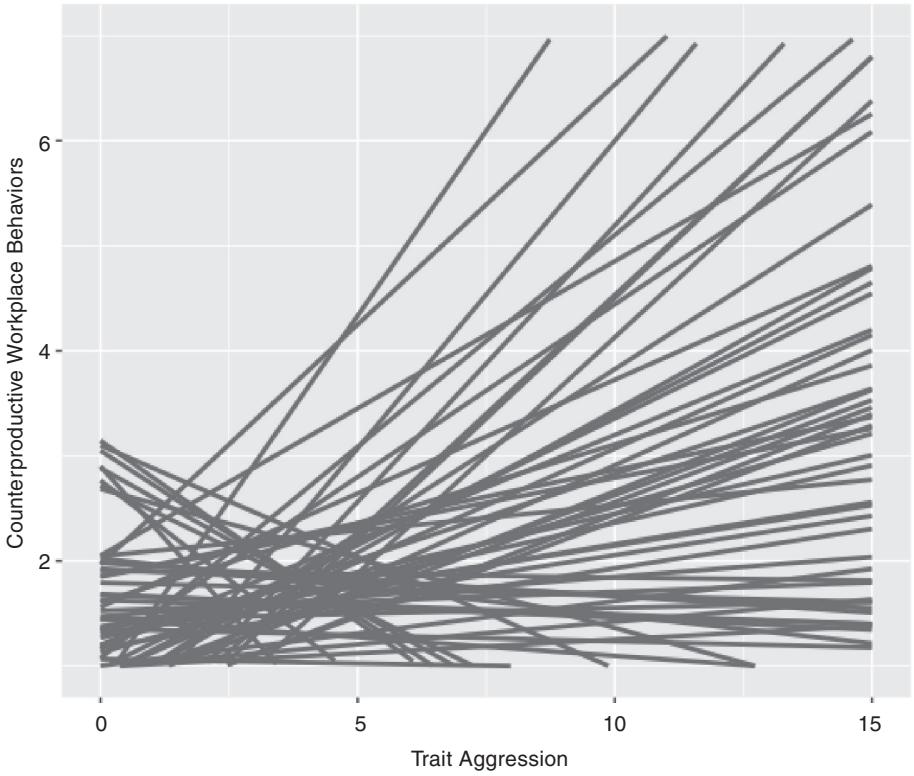


FIGURE 17.5. Simple linear regression of counterproductive workplace behaviors onto trait aggression for each of the 60 groups.

Figure 17.5 may also be related to COH. Specifically, we hypothesize weaker (positive) relationships between CWBs and AGG for highly cohesive groups, but this (positive) relationship will grow stronger as teams become less cohesive. Conceptually, we are trying to use a group-level variable (COH) to explain some of the between-groups variability in the regression coefficients (see Table 17.1 and Figure 17.5).

Stated formally, we plan to test the following hypotheses:

1. There is a positive relationship between individual-level trait AGG and individual-level CWBs.
2. There is a negative relationship between group-level COH and individual-level CWBs.
3. Group-level COH will moderate the strength of the relationship between individual-level trait AGG and individual-level CWBs, such that the relationship will become weaker as group-level COH increases.

For this example, we will assume that cohesion is operationalized as the shared perception among employees nested in work groups. It is presumed that the cohesion data were collected at the individual level and aggregated (i.e., averaged) to the group level after establishing that sufficient agreement among employees existed within the work groups (see Chapter 12, this volume; also cf. James, Demaree, & Wolf, 1984; 1993; LeBreton & Senter, 2008). The remainder of this chapter discusses how a researcher could use our sample data to test the above hypotheses. Like other treatments of MLR, we first explain the necessity of MLR for testing these hypotheses and then adopt a model-building procedure to formally test our hypotheses.

NONINDEPENDENCE AND MULTILEVEL REGRESSION

One might speculate whether it is necessary to use more complicated multilevel techniques when it appears that simpler techniques (e.g., ordinary

least squares [OLS] regression or analysis of covariance [ANCOVA]) may be suited for testing hypotheses such as those noted earlier in the chapter. For example, we could test a regression model that includes a term representing AGG, a term representing COH, and a term representing the cross-product of AGG with COH (AGG∗COH). Alternatively, we could simply add the group-level cohesion scores to the data in Table 17.1 and then use those scores to predict the regression coefficients. That is, we could examine whether the variability in COH overlaps with the variability in intercepts (β_0) and slopes (β_1). At first blush, either of these strategies seems to be a reasonable approach to testing our hypotheses. However, as we will see, these approaches fail to properly account for the fact that employees are nested within groups.

OLS is a very powerful, simple, and effective estimation procedure. For example, when errors (a) have a mean of zero, (b) are identically distributed, and (c) are uncorrelated with one another, the OLS estimates are said to be "BLUE"—the Best Linear Unbiased Estimates (Cohen, Cohen, West, & Aiken, 2003; Myers, 1990). If the errors are also normally distributed, the OLS estimates are said to be "MVUE"—the minimum variance unbiased estimates (Cohen et al., 2003; Myers, 1990). Unfortunately, when data are nested, the errors are frequently nonindependent. This nonindependence is visually depicted in Figure 17.6, which contains an analysis of the data from the first two groups in our example. When we separately estimate regression lines for each group, we see that Group 1 (thin dashed line) has a very steep slope, whereas Group 2 (thin dot-dash line) is flatter. However, if we ignored group membership and simply analyzed the data from all 20 employees, we would obtain a common regression line (thick solid line). The fact that these data are nonindependent is easily observed when considering the implications of using the common regression line to predict CWBs using AGG scores. Essentially, using the common line will tend to underestimate the predicted scores for employees working in Group 1 and overestimate the predicted scores for employees working in Group 2. The problem of nonindependence may also be

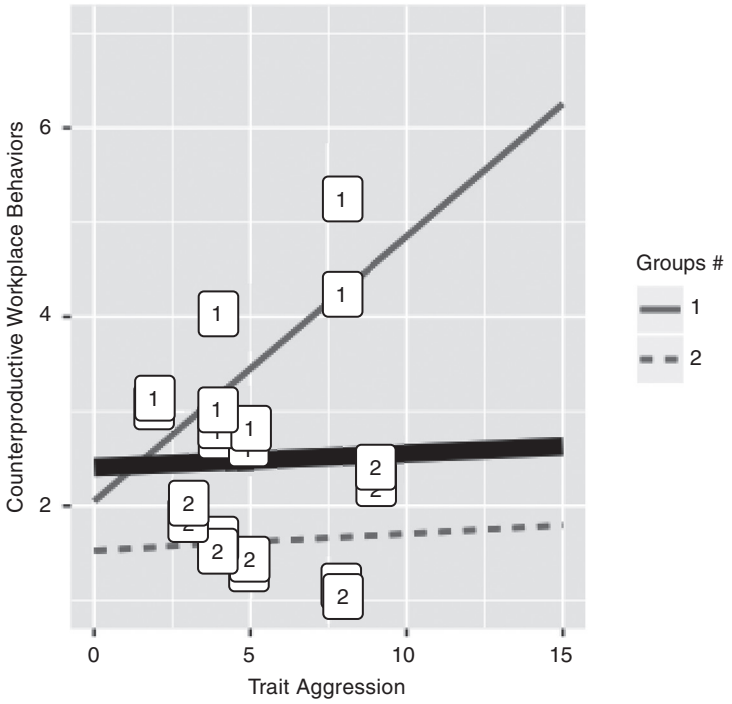


FIGURE 17.6. Simple linear regression of counterproductive workplace behaviors onto trait aggression using data from the first two groups.

observed by comparing the equations that would be tested using OLS to the equation that we will eventually test using MLR.

Single-Level Regression Approach

To illustrate the differences between the single-level OLS approach and the MLR approach we will examine the equations that each approach would use for testing a common "main effects" model. The OLS approach essentially relies on an analysis of covariance (James & Williams, 2000; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002):

$$CWB_{ij} = \beta_0 + \beta_1 (AGG_{ij}) + \beta_2 (COH_j) + e_{ij}, \quad (17.1)$$

where $i = 1$ to n_j Level-1 units (e.g., employees; children) nested within the $j = 1$ to J Level-2 units (e.g., teams; classrooms). Thus, in our example, AGG_{ij} refers to the trait aggression score for the ith employee working in the jth group and COH_j refers to the cohesion score for group j, which is computed as the mean taken over the n_j employees nested in that group. The βs represent unstandardized regression coefficients and e_{ij} represent the unique error for employee i nested in group j.

To summarize, the OLS approach is represented by Equation 17.1 and consists of our outcome variable, a single (fixed) intercept coefficient, two fixed slope coefficients, and a single error term. As we will see, reliance on a single error term is problematic because we are unable to properly model potential dependencies in the data attributed to the nesting of the Level-1 units within the Level-2 units. As a result, estimates of error variance may be biased, resulting in biased standard errors and tests of statistical significance. To address this issue, the MLR approach attempts to disentangle or partition the error variance into variance that resides at the individual level (i.e., within-group variance) and variance that resides at the group level (i.e., between-group variance).

Multilevel Regression Approach

The main effects model using MLR is a bit different (Hox, 2010; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002):

$$\text{Level 1: } CWB_{ij} = \beta_{0j} + \beta_{1j} (AGG_{ij}) + r_{ij} \quad (17.2)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01} (COH_j) + U_{0j} \quad (17.3)$$

$$\beta_{1j} = \gamma_{10}, \quad (17.4)$$

where β represent Level-1 *random* regression coefficients (i.e., coefficients that may vary across groups), γ represent the Level-2 *fixed* regression coefficients (i.e., coefficients that are invariant across groups), r_{ij} is the Level-1 random effect, and U_{0j} is the Level-2 random effect. Substituting the Level-2 equations into the Level-1 equations, we obtain the mixed equation

$$CWB_{1j} = [\gamma_{00} + \gamma_{01} (COH_j) + U_{0j}] + [\gamma_{10} (AGG_{ij})] + r_{ij}. \quad (17.5)$$

Returning to Equation 17.5, the elements in the first set of brackets provide information about the unique intercept for group j (i.e., β_{0j}). This intercept is a function of (a) a fixed intercept coefficient (γ_{00}) representing a common or pooled intercept applied to all individuals in all groups, (b) the effect of group j's score on COH_j (γ_{01}), and (c) a unique effect (i.e., error) that is applied to only the members of group j (U_{0j}). Because each group will have a unique, group-level effect (i.e., each group has a different score on U_{0j}), the variability in these group-level scores represents the between-groups variability in intercepts (i.e., the extent that groups assume different intercept values, even after modeling COH_j and AGG_{ij}).

The elements in the second bracket provide information about the unique slope for group j (i.e., β_{1j}). In the "main effects only" model, the unique slope for group j isn't really unique, but instead is fixed to common value across all groups: γ_{10}. Later in the chapter, we illustrate how it is possible to estimate separate group-level slopes. To further the comparison with OLS, Equation 17.5 may be rewritten as

$$CWB_{ij} = \gamma_{00} + \gamma_{10} (AGG_{ij}) + \gamma_{01} (COH_j) + U_{0j} + r_{ij}. \quad (17.6)$$

Comparison of Equations 17.1 and 17.6 reveals the similarities and differences between these approaches. In terms of similarities, both approaches furnish estimates of three (fixed) regression coefficients. These coefficients represent (a) a fixed intercept (i.e., β_0 for OLS; γ_{00} for MLR), (b) a fixed slope representing the

strength of the relationship between CWB and AGG (i.e., β_1 for OLS; γ_{10} for MLR), and (c) a fixed slope representing the cross-level effect of COH on CWB (i.e., β_2 for OLS; γ_{01} for MLR).

The fundamental difference is found in the error terms. The OLS approach contains a single error term (i.e., e_{ij}), whereas the MLR approach includes two error terms—one that is unique to the employee (i.e., r_{ij}) and one that is applied to all of the employees nested within a common group (i.e., U_{0j}). Thus, the MLR approach partitions errors into a component that resides within groups (i.e., r_{ij}) and a component that resides between groups (i.e., U_{0j}). By appropriately partitioning these errors, we are able to obtain unbiased estimates of the standard errors and thus unbiased tests of statistical significance (Bliese & Hanges, 2004; Raudenbush & Bryk, 2002).

We now turn to a detailed, step-by-step tutorial on how researchers may go about building and testing models within an MLR framework. To facilitate our tutorial, we will analyze and interpret data corresponding to the illustrative example presented earlier in the chapter.

MULTILEVEL REGRESSION: A MODEL-BUILDING APPROACH

With a basic grasp of the tenets underlying MLR in place, we now turn to our illustrative example and our tutorial on MLR. Our tutorial adopts a model-building/model comparison perspective (Aguinis & Culpepper, 2015; Bliese & Ployhart, 2002; Hofmann, 1997; Hofmann, Griffin, & Gavin, 2000) and is structured as follows. First, for each model, we provide a general introduction/overview of the structural equations to be tested. Next, we discuss the interpretation of the regression coefficients and variance components associated with each model. Finally, we test each model using our illustrative example data, and provide a brief interpretation of the results.

Model 1: Null Model or the Unconditional ANOVA Model

Overview. Researchers with nested data are encouraged to test the extent to which their data may violate the independence assumption underlying the use of OLS. This is accomplished by conducting a simple, one-way ANOVA on the dependent variable (e.g., CWB_{ij}), where we treat group membership as a "grouping" variable or factor in a one-way random effects ANOVA. The Null model is so named because it does not include any Level-1 or Level-2 predictor variables. As we illustrate, the Null model partitions the variance in the outcome or dependent variable into a component that resides within groups and a separate component that resides between groups. If the nesting of Level-1 units (e.g., employees) within Level-2 units (e.g., groups or teams) violates the independence assumption, we would expect to see that group membership explains a nontrivial amount of variance in our dependent variable. Using the notation of Raudenbush and Bryk (2002), the Null model may be presented as

$$\text{Level 1: } CWB_{ij} = \beta_{0j} + r_{ij} \qquad (17.7a)$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + U_{0j}, \qquad (17.8a)$$

where, β_{0j} = the mean CWB score for group j, γ_{00} = grand mean score on CWB based on data from all individuals in all groups, r_{ij} = Level-1 residual for person *i* in group *j* (i.e., the deviation between a person's CWB score and his or her group's mean), and U_{0j} = Level-2 residual for group *j* (i.e., the deviation between the mean for group *j* and the grand mean). Substituting Equation 17.8 into Equation 17.7 yields the "mixed" equation:

$$\text{Mixed: } CWB_{ij} = \gamma_{00} + U_{0j} + r_{ij}. \qquad (17.9a)$$

This equation states that the CWB score for person *i* nested in group *j* is a function of (a) a common or grand mean, (b) the extent to which their group mean deviates from the grand mean, and (c) the extent to which their individual score deviates from the group mean.

Raudenbush and Bryk (2002) noted that Equation 17.8 represents a one-way random effects ANOVA model because "the group effects are construed as random" (p. 24). And the

total variance (VAR) in CWBs may be partitioned thusly:

$$\mathrm{VAR}(\mathrm{CWB}_{ij}) = \mathrm{VAR}(\gamma_{00} + U_{0j} + r_{ij}) \quad (17.10a)$$

$$\mathrm{VAR}(\mathrm{CWB}_{ij}) = \mathrm{VAR}(\gamma_{00}) + \mathrm{VAR}(U_{0j}) + \mathrm{VAR}(r_{ij}). \quad (17.10b)$$

Because the grand mean is a constant (i.e., it has no variance), it drops out of the equation:

$$\mathrm{VAR}(\mathrm{CWB}_{ij}) = \mathrm{VAR}(U_{0j} + r_{ij}). \quad (17.11)$$

And, assuming that the data meet requisite assumptions concerning the independence of Level-1 and Level-2 errors (see Raudenbush & Bryk, 2002):

$$\mathrm{VAR}(\mathrm{CWB}_{ij}) = \tau_{00} + \sigma^2, \quad (17.12a)$$

where τ_{00} = the variance in CWB scores that resides between groups (i.e., the variance in U_{0j}) and σ^2 = to the variance in CWB scores that resides within groups (i.e., the variance in r_{ij}). If you are like us, the first time you see these equations, you may have a difficult time making the connection between the MLR notation and the basic ANOVA model that is purportedly being tested using Equations 17.7a–17.9a). However, the connection becomes clearer with a slightly different representation of these equations.

Alternative notation. First, we will replace CWB_{ij} with the more commonly used Y_{ij} and rewrite Equations 17.7 and 17.8 using a conventional ANOVA notation:

$$\text{Level 1: } Y_{ij} = \bar{Y}_j + r_{ij} \quad (17.7b)$$

$$\text{Level 2: } \bar{Y}_j = \bar{\bar{Y}} + U_{0j}, \quad (17.8b)$$

where, $\bar{Y}_j$ = the mean outcome score for group j, $\bar{\bar{Y}}$ = grand mean score on the outcome based on scores from all individuals in all groups, r_{ij} = Level-1 residual for person i in group j (i.e., the deviation between a person's outcome score and his or her group's mean; $(Y_{ij} - \bar{Y}_j)$), and U_{0j} = Level-2 residual for group j (i.e., the deviation between the mean for group j and the grand mean; $(\bar{Y}_j - \bar{\bar{Y}})$. Substituting Equation 17.8a into Equation 17.7a yields the "mixed" equation:

$$\text{Mixed: } Y_{ij} = \bar{\bar{Y}} + U_{0j} + r_{ij}, \quad (17.9b)$$

which is equivalent to

$$\text{Mixed: } Y_{ij} = \bar{\bar{Y}} + \left(\bar{Y}_j - \bar{\bar{Y}}\right) + \left(Y_{ij} - \bar{Y}_j\right). \quad (17.9b2)$$

The first step in estimating the variance in Y_{ij} is to compute the sum of squared deviations between Y_{ij} and the grand mean, $\bar{\bar{Y}}$:

$$\sum\left(Y_{ij} - \bar{\bar{Y}}\right)^2 = \sum\left(\left[\bar{\bar{Y}} + \left(\bar{Y}_j - \bar{\bar{Y}}\right) + \left(Y_{ij} - \bar{Y}_j\right) - \bar{\bar{Y}}\right]\right)^2 \quad (17.9b3)$$

$$\sum\left(Y_{ij} - \bar{\bar{Y}}\right)^2 = \sum\left(\bar{Y}_j - \bar{\bar{Y}}\right)^2 + \sum\left(Y_{ij} - \bar{Y}_j\right)^2. \quad (17.9b4)$$

We then divide these sums by the appropriate degrees of freedom to obtain estimates of means squares and eventually variance components:

$$\frac{\sum\left(Y_{ij} - \bar{\bar{Y}}\right)^2}{N-1} = \frac{\sum\left(\bar{Y}_j - \bar{\bar{Y}}\right)^2}{J-1} + \frac{\sum\left(Y_{ij} - \bar{Y}_j\right)^2}{N-J} \quad (17.9b4)$$

$$\mathrm{VAR}(Y_{ij}) = \tau_{00} + \sigma^2. \quad (17.12b)$$

Interpreting the regression coefficients. Another way to approach the interpretation of the multilevel notation is to consider the meaning of the regression coefficients presented in Equations 17.7a through 17.9a. Equation 17.7a is simply stating that, absent any other information about the i individuals nested in group j, the best "estimate" for each person's outcome scores (i.e., CWB_{ij}) will be the mean CWB score for group j (β_{0j} in Equation 17.7a and $\bar{Y}_j$ in Equation 17.7b). The group mean will be an imperfect estimate of the individual scores, and this imperfection is manifested as error (r_{ij}). Equation 17.8a is simply stating that, absent any other information about the j groups, the best "estimate" for each group mean (denoted β_{0j} in Equation 17.8a and $\bar{Y}_j$ in Equation 17.8b) is the grand mean (denoted γ_{00} in Equation 17.8a and $\bar{\bar{Y}}$in Equation 17.8b). The grand

mean will be an imperfect estimate of the group means and this imperfection is manifested as error (U_{0j}).

Variance decomposition and interpretation. The variance components computed from the Null model allow us to estimate the proportion of variance in CWBs that resides between and within groups. This variance ratio will enable us to determine if the nesting of Level-1 units within Level-2 units has resulted in a violation of the independence assumption. Formally, this variance ratio is defined as the intraclass correlation obtained from a one-way random effects ANOVA (Bliese, 2000; James, 1982; Raudenbush & Bryk, 2002):

$$ICC(1) = \frac{\tau_{00}}{\tau_{00} + \sigma^2}. \qquad (17.13)$$

Returning to our original hypotheses, we are hoping to predict a portion of the within-groups (i.e., employee-level) variance in CWBs using our employee-level predictor (AGG) and we are hoping to predict some of the variability in group-level intercepts and slopes using our group-level predictor (COH). Thus, it is important that we establish that the variance in CWBs exists both within and between groups. In other words, if there is no meaningful within-group variance in CWBs, it is unlikely that AGG will emerge as a significant predictor of CWBs. Likewise, if CWB scores do not vary between groups, it is unlikely that COH will emerge as a significant predictor of CWBs.

Model 1: Illustrative example. Appendix 17.1 contains the R code used to analyze the data in our illustrative example (R Development Core Team, 2017). The opening lines of Appendix 17.1 address data management issues and the installation of the R packages we use in our data analysis. For most of our analyses, we use predefined functions that are part of the *multilevel* package and the *nlme* package. Specifically, we rely heavily on the linear mixed effects (*lme*) function that we use to test the Null model presented in Equations 17.7 and 17.8. The *lme* function requires the user to identify the outcome variable, specify which effects will be treated as fixed and which will be treated as random, and to specify the grouping (nesting) variable corresponding to those random effects.

Fixed effects refer to the effects in Equations 17.7a and 17.8a that do not vary and reside at Level 2. In our example, the fixed effects are represented using γ. In contrast, the random effects are the effects that are allowed to vary across groups (e.g., U_{0j}) and within groups (e.g., r_{ij}). The R code to estimate the Null model is

```
Null.Model = lme(CWB~1, random =
    ~1|GROUP, data = mlr)
```

The above code accomplishes the following:

- It creates a new "object" in the R environment called Null.Model. And it assigns to this new object the results of the *lme* function.
- The first argument needed for the *lme* function is the name of our outcome variable, CWB, which is regressed on our mixed effects model. This regression is denoted by ~.
- The second argument identifies the fixed component of the mixed effects regression. In the case of the Null model, we have a single fixed effect (i.e., the grand mean of CWB). Referring back to Equation 17.8a, we see that this grand mean is represented by the fixed regression coefficient, γ_{00}. Readers familiar with the matrix algebra representation of multiple regression may recall that it is necessary to append a vector of "1s" to the matrix of predictor variables in order to estimate the intercept. Thus, R uses the number "1" to represent intercepts.

The third argument identifies the random component of the mixed effects regression. Here we set this component equal to the regression of CWBs onto random intercepts. In addition, it is necessary to identify the grouping variable for which unique intercepts are estimated. In our case, this is simply the group identifier variable. Thus, the random component is given by using the syntax "random=~1|GROUP".

Finally, we identify the dataframe we are analyzing. In our case, our data are stored in a file called mlr (consistent with our chapter title, multilevel regression) thus the final argument for the *lme* function is simply "data=mlr".

Essentially, the above code runs the Null model and saves the output to a new object called Null.Model. To see a summary of the results we simply apply the *summary* function to the Null.Model. We present a portion of this output here:

summary(Null.Model)

Random effects:
Formula: ~1 | GROUP

	(Intercept)	Residual
StdDev:	0.50181	0.7770552

Fixed effects: CWB ~ 1

	Value	Std.Error	DF
(Intercept)	1.858667	0.07213353	540

	t-value	p-value
(Intercept)	25.76703	0

Number of Observations: 600
Number of Groups: 60

The above output includes information about the single fixed effect that we estimated—the grand mean on CWBs ($\gamma_{00} = 1.86$); this number is statistically significantly different from zero (which may or may not have substantive meaning, depending on the specific research questions being addressed). We interpret this coefficient as indicating that, on average, employees engaged in limited acts of CWB (i.e., a 1.86 is a relatively low average given a 7-point Likert-type scale). In addition, information is presented about the random effects. Specifically, τ_{00} (i.e., the variance in intercepts/the variance in CWBs residing between groups) and σ^2 (i.e., the variance within groups). Unfortunately, this information is presented as standard deviations rather than variance components. Thus, one either needs to manually convert these estimates to variance components by squaring them or apply the *VarCorr* function to the Null.Model object.

VarCorr(Null.Model)

GROUP = pdLogChol(1)

	Variance	StdDev
(Intercept)	0.2518133	0.5018100
Residual	0.6038148	0.7770552

We can use these variance components to compute the ICC(1) using Equation 17.13. Here we see that approximately 29% of the variance in CWBs resides between groups (i.e., .2518133/(.2518133 + .6038148) = 0.2943023), and thus 71% of the variance resides within groups. LeBreton and Senter (2008) noted that ICC(1) values of .05 or larger may be interpreted as indicative of practically significant nesting effects. In addition, we can test whether the between-groups variability is statistically significantly different from zero by comparing the Null model to a model that constrains the intercepts to be fixed. This is accomplished by estimating the fixed regression using the *gls* function in R and then comparing the overall fit of the two models using the *anova* function with a likelihood ratio test.

Null.Model.2 = gls(CWB~1, data=mlr)
anova(Null.Model, Null.Model.2)

Model	df	AIC	BIC	logLik	Test
Null.Model	1	3	1507.031	1520.217	
Null.Model.2	2	2	1614.235	1623.025	

Model	df	L.Ratio	p-value
Null.Model		–750.5154	
Null.Model.2		–805.1175	1 vs 2 109.2041 < .0001

The above results indicate that constraining the intercepts to be equal results in a statistically worse-fitting model. Stated alternatively, the variance in intercepts that was estimated as part of the Null model is statistically significantly different from zero—or simply stated, there is significant between-groups variance in CWB scores. Overall, we conclude that there is sufficient between-groups variance to warrant adopting MLR to test our hypotheses.

Model 2: Random Coefficient Regression Model

Overview. After establishing that the CWB scores vary both within and between groups, we may begin

adding predictors to our model to try and explain some of this variability. This new model is denoted the random coefficient regression (RCR) model. We typically proceed by adding lower level predictors into the model before adding in the higher level predictors. Hypothesis 1 states that individual-level AGG will be positively related to individual-level CWBs. Thus, we will proceed by including AGG scores as a Level-1 predictor of CWBs. However, before adding these variables to our model, there are two important decisions that must be made.

First, we must decide how to "scale" the Level-1 predictor, AGG. Although our measure of AGG has a meaningful zero point, many measures used in the social sciences do not. A meaningful zero point is important because it allows for a meaningful interpretation of intercepts. To provide scales with a meaningful zero point, researchers typically center scores around a mean—either a grand mean or the group mean. Grand mean centering yields results that are identical to using the raw data (with the exception of the intercepts, because one is simply adding or subtracting the same constant value from every score; Hofmann & Gavin, 1998; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002). Thus, the variance does not change nor does the covariance between a grand mean-centered variable and the other variables in the analysis. In contrast, group mean centering (also referred to as centering within context; Zhang, Zyphur, & Preacher, 2009) will typically yield results that differ from both the raw data and the grand mean-centered data. In the case of Hypothesis 1, we are interested in the individual-level effect of AGG and thus we want to purge any possible group-level effect from our data. This is accomplished by centering each of the AGG_{ij} scores around their respective group means, $\overline{AGG}_j$. To verify that we have eliminated all between-group differences in aggression by group mean-centering the scores, we can run a one-way ANOVA on the centered scores or simply request a summary of group means on the centered scores. The ANOVA returns an F-value of 1 and the summary of group means reveals that each group now has a mean of zero; thus, because all groups have identical group means, one can verify that there is no between-groups variability in the group mean-centered AGG scores.

Second, we must decide which regression coefficients to treat as random. Looking ahead, we see that Hypothesis 2 is essentially a "main effect" hypothesis suggesting that fewer CWBs will be observed as group COH scores increase. This main effect is manifested as variability in Level-1 intercepts. Thus, we will continue to treat the Level-1 intercepts as random and allow these intercepts to vary between groups. In addition, Hypothesis 3 represents a cross-level moderation hypothesis; it states that the strength of the relationship between AGG and CWB (i.e., Level-1 slope) will vary as a function of group COH (i.e., Level-2 variable). The strength of the AGG→CWB relationship is manifested in the Level-1 slope coefficients. Thus, if we wish to establish that COH explains variance in the Level-1 relationship, we must also treat the Level-1 slopes as random.

With those decisions made, we may specify the structural equations for the random coefficient regression model:

$$\text{Level 1: } CWB_{ij} = \beta_{0j} + \beta_{1j}\left(AGG_{ij} - \overline{AGG_j}\right) + r_{ij} \tag{17.14}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + U_{0j} \tag{17.15a}$$

$$\beta_{0j} = \gamma_{10} + U_{1j}, \tag{17.15b}$$

where β_{0j} = unique intercept for group j, β_{1j} = the unique slope for group j, γ_{00} = the average or fixed intercept (pooled within groups), γ_{10} = the average or fixed slope (pooled within groups), r_{ij} = Level-1 residual for individual i in group j, U_{0j} = Level-2 intercept residual for group j (i.e., the difference between the fixed intercept and the unique intercept assigned to group j), U_{1j} = the Level-2 slope residual for group j (i.e., the difference between the fixed slope and the unique slope assigned to group j).

As can be seen, there are two major differences between the Null model and the RCR model. First, a Level-1 predictor and its corresponding β_{1j} coefficient are introduced in the Level-1 equation. Thus, we have added the β_{1j} coefficients as a second, Level-2 outcome variable. This model is somewhat similar to the traditional OLS regression; however, both the intercept and slope are allowed to vary across groups. This model is called the random

coefficient regression model because we are allowing the intercept and slope coefficients to vary across groups.

Interpreting the regression coefficients. Whereas the Null model consists of a single fixed effect (fixed intercept), the RCR model consists of two fixed effects: a fixed intercept and a fixed slope. In the Null model, γ_{00} is interpreted as the grand mean of CWB; however, in the RCR model the intercept takes on a new meaning: it is interpreted as a common (or pooled within-groups) intercept. Our fixed slope, γ_{10}, is interpreted as a common (or pooled within-groups) slope –; it represents the (average) relationship between individual-level trait aggression and individual-level CWBs. Remember, by group mean centering AGG (i.e., removing the between-group variance) we obtain a pure estimate of the (individual-level) covariance between AGG and CWBs. Thus, we test Hypothesis 1 by examining the significance of γ_{10}. In addition to these two fixed coefficients, separate intercepts (β_{0j}) and slopes (β_{1j}) are also estimated for each group.

Variance decomposition and interpretation. Whereas the Null model consisted of two random effects corresponding to the Level-1 error variance (i.e., within-groups variance in CWB) and the Level-2 error variance (i.e., between-groups variance in CWB), the RCR model consists of four random effects: σ^2, τ_{00}, τ_{11}, *and* τ_{01}.

Variance within groups. First, we obtain an estimate of the variance in r_{ij} (i.e., σ^2). This variance component is now interpreted as the *residual* within-group variance in CWBs (i.e., the within-groups variance in CWBs that exists *after* adding centered AGG scores as a Level-1 predictor variable). We can compare the estimate obtained for σ^2 from the RCR model with the one obtained from the Null model, to determine how much of the within-groups variance in CWBs was accounted for by our Level-1 predictor. Specifically, we can compute an effect size that is interpreted as a proportional reduction in error (see Chapter 15 for additional information about estimating effect sizes in MLR):

$$\text{pseudo-}R^2 = \frac{\sigma^2_{ANOVA} - \sigma^2_{RCR}}{\sigma^2_{ANOVA}}. \qquad (17.16)$$

Variance in Level-1 intercepts. Second, we obtain an estimate of the variance in U_{0j}, denoted τ_{00}. This variance component is now interpreted as the variance in the Level-1 intercepts across groups. Recall that Hypothesis 2 states COH will explain variance in CWBs. Because COH is a variable that resides between groups, it is only able to predict the portion of variance in CWBs that resides between groups. Thus, we want to confirm that there is statistically significant variance in intercepts. This is accomplished by testing two nested models: one where group-level intercepts are fixed to a common value, the other where they are allowed to take on unique values.

Variance in Level-1 slopes. Third, because we decided to treat the Level-1 predictor (i.e., $AGG_{ij} - \overline{AGG}_j$) as a random effect, we also obtained an estimate of the variance in U_{1j}, denoted τ_{11}. This variance component is interpreted as the variance in group-level slopes. Recall that Hypothesis 3 states COH will moderate the strength of the Level-1 relationship between AGG and CWBs. Because COH is a variable that resides between groups, it is only able to predict variance in slopes that reside between groups. Thus, we want to confirm that there is statistically significant variance in slopes. Again, this is accomplished by testing two nested models: one where group-level slopes are fixed to a common value, and the other where they are allowed to take on unique values.

However, Aguinis and Culpepper (2015) noted several problems with relying solely on the test of nested models. Specifically, they noted that (a) the likelihood ratio test used to test for significant variance in slopes is asymptotically too conservative and (b) it is possible to conclude that there is statistically significant variance in slopes when the magnitude of slope variance is practically nonsignificant/trivial. To address these concerns, Aguinis and Culpepper offered a new intraclass correlation: ICC_β. This new statistic provides an effect size representing the proportion of "within-group outcome variance attributed to slope differences" (p. 162). Thus, the ICC_β provides an effect size for Model 2 (i.e., the RCR model) that is akin to the ICC(1) effect size estimated for Model 1 (i.e., the Null model). One important implication of Aguinis and Culpepper's new statistic is that researchers interested in testing cross-level

interactions are encouraged to include the estimation of ICC_β in their model testing process, even if the ICC(1) does not suggest practically significant nesting effects. As Aguinis and Culpepper noted,

> We suggest that researchers contemplating the use of multilevel [regression], as well as those who suspect nonindependence in their data structure, expand the decision criteria for using such data analytic approach to include both types of intraclass correlations. Continued use of [ICC(1)] as the sole decision criteria may lead to inappropriate use of data analytic approaches that require independence across observations and also lead to opportunity cost in terms of testing precise and specific cross-level interaction effect hypotheses. (p. 170)

We concur and encourage researchers, especially those testing cross-level interactions, to estimate both types of ICCs.

Covariance between Level-1 intercepts and slopes. Finally, because we allowed both intercepts and slopes to vary, we obtain an estimate of the covariance between U_{0j} and U_{1j}, denoted τ_{01}. This final variance component is interpreted as the covariance between group-level intercepts and group-level slopes. Although this covariance is not of particular interest in the current study, the covariance between intercepts and slopes is frequently of interest in longitudinal studies examining growth/decline over time.

Model 2: Illustrative example. The test of Hypothesis 1, which is organic to the RCR model, is the significance test for the Level-2 fixed effect γ_{10}. The γ_{10} coefficient is the mean (pooled within-groups) slope coefficient across all groups. A significant γ_{10} coefficient tells us that, on average across all groups, the slope describing the individual relationship between CWB and AGG is significantly different from zero. The significance of the γ_{10} parameter tells us that there is a significant individual-level relationship between CWB and AGG; however, this significance test does not provide information about the magnitude of the relationship, which may be estimated using Equation 17.16. Before using the RCR Model to test our example hypotheses, we will recap the three important pieces of information that this specific model provides:

- An estimate and significance test for Level-2 intercept and slope variance; along with effect size estimates for slope variation (i.e., ICC_β),
- An estimate for the Level-1 relationships (direct test of H1), and
- A pseudo-R^2 to evaluate the magnitude of the Level-1 relationships.

Prior to specifying the RCR model, we will first compute group-mean centered scores on our measure of trait aggression and add these scores to our data frame. To accomplish this process, we use the *aggregate* function to compute group means and save these values into a new data frame. We then use the *rename* function to simply assign meaningful names to the variables, followed by the *merge* function to combine the two data frames. Finally, we compute a new variable by subtracting group means from raw scores and saving the centered scores as a new variable in our data frame.

```
AGG.Aggregated=aggregate(mlr[,c(2)],
    list(mlr$GROUP), mean, na.rm=T)
AGG.Aggregated=rename(AGG.Aggregated,
    replace=c("Group.1"="GROUP",
    "x"="AGG.GroupMean"))
mlr=merge(mlr,AGG.Aggregated,by="GROUP")
mlr$AGG.Group=mlr$AGG-mlr$AGG.
    GroupMean
```

Now that we have properly scaled our measure of trait aggression (AGG), we can write and execute the R code to test the RCR model:

```
RCR.Model=lme(CWB~1+AGG.Group,
    random=~1+AGG.Group|GROUP,data=mlr)
```

The above code accomplishes the following:

- It creates a new "object" in the R environment called RCR.Model, and it assigns to this new object the results of the linear mixed effects (*lme*) function.
- The first argument needed for the *lme* function is the name of our outcome variable, CWB, which is to be regressed on our mixed effects model. This regression is denoted by ~ .

- The second argument identifies the fixed component of the mixed effects regression. In the case of the RCR model, we have two fixed effects (i.e., a fixed intercept and a fixed slope for the group-mean centered aggression scores; see Equations 17.15a and 17.15b).
- The third argument identifies the random component of the mixed effects regression. Here we set this component equal to the regression of CWBs onto both random intercepts and slopes (see Equation 17.14). In addition, we also must identify the grouping variable. Thus, the random component is given by using the syntax random = ~1+AGG.Group|GROUP.
- Finally, we identify our dataframe using data=mlr.

After running the above code, we can request a summary of the results and apply the *VarCorr* function to the RCR.Model and compare the results to the *VarCorr* function applied to our original Null. Model. These variance components are used to compute the pseudo-R^2 statistics.

summary(RCR.Model)

Linear mixed-effects model fit by REML
Data: mlr

AIC	BIC	logLik
1427.714	1454.075	–707.8569

Random effects:
Formula: ~1 + AGG.Group | GROUP
Structure: General positive-definite, Log-Cholesky parametrization

	StdDev	Corr
(Intercept)	0.5137365	(Intr)
AGG.Group	0.1414664	0.751
Residual	0.6947630	

Fixed effects: CWB ~ 1 + AGG.Group

	Value	Std.Error	DF
(Intercept)	1.858667	0.07213353	539
AGG.Group	0.107824	0.02484711	539

	t-value	p-value
(Intercept)	25.767026	0
AGG.Group	4.339498	0

Correlation:

	(Intr)
AGG.Group	0.508

Standardized Within-Group Residuals:

Min	Q1	Med
–2.9409451	–0.6694657	–0.1675953

Q3	Max
0.5779086	3.8372834

Number of Observations: 600
Number of Groups: 60

VarCorr(RCR.Model)

GROUP = pdLogChol(1 + AGG.Group)

	Variance	StdDev	Corr
(Intercept)	0.26392523	0.5137365	(Intr)
AGG.Group	0.02001274	0.1414664	0.751
Residual	0.48269566	0.6947630	

Starting with the direct test of Hypothesis 1, the fixed effect for the Level-1 relationship between CWB and group-mean centered AGG (γ_{10}) is significant and positive; $\gamma_{10} = 0.11$, $p < .05$. This significant fixed effect provides support for Hypothesis 1 and is interpreted as "for every unit increase in individual-level aggression, CWB scores are predicted to increase by .11 units." Next, we compute a pseudo-R^2 to investigate the magnitude of this relationship by plugging the appropriate variance components into Equation 17.16.

VarCorr(Null.Model) #displays variance/ covariance of parameters from ANOVA

GROUP = pdLogChol(1)

	Variance	StdDev
(Intercept)	0.2518133	0.5018100
Residual	0.6038148	0.7770552

VarCorr(RCR.Model) #displays variance/ covariance of parameters from RCR

GROUP = pdLogChol(1 + AGG.Group)

	Variance	StdDev	Corr
(Intercept)	0.26392523	0.5137365	(Intr)
AGG.Group	0.02001274	0.1414664	0.751
Residual	0.48269566	0.6947630	

The estimate of the within-groups variance in CWBs from the Null model (i.e., σ^2_{NULL}) is 0.60. The estimate of the residual within-groups variance in CWBs, after controlling for group-mean centered AGG (i.e., σ^2_{RCR}) is 0.48. Therefore, using Equation 17.16, the pseudo-R^2 for the relationship between CWB and AGG is approximately 0.20, suggesting that group-mean centered AGG accounts for roughly 20% of the within-group variance in CWB. Thus, not only was our predictor statistically significantly related to the outcome, but it appeared to explain a practically meaningful proportion of the within-groups variance in CWBs. At this point, if we had additional Level-1 predictors (e.g., age, sex, other personality traits), we could add them into the Level-1 equation and try to explain even more of the within-groups variance in CWBs.

Our next two hypotheses require that we have significant variance in both intercepts (Hypothesis 2) and slopes (Hypothesis 3). It is possible to conduct a significance test on the variance estimates in the RCR.Model output by comparing the RCR.Model to a model that constrains the intercept and slope to assumed fixed values (i.e., sets the variance in intercepts = 0 and the variance in slopes = 0). First, we use *gls* to estimate a model with both fixed intercept and fixed slope:

```
RCR.Model.2=gls(CWB~1+AGG.Group,data=mlr)
```

Next we use *lme* to estimate a model with random intercepts, but with a fixed slope and compare this to the prior model with fixed intercept and fixed slope.

```
RCR.Model.3=lme(CWB~1+AGG.Group,
   random=~1|GROUP,data=mlr)
   anova(RCR.Model.2,RCR.Model.3)
```

Model df	AIC	BIC	logLik	Test
RCR.Model.2	1	3	1587.21	1600.390
RCR.Model.3	2	4	1465.15	1482.724

Model df	L.Ratio	p-value
RCR.Model.2	790.6048	
RCR.Model.3	–728.5749	1 vs 2 124.0599 < .0001

The results indicate that the random intercepts model (RCR.Model.3) is a better fit to the data; thus, there is statistically significant variance in intercepts. We can repeat this process by comparing RCR. Model.3 (i.e., random intercepts and a fixed slope) to the original RCR.Model (i.e., random intercepts and random slopes).

```
anova(RCR.Model.3,RCR.Model)
```

Abbreviated R Output:

Model df	AIC	BIC	logLik	Test
RCR.Model.3	1	4	1465.150	1482.724
RCR.Model.	2	6	1427.714	1454.075

Model df	L.Ratio	p-value
RCR.Model.3	–728.5749	
RCR.Model.	–707.8569	1 vs 2 41.43588 < .0001

Results confirm there is also statistically significant variance in slopes. Stated alternatively, there is significant between-groups variance in both intercepts ($\tau_{00} = 0.26$, $p < .0001$) and slopes ($\tau_{11} = 0.02$, $p < .0001$). Consequently, we are justified in moving forward with our model-building/comparison approach to test Hypotheses 2 and 3. In addition, we could estimate the ICC_β using the *iccbeta* package in R. Unfortunately, this package relies on a different package (*lme4*) and a different function (*lmer*) for estimating the variance components used to estimate ICC_β Introducing and explaining this alternative package and the accompanying functions is beyond the scope of the current chapter. However, the interested reader is directed to Aguinis and Culpepper (2015) for an excellent discussion underlying the use and interpretation of ICC_β. Hopefully, the *iccbeta* package will be updated to allow output from the *lme* function that is part of the *nlme* package.

Model 3: Intercepts-as-Outcomes Model

Overview. Hypothesis 2 proposes that our Level-2 predictor, COH, will explain a portion of the between-groups variance in CWBs. This test is manifested as a cross-level direct or main effect in our third model, the Intercepts-as-Outcomes (IAO) model:

Level 1: $CWB_{ij} = \beta_{0j} + \beta_{1j}\left(AGG_{ij} - \overline{AGG_j}\right) + r_{ij}$ (17.14, revisited)

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}(COH_j) + U_{0j} \quad (17.17a)$$

$$\beta_{0j} = \gamma_{10} + U_{1j}, \quad (17.17b)$$

where β_{0j} = unique intercept for group j, β_{1j} = the unique slope for group j, γ_{00} = the average or fixed intercept (pooled within-groups), γ_{10} = the average or fixed (pooled within-groups) slope corresponding to the group-mean centered aggression scores, γ_{01} = fixed slope corresponding to the effect of group-level COH, r_{ij} = Level-1 residual for individual i in group j, U_{0j} = Level-2 intercept residual for group j (i.e., the difference between the fixed intercept and the unique intercept assigned to group j), U_{1j} = the Level-2 slope residual for group j (i.e., the difference between the fixed slope and the unique slope assigned to group j). The IAO model is similar to the RCR model. The primary difference is the introduction of a Level-2 predictor (COH_j) in the Level-2 intercept equation (17.17a). There are two important consequences that occur when we introduce a Level-2 predictor of β_{0j}. Adding COH to the Level-2 intercept equation (a) changes the interpretation of τ_{00} and (b) introduces a new fixed effect (γ_{01}) into the Level-2 intercept equation.

Interpreting the regression coefficients. Our IAO model consists of three fixed effects: a fixed intercept and two fixed slopes. The fixed intercept, γ_{00}, is still interpreted as a common (or pooled within-groups) intercept. Our fixed slope, γ_{10}, is interpreted as a common (or pooled within-groups) slope—it represents the (average) relationship between individual-level trait aggression and individual-level CWBs. The final fixed effect, γ_{01}, is interpreted as a common (or pooled within-groups) slope—it represents the (average) relationship between group-level cohesion scores and (the between-groups portion of) individual-level CWBs.

Variance decomposition and interpretation.

Variance within groups. This model will continue to generate an estimate of within-groups residual error variance (i.e., the within-groups variance in CWBs that exists after including group-mean centered trait aggression scores in the model). Because we have not made any changes to the Level-1 equation, the σ^2 estimate from the IAO model is typically the same as what was observed in the RCR model. However, it is possible for these estimates to change slightly.

Variance in Level-1 intercepts. In contrast, our estimate of the variance in U_{0j}, denoted τ_{00}, typically will change when compared to the estimate that was obtained using the RCR model. This variance component is now interpreted as the residual variance in the Level-1 intercepts that remains after including COH in the model. It is possible to use the variance components from the RCR and IAO models to obtain a pseudo-R^2 for our Level-2 predictor (again, we encourage readers to refer to Chapter 15 for additional information about estimating effect sizes in MLR). Although there are several options for estimating effect sizes, we opted to compute a pseudo-R^2 using the τ_{00} estimate from the RCR model and the τ_{00} estimate from the IAO model:

$$\text{Level-2 intercept model pseudo-}R^2 = \frac{\tau_{00(RCR)} - \tau_{00(IAO)}}{\tau_{00(RCR)}}. \quad (17.18)$$

Finally, we can also test whether the remaining variance in intercepts is statistically significant by comparing two nested models: one where group-level intercepts are fixed to a common value, and the other where they are allowed to take on unique values. If τ_{00} from the IAO model is statistically significant, we could try to explain this residual between-groups variance in intercepts by adding in additional group-level variables (e.g., group size, group age).

Variance in Level-1 slopes. Because we continued to treat the Level-1 predictor (*i.e.*, $AGG_{ij} - \overline{AGG_j}$) as a random effect, we also obtain an estimate of the variance in U_{1j}, denoted τ_{11}. This variance component is interpreted as the variance in group-level slopes. Recall that Hypothesis 3 states COH will moderate the strength of the Level-1 relationship between AGG and CWBs. Thus, we want to confirm that there is statistically significant variance in slopes across groups. Again, this is accomplished by comparing two nested models: one where group-level slopes are fixed to a common value, the other where the group-level slopes are allowed to take on unique values.

Covariance between Level-1 intercepts and slopes. Finally, because we allowed both intercepts and slopes to vary, we will continue to obtain an estimate of the covariance between U_{0j} and U_{1j}, denoted τ_{01}. This final variance component is interpreted as the covariance between group-level intercepts and group-level slopes.

Model 3: Illustrative example. The test of Hypothesis 2 is furnished by the significance test on the fixed regression weight, γ_{01}. We specify the IAO model in R:

```
IAO.Model=lme(CWB~1+AGG.Group+COH,
    random=~1+AGG.Group|GROUP,data=mlr)
```

The above code simply augments the fixed portion of the mixed regression by requesting a slope coefficient for the COH variable. After requesting a summary of the results, we find that there is a significant main effect of COH on CWB ($\gamma_{01} = -0.61$, $p < .05$), suggesting that for every unit increase in group-level cohesion scores, CWBs decrease by .61 units. Thus, as groups become more cohesive, individuals within those groups are (on average) less likely to engage in CWBs.

```
summary(IAO.Model)
```

```
Linear mixed-effects model fit by REML
 Data: mlr
      AIC      BIC    logLik
 1409.822 1440.565 -697.9109

Random effects:
 Formula: ~1 + AGG.Group | GROUP
 Structure: General positive-definite, Log-
      Cholesky parametrization
             StdDev    Corr
(Intercept) 0.3754431 (Intr)
AGG.Group   0.1426068 0.63
Residual    0.6950287

Fixed effects: CWB~ 1 + AGG.Group + COH
             Value  Std.Error  DF
(Intercept) 5.237037 0.6025813 539
AGG.Group   0.107299 0.0251456 539
COH        -0.613264 0.1089084  58
             t-value p-value
(Intercept) 8.691006       0
AGG.Group   4.267102       0
COH        -5.631011       0
```

We obtain estimates of the variance components using the *VarCorr* function and compare the between-groups variance in intercepts from the RCR model to the residual between-groups variance in intercepts from the IAO model.

```
VarCorr(RCR.Model)
```

```
GROUP = pdLogChol(1 + AGG.Group)
            Variance    StdDev    Corr
(Intercept) 0.26392523 0.5137365 (Intr)
AGG.Group   0.02001274 0.1414664 0.751
Residual    0.48269566 0.6947630
```

```
VarCorr(IAO.Model)
```

```
GROUP = pdLogChol(1 + AGG.Group)
            Variance    StdDev    Corr
(Intercept) 0.14095753 0.3754431 (Intr)
AGG.Group   0.02033669 0.1426068 0.63
Residual    0.48306484 0.6950287
```

Specifically, we estimate a pseudo-R^2 using Equation 17.18 and obtain a value of .46 (i.e., [0.26–0.14]/0.26 = .46), which indicates that by adding COH to the model, we were able to explain roughly 46% of the between-groups variance in intercepts. It is possible to test whether the remaining variance in intercepts is statistically significant by comparing a model that constrains the intercepts to be fixed to a model that frees them to vary. First, we use *gls* to estimate a model with both fixed intercept and fixed slope:

```
IAO.Model.2=gls(CWB~1+AGG.Group+COH,
    data=mlr)
```

Next, we use *lme* to estimate a model with random intercepts, but with fixed slopes:

```
IAO.Model.3=lme(CWB~1+AGG.Group+COH,
    random=~1|GROUP,data=mlr)
```

Comparing the fit of these models indicates that even after including our significant predictor (COH), the remaining (i.e., residual) variance in intercepts is statistically significant. Thus, if we had additional predictors of group intercepts, we could attempt to predict some of the remaining variance by including those predictors in Equation 17.17a.

anova(IAO.Model.2,IAO.Model.3)

Model df	AIC	BIC	logLik	Test
IAO.Model.2	1	4	1488.774	1506.341
IAO.Model.3	2	5	1437.808	1459.767

Model df	L.Ratio	p-value
IAO.Model.2	–740.3869	
IAO.Model.3	–713.9038	1 vs 2 52.96621 < .0001

Finally, we can compare the fit of IAO.Model.3 to the original IAO.Model to confirm that the variance in slopes continues to be significant, and thus we are justified in moving forward to test our third and final hypothesis. Not surprisingly, the variance in slopes is statistically significant. Thus, we will proceed to test Hypothesis 3, which states that COH will help to explain some of the variance in slopes (i.e., group-level COH will moderate the strength of the relationship between employee-level AGG and employee-level CWB).

anova(IAO.Model.3,IAO.Model)

Model df	AIC	BIC	logLik	Test
IAO.Model.3	1	5	1437.808	1459.767
IAO.Model.	2	7	1409.822	1440.565

Model df	L.Ratio	p-value
IAO.Model.3	–713.9038	
IAO.Model	–697.9109	1 vs 2 31.98579 < .0001

Model 4: Slopes-as-Outcomes Model

Overview. Hypothesis 3 proposes that our Level-2 variable, COH, will moderate the relationship between our Level-1 variables, AGG and CWB. This test is manifested as a cross-level interaction in our fourth model, the Slopes-as-Outcomes (SAO) model:

$$\text{Level 1: } CWB_{ij} = \beta_{0j} + \beta_{1j}\left(AGG_{ij} - \overline{AGG_j}\right) + r_{ij} \quad (17.14, \text{revisited})$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}\left(COH_j\right) + U_{0j} \quad (17.17a)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\left(COH_j\right) + U_{1j}, \quad (17.19)$$

where β_{0j} = unique intercept for group j, β_{1j} = the unique slope for group j, γ_{00} = the average or fixed intercept (pooled within-groups), γ_{10} = the average or fixed (pooled within-groups) slope corresponding to the group-mean centered aggression scores, γ_{01} = the fixed slope corresponding to the effect of group-level COH, γ_{11} = fixed slope corresponding to the relationship of group-level COH to the Level-1 slopes (β_{1j}), r_{ij} = Level-1 residual for individual i in group j, U_{0j} = Level-2 intercept residual for group j (i.e., the difference between the fixed intercept and the unique intercept assigned to group j), U_{1j} = the Level-2 slope residual for group j (i.e., the difference between the fixed slope and the unique slope assigned to group j). The SAO model is similar to the IAO model. The primary difference is the introduction of a Level-2 predictor (COH_j) in the Level-2 slope equation (17.19). There are two important consequences that occur when we introduce a Level-2 predictor of β_{1j}. Adding COH to the Level-2 slope equation (a) changes the interpretation of τ_{11}, and (b) introduces a new fixed effect parameter (γ_{11}).

Interpreting the regression coefficients. Our SAO model consists of four fixed effects: a fixed intercept and three fixed slopes. The fixed intercept, γ_{00}, is still interpreted as a common (or pooled within-groups) intercept. Our fixed slope, γ_{10}, is interpreted as a common (or pooled within-groups) slope—it represents the (average) relationship between individual-level trait aggression (AGG) and individual-level CWBs. The fixed slope, γ_{01}, is interpreted as a fixed slope representing the (average) relationship between group-level cohesion scores and the between-groups portion of individual-level CWBs. Finally, γ_{11}, is interpreted as a fixed slope representing the relationship between group-level cohesion scores and the Level-1 slopes (β_{1j}). If γ_{11} is positive, it indicates that as COH increases (i.e., groups become more cohesive), the bivariate relationship between employee-level AGG and CWB also increases. If γ_{11} is negative, it indicates that as COH increases (i.e., groups become more cohesive),

the relationship between individual-level AGG and CWB decreases. Based on Hypothesis 3, we are hoping to see a significant negative γ_{11} coefficient.

Variance decomposition and interpretation.

Variance within groups. This model will continue to generate an estimate of within-groups residual error variance (i.e., the within-groups variance in CWBs that exists after including group-mean centered AGG scores in the model). Because we have not made any changes to the Level-1 equation, the σ^2 estimate from the SAO model is typically the same as what was observed in the RCR and IAO models. However, it is possible for these estimates to change slightly.

Variance in Level-1 intercepts. Similarly, our estimate of the variance in U_{0j}, denoted τ_{00}, is unlikely to change dramatically from what was observed in the IAO model, because we have not modified Equation 17.17a. This variance component is still interpreted as the residual variance in the Level-1 intercepts that remains after including COH in the model.

Variance in Level-1 slopes. In contrast, our estimate of the variance in U_{1j}, denoted τ_{11}, takes on new meaning in the SAO model. Specifically, this variance component is interpreted as the *residual* variance in group-level slopes that remains after including COH_j as a predictor of slopes (see Equation 17.19). It is possible to use the variance components from the IAO and SAO models to obtain a pseudo-R^2 for our Level-2 predictor. For our estimate of pseudo-R^2, we rely on the τ_{11} estimate from the IAO model and the τ_{11} estimate from the SAO model:

Level-2 intercept model pseudo-R^2

$$= \frac{\tau_{11(IAO)} - \tau_{11(SAO)}}{\tau_{11(IAO)}}. \qquad (17.20)$$

This coefficient is interpreted as an effect size, representing the proportion of variance in slopes that is explained using group-level COH scores. Finally, we can also test whether the remaining variance in slopes is statistically significant by comparing two nested models, one where group-level slopes are fixed to a common value the other where they are allowed to take on unique values. If the residual variance, τ_{11}, from the SAO model is statistically significant, we could try to explain this variance by adding in additional group-level variables (e.g., group size, group structure, group age) as predictors of Level-1 slopes.

Covariance between Level-1 intercepts and slopes. Finally, because we allowed both intercepts and slopes to vary, we will continue to obtain an estimate of the covariance between U_{0j} and U_{1j}, denoted τ_{01}. This final variance component is interpreted as the covariance between group-level intercepts and group-level slopes.

Model 4: Illustrative example. The test of Hypothesis 3 is furnished by the significance test on the fixed regression weight, γ_{11}. We specify the SAO model in R:

```
SAO.Model=lme(CWB~1+AGG.Group+COH+
    AGG.Group:COH,random=~1+AGG.
    Group|GROUP,data=mlr)
```

The above code simply augments the fixed portion of the mixed regression by including a cross-product term between employee-level (group-mean centered) aggression and group-level cohesion. In R, cross-products between specific variables are specified using a colon (:) rather than an asterisk (*), which is more commonly used in other software packages. After requesting a summary of the results, we find that COH is a significant moderator of the relationships between individual-level AGG and CWB ($\gamma_{11} = -0.17$, $p < .05$). Specifically, for every unit increase in group-level cohesion scores, the group-level relationships (i.e., β_{1j}) are predicted to decrease by .17 units. This significant cross-level interaction is visually depicted in Figure 17.7.

```
summary(SAO.Model)
```

```
Linear mixed-effects model fit by REML
 Data: mlr
      AIC      BIC    logLik
 1405.847 1440.969 -694.9234
```

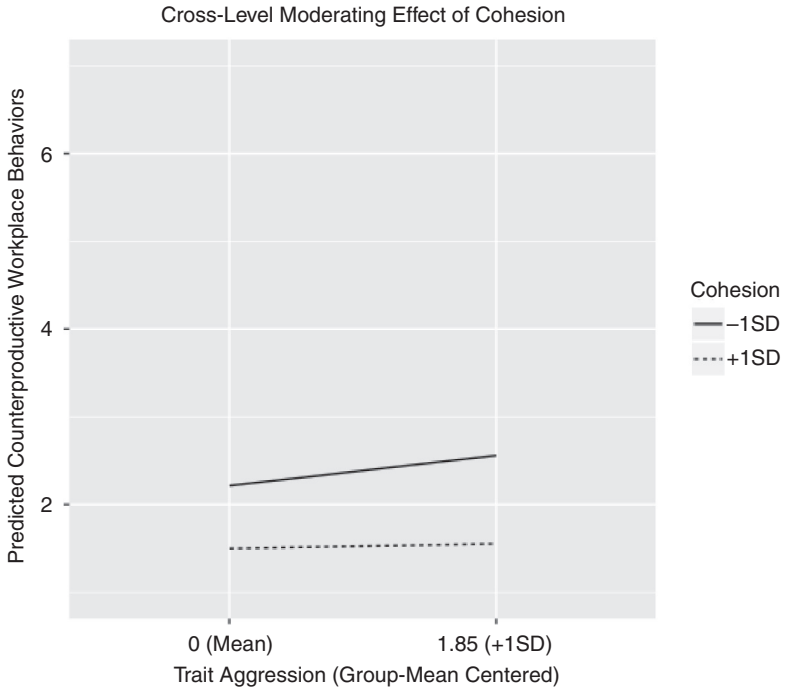


FIGURE 17.7. Cross-level moderating effect of group-level cohesion on the individual-level relationship between counterproductive workplace behaviors and trait aggression.

Random effects:
Formula: ~1 + AGG.Group | GROUP
Structure: General positive-definite, Log-Cholesky parametrization

	StdDev	Corr
(Intercept)	0.3701347	(Intr)
AGG.Group	0.1209911	0.601
Residual	0.6945231	

Fixed effects: CWB ~1 + AGG.Group + COH + AGG.Group:COH

	Value	Std.Error	DF
(Intercept)	6.002650	0.6461452	538
AGG.Group	1.020849	0.2761603	538
COH	–0.752243	0.1168581	58
AGG.Group:COH	–0.165673	0.0499978	538

	t-value	p-value
(Intercept)	9.289939	0e+00
AGG.Group	3.696583	2e-04
COH	–6.437237	0e+00
AGG.Group:COH	–3.313596	1e-03

We obtain estimates of the variance components using the *VarCorr* function and compare the between-groups variance in slopes from the IAO model to the residual between-groups variance in slopes from the SAO model:

VarCorr(IAO.Model)

GROUP = pdLogChol(1 + AGG.Group)

	Variance	StdDev	Corr
(Intercept)	0.14095753	0.3754431	(Intr)
AGG.Group	0.02033669	0.1426068	0.63
Residual	0.48306484	0.6950287	

VarCorr(SAO.Model)

GROUP = pdLogChol(1 + AGG.Group)

	Variance	StdDev	Corr
(Intercept)	0.13699971	0.3701347	(Intr)
AGG.Group	0.01463884	0.1209911	0.601
Residual	0.48236236	0.6945231	

Specifically, we estimate a pseudo-R^2 using Equation 17.20 and obtain a value of 0.28

(i.e., [0.020–0.014]/0.020 = .28), which indicates that by adding COH to the model, we were able to explain roughly 28% of the between-groups variance in slopes. It is possible to test whether the remaining variance in slopes is statistically significant by comparing a model that constrains the slopes to be fixed to a model that allows them to vary. Recall that the residual variance in intercepts from the IAO model was statistically significant. Thus, we will continue to model intercepts as random while testing whether significant variance remains in slopes:

```
SAO.Model.2=lme(CWB~1+AGG.Group+COH+
    AGG.Group:COH,random=~1|GROUP,data=mlr)
```

We then compare the fit of the original SAO model (random intercepts and random slopes) with the above model (random intercepts and fixed slope):

```
anova(SAO.Model.2, SAO.Model)
```

Model df	AIC	BIC	logLik	Test
SAO.Model.2	1	6	1422.043	1448.384
SAO.Model	2	8	1405.847	1440.969

Model df	L.Ratio	p-value
SAO.Model.2	–705.0212	
SAO.Model	–694.9234	1 vs 2 20.19572 < .0001

The results indicate that the remaining (i.e., residual) variance in slopes is statistically significant. Thus, if we had additional predictors of group slopes, we could attempt to predict the remaining variance by including those predictors in Equation 17.19 (and, as a main effect in Equation 17.17a, too).

EXPANDING THE MODEL-BUILDING/ COMPARISON APPROACH TO MULTILEVEL REGRESSION

A disclaimer is in order after walking the reader through the process of developing and comparing a series of models in our quest to test three example hypotheses. The hypotheses and models that we tested were extremely simple. However, after mastering the basics of MLR, the reader will be better positioned to articulate and test more sophisticated models. For example, consider how you would update the four basic models (Null, RCR, IAO, SAO) to test a few additional hypotheses:

- Hypothesis 4: Employee sex will moderate the strength of the relationship between employee aggression and counterproductive workplace behaviors, such that the strength of the relationship will be stronger for male employees than for female employees.
- Hypothesis 5: Group size will moderate the strength of the relationship between group cohesion and counterproductive workplace behaviors, such that the negative effect will become stronger as group size gets smaller.

Essentially, Hypothesis 4 is adding an additional Level-1 predictor of CWBs and treating it as a moderator variable for the Level-1 relationship, and Hypothesis 5 is adding an additional Level-2 predictor of CWBs and treating it as a moderator variable of the Level-2 relationship. These seem like simple enough hypotheses, but how would one go about using the model-building approach to test these hypotheses?

We encourage you to take a few minutes to think through (and write down) the model equations and the R code that would be required to test these additional hypotheses. Did you come up with something along the lines of the following?

Null Model

$$\text{Level 1: } CWB_{ij} = \beta_{0j} + r_{ij}$$

$$\text{Level 2: } B_{0j} = \gamma_{00} + U_{0j}$$

The Null model doesn't change from what was presented in the chapter. Thus, the Null model would have a single fixed effect (γ_{00}) and two random effects corresponding to the variance in CWBs that reside within (σ^2) and between groups (τ_{00}).

RCR Model

In contrast, the RCR model becomes appreciably more complex. Specifically, we have added employee sex to the Level-1 equation, along with the cross-product between sex and aggression. In addition, for each of these additional terms, we had to decide whether to treat them as random or fixed (or both).

In our case, we elected to treat each of them as both fixed and random (hence the separate error terms for each of the Level-2 equations):

$$\text{Level 1: } CWB_{ij} = \beta_{0j} + \beta_{1j}\left(AGG_{ij} - \overline{AGG_j}\right) + \beta_{2j}\left(SEX_{ij}\right) + \beta_{3j}\left(AGG_{ij} - \overline{AGG_j}\right) * \left(SEX_{ij}\right) + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + U_{1j}$$

$$\beta_{2j} = \gamma_{20} + U_{2j}$$

$$\beta_{3j} = \gamma_{30} + U_{3j}.$$

The test of Hypothesis 4 is provided by a test on the significance of γ_{30}. If this coefficient was significant, then we would likely want to estimate the proportion of Level-1 variance that is attributed to the interaction effect. This effect size would be computed by comparing the σ^2 from the above RCR model to the σ^2 obtained from a main effects model. Formally, we could estimate the effect size for the interaction between two Level-1 variables as

$$\text{Level 1 pseudo-}R^2 = \frac{\sigma^2_{MainEffects} - \sigma^2_{MainEffects+Cross-Product}}{\sigma^2_{MainEffects}}.$$

To properly specify this RCR model, we need to know which effects to treat as fixed and which to treat as random. In the above equation, there are four fixed effects corresponding to the Level-2 γ coefficients. In addition, the above equations indicate that random effects are to be estimated for each of the predictors. Thus, there are four random effects corresponding to the variability in intercepts (τ_{00}), the slopes for aggression (τ_{11}), the slopes for sex (τ_{22}), and the slopes for the cross-product terms (τ_{33}). And, there is a Level-1 random effect corresponding to the within-groups variance (σ^2). Finally, the four random coefficients from the Level-1 equations are also allowed to covary with one another, resulting in six additional covariance terms (i.e., τ_{01},τ_{02},τ_{03},τ_{12},τ_{13},τ_{23}).

Thus, moving from our original RCR model to a model that includes an additional Level-1 variable and its cross-product with our original Level-1 variable yields a substantially more complex model. Adding to this complexity would be decisions concerning how to scale sex (e.g., dummy codes, effect codes, group or grand-mean centered). We do not delve into the various interpretations of different centering strategies; rather, we direct the reader to other sources that are specifically focused on scaling/centering of predictors (Hofmann & Gavin, 1998; Raudenbush & Bryk, 2002). Whereas our original RCR model contained two fixed effects and four random effects, our revised RCR model now contains four fixed effects and 11 random effects. The R code for this revised RCR model would look something like:

```
RCR.Model = lme(CWB~1+AGG.Group+
    SEX+AGG.Group:SEX,random=~1+AGG.
    Group+SEX+AGG.Group:SEX|GROUP,
    data=mlr)
```

IAO Model

To test Hypothesis 5, we will need to also revise the IAO model. Specifically, we have added group size to the Level-2 equation predicting intercepts, along with the cross-product between group size and cohesion:

$$\text{Level 1: } CWB_{ij} = \beta_{0j} + \beta_{1j}\left(AGG_{ij} - \overline{AGG_j}\right) + \beta_{2j}\left(SEX_{ij}\right) + \beta_{3j}\left(AGG_{ij} - \overline{AGG_j}\right) * \left(SEX_{ij}\right) + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}\left(COH_j\right) + \gamma_{02}\left(SIZE_j\right) + \gamma_{03}\left(COH_j\right) * \left(SIZE_j\right) + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + U_{1j}$$

$$\beta_{2j} = \gamma_{20} + U_{2j}$$

$$\beta_{3j} = \gamma_{30} + U_{3j}.$$

The test of Hypothesis 5 is provided by a test on the significance of γ_{03}. If this coefficient were significant, we would need to estimate the proportion of intercept variance that is attributed to this interaction effect. This effect size would be computed by comparing the τ_{00} from the above IAO model with the τ_{00} obtained from a main effects

model. Formally, we could estimate the effect size for the interaction between two Level-2 variables as

$$\text{Level 2 pseudo-}R^2 = \frac{\tau_{00(MainEffects)} - \tau_{00(MainEffects+Cross-Product)}}{\tau_{00(MainEffects)}}.$$

To properly specify this IAO model, we need to know which effects to treat as fixed and which to treat as random. In the above equation, there are seven fixed effects corresponding to the Level-2 γ coefficients. In addition, the above equations indicate that random effects are to be estimated for each of the predictors. Thus, we will continue to have the 11 random effects that were estimated in our revised RCR model.

The R code for this revised IAO model would look something like:

```
IAO.Model = lme(CWB~1+AGG.Group+SEX+
    AGG.Group:SEX+COH+SIZE+COH:SIZE,
    random=~1+AGG.Group+SEX+AGG.Group:
    SEX|GROUP, data=mlr)
```

Finally, in order to test our original Hypothesis 3, we would specify the SAO model by adding COH as a Level-2 predictor of slope coefficients representing the regression of CWBs onto aggression:

$$\text{Level 1: } CWB_{ij} = \beta_{0j} + \beta_{1j}\left(AGG_{ij} - \overline{AGG_j}\right) + \beta_{2j}\left(SEX_{ij}\right) + \beta_{3j}\left(AGG_{ij} - \overline{AGG_j}\right) * \left(SEX_{ij}\right) + r_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}\left(COH_j\right) + \gamma_{02}\left(SIZE_j\right) + \gamma_{03}\left(COH_j\right) * \left(SIZE_j\right) + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\left(COH_j\right) + U_{1j}$$

$$\beta_{2j} = \gamma_{20} + U_{2j}$$

$$\beta_{3j} = \gamma_{30} + U_{3j}.$$

Thus, we need to estimate a final, eighth, fixed effect (γ_{11}). This is accomplished in R:

```
SAO.Model = lme(CWB~1+AGG.Group+SEX+
    AGG.Group:SEX+COH+SIZE+COH:SIZE+
    COH:AGG.Group,random=~1+AGG.Group+
    SEX+AGG.Group:SEX|GROUP,data=mlr)
```

MULTILEVEL (RANDOM COEFFICIENT) REGRESSION MODELS: CONCLUDING COMMENTS

Given the ubiquitous nature of nested data structures, we hope that our chapter provides researchers with an initial (albeit basic) introduction to how MLR models may be used to test hypotheses using data conforming to a nested or multilevel structure. Table 17.2 provides a brief summary of our model-building/comparison approach using MLR. In addition to our basic introduction (and the other helpful chapters in this handbook), we encourage readers to explore issues including (a) implications for the scaling and centering of data (e.g., Hofmann & Gavin, 1998; Raudenbush & Bryk, 2002 [especially Chapters 2 and 5]; Zhang, Zyphur, & Preacher, 2009), (b) extensions of the MLR model to longitudinal applications (i.e., temporal nesting of repeated Level-1 observations within Level-2 units; Bliese & Ployhart, 2002; Bryk & Raudenbush, 1987; Raudenbush & Bryk, 2002), (c) strategies for testing hypotheses involving multilevel mediation (see Chapter 20; also Bauer, Preacher, & Gil, 2006; Preacher, Zyphur, & Zhang, 2010; Zhang et al., 2009), (d) issues associated with the analysis of dyadic data (i.e., individuals nested in couples or pairs; see Chapter 18; also Atkins, 2005; Kenny, Kashy, & Cook, 2006; Krasikova & LeBreton, 2012; Lyons & Sayer, 2005), and (e) how to avoid fallacies of inference when interpreting the results from a multilevel analysis (Firebaugh, 1978; Greenland, 2002; James, 1982; Mossholder & Bedeian, 1983; Ostroff, 1993).

Finally, just like the traditional, single-level OLS regression model carries with it certain underlying assumptions, so too does the more complicated MLR model (Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). One's data and theory should be closely examined to determine whether these assumptions are likely met or violated:

- The Level-1 residuals (r_{ij}) have a mean of zero, are normally and independently distributed, and have a constant (homoscedastic) variance (σ^2).
- The Level-1 residuals are uncorrelated with the Level-1 predictors.
- The Level-1 residuals are uncorrelated with the Level-2 residuals (U_{0j}, U_{1j}, etc.).

TABLE 17.2

Multilevel Regression: A Model Comparison Approach

Model	Specific steps	Associated equations	What to look for
1. Null	A. Estimate Null model with fixed and random intercepts	Level 1: $CWB_{ij} = \beta_{0j} + r_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + U_{0j}$	
	B. Estimate variance components		τ_{00}: variance in Level-1 outcome variable that resides between Level-2 units. σ^2: variance in Level-1 outcome variable that resides within Level-2 units.
	C. Compute and interpret ICC(1)	$ICC(1) = \frac{\tau_{00}}{\tau_{00} + \sigma^2}$	Interpreted as the proportion of variance in Level-1 outcome variable that is attributed to the nesting of Level-1 units in Level-2 units.
	D. Estimate alternative model with fixed intercept	Level 1: $CWB_{ij} = \beta_{0j} + r_{ij}$ Level 2: $\beta_{0j} = \gamma_{00}$	
	E. Test for significant variance in intercepts by comparing fit of the Null model (Step A) and the alternative model (Step D)		If the less restrictive model (Step A) is better fitting, then there is significant variance in Level-1 outcome scores across Level-2 units.
2. RCR	A. Add Level-1 predictor variables and estimate the RCR model with fixed and random intercepts and slopes	Level 1: $CWB_{ij} = \beta_{0j} + \beta_{1j}(AGG_{ij} - \overline{AGG}_j) + r_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + U_{0j}$ $\beta_{1j} = \gamma_{10} + U_{1j}$	
	B. Estimate variance components for RCR		τ_{00}: variance in intercepts τ_{11}: variance in slopes τ_{01}: covariance between intercepts and slopes σ^2: residual variance in Level-1 outcome (variable that exists after including Level-1 predictor variables in the model)
	C. Compute and interpret ICC_β	ICC_β (See Aguinis & Culpepper, 2015, for equations or use the *iccbeta* package in R to estimate this intraclass correlation.)	Interpreted as the proportion of variance in Level-1 outcome variable that is attributed to group slope differences. This statistic is especially important if researchers have hypotheses about cross-level interactions/ moderators.
	D. Interpret fixed coefficients		γ_{00}: pooled or average intercept γ_{10}: pooled or average slope (effect of Level-1 predictor on Level-1 outcome)
	E. Compute and interpret pseudo-R^2	pseudo-$R^2 = \frac{\sigma^2_{Null} - \sigma^2_{RCR}}{\sigma^2_{Null}}$	Interpreted as the proportion of Level-1 outcome variance that may be explained using the Level-1 predictor variable.
	F. Estimate alternative model with fixed intercept and fixed slope	Level 1: $CWB_{ij} = \beta_{0j} + \beta_{1j}(AGG_{ij} - \overline{AGG}_j) + r_{ij}$ Level 2: $\beta_{0j} = \gamma_{00}$ $\beta_{1j} = \gamma_{10}$	
	G. Estimate alternative model with random intercept and fixed slope.	Level 1: $CWB_{ij} = \beta_{0j} + \beta_{1j}(AGG_{ij} - \overline{AGG}_j) + r_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + U_{0j}$ $\beta_{1j} = \gamma_{10}$	

TABLE 17.2

Multilevel Regression: A Model Comparison Approach (*Continued*)

Model	Specific steps	Associated equations	What to look for
	H. Test for significant variance in intercepts by comparing fit of models from Steps F and G.		If the model from Step G is a better-fitting model than the one from Step F, then there is significant variance in intercepts. Note: this result is typically consistent with the results of Step E from the previous Null model testing sequence.
	I. Test for significant variance in slopes by comparing fit of models from Steps A and F.		If the model from Step A is a better-fitting model the one from Step G, then there is significant variance in slopes.
3. IAO	A. Add Level-2 predictors of intercepts and estimate IAO model with fixed and random intercepts and slopes.	Level 1: $CWB_{ij} = \beta_{0j} + \beta_{1j}(AGG_{ij} - \overline{AGG}_j) + r_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(COH_j) + U_{0j}$ $\beta_{1j} = \gamma_{10} + U_{1j}$	
	B. Estimate variance components for IAO		τ_{00}: residual variance in intercepts that exists after including Level-2 predictor variable in the model τ_{11}: variance in slopes τ_{01}: covariance between intercepts and slopes σ^2: residual variance in Level-1 outcome variable that exists after including Level-1 predictor variables in the model.
	C. Interpret fixed coefficients	γ_{00}; γ_{10}; γ_{01}	γ_{00}: fixed (pooled) intercept γ_{10}: fixed (pooled) slope (relationship between Level-1 predictor variable and Level-1 outcome variable) γ_{01}: fixed slope (relationship between Level-2 predictor variable on the Level-1 outcome variable)
	D. Compute and interpret pseudo-R^2	pseudo-$R^2 = \frac{\tau_{00(RCR)} - \tau_{00(IAO)}}{\tau_{00(RCR)}}$	Interpreted as the proportion of intercept variance that may be explained using the Level-2 predictor variable.
	E. Estimate alternative model with fixed intercept and fixed slope.	Level 1: $CWB_{ij} = \beta_{0j} + \beta_{1j}(AGG_{ij} - \overline{AGG}_j) + r_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(COH_j)$ $\beta_{1j} = \gamma_{10}$	
	F. Estimate alternative model with random intercept and fixed slope.	Level 1: $CWB_{ij} = \beta_{0j} + \beta_{1j}(AGG_{ij} - \overline{AGG}_j) + r_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(COH_j) + U_{0j}$ $\beta_{1j} = \gamma_{10}$	
	G. Test for significant residual variance in intercepts by comparing fit of models from Steps E and F.		If the model from Step E is the better-fitting model, that suggests the Level-2 predictor has explained all of the meaningful variance in intercepts. If the model from Step F is the better-fitting model, which suggests there is still significant variance in intercepts (could add additional Level-2 predictors to try and explain this variance).

(*continues*)

TABLE 17.2

Multilevel Regression: A Model Comparison Approach (*Continued*)

Model	Specific steps	Associated equations	What to look for
	H. Test for significant variance in slopes by comparing fit of models from Steps A and F.		If the model from Step A is the better-fitting model, that suggests there is significant variance in slopes, and it makes sense to proceed with tests of potential cross-level moderators.
4. SAO	A. Add level-2 predictors of slopes and estimate SAO model with fixed and random intercepts and slopes.	Level 1: $CWB_{ij} = \beta_{0j} + \beta_{1j}(AGG_{ij} - \overline{AGG}_j) + r_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(COH_j) + U_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11}(COH_j) + U_{1j}$	
	B. Estimate variance components for SAO		τ_{00}: residual variance in intercepts that exists after including Level-2 predictor variable in the model τ_{11}: residual variance in slopes that exists after including Level-2 predictor variable in the model τ_{01}: covariance between intercepts and slopes σ^2: residual variance in Level-1 outcome variable that exists after including Level-1 predictor variables in the model.
	C. Interpret fixed coefficients		γ_{00}: fixed (pooled) intercept γ_{10}: fixed (pooled) slope (relationship between Level-1 predictor variable and Level-1 outcome variable) γ_{01}: fixed slope (relationship between Level-2 predictor variable on the Level-1 outcome variable) γ_{11}: fixed slope (relationship between the Level-2 predictor variable and the relationship between the Level-1 predictor and outcome variables)
	D. Compute and interpret pseudo-R^2	$\text{pseudo-}R^2 = \frac{\tau_{11(IAO)} - \tau_{11(SAO)}}{\tau_{11(IAO)}}$	Interpreted as the proportion of slope variance that may be explained using the Level-2 predictor variable.
	E. Estimate alternative model with fixed slope.	Level 1: $CWB_{ij} = \beta_{0j} + \beta_{1j}(AGG_{ij} - \overline{AGG}_j) + r_{ij}$ Level 2: $\beta_{0j} = \gamma_{00} + \gamma_{01}(COH_j) + U_{0j}$ $\beta_{1j} = \gamma_{10} + \gamma_{11}(COH_j)$	
	F. Test for significant residual variance in slopes by comparing fit of models from Steps A and E.		If the model from Step E is the better-fitting model, which suggests the Level-2 predictor has explained all of the meaningful variance in slopes. If the model from Step A is the better-fitting model, that suggests there is still significant variance in slopes.
	G. Test for significant variance in slopes by comparing fit of models from Steps A and F.		If the model from Step A is the better fitting model, then that suggests there is significant variance in slopes and it makes sense to proceed with test of potential cross-level moderators.
	H. Graph cross-level moderators		

Note. AGG = aggression; COH = cohesiveness; CWB = counterproductive workplace behaviors; IAO = Intercepts-as-Outcomes; ICC = intraclass correlation; RCR = random coefficients regression; SAO = Slopes-as-Outcomes.

- The Level-2 residuals each have a mean of zero, follow a multivariate normal distribution, and are independent among Level-2 groups.
- The Level-2 residuals are uncorrelated with the Level-2 predictors.

APPENDIX 17.1: ANNOTATED R CODE

```
# Code corresponding to:
# Shiverdecker, L. K., & LeBreton, J. M. (2018).
    Multilevel (random coefficient) regression
    modeling.
# In S. E. Humphrey & J. M. LeBreton
    (Eds.), Handbook of multilevel theory,
    measurement, and
# analysis (Chapter 17). Washington, DC:
    American Psychological Association.

# Step 1: Read in Data File for Multilevel
    Regression Examples (mlr.csv)
mlr <- read.csv("mlr.csv", header=T, sep=",")

# Step 2: Install packages and load into working
    library of tools
install.packages("multilevel"); library(multilevel)
install.packages("ggplot2"); library(ggplot2)
install.packages("data.table"); library(data.table)
install.packages("plyr"); library(plyr)

# Step 3: Create function to plot figures 1:4
ggplotRegression <- function (fit) {
require(ggplot2)
ggplot(fit$model, aes_string(x =
    names(fit$model)[2], y = names(fit$model)
    [1])) + geom_point() + xlim(0,15) +
    ylim(1,7)+
stat_smooth(method = "lm", se=FALSE, col =
    "red", fullrange=T) +
labs(title = paste("R-Square =
    ",signif(summary(fit)$r.squared, 5),
    "Intercept =",signif(fit$coef[[1]],5),
    "Slope =",signif(fit$coef[[2]], 5),
    "P =",signif(summary(fit)$coef[2,4], 5)))+
labs(x="Trait Aggression", y="Counterproductive
    Workplace Behaviors") }

#Step 4: Plot Figures
###Create Subsets of Data Used in Figures 1–4
    (note: mlr dataframe was already sorted by
    GROUP variable)
group1 <- lm(CWB~AGG,data=mlr[01:10,])
group2 <- lm(CWB~AGG,data=mlr[11:20,])
group3 <- lm(CWB~AGG,data=mlr[21:30,])
group4 <- lm(CWB~AGG,data=mlr[31:40,])
group1234 = lm(CWB~AGG,data=mlr[01:40,])

###Generate Figures 1–4
ggplotRegression(group1)
ggplotRegression(group2)
ggplotRegression(group3)
ggplotRegression(group4)
ggplotRegression(group1234)

###Generate Figure 5
all.groups=ggplot(mlr,aes(x=AGG,y=CWB,grou
    p=GROUP))+ xlim(0,15) + ylim(1,7)+stat_
    smooth(method = "lm", se=FALSE,
    fullrange=T) + labs(x="Trait Aggression",
    y="Counterproductive Workplace
    Behaviors")
all.groups

###Figure 6
groups12=ggplot(mlr[1:20,],aes(y = CWB, x
    = AGG)) + geom_point(size = 3, alpha
    = .8) + geom_smooth(method="lm",
    fullrange=T, se= F, size = 1, aes(linetype
    = as.factor(GROUP), group = GROUP))
    + geom_smooth(method = "lm",size =
    3, colour = 'black', se = F, fullrange=T)
    + xlim(0,15) + ylim(1,7)+ labs(x="Trait
    Aggression", y="Counterproductive
    Workplace Behaviors", linetype="Group #")+
    geom_label(aes(label =GROUP),color='blue')
groups12

#Step 5: Generate the Coefficients in Table 1
set.seed(1)
dat <- data.table(x=mlr$AGG, y=mlr$CWB,
    grp=mlr$GROUP)
OLS = dat[,list(intercept=coef(lm
    (y~x))[1], coef=coef(lm(y~x))
    [2],rsq=summary(lm(y~x))$r.squared)
```

```
,by=grp]
OLS
min(OLS[,2]); max(OLS[,2])
min(OLS[,3]); max(OLS[,3])
min(OLS[,4]); max(OLS[,4])

# Step 6: Multilevel Regression Models
###Null Model(One-Way Random Effects
    ANOVA)
# CWB~1, Regressing CWB onto a fixed
    intercept
# random=~1 Regressing CWB onto random
    intercepts
# |GROUP, Identifies the level 2 variable named
    as "group"
# data=mlr Identifies the data set where "CWB"
    and "group" are located
Null.Model = lme(CWB~1, random=~1|GROUP,
    data=mlr)
summary(Null.Model)

#Estimating the ICCs from the Null Model
VarCorr(Null.Model)
Null.ICC=GmeanRel(Null.Model)
names(Null.ICC) #returns the names of the
    variables in the object "Null.ICC"
Null.ICC$ICC #returns the value of the ICC
    variable in the Null.ICC object which is the
    ICC(1)
Null.ICC$MeanRel #returns the 60 ICC(2)
    values corresponding to each group
    intercept
#Note: values identical b/c all groups are the
    exact same size . . . so sigma^2/nj is the same
    for all groups.
mean(Null.ICC$MeanRel) #estimates the mean
    across the groups which is the ICC(2) or
    ICC(k) [where k = 60]

Null.Model.2=gls(CWB~1, data=mlr)
    #Estimating Null Model with Fixed
    Intercepts
logLik(Null.Model.2)*–2 #Manually estimating
    –2*loglikelihood of Null Models
logLik(Null.Model)*–2
–2*(–805.1175)—2*(–750.5154) #Manually
    estimating the difference in likelihood ratios
anova(Null.Model, Null.Model.2) #Manually
    comparing fixed vs. fixed + random
    intercepts
summary(Null.Model) #Displaying summary of
    Null.Model
VarCorr(Null.Model) #Displaying the variance/
    covariance matrix

### Grand Mean Center the Level 1 Predictor
    (NOTE: Not used in this chapter; included
    only for illustrative purposes)
mean(mlr$AGG)
mlr$AGG.Grand=mlr$AGG-4.035167
# Aggregate employee-level trait aggression
    scores to the group level
AGG.Aggregated=aggregate(mlr[,c(2)],
    list(mlr$GROUP), mean, na.rm=T)
names(AGG.Aggregated)
AGG.Aggregated=rename(AGG.Aggregated,
    replace=c("Group.1"="GROUP",
    "x"="AGG.GroupMean")) #renames
    specific variables
names(AGG.Aggregated)
mlr=merge(mlr,AGG.Aggregated,by="GROUP")
names(mlr)

### Group Mean Center the Level 1 Predictor
mlr$AGG.Group=mlr$AGG-mlr$AGG.
    GroupMean
names(mlr)

#Random Coefficient Regression Model
# CWB~1+AGG.group regressing CWB onto
    fixed effect for intercept
# random=~1+AGG.group regressing CWB onto
    random intercepts and slopes (AGG.group)
# |GROUP, identifying the level 2 (grouping)
    variable
# data=mlr name of data file

RCR.Model=lme(CWB~1+AGG.Group,
    random=~1+AGG.Group|GROUP,data=mlr)
summary(RCR.Model)
VarCorr(Null.Model) #displays variance/
    covariance of parameters from ANOVA
VarCorr(RCR.Model) #displays variance/
    covariance of parameters from RCR
```

```
#Estimating the ICC_beta from the RCR Model
###NOTE: As of August 9, 2017 the icc_beta
    package was not available for version 3.4.0
    of R

#Use gls to estimate model with fixed intercept
    and fixed slope
RCR.Model.2=gls(CWB~1+AGG.
    Group,data=mlr)

#Use lme to estimate a model with random
    intercept and fixed slope
RCR.Model.3=lme(CWB~1+AGG.Group,
    random=~1|GROUP,data=mlr)
#Test if RCR.Model.3 is a better fit to the data
    than RCR.Model.2
anova(RCR.Model.2,RCR.Model.3)

#Test if RCR.Model (Random Intercepts &
    Slopes) is a better fit to the data than RCR.
    Model.3
anova(RCR.Model.3,RCR.Model)

#Intercepts-as-Outcomes Model
#Adding Cohesion as a level 2 predictor
IAO.Model=lme(CWB~1+AGG.
    Group+COH,random=~1+AGG.
    Group|GROUP,data=mlr)
summary(IAO.Model)
VarCorr(RCR.Model)
VarCorr(IAO.Model)

#Use gls to estimate model with fixed intercept
    and fixed slopes
IAO.Model.2=gls(CWB~1+AGG.
    Group+COH,data=mlr)

#Use lme to estimate a model with random
    intercept and fixed slopes
IAO.Model.3=lme(CWB~1+AGG.Group+COH,
    random=~1|GROUP,data=mlr)

#Test if IAO.Model.2 (Fixed Intercept & Slopes)
    is a better fit than IAO.Model.3 (Random
    Intercepts & Fixed Slopes)
anova(IAO.Model.2,IAO.Model.3)

#Test if IAO.Model3 (Random Intercepts &
    Fixed Slopes) is a better fit original IAO.
    Model (Random Intercepts & Slopes)
anova(IAO.Model.3,IAO.Model)

#Slopes-as-Outcomes Model
#Testing the cross-level moderating effect of L2
    cohesion on the relationship
###between L1 CWBs and L1 group-mean
    centered aggression scores
#Cross-Product Effect is denoted using ":"
#helping~1+AGG.Group+COH + AGG.
    Group:COH, Fixed coef. for intercept,
    aggression, cohesion, cross-product
#random=~1+AGG.Group Random coefficients
    for intercept and aggression
#|GROUP Name of level 2 grouping variable
    "GROUP"
#data=mlr Data containing variables in SAO

SAO.Model=lme(CWB~1+AGG.
    Group+COH+AGG.
    Group:COH,random=~1+AGG.
    Group|GROUP,data=mlr)
summary(SAO.Model)

#Request variance components to estimate
    quasi-R2 for slope variance
VarCorr(IAO.Model)
VarCorr(SAO.Model)
SAO.Rsq=(0.02033669–0.01463884)/
    0.02033669
SAO.Rsq

#Test whether the remaining variance in slopes is
    significant
SAO.Model.2=lme(CWB~1+AGG.
    Group+COH+AGG.Group:COH,random=
    ~1|GROUP,data=mlr)
anova(SAO.Model.2, SAO.Model)

# Step 7: Plotting the Cross-Level Interaction
    (Figure 7)
#Descriptive statistics for cross-level moderator
mean(mlr$COH)
sd(mlr$COH)
```

```
#manual estimation of moderator values at +1
    and –1 sd
5.508833 – 0.4758722; 5.508833 + 0.4758722

mean(mlr$AGG.Group) #estimating mean of
    trait aggression
sd(mlr$AGG.Group) # obtaining sd of trait
    aggression

#creating a new data frame with high-low
    values for (AGG.Group (M; +1SD),
    COH(-1SD; +1SD))
inter.data=data.frame(AGG.
    Group=c(0.0000,0.0000, 1.835865,
    1.835865),
COH=c(5.032961,5.984705,5.032961,5.984705))
inter.data #displaying the data frame

#adding a new variable called "predicted" to
    this new data frame using the 'predict'
    function
inter.data$predicted=predict(SAO.Model,inter.
    data,level=0)
inter.data #confirming new variable added to
    data frame

#plotting the cross-level interaction (Option 1:
    Use the 'interaction.plot' function)
interaction.plot(inter.data$AGG.Group,inter.
    data$COH,inter.data$predicted)
interaction.plot(inter.data$AGG.Group,inter.
    data$COH,inter.data$predicted,
    ylab="Counterproductive Work Behaviors",
    xlab="Aggression (Group-Mean Centered)",
    trace.label="Cohesion")
###cleaning up the plot
inter.data$AGG.Group2=c("0 (Mean)","0
    (Mean)","1.85 (+1SD)","1.85 (+1SD)")
inter.data$COH2=c("-1SD","+1SD",
    "-1SD","+1SD")
inter.data
interaction.plot(inter.data$AGG.Group2,inter.
    data$COH2,inter.data$predicted,
    ylab="Counterproductive Work Behaviors",
    xlab="Aggression (Group-Mean Centered)",
    trace.label="Cohesion")
#plotting the cross-level interaction (Option 2:
    Use the 'ggplot' function)
ggplot(inter.data,aes(x=AGG.Group2,
    y=predicted,group=COH2)) +
geom_line(aes(linetype=factor(COH2)))+ #
    Graph separate lines for high and low levels
    of cohesion
labs(title="Cross-Level Moderating Effect
    of Cohesion", x="Trait Aggression
    (Group-Mean Centered)", y="Predicted
    Counterproductive Work Behaviors")+
scale_linetype_discrete(name="Cohesion")+
ylim(1,7)

#Step 8: Concluding Examples
####SEX as level-1 moderator of cwb-agg
####SIZE as level-2 moderator of cwb-coh

#Specification of Null Model
Null.Model = lme(CWB~1, random=~1|GROUP,
    data=mlr)

#Specification of RCR
RCR.Model = lme(CWB~1+AGG.
    Group+SEX+AGG.Group:SEX,
random=~1+AGG.Group+SEX+AGG.
    Group:SEX|GROUP, data=mlr)

#Specification of IAO
IAO.Model = lme(CWB~1+AGG.
    Group+SEX+AGG.
    Group:SEX+COH+SIZE+COH:SIZE,
    random=~1+AGG.Group+SEX+AGG.
    Group:SEX|GROUP, data=mlr)
#Specification of SAO
SAO.Model = lme(CWB~1+AGG.
    Group+SEX+AGG.Group:SEX+COH
    +SIZE+COH:SIZE+COH:AGG.Group,
    random=~1+AGG.Group+SEX+AGG.
    Group:SEX|GROUP, data=mlr)
```

References

Aguinis, H., & Culpepper, S. A. (2015). An expanded decision-making procedure for examining cross-level interaction effects with multilevel modeling. *Organizational Research Methods*, *18*, 155–176. http://dx.doi.org/10.1177/1094428114563618

Atkins, D. C. (2005). Using multilevel models to analyze couple and family treatment data: Basic and advanced issues. *Journal of Family Psychology*, *19*, 98–110. http://dx.doi.org/10.1037/0893-3200.19.1.98

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods*, *11*, 142–163. http://dx.doi.org/10.1037/1082-989X.11.2.142

Bennett, R. J., & Robinson, S. L. (2000). Development of a measure of workplace deviance. *Journal of Applied Psychology*, *85*, 349–360. http://dx.doi.org/10.1037/0021-9010.85.3.349

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

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

Bliese, P. D., & Ployhart, R. E. (2002). Growth modeling using random coefficient models: Model building, testing, and illustrations. *Organizational Research Methods*, *5*, 362–387. http://dx.doi.org/10.1177/109442802237116

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*, 147–158. http://dx.doi.org/10.1037/0033-2909.101.1.147

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). New York, NY: Routledge (Lawrence Erlbaum Associates).

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review*, *43*, 557–572. http://dx.doi.org/10.2307/2094779

Greenland, S. (2002). A review of multilevel theory for ecologic analyses. *Statistics in Medicine*, *21*, 389–395. http://dx.doi.org/10.1002/sim.1024

Hofmann, D. A. (1997). An overview of the logic and rationale of hierarchical linear models. *Journal of Management*, *23*, 723–744. http://dx.doi.org/10.1177/014920639702300602

Hofmann, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research in organizations. *Journal of Management*, *24*, 623–641. http://dx.doi.org/10.1177/014920639802400504

Hofmann, D. A., Griffin, M. A., & Gavin, M. B. (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, CA: Jossey-Bass.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). In G. A. Marcoulides (Ed.), *Quantitative Methodology Series*. New York, NY: Routledge.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology*, *67*, 219–229. http://dx.doi.org/10.1037/0021-9010.67.2.219

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. http://dx.doi.org/10.1037/0021-9010.69.1.85

James, L. R., Demaree, R. G., & Wolf, G. (1993). r_{wg}: An assessment of within-group interrater agreement. *Journal of Applied Psychology*, *78*, 306–309. http://dx.doi.org/10.1037/0021-9010.78.2.306

James, L. R., & LeBreton, J. M. (2010). Assessing aggression using conditional reasoning. *Current Directions in Psychological Science*, *19*, 30–35. http://dx.doi.org/10.1177/0963721409359279

James, L. R., & LeBreton, J. M. (2012). *Assessing the implicit personality through conditional reasoning*. Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/13095-000

James, L. R., & Williams, L. J. (2000). The cross-level operator in regression, ANCOVA, and contextual analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 382–424). San Francisco, CA: Jossey-Bass.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. New York, NY: Guilford Press.

Krasikova, D. V., & LeBreton, J. M. (2012). Just the two of us: Misalignment of theory and methods in examining dyadic phenomena. *Journal of Applied Psychology*, *97*, 739–757. http://dx.doi.org/10.1037/a0027962

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781849209366

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*, 815–852. http://dx.doi.org/10.1177/1094428106296642

Lyons, K. S., & Sayer, A. G. (2005). Longitudinal dyad models in family research. *Journal of Marriage and Family*, *67*, 1048–1060. http://dx.doi.org/10.1111/j.1741-3737.2005.00193.x

Mossholder, K. W., & Bedeian, A. G. (1983). Cross-level inference and organizational research: Perspectives on interpretation and application. *The Academy of Management Review*, *8*, 547–558. http://dx.doi.org/10.2307/258256

Myers, R. H. (1990). *Classical and modern regression with applications* (2nd ed.). Belmont, CA: Duxbury Press.

Ostroff, C. (1993). Comparing correlations based on individual-level and aggregated data. *Journal of Applied Psychology*, *78*, 569–582. http://dx.doi.org/10.1037/0021-9010.78.4.569

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*, 209–233. http://dx.doi.org/10.1037/a0020141

R Development Core Team. (2017). *The R Project for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org

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

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Washington, DC: Sage.

Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods*, *12*, 695–719. http://dx.doi.org/10.1177/1094428108327450