CHAPTER 3

Methods for Studying Leadership

John Antonakis

Chester A. Schriesheim

Jacqueline A. Donovan

Kishore Gopalakrishna-Pillai

Ekin K. Pellegrini

Jeanne L. Rossomme

Understanding and conducting leadership research requires knowledge of research methodology. In this chapter, we provide an overview of important elements relating to the use of research methods and procedures for studying and making inferences about leadership. We anticipate that our chapter will provide readers with useful pointers for conducting and for assessing the quality of leadership research. Apart from a basic review, we also cover some advanced topics (e.g., use of levels of analysis, structural-equation modeling) that should be interesting to researchers and graduate students. Where possible, we have attempted to write the chapter using mostly nontechnical and nonstatistical terminology so that its contents are accessible to a comparatively large audience.

In the sections that follow, first we define science and differentiate it from intuition. Then we discuss the nature of theory and how it is tested and developed. Next, we discuss methods of research, differentiating between quantitative and qualitative modes of inquiry. We also discuss the major categories of research, including experimental and non-experimental methods. Finally, we discuss potential boundary conditions of theories and how boundaries can be detected.

Our Knowledge of the World: Science Versus Intuition

As readers have ascertained from the first two chapters of this book, leadership is a complex concept; however, its complexity does not render it immune from scientific study. To appreciate or conduct research in the field of leadership, one must first understand some fundamentals about science, scientific theories, and the way in which theories are tested.

Science differs from common knowledge in that it is systematic and controlled (Kerlinger & Lee, 2000). Scientists first build theories, then conduct research in a systematic fashion that subjects their theories to controlled empirical testing. Lay or nonscientific persons tend to use theory loosely and to accept theories or explanations that have not been subjected to rigorous testing. Scientists also use controls in an effort to isolate relationships between variables so as to uncover causality and ensure that other presumed causes, as well as chance, have been ruled out. In contrast, lay people tend to accept causal explanations that may be based merely on associations—whether putative, spurious, or actual—between variables. Scientists also discard untestable explanations of phenomena, whereas nonscientists may accept such explanations (Kerlinger & Lee, 2000).

In our day-to-day lives, we are all "commonsense" or "intuitive" psychologists in the sense that, by using data from personal observations or secondary sources, we try to figure out the world and people around us (L. Ross, 1977). In our interactions with the world, we try to make sense of events or causes of outcomes and often use explanations that are intuitively appealing. We all have "theories" about how the world works and "test" these theories daily through our observations. Sometimes our theories are confirmed, either because our theories are correct, because we have created a self-fulfilling prophecy, or perhaps even as a result of chance; other times, we process information consistent with our theories (see S. T. Fiske, 1995; Merton, 1948b). As intuitive psychologists, however, we often are wrong (L. Ross, 1977; Tversky & Kahneman, 1974). We probably are more likely than others to fall prey to our false beliefs in trying to understand the phenomenon of leadership because many of us, as followers or leaders, have experienced leadership first hand. Thus, it is important that we establish our knowledge of the world using the scientific method.

Readers of this book probably would not trust a "witch doctor"—offering a concoction of crushed rhinoceros nose, bat claws, and frog eyes, mixed in python blood—to cure a deadly disease, no matter how much experience or how many satisfied clients the witch doctor may have or claims to have. Readers would rightly

expect that a medical treatment has been subjected to rigorous scientific tests to determine whether the treatment is safe and efficacious in fighting a particular illness.

Similarly, one would expect that readers, as users of leadership theory or as consumers of leadership training programs, self-help guides, books, and the like, would select models and methods whose validity has been established. Unfortunately, consumers of leadership products cannot turn to an authority—akin, for example, to the Food and Drug Administration—to determine whether a particular product is "safe for consumption" (i.e., valid). Whether intentionally or unintentionally, producers of leadership products may cut a few corners here or there knowing that there is no administrative body that keeps a watchful eye on what they peddle to unsuspecting consumers. Thus, we believe that the need to understand the methods and procedures used in leadership research is imperative not only for those who study leadership, but for those who consume its products as well.

The Scientific Method

Slife and Williams (1995) stated that the goal of science is to "establish with some authority the causes of events, and provide an understanding of phenomena that is objective and uncontaminated by traditions and subjective speculations" (p. 173). The scientific approach involves four steps (Kerlinger & Lee, 2000). The first is the expression of a problem, obstacle, or idea (e.g., why do some leaders motivate followers better than do other leaders?). In this step, the idea is not refined; it may be vague or based on unscientific "hunches." The second step is the development of conjectural statements about the relationship between two or more phenomena (e.g., behavioral leadership style is associated with motivation for reason X). The next step is the reasoning or deduction step. During this step, the consequences of the conjectural statements are developed. Specifically, the types and results of hypothesized relationships between variables are speculated (e.g., democratic leadership is positively associated to motivation, whereas autocratic leadership is negatively associated with motivation, under conditions Y). The last step includes observation, testing, and experimentation to put the scientist's reasoning to empirical test to determine if the hypothesized relation (or difference) between the variables is statistically probable and not merely attributable to chance (i.e., the scientist measures the leadership styles of the leader and the corresponding degree of motivation in followers to determine if there is a statistically significant relation between leader style and motivation).

Rigorous scientific research adheres to a specific set of standards and guidelines at every step in the process (Filley, House, & Kerr, 1976). The process must be rigorous in order for drawn inferences to be valid. Rigor is defined as "methodology and argument that is systematic, sound, and relatively error free" (Daft, 1984, p. 10). Rigor deals with the process and not the outcome, whereas quality deals with the outcome of the process, which includes the overall contribution of the work and the degree to which findings mirror the real world (Kerlinger & Lee, 2000). There may be a trade-off between the rigor of the process, on one hand, and the real-world

contribution and generalizability of the study, on the other. This trade-off occurs because there are many interactions among phenomena in real-world settings, and studying specific, isolated relationships in scientific research may yield findings that do not mirror the relationships that occur in their natural settings. Because adherence to the scientific process creates these types of generalizability issues, this trade-off must be considered when any type of scientific research is planned or evaluated. We revisit this issue later.

What Is Theory?

The creation of theoretical frameworks that can explain a practical phenomenon is the ultimate aim of science (Kerlinger, 1986). Unfortunately, nonscientists may not necessarily have a positive perspective of theory. One often hears definitions of theory that include such phrases as "something that is not testable," "something unproven," "something hypothetical," and "an ideal." These definitions of theory are incorrect. Lewin (1945) once stated, "Nothing is as practical as a good theory" (p. 129). A theory must, therefore, reflect reality and be applicable to practice. If it does not reflect reality, it cannot be applicable to practice; hence, it is not a good theory.

Theories advance our knowledge of social life by "proposing particular concepts (or constructs) that classify and describe [a] phenomenon: then they offer a set of interrelated statements using these concepts" (Pettigrew, 1996, p. 21). Theories are not absolute laws; they are always tentative and are based on probabilities of events occurring (Pettigrew, 1996; see also Popper, K. R. (1965)). According to J. P. Campbell (1990), the functions of theory are to (a) differentiate between important and less important facts, (b) give old data new understanding, (c) identify new and innovative directions, (d) provide for ways to interpret new data, (e) lend an understanding to applied problems, (f) provide a resource for evaluating solutions to applied problems, and (g) provide solutions to previously unsolvable problems.

A theory is a collection of assertions that identify which elements are important in understanding a phenomenon, for what reasons they are important, how they are interrelated, and *under what conditions* (i.e., boundary conditions) the elements should or should not be related (Dubin, 1976). Specifically, Kerlinger (1986) defined theory as being "a set of interrelated constructs (concepts), definitions, and propositions that present a systematic view of phenomena by specifying relations among the variables, with the purpose of explaining and predicting the phenomena" (p. 9). At this point, we should clarify the distinction made between the terms *construct* and *variable* (Bacharach, 1989). A construct cannot be observed directly, whereas a variable can be observed and measured. Based on this distinction, relationships among constructs are set forth by propositions, and relationships among variables are described by hypotheses. A theory thus includes propositions regarding constructs and their hypothesized relations. The consequence of the hypothesized relation, in the form of measured variables and their interrelations (or differences), is what researchers directly test (Kerlinger, 1986). The types of statistical methods to test for relations and differences are briefly introduced later.

Following Filley et al. (1976), we make two distinctions concerning theory: (a) the method of theory development and its implications for testing and (b) the purpose of the theory. Theories may be inductive or deductive (Dubin, 1976). Inductive theories flow from data to abstraction. Here, the emphasis is on theory building, whereby a researcher uses observations, data, or the conclusions and results of other studies to derive testable theoretical propositions. The flow of development is from specific observations to more general theory. Deductive theory focuses on theory testing by beginning with a strong theory base. Deductive theory building flows from abstraction to specific observation; it places the emphasis on testing hypotheses derived from propositions and thus attempts to confirm hypotheses derived from theory.

The second distinction is the purpose of the theory, which may be descriptive or prescriptive. Descriptive theories concern actual states—they describe what "is" or "was." For example, the full-range leadership theory is purported to include nine leadership dimensions, ranging from the passive-avoidant leader to the inspiring and idealized leader, and each of the dimensions is hypothesized to predict certain leader outcomes (see Avolio, 1999; Bass, 1998). In contrast, prescriptive theories are normative and describe what "should occur." For example, Vroom and associates (Vroom & Jago, 1988; Vroom & Yetton, 1973) have developed a normative decision-making model of leadership that suggests which leadership style should be used in certain environmental contingencies. Of course, a descriptive theory also can be prescriptive, and prescriptive theory must be built on a description of sorts. For example, based on empirical studies and theoretical reasoning, Bass (1998) and Avolio (1999) stated that passive-avoidant leadership should be used least often, because it is associated with negative follower outcomes, whereas active and idealized leadership should be used most often.

Filley et al. (1976) developed five evaluative criteria for judging the acceptability of theory. First, a good theory should have internal consistency. That is, it should be free of contradictions. Second, it should be externally consistent. That is, it should be consistent with observations and measures found in the real world (i.e., data). Third, the theory must be testable. The theory must permit evaluation of its components and major predictions. Fourth, the theory must have generality. That is, it must be applicable to a range of similar situations and not just a narrow set of unique circumstances. Finally, a good theory must be parsimonious. A simple explanation of a phenomenon is preferred over a more complex one.

Hypotheses

Once a theory and its propositions have been developed, it needs to be tested for external consistency. To do this, constructs need to be associated with measured variables (i.e., empirical indicators), and propositions must be converted into testable hypotheses. A hypothesis is a conjectural statement about one or more variables that can be tested empirically. Variables are defined as observable elements that are capable of assuming two or more values (Schwab, 1980). Whereas variables,

as manifest indicators, are observable, constructs are not directly observable (Maruyama, 1998). Hypotheses are the working instruments of theory because they help the scientist disconfirm (or fail to disconfirm) a theory by not supporting (or supporting) predictions that are drawn from it.

Good hypotheses must contain variables that are measurable (i.e., that can assume certain values on a categorical or continuous scale) and must be testable (Kerlinger & Lee, 2000). Hypotheses should not be in the form of a question. For example, "Does leader supportiveness have an effect on group performance?" is not testable and therefore is not a hypothesis but instead a research question, which may be valuable for guiding research per se. An acceptable hypothesis would be that "Leader supportiveness is positively associated with group performance." Furthermore, hypotheses should not contain value statements such as "should," "must," "ought to," and "better than." Finally, hypotheses should not be too general or too specific. If they are too vague and general, they cannot be tested. If they are too specific, research findings may not be generalizable and therefore may make less important substantive contributions to knowledge.

Theory Development

The process of developing a good theory involves four steps (Dubin, 1976). The first step is the selection of elements whose relationships are of interest (e.g., the relation between leader style and follower motivation). All relevant elements should be included in the theory, but those that may be viewed as extraneous should be left out.

Once the above elements are chosen, the second step is to specify how the elements are related. In particular, what impact do the elements have on one another? In this step, specific relationships must be articulated (e.g., democratic leadership is positively associated with motivation).

The third step specifies the assumptions of the theory. These include justifications for the theory's elements and the relationships among them. The step also involves the specification of boundary conditions within which the elements interact and are constrained (Dubin, 1969). The boundary conditions set forth the circumstances to which the theory may be generalized and are necessary because no theory is without boundaries. Bacharach (1989) noted that theories are bounded by the implicit values of the theorist, as well as by space and time boundaries. Boundaries refer to the limits of the theory; that is, the conditions under which the predictions of the theory hold (Dubin, 1976). The more general a theory is, the less bounded it is (Bacharach, 1989).

To revisit our democratic-autocratic example, a potential boundary could be national culture or risk conditions. That is, in some cultural contexts, autocratic leadership may be more prevalent because of large power differentials between those who have power (e.g., leaders) and those who do not (e.g., followers). Thus, in that context (i.e., boundary), the nature of the relation between a leader's style and follower motivation changes, and we may find that autocratic leadership is positively related to motivation. Thus, cultural context can be referred to as a

moderator of leadership effectiveness but also as a contextual factor affecting the type of leadership that may emerge, as we will discuss later.

In the fourth step, there must be specification of the system states in which the theory operates. A system state is "a condition of the system being modeled in which the units of that system take on characteristic values that have a persistence through time, regardless of the length of time interval" (Dubin, 1976, p. 24). In other words, the system states refer to the values that are exhibited by units constituting the theoretical system and the implications associated with those values.

Methods of Research

Qualitative and Quantitative Distinctions

According to F. Williams (1992), there are two major methodological research traditions: qualitative and quantitative methods. Quantitative methods should be utilized when the phenomenon under study needs to be measured, when hypotheses need to be tested, when generalizations are required to be made of the measures, and when generalizations need to be made that are beyond chance occurrences. As noted by Williams, "if measures are not apparent or if researchers cannot develop them with confidence, then quantitative methods are not appropriate to the problem" (p. 6). Thus, choice of which approach to use will depend on a number of factors.

Leadership researchers typically have used quantitative approaches; however, to better understand complex, embedded phenomena, qualitative approaches to studying leadership are also necessary (see Alvesson, 1996; Bryman, Stephens, & Campo, 1996; Conger, 1998). However, "qualitative studies remain relatively rare . . . [and should be] the methodology of choice for topics as contextually rich as leadership" (Conger, 1998, p. 107). Given the contextual and complex nature of leadership, it is important that qualitative methods—as a theory-generating approach—complement quantitative methods, whose strengths are in theory testing. The reasons for using qualitative methods to study contextually rich and holistically embedded phenomena are discussed below.

Quantitative approaches to scientific inquiry generally rely on testing theoretical propositions, as previously discussed, in regard to the scientific method. Contrarily, qualitative approaches focus on "building a complex, holistic picture, formed with words, reporting detailed views of informants, and conducted in a natural setting" (Creswell, 1994, p. 2). A qualitative study focuses on meanings as they relate in context. Lincoln and Guba (1985), referred to the qualitative approach as a post-positivist [i.e., postquantitative] *naturalistic inquiry* method of inquiry. Unlike the positivist, the naturalist "imposes no a priori units on the outcome" (Lincoln & Guba, 1985, p. 8).

Lincoln and Guba (1985) also argued that positivism constrains the manner in which science is conceptualized, is limited in terms of theory building, relies too much on operationalizing, ignores meanings and contexts, and attempts to reduce

phenomena to universal principles. Lincoln and Guba stated that the weakness of the positivist approach includes the assumption that phenomena can be broken down and studied independently, while ignoring the whole. Moreover, the distancing of the researcher from what is being researched, the assumption that sampling observations in different temporal and spatial dimensions can be invariant, and the assumption that the method is value free further limit this approach. Lastly, Lincoln and Guba stated that instead of internal and external validity, reliability, and objectivity, naturalists strive for credibility, transferability, dependability, and confirmability of results.

Qualitative research has often been criticized for being biased. Because qualitative analysis is constructive in nature, the data can be used to construct the reality that the researcher wishes to see—thus creating a type of self-fulfilling prophecy stemming from expectancy-based information processing. The result, therefore, may be in "the eye of the beholder," in a manner of speaking. That is, the observer may see evidence when he or she is looking for it, even though contrary evidence also is present (see S. T. Fiske, 1995). Thus, special controls—or triangulation—must be used to ensure that results converge from different types or sources of evidence (Maxwell, 1996; Stake, 1995; Yin, 1994).

As can be seen from the above discussion, qualitative and quantitative paradigms of research differ on a number of assumptions, and the selection of qualitative over quantitative approaches will depend entirely on the purpose and nature of the inquiry (Creswell, 1994). Similar to quantitative methods, qualitative methods employ a variety of techniques to acquire knowledge and may include (Creswell, 1994): (a) ethnographies, in which a cultural group is studied over a period of time; (b) grounded theory, where data are continually used, categorized, and refined to inform a theory; (c) case studies, in which a single bounded case or multiple cases are explored using a variety of data-gathering techniques; and (d) phenomenological studies, which examine human experiences and their meanings. Qualitative research also can include the active participation of the researcher in an intervention, a practice normally labeled as participatory action research (Greenwood & Levin, 1998; W. F. Whyte, 1991).

For some specific examples relating qualitative and quantitative research-gathering methods for leadership research, refer to Kroeck, Lowe, and Brown (Chapter 4, this volume). Because the vast majority of research that is conducted in the leadership domain is quantitative in nature and because theory can be tested appropriately only with quantitative methods, we will focus the rest of the chapter on the quantitative paradigm and its associated methods.

Categories of Research

Kerlinger (1986) stated that although a large part of research using the scientific method was generally experimental in nature, much research nowadays in the behavioral sciences involves non-experimental research. Because there are many possible causes for behavior, Kerlinger argued that the most dangerous fallacy in

science is the "*post hoc, ergo propter hoc* [fallacy]: after this, therefore caused by this" (p. 347). This fallacy is most likely to occur in non-experimental conditions; that is, conditions in which the scientist does not have control over the independent variables "because their manifestations have already occurred or because they are inherently not manipulable. Inferences about relations among variables are made, without direct intervention, from concomitant variation of independent and dependent variables" (p. 348). Based on the above distinction between experimental and non-experimental research Kerlinger (1986) classified research into four major categories or types, namely (a) laboratory experiments, (b) field experiments, (c) field studies, and (d) survey research.

Laboratory Experiments

According to Kerlinger (1986), laboratory experiments are useful for manipulating independent measures in an isolated situation, thus yielding precise and accurate measures. Variables are studied in pure and uncontaminated conditions, because participants are assigned randomly to groups to ensure that the resultant variance in dependent outcomes is accounted for by the independent measures. Laboratory experiments are useful for testing predictions and for refining theories and hypotheses. Kerlinger, however, noted that some disadvantages exist in laboratory experiments; namely that the effect of the independent measures may be weak because they do not exist in real-life conditions or have been artificially created. Thus, the results of a laboratory experiment may not be generalizable to real-life situations and may lack external validity even though they have high internal validity.

Binning, Zaba, and Whattam (1986), for example, experimentally manipulated group performance cues (i.e., indicators of effectiveness) to determine whether they influenced ratings of the group's leader. Binning et al. exposed subjects to a video of a problem-solving group and then obtained ratings of the leader's behavior and effectiveness under either good or bad group performance cues. They found that leader behavior and global effectiveness ratings were susceptible to performance cues, especially leader behavior that was not richly described (i.e., for which raters did not have sufficient information to construct an accurate schema). Thus, important here is the identification of the conditions (i.e., low information conditions) under which ratings are susceptible to performance-cue effects.

A second example is a study by J. M. Howell and Frost (1989), who experimentally manipulated three leadership styles using trained confederate leaders (i.e., individuals working for the researcher) and who exposed participants to either a high or low productivity group, also including confederates. Their results showed that participants working under a charismatic leader generally had better work outcomes and were more satisfied than were participants under a considerate or structuring leader. Furthermore, followers under charismatic leaders were equally influenced irrespective of group productivity standards. An important element of this study is that leader charisma was manipulated in a laboratory setting.

Laboratory studies thus can provide valuable insights about leader processes and perceptions and are particularly useful for studying research from many

perspectives (e.g., information processing, trait, behavioral, contingency). Indeed, laboratory studies of leadership have been conduced as far back as the 1930s (e.g., Lewin & Lippitt, 1938) to assess the impact of leadership behavioral style (e.g., democratic, autocratic). However, according to D. J. Brown and Lord (1999), leadership researchers, especially from the "new leadership" paradigm, have not yet made full use of experimental methods. They argued that nonconscious information-processing functions—integral to how leaders are legitimized and how followers perceive leaders—can be studied only in controlled laboratory settings. Furthermore, Brown and Lord stated that laboratory research is useful for studying unique interactions of independent variables and for untangling the unique effects of highly correlated leader dimensions that are not easily distinguishable in the field because of their co-occurrence.

Field Experiments

Kerlinger (1986) defined a field experiment as a study that occurs in a real setting in which both the independent variables and group assignment are under the control of the researcher. The degree of control that the researcher has in comparison to laboratory conditions is less, suggesting that other independent variables that are not under the control of the researcher may have a bearing on the dependent measures. Like laboratory experiments, field experiments are useful for testing theory and are applicable to studying leadership from many points of view; however, because field experiments occur in real settings, they are also useful for finding answers to practical problems. Similar to laboratory studies, field experiments unfortunately are not very common in leadership domain. We discuss two examples below, focusing on the effects of leadership training interventions.

In a field experiment, Barling, Weber, and Kelloway (1996) demonstrated that training interventions can change managers' leadership styles (including their charisma), resulting in increased organizational commitment of subordinates (i.e., a subjective dependent measure) and improved organizational financial outcomes (i.e., an objective dependent measure). In another example, using a military sample, Dvir, Eden, Avolio, and Shamir (2002) tested the efficacy of a leadership training intervention on leaders' direct and indirect followers. They found that transformational leadership training had a more positive effect on direct follower development and on indirect follower performance than did eclectic leadership training. A major implication of these studies is that leadership can be taught through interventions by making leaders aware of their leadership style as they and others perceive it and by facilitating a learning and goal-setting process conducive to the development of a more effective leadership style.

Field Studies

According to Kerlinger (1986), field studies are non-experimental, occur in real-life situations, and attempt to discover important variables and their interrelationship by studying "the relations among the attitudes, values, perceptions,

and behaviors of individuals and groups in the situation" (p. 372). The major weakness of field studies is that the researcher has no control over the independent variables and, hence, the variance that is exhibited in the outcome measures. As with laboratory and field experiments, many types of research questions can be answered using this type of research.

For example, Atwater, Dionne, Avolio, Camobreco, and Lau (1999) linked individual characteristics of military cadets to their leadership effectiveness/emergence both cross-sectionally (i.e., at one point in time) and longitudinally (i.e., where subjects were tracked over a period of time—another rarity in leadership research). As regards to the longitudinal results, they found that measures of physical fitness and prior leadership experience (measured at Year 1) were significantly associated with leader effectiveness and emergence (measured at Year 4). These results suggest that the potential to lead can be predicted using individual-difference measures.

Bass, Avolio, Jung, and Berson (2003) studied the relations between the leadership style (i.e., transformational or transactional leadership, or nonleadership) of military commanders and the simulated performance of their units. Leadership style of platoon lieutenants and sergeants was measured several weeks before their units participated in a training simulation. Bass et al. found that the leadership styles of the lieutenants and the sergeants were significantly associated with platoon performance in the field and that the impact of the leadership styles was partially mediated by the potency and cohesion of the platoon.

Survey Research

Researchers use survey research to determine the characteristics of a population so that inferences about populations can be made. Surveys generally focus on "the vital facts of people, and their beliefs, opinions, attitudes, motivations, and behavior" (Kerlinger, 1986, p. 378) and generally are cross-sectional in nature. Data for surveys can be gathered in a number of ways, including in-person interviews, mail questionnaires, telephone surveys, and the Internet. Survey research is practical to use because it is cheap and relatively easy to conduct. Although this type of research is limited by the questionnaire or interview format used, it can be quite useful if properly designed. Survey methods have been used to answer many types of research questions emanating from all leadership perspectives. Indeed, the leadership field is replete with survey research; thus, in the interest of saving space, we do not list any studies here (refer to Kroeck et al., Chapter 4, this volume, for examples of survey instruments).

In conclusion, the type of research used will depend on the type of research questions to be answered. For example, if a researcher wishes to determine whether there is cross-situational consistency in individuals' leadership styles and whether certain traits are predictive of leadership in different situations, then that researcher would be advised to use rotation designs, similar to those used by Kenny and Zaccaro (1983) and Zaccaro, Foti, and Kenny (1991), in which leaders' situational conditions (e.g., group composition and task demands) vary (refer to Zaccaro, Kemp, & Bader, Chapter 5, this volume, for further discussion of these studies). The

remaining chapters in this book—especially those in Part III—provide interesting examples of the different types of research methods used.

Research Design and Validity Issues

The design of a study is integrally linked to the type of research used. There are three important facets of design (see Kerlinger, 1986). First, the design should ensure that the research question is answered. Second, the design should control for extraneous variables, which are independent variables not intended to be measured by the study that may have an effect on the dependent variables. This is referred to as internal validity. Third, the study should be able to generalize its results to a certain extent to other subjects in other conditions; that is, it should have some bearing on theory. This may also be referred to as external validity. Thus, Kerlinger stated that congruence must exist between the research problem and the design, what is to be measured, the analysis, and the inferences to be drawn from the results (refer also to Cook, Campbell, & Peracchio, 1990, for issues relating to validity). By adequately taking the above points into account in the planning stages of research, the researcher increases the likelihood of drawing accurate inferences and conclusions about the nature of the problem investigated. This is because the design dictates what is to be observed, how it is to be observed, and how the data must be analyzed.

As suggested previously, external validity may be an issue with experiments; however, according to C. A. Anderson, Lindsay, and Bushman (1999), the external validity threats of experiments are exaggerated, given that findings in laboratory and field settings generally converge. To avoid problems associated with external validity, researchers should draw from a sampling frame that is representative of the population and control as well as possible the level of artificiality, recognizing that some artificiality may be inherent in laboratory experiments. Furthermore, the researcher must ensure that the experimental manipulation has had the required effect (i.e., the manipulation was perceived as intended). A manipulation check usually is used to address this issue.

As for non-experimental research, causation is an important caveat. One of the dicta of scientific research is that correlation does not imply causation. Furthermore, typical of survey methods is the assessment of two variables that are intended to represent different constructs, with the data gathered using the same method from the same source (e.g., a self-report questionnaire) (see Avolio, Yammarino, & Bass, 1991). The observed relationship between the variables may be attributable to a true relationship between the constructs or to bias resulting from the use of a common source and a common data collection method. This issue has long been of concern to psychologists (D. T. Campbell & Fiske, 1959; D. W. Fiske, 1982), particularly with respect to self-report measures (Sackett & Larson, 1990).

Finally, there are other ways in which to go about doing research, which cannot be described by the four categories of research discussed above but which can shed interesting perspectives on the study of leadership. As an example, the analysis of historical data, which is not traditionally used in applied psychological settings and

which is usually associated with the qualitative research paradigm (e.g., case studies) (see Yin, 1994), cannot be placed in the above classification scheme. Refer to Simonton (2003) for a discussion on the quantitative and qualitative analyses of historical data, and to House, Spangler, and Woycke (1991) for an example in leadership research.

Methods of Statistical Analysis

Analytical techniques (e.g., regression analysis) are nothing more than tools by which the researcher is able to test and refine research hypotheses and, subsequently, theories. The hypotheses and characteristics of the data should drive the types of analysis conducted, not vice versa. Readers who have not had a thorough introduction to statistics and psychometrics should consult one of the many available books (see, e.g., Nunnally & Bernstein, 1994; Sharma, 1996; Tabachnik & Fidell, 2001). Briefly, there are three major classes of methods: (a) *dependence methods* of analysis, which are used in order to test the presence and strength of an effect; (b) *interdependence methods,* which are used to determine relationships between variables and the information they contain; and (c) *structural-equation models*, which are the newest generation of analytical techniques useful for testing complex models of both presumed cause and effect that include latent/unobserved variables and the effects of measurement error. Structural-equation modeling (SEM) is based on, and simultaneously uses principles of, path analysis, regression analysis, and factor analysis. It is thus a very powerful and flexible statistical methodology for testing theoretical frameworks (see Bollen, 1989; Kline, 1998; Maruyama, 1998). Applications of SEM in leadership are discussed below.

Context and the Boundary Conditions of Leadership Theories

The context in which leadership is enacted has not received much attention, leading House and Aditya (1997) to state that "it is almost as though leadership scholars . . . have believed that leader-follower relationships exist in a vacuum" (p. 445). Further calls have been made to integrate context into the study of leadership (Lowe & Gardner, 2000) and organizational behavior (Johns, 2001; Rousseau & Fried, 2001). Context constrains the variability that potentially can be measured (Rousseau & Fried, 2001), is linked to the type of leadership that is considered as being prototypically effective (Lord, Brown, Harvey, & Hall, 2001), and determines the dispositional antecedents of effective leadership (Zaccaro, 2001).

Zaccaro and Klimoski (2001) argued that "unlike the situation as a moderator [as in contingency theories of leadership], we view situation or context as boundary conditions for theory building and model specification" (p. 13). According to R. M. Baron and Kenny (1986), a moderator is a categorical or continuous "variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable" (p. 1174). Thus, in addition to finding moderators that alter the strength or the direction of a

relationship between an independent (e.g., leader behavior) and dependent (e.g., leader outcomes) variable, situations also could be conceived as range restrictors of the types of independent variables that emerge. In other words, context should be considered in attempts to understand how phenomena like leadership emerge, and not only the extent to which or how context may affect the strength of relations between independent (e.g., leadership) and dependent (e.g., organizational effectiveness) variables (see Shamir & Howell, 1999).

In striving to generalize phenomena, leadership researchers often overlook context and draw incorrect conclusions about how a phenomenon is modeled (Antonakis, Avolio, & Sivasubramaniam, 2003). For example, a leadership model (e.g., the transformational, transactional, and laissez-faire models) may be universal in the sense that the constructs of which it is composed can be measured across different contexts; however, if the phenomenon is by nature contextually sensitive (i.e., leader prototypes and expected behaviors vary by context), the phenomenon may "work" in a similar manner only when sampling units are homogeneous and not when sampling units are heterogeneous (Antonakis, Avolio, et al., 2003). In other words, the emergence and enactment of a behavior may vary by context, which includes the following, among others:

1. National culture—some leader behaviors and their enactment may be universal or may vary systemically as a function of national culture (e.g., refer to Den Hartog & Dickson, Chapter 11, this volume; see also Brodbeck et al., 2000; Hofstede, 1980; Koopman et al., 1999; Meade, 1967).
2. Hierarchical leader level—leadership differs qualitatively depending on whether leadership is exercised at a high (e.g., strategic) or low (e.g., supervisory) hierarchical level (e.g., refer to Sashkin, Chapter 8, this volume; see also Antonakis & Atwater, 2002; D. J. Brown & Lord, 2001; J. G. Hunt, 1991; House & Aditya, 1997; Lord, Brown, Harvey, et al., 2001; Lowe, Kroeck, & Sivasubramaniam, 1996; Sashkin, 1988a; Waldman & Yammarino, 1999; Zaccaro, 2001; Zaccaro & Klimoski, 2001).
3. Organizational characteristics—organizational stability, organizational structure (e.g., bureaucratic vs. organic environments), and so forth affect the type of leadership that may be necessary (refer to Brown, Scott, & Lewis, Chapter 6, this volume; see also Antonakis & House, 2002; Avolio, 1999; Bass, 1998; Lord & Emrich, 2000).
4. Leader and/or follower gender—leader behaviors may vary systematically as a function of leader gender or follower gender because of gender-role expectations and other factors (refer to Eagly & Carli, Chapter 12, this volume; see also Eagly & Johnson, 1990; Lord, Brown, Harvey, et al., 2001).
5. Leadership mediated by electronic means (i.e., e-leadership; see Avolio, Kahai, & Dodge, 2001; Avolio, Kahai, Dumdum, & Sivasubramaniam, 2001; see also Antonakis & Atwater, 2002; Shamir, 1999; Shamir & Ben-Ari, 1999)—the nature of the influence process may vary in conditions where leaders and followers are not face-to-face.

Types of leader behaviors that are enacted may vary as a function of context; however, so may the dispositional (i.e., leader characteristics) antecedents linked to those behaviors. For example, cognitive requirements differ as a function of leader hierarchical level (Zaccaro, 2001; see also Zaccaro et al., Chapter 5, this volume, and Sashkin, Chapter 8, this volume). Finally, the level of analysis at which a leadership phenomenon may hold will also vary by context (Antonakis & Atwater, 2002; Waldman & Yammarino, 1999). That is, the level at which leadership operates may vary from individual to group, or organizational level, depending on the context (e.g., CEO-level vs. supervisory-level leadership) in which it is enacted.

It thus becomes evident that researchers need to consider contextual factors and levels of analysis in theory development and measurement (see Antonakis, Avolio, et al., 2003; Dansereau, Alutto, & Yammarino, 1984; Schriesheim, Castro, Zhou, & Yammarino, 2001; Zaccaro & Klimoski, 2001) and then use the appropriate methods to gauge these factors and levels (i.e., boundaries). This perspective should complement traditional notions of context, in which context is seen as a moderator of the relation between leader characteristics (e.g., traits, behaviors) and leader outcomes.

The Detection of Boundary Conditions of Theories

How are boundary conditions of theories detected? In this section, we discuss four quantitative approaches that are useful in detecting moderators or contextual factors, which can be considered as boundary conditions of theories (see James, Mulaik, & Brett, 1982). For each of the four methods discussed below (i.e., levels of analysis, structural-equation modeling, moderated regression, and meta-analysis), we provide relevant examples from the leadership literature to which readers may refer. We do not cover interaction effects—as in traditional factorial ANOVA designs, which are useful for the detection of moderation (see R. M. Baron & Kenny, 1986), especially in experimental designs—because we assume that most readers are familiar with the technique. In our discussions below, we focus extensively on levels-of-analysis issues, because many, if not most, researchers in psychology and management generally ignore levels-of-analysis concerns. We also extensively discuss structural-equation modeling, a very powerful but underutilized statistical method in leadership research.

Levels of Analysis

In recent years, level of analysis has emerged as an important issue in organizational research (e.g., Klein, Dansereau, & Hall, 1994; Rousseau, 1985). Organizations are inherently multilevel, because individuals work in dyads and groups. Furthermore, several groups or departments make up an organization, organizations make up an industry, and so on. Few, if any, constructs are inherently restricted to one level of analysis. Typically, constructs may operate at one or more levels, such as individuals, dyads, groups, departments, organizations, and industries (Rousseau, 1985).

Thus, it is now increasingly recognized that good theory should also specify the level of analysis at which its phenomena operate (see Dansereau, Alutto, et al., 1984), because level of analysis can be conceived as a boundary condition of a theory (Dubin, 1976). Levels of analysis might even be considered as a sixth criterion that should be assessed when one considers a theory's adequacy (in addition to Filley et al.'s, 1976, original five criteria discussed previously).

Levels of analysis make studying leadership in organizations very complex because leadership phenomena may operate on any or all levels. Consider, for example, performance. This variable can be examined on an individual, dyadic, group, departmental, organizational, or industry level. The specification and testing of the levels at which constructs of a theory operate is therefore important because research conclusions may differ as a function of the level of analysis that is employed (Dansereau, Alutto, et al., 1984; Klein et al., 1994).

The correlates or causes of individual performance may be very different from the correlates or causes of group or organizational performance. Because obtained results may vary depending on the level at which the data are analyzed, erroneous conclusions can be drawn if the level tested is incongruent with the level specified by a theory (Dansereau, Alutto, et al., 1984; W. H. Glick, 1985). These erroneous conclusions include, among others, ecological fallacies (i.e., using aggregated data to make inferences about individuals) or individualistic fallacies (i.e., using individual-level data to make inferences about groups; see Pedhazur, 1997, for examples) and other types of misspecifications (Rousseau, 1985).

A recent special issue of *Leadership Quarterly* (2002, Vol. 13, No. 1) highlighted problems of failing to take levels into account. For example, a data set purposefully analyzed at the individual level of analysis using regression techniques failed to detect a moderation effect of leadership climate on task significance of followers, because the moderation effect did not operate at the individual level of analysis (Bliese, Halverson, & Schriesheim, 2002). However, when using various methods designed to test for multilevel effects, evidence of a moderator (i.e., leadership climate) was evident (see Bliese & Halverson, 2002; Gavin & Hofmann, 2002; Markham & Halverson, 2002) and leadership climate exhibited group-level affects that were ignored by other methods.

Thus, if theory, analysis, and measurement level are not correctly specified and aligned, "we may wind up erecting theoretical skyscrapers on foundations of empirical jello" (Schriesheim, Castro, Zhou, et al., 2001, p. 516). For example, if a leader's behavior—as perceived by followers—is not homogeneously viewed by followers, then the leader's behavior operates at the individual level of analysis. Therefore, any inferences that are made should be based on the individual and use individual-level data, because individual responses are independent. However, if the leader's behavior is viewed homogeneously, then it is justifiable to aggregate the individual data to the group level and make inferences at the group level of analysis, because individual responses are dependent on group membership.

There are several examples of leadership research incorporating levels-of-analysis perspectives. However, in leadership research only a small group of researchers have used methods developed to test levels-of-analysis effects. We cite three examples

below of levels-of-analysis perspectives regarding leadership theories and show that consideration of the boundary condition of levels leads to better theories that are more applicable to practice. Although there is now a broader range of researchers testing theories using multilevel methods, we stress that there is *still insufficient attention paid to levels-of-analysis* issues by mainstream leadership, organizational behavior, and management researchers.

In the first example, Yammarino (1990) demonstrated that correlations between leader behavior and outcome variables may exhibit differential levels-of-analysis effects. For instance, the correlation between a leader's group-directed initiating-structure behavior and role ambiguity perceptions of followers may be valid at the individual level of analysis; however, the correlation between a leader's group-directed consideration behavior and role ambiguity perceptions of followers may be valid at the group level of analysis. The practical implications of such results suggest that leaders enact their initiating-structure behaviors in an individualized manner with respect to follower role ambiguity (because individual perceptions of these behaviors are not homogeneously held by followers). In contrast, when follower perceptions with an outcome (e.g., role ambiguity) are homogeneous, then the leader can enact group-wide behaviors. Yammarino (1990) also demonstrated that raw (i.e., total) correlations may be ambiguous because correlations of similar magnitudes may have different levels of analysis effects. Finally, Yammarino's study is notable for showing that wording questions in a way that uses the group as a referent does not ensure that the group will be the resulting level of analysis. This finding has implications for question or item design.

In the second example, Schriesheim, Castro, Zhou, et al. (2001) argued that although leader-member exchange (LMX) theory was originally conceived of as being a dyadic theory (i.e., that the level of analysis is the leader-follower dyad), current conceptualizations of the theory basically ignore the theory's dyadic roots and generally analyze LMX data either from the leaders' or followers' perspectives. Unfortunately, the theory is still not well understood, because level-of-analysis effects associated with the theory have hardly been tested (Schriesheim, Castro, & Cogliser, 1999; Schriesheim, Castro, Zhou, et al., 2001). After testing the levels-of-analysis effects of the theory, Schriesheim, Castro, Zhou, et al. (2001) found relatively strong support for the dyadic nature of the theory, a finding with important implications for data gathering and analysis practices. That is, matching data (i.e., using both followers and corresponding leaders) should be collected, and levels-of-analysis effects should be examined on the dyadic relation (while ruling out competing levels). Until we have more research that tests the level of analysis at which the theory is valid, "all the extant [LMX] research is fundamentally uninformative" (Schriesheim, Castro, Zhou, et al., 2001, p. 525).

In the third example, Yammarino and Dubinsky (1994) showed that even though the literature explicitly or implicitly asserted that transformational leadership would be evident at a higher level of analysis than the individual level, empirical data suggest that most of the effects are evident at the individual level of analysis. That is, "what one individual perceives differs from what others perceive" (Yammarino & Dubinsky, 1994, p. 805), even though these individuals may be perceiving the same

social object (i.e., their leader). These results have been demonstrated repeatedly (e.g., Yammarino & Bass, 1990; Yammarino, Spangler, & Dubinsky, 1998), suggesting that transformational (and transactional) leaders have differential effects on followers and that they do not cause similar outcomes in groups of followers (refer to Antonakis & Atwater, 2002, for theoretical arguments concerning when level of analysis may be expected to change).

Multilevel Analysis Methods. There are many ways in which one can test for levels effects, from whether it is appropriate to aggregate data to a higher level of analysis (e.g., aggregating a group of followers' ratings of a leader) to partitioning variance in outcome measures (i.e., determining the extent to which outcomes are accounted for by various independent variables). We briefly discuss four statistical procedures that are useful for conducting research at multiple levels of analysis. For more technical discussions regarding the strengths and weaknesses of the procedures, refer to Castro, (2002), George and James (1993), Schriesheim (1995), Yammarino (1998), and Yammarino and Markham (1992).

The r_{wg} coefficient (James, Demaree, & Wolf, 1984) was developed primarily as an index of interrater agreement within a single group (Castro, 2002). As such, it is useful to determine whether responses of group members can be aggregated to the group level; however, it is less useful for making comparisons between groups (i.e., to determine whether variability exists between groups) because it was not designed to perform such a function (see Castro, 2002; Yammarino & Markham, 1992). Thus, it may be more lenient than the other aggregation-testing methods discussed below.

Intraclass correlation coefficients (ICCs) are useful for determining the extent to which variance of individual responses is attributed to group membership and whether means of groups are reliably differentiated (see Bliese, Halverson, et al., 2002; Castro, 2002). Thus, by design, ICCs may be more generally useful than is r_{wg}, especially if between-group variability is theoretically important and requires detailed examination.

Within and between analysis (WABA) assess the extent to which variables—and their covariation—exhibit variation within or between groups by testing the within- and between-group variation for statistical and practical significance (Dansereau, Alutto, et al., 1984). WABA should be used when the assumption for aggregating to a higher level requires testing, especially if between-group variation is theoretically important (see Yammarino, Dubinsky, Comer, & Jolson, 1997). WABA also can be used to test certain types of cross-level effects (Castro, 2002; Yammarino, 1998).

WABA has been criticized for being too conservative, especially in situations where range restrictions may exist between groups. For example, in many of the studies cited previously that used WABA (e.g., Yammarino & Bass, 1990; Yammarino & Dubinsky, 1994; Yammarino, Spangler, et al., 1998) and found individual-level effects for leadership, sampling units were restricted to the same or similar contexts. Thus, it is possible that because of similar between-group contextual conditions and the resulting range restriction that occurs, higher levels of analysis effects were

attenuated and thus not detected by WABA (see George & James, 1993; Schriesheim, 1995). Therefore, one must very carefully choose heterogeneous sampling units if one is looking for between-group differences (Yammarino & Markham, 1992). Practically speaking, though, as an omnibus test, WABA is particularly useful for determining the extent of within-group homogeneity, whether higher-level relations are evident, and whether higher-level entities are useful as predictors of outcome variables (Schriesheim, 1995; Yammarino, 1998).

Hierarchical linear modeling (HLM) is a regression-based procedure useful for testing relationships that exists at different levels of analysis (i.e., cross-level hierarchical relationships) (Hofmann, 1997). Specifically, HLM can estimate different sources of variance in dependent variables attributable to independent variables at the same or higher level of analysis and can be used for detecting cross-level moderators (Castro, 2002; Gavin & Hofmann, 2002). HLM does not test whether it is appropriate to aggregate data at a higher level of analysis (Castro, 2002; Hofmann, 1997; Yammarino, Dubinsky, et al., 1997). Thus, HLM should be complemented by other procedures designed to test whether aggregation is appropriate (Castro, 2002). For extensions of HLM-type procedures using structural-equation modeling (i.e., MLM or multilevel modeling), refer to Heck and Thomas (2000).

Structural-Equation Modeling

Multiple-groups structural-equation modeling (SEM) can be useful for the detection of moderators (R. M. Baron & Kenny, 1986; James, Mulaik, et al., 1982; Kline, 1998). Although this is a potentially powerful and useful procedure, unfortunately, it is "not used all that frequently . . . in social science literature" (Maruyama, 1998, p. 257) and in leadership research in particular. Indeed, we found it difficult to locate a good example (see Dorfman, Howell, Hibino, Lee, Tate & Bautista, 1997) of the procedure in the leadership literature.

Essentially, the question asked in this type of analysis is "Does group membership moderate the relations specified in the model[?]" (Kline, 1998, p. 181). In this type of analysis, one is thus interested in testing whether structural parameters are equivalent across groups. For example, one can test whether the relation between an independent variable (e.g., leader behavior) and an outcome factor (e.g., follower motivation) is the same across groups. In this case, the grouping variable is the moderator. If the constraint of equivalence in the parameter does not result in a significant decrement in model fit, then one can conclude that group membership does not moderate the relation between the independent and dependent variable (Kline, 1998). If, however, the fit of the constrained model is significantly worse, then one can conclude that the grouping variable moderates the particular structural relation (Kline, 1998).

Using the procedures discussed above, Dorfman, Howell, Hibino, et al. (1997) found that the relations between leader behaviors and leader outcomes varied when using several national cultures as moderator groups (for further discussion of their results, refer to Den Hartog & Dickson, Chapter 11, this volume). Note that in this example, the moderator variable was a grouping factor (i.e., a categorical variable).

However, interaction effects using continuous moderator variables also can be specified in SEM (see Kenny & Judd, 1984).

On a more basic level, SEM also is useful for determining whether constructs are equivalent across theoretically distinct groups (e.g., based on national culture). There are various degrees of equivalence or invariance. For example, one can determine whether the same indicators for a factor are associated in the same way across groups (i.e., factor pattern loadings are the same). This condition is referred to as configural invariance (Vandenberg & Lance, 2000). Establishing configural invariance is absolutely critical so that further tests between groups can be conducted, because if the factors are not associated with the same indicators, then the "respondent groups were [not] employing the same conceptual frame of reference" with respect to how the factor is perceived (Vandenberg & Lance, 2000, p. 37). Once configural invariance is established, further intergroup restrictions can be placed to determine if the model fit becomes significantly worse based on the chi-squared difference test (i.e., likelihood ratio test between the configural condition and the more restrictive condition). An example of a study in which a configural invariance condition was tested is that of Gillespie (2002), who found that a 360-degree leadership feedback survey did not satisfy the criteria for configural invariance across four cultures. This result suggests that the raters in the various countries conceptualized the factors in different ways, indicating that further cross-country comparisons on the factors would be unwarranted.

A more stringent condition of equivalence is metric invariance, which, if found, suggests that a unit change in the factor can affect the indicators in the same way across different groups (Vandenberg & Lance, 2000). If metric invariance is found, one is confident that the loadings of the respective indicators of the factor are equivalent between groups (e.g., see Dorfman, Howell, Hibino, et al., 1997). Finally, scalar invariance tests whether the intercepts of the indicators are equivalent across groups, a result that has implications regarding systematic bias in responses (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000). If a model has configural, metric, and scalar invariance, or at least partial configural, metric, and scalar invariance, then further substantive tests (e.g., latent mean difference tests, see Byrne, Shavelson, & Muthén, 1989; Steenkamp & Baumgartner, 1998) can be performed (e.g., Antonakis, Avolio, et al., 2003, conducted a latent mean difference test on a range of leadership dimensions using leader-follower gender as a grouping factor).

Unfortunately, however, Cheung and Rensvold (2000) noted that oftentimes researchers take what they called a naive approach, especially when conducting cross-cultural research, by simply comparing scales (i.e., item composites) across groups using a *t* test or an ANOVA-type procedure, without having verified whether what is being compared is equivalent (see X. Zhou, Schriesheim, & Beck, 2001, for an illustration of how invariance may be tested). Thus, the "observed difference, if there is one, may be due to a [grouping] difference; however, it may also be due to one or more invariance failures in the measurement model" (p. 188). As an example, Kakar, Kets de Vries, Kakar, and Vrignaud (2002) compared Indian and U.S. managers on a range of leadership factors and found significant mean

differences between the groups using a *t* test. Thus, because they did not establish model equivalence before they made comparisons, their results should be viewed as highly suspect.

Moderated Regression

As noted by Schriesheim, Cogliser, and Neider (1995), testing for moderation using regression was (and often still is) conducted by using a median split in the moderator variable and then running separate regressions for the groups to determine whether parameter estimates differed. A more powerful procedure, moderated multiple regression (see Cohen & Cohen, 1983), treats the moderator variable as a continuous variable, thus retaining more information on the interaction effect (Schriesheim, Cogliser, et al., 1995). Moderated multiple regression uses a hierarchical regression procedure in which the independent variable is added first to the regression model, followed by the moderator variable. In the final step, the interaction (i.e., the cross-product) is added and assessed to determine the unique variance that is predicted by the interaction term (see also Schriesheim, 1995).

As demonstrated by Schriesheim, Cogliser, et al. (1995), who reanalyzed the data of a study conducted by Schriesheim and Murphy (1976), major discrepancies can arise between the split-group and moderated regression procedures. For example, although Schriesheim and Murphy originally found that role clarity did not moderate the relation between leader behaviors and outcomes (e.g., follower performance and satisfaction), Schriesheim, Cogliser, et al. (1995) actually found moderation effects. Specifically, initiating structure was positively associated with performance under high role clarity conditions but negatively associated with performance under low role clarity conditions. However, initiating structure (and consideration) was negatively associated with satisfaction under high role clarity conditions but positively associated with performance under low role clarity conditions. The results relating to performance were opposite to what was hypothesized, as based on the prevailing literature. Given that previous studies testing for moderation have used the weaker procedure, Schriesheim, Cogliser, et al. (1995) concluded "we do not believe that sound conclusions can be drawn from much of the existing leadership literature—at least not without making heroic (and quite probably unfounded) assumptions about the equivalency of analytic procedures and, of course, the results they produce" (p. 135). For further information about moderated regression analysis and how to improve the likelihood of finding moderators in leadership research, refer to Villa, Howell, Dorfman, and Daniel (2003).

Meta-Analysis

Glass (1976) differentiated three levels of analysis in research: *primary analysis*, in which researchers analyze the results of data they have gathered; *secondary analysis*, which refers to the reanalysis of primary data to better answer the original research question or to answer new questions; and *meta-analysis*, which is the analysis and synthesis of analyses of independent studies. This technique is useful where a domain

needs to be synthesized, by integrating the results of various studies and reconciling their diverse or conflicting findings (for various meta-analytic techniques, refer to Hedges & Olkin, 1985; Hunter & Schmidt, 1990; Rosenthal, 1991). Essentially, with meta-analysis, the researcher can determine the population correlation coefficient—or other indicators of effect—between independent and dependent measures by controlling for measurement and statistical artifacts (i.e., errors). Meta-analysis also is useful for detecting moderator effects (see Sagie & Koslowsky, 1993).

Lord, De Vader, and Alliger (1986) reanalyzed Mann's (1959) data and found that intelligence showed a mean corrected correlation of .52 with leadership emergence (whereas Mann had declared that intelligence was not strongly associated with leadership). As mentioned by Antonakis, Cianciolo, and Sternberg (Chapter 1, this volume) and Zaccaro et al. (Chapter 5, this volume), the Lord et al. study was instrumental in refocusing efforts on individual-difference predictors of leadership, and the main the reasons why Lord et al. were able to show strong links between intelligence and leadership was that they used a more sophisticated and applicable method to synthesize research findings.

An example of a meta-analysis in which moderators were detected is the study of Lowe, Kroeck, et al. (1996), who analyzed the results of studies using the Multifactor Leadership Questionnaire (MLQ). They found that the type of organization (i.e., public or private) in which leadership measures were gathered and the type of criterion used (i.e., subjective or objective) moderated the relations between leadership measures and outcome variables. For example, Lowe et al. reported that the mean corrected correlation between charisma and effectiveness was .74 and .59 in public and private organizations, respectively, and that the difference between the correlations was significant. Many other interesting examples of meta-analytic studies have been conducted in the leadership field (e.g., DeGroot, Kiker, & Cross, 2000; Eagly & Johnson, 1990; Judge, Bono, Ilies, & Gerhardt, 2002; Schriesheim, Tepper, & Tetrault, 1994), to which readers are encouraged to refer.

Conclusion

In conclusion, we have presented a summary of key research-method subdomains and other important research-related issues that are likely to have value for persons interested in reading about or conducting leadership research. Because competency in research methods requires extensive knowledge, we could only brush the surface in many of the areas that we have discussed. We therefore would like to encourage readers to pursue further knowledge in the different areas by consulting and reading the reference citations that appear throughout this chapter and in the other chapters that have discussed research findings. By doing so, they are likely to improve the quality of future leadership research, as consumers demand higher-quality research and producers take the necessary precautions to ensure the validity of their findings.

Prior to concluding this chapter, we would like to address one final issue—that of judging the quality of the "concluding" or "discussion" section of a paper, article,

or chapter. The discussion section of any of these addresses the implications of the results for future theory and, perhaps, leadership practice. Discussion sections sometimes get speculative, with implications and conclusions being drawn that go well beyond what the data actually support. Readers therefore need to evaluate such discussions carefully and in the light of the actual study findings or the findings of other research in the same domain. Discussion sections oftentimes talk about needed new directions for future research and comment on limitations of the reported research, such as a study's use of convenience sampling, which can be helpful in highlighting concerns that others can address in future research. The practice of acknowledging study shortcomings is also congruent with the philosophy of science position that all research is flawed in one way or another and that knowledge can be built only through multiple investigations of the same phenomena, using different samples, methods, analytic practices, and so forth. The analogy that we sometimes use is that doing research is like fishing with a net that has one or more holes (shortcomings). Using multiple nets, each with holes in different places, enables us to catch and hold that which we seek (knowledge), whereas the use of a single net is much less likely to do so.

To conclude, we will ask a question that probably crossed the minds of many readers. How is it possible that we need to make exhortations about raising the quality of research? Do journals not ensure that the research they publish is of a high quality? Indeed, most research published in top-quality journals is well done; however, from time to time, even top journals have holes in their review nets. Furthermore, there are journals and then there are *journals*! In other words, journal quality is variable, and not all journals publish high-quality scientific work. Thus, research that is not published in high-quality journals or that does not meet the necessary standards of scientific research must be looked at skeptically, especially if this research is published in popular books that address fashionable topics (e.g., traits that have not been tested scientifically and that have not demonstrated incremental predictive validity over and above established psychological traits) (see Antonakis, in press; Zeidner, Matthews, & Roberts, in press).

We trust that we have made the point that to generate scientific knowledge that is useful for society is not an easy task and requires extensive methodological expertise. If one chooses to drink the elixirs of witch doctors, one does so at one's own peril!