PART II

Perspectives on Inquiry

CHAPTER 2

Quantitative Social Science Methods of Inquiry

Timothy R. Levine

As you are reading this chapter, it can be inferred that (a) you are interested in the topic of interpersonal communication and (b) you want to learn more about it. Why else would you be reading a chapter in the *Handbook of Interpersonal Communication*? Obviously, you want good information. No reasonable person would want to waste his or her time learning useless, misinformed, or misleading stuff. But how can you obtain this new knowledge, and when you do, how will you know if the knowledge is reliable? Will what you think you know stand the tests of replication and application? How can we sort out what is real insight and what is just high-sounding "bullshit."[1] There is no shortage of quackery out there. So how can you meaningfully judge the worth of the research findings you read? If you want to have a sound understanding of interpersonal communication, you need answers to questions such as these. This chapter is for you.

Some may hold the misconception that quantitative researchers are, of necessity, statistics nerds who have a natural affinity for math and computer programs. But statistics may be less central to really good quantitative interpersonal research than you may think. Knowing statistics, at least as it is most often taught, is only a small and relatively unimportant part of good quantitative social science methods. Anyway, statistics are just tools for getting at and understanding the real topic of interest. Knowledge of methods is necessary to critically assess the utility and plausibility of knowledge claims, but methods and statistics are not the same thing, and both are only a means to an end. In fact, too much focus on the rules and details of statistics can lead to boring work that lacks insight and even, ironically, to research findings that are erroneous or misleading. The path to enlightenment is not through ever more complex multivariate analyses, random effects models, hierarchical linear modeling, or structural equation models with correlated error terms, and others ad nauseam. With a few notable exceptions, simple statistics are almost always better (Boster, 2002; Cohen, 1990; Smith, Levine, Lachlan, & Fediuk, 2002). A good picture, plot, or graph of the data is often more informative than a statistical probability, such as $p < .05$. Thoughtful research questions coupled with sound measurement and tight research design typically reduce the need for complex inferential statistics. If a

finding cannot be seen with a careful "eyeballing" of a descriptive presentation of the data and instead is only apparent from asterisks on a printout, there is a good chance that the finding is not very meaningful.

I challenge those skeptical of the above claims to read my work, especially my more recent work that appears in top journals.[2] The reader will never find a MANOVA (multivariate analysis of variance) or similar statistic in any of my articles. Although I can do path analysis by hand, you won't see path analysis or structural equation modeling (SEM) in many of my articles. The reader is much more likely to see a graph of my raw data than an HLM (hierarchical linear modeling) analysis. If I could get away with not reporting *p* values at all, I would in many instances. This preference for simplicity and design and measurement over statistical analysis has not kept me out of leading communication journals. On the contrary, I have enjoyed much success and over the years. My rejection rates are much lower when I submit to *Communication Monographs* or *Human Communication Research* than to the regional journals. I take this as evidence that I am doing some things right.

Robert Abelson (1995) wrote that a good social scientist, and this includes quantitative interpersonal communication scholars, needs the skills of a good detective, an ethical trail lawyer, and an engaging storyteller. A good interpersonal detective gathers, sorts, and integrates evidence to solve some social, theoretical, or communicative mystery. The metaphorical ethical trail lawyer uses the evidence to make a persuasive case about the facts without distorting them, and the storyteller weaves those facts and arguments into an engaging narrative.

To these three skill sets, I will add two more—"substantive feel" and quantitative-scientific literacy. By substantive feel, I mean that to really understand interpersonal communication, you need to understand people, social interaction, and social context. You must be a curious and an insightful "people watcher" who wonders, "Why did a person do that?" and who can convert observations into questions, hypotheses, or theories that can be tested. But this needs to be done in a systemic, disciplined way if the new knowledge is going to be convincing and withstand scrutiny. This is where quantitative-scientific literacy comes in. But quantitative-scientific literacy does not demand or imply statistical complexity (Boster, 2002).

One of the best contemporary examples of these points is observed in the work of Bob Cialdini (1980)—and the idea of "full-cycle social psychology," which can be aptly changed to "full-circle interpersonal scholarship." Read almost any of the many influential Cialdini research articles, and you can see this in practice. Notable examples include his work on norms and littering (e.g., Cialdini, Kallgren, & Reno, 1991), football studies on basking in reflected glory (Cialdini et al., 1976), and the effectiveness of various compliance-gaining strategies (e.g., Cialdini et al., 1975). Cialdini's research starts with an insightful observation of human interaction. First, he identifies the principle(s) of human nature that seem to be at play. Then, he designs experiments to document the phenomenon he has observed and to distinguish between plausible mechanisms that explain the phenomenon. Finally, he applies the understanding that he had gained from his field experiments to the original observation and context. When you read Cialdini's work, it is clear not only that he is a skilled detective who tells convincing and engaging stories but also that he has both a keen substantive feel and a serious scientific mojo.

Abelson (1995) also draws a useful distinction between the record and the lore. The record is the set of all research findings on a topic, and the lore is the social understanding of those findings. The two are not the same, and they can, in practice, be very different. Reading the record requires quantitative-scientific literacy and relying on the lore, which, often passed down in textbooks, literature reviews, or article discussion sections, can lead to the propagation of persistent myths. A few examples are provided in the following paragraphs.

Have you ever heard of Solomon Asch's (1956) famous conformity experiments? Briefly,

participants are shown a series of lines and asked which of three lines is of the same length as a comparison line. Participants make judgments either individually (control group) or in groups of eight people. When in groups, the other participants are actually research confederates, who sometimes all give the same wrong answer. Assuming that you learned about this study before, what percentage of Asch's participants always conformed? What percentage never conformed, always giving the correct answer regardless of others' incorrect answers? Which percentage is greater?

Next, have you ever heard of the tendency of people to explain their own behavior with situational causes but to attribute others' behavior to personal causes? Does research support this claim?

Now, the answers. In Asch (1956), 29 out of 123 (24%) participants never conformed, whereas only 6 out of 123 (5%) always conformed. The results are clearly presented in the original write-up. In attribution research, across 173 studies, research does not find evidence for actor–observer asymmetry in attributions (and this is not the same as the so-called fundamental attribution error; see Malle, 2006). Surprised?

The lore is fallible. Textbook presentations, and literature reviews and discussion sections in articles cannot always be trusted. This, by the way, is true of methods education too. A lot of information presented in methods texts "just ain't so" (Cohen, 1990).

This chapter is about creating and understanding the research record so that you do not have to rely on the lore. It is about methods of generating new knowledge and assessing the merit of existing knowledge claims involving interpersonal communication. Whether the reader is a cutting-edge interpersonal researcher, a student taking his or her first graduate class in interpersonal communication, or a practitioner of interpersonal skills, the ideas described here are useful for both the creators and the users of new knowledge. The focus is on methods of ensuring that the knowledge claims carry with them both credentials and real utility while at the same time being innovative, convincing, and engaging. This is not a chapter about the dry, boring, conventional, and intellectually barren approach to social science statistics that unfortunately populates the university curriculum, especially statistics textbooks. It is not about mindlessly following the conventions associated with testing data against implausible statistical distributions that provide the basis for the null hypothesis used in significance testing. It is not about conventional, philosophically illiterate takes on metatheory or the philosophy of science. Instead, it is about the disciplined thinking and the sophisticated approach to understanding that are needed to unlock the mysteries behind why people do what they do when they communicate interpersonally.

Social Science and the Quantitative–Qualitative (False?) Dichotomy

Quantitative, social-scientific communication research involves the application of a set of social-scientific methods for testing defensible knowledge claims about human communication based on empirical data, statistical description, and/or statistical inference. The term *quantitative approach* implies a contrast between quantitative research and qualitative research. The former seeks to quantify constructs of interest, whereas the latter does not. Qualitative research is sometimes portrayed as more exploratory, being useful in generating new ideas and understandings, while quantitative research is often seen as involving formal hypothesis testing. Both quantitative and qualitative research, however, can serve either function.

Qualitative methods can be used to formally test theories (e.g., Festinger, Riecken, & Schachter, 1956), and quantitative research does not need hypotheses (Rozin, 2001). Generally, quantitative research is useful when the phenomena of interest can either be classified as present/absent or when the phenomena have measurable attributes that

vary in degrees or amounts. If something can be scaled or counted, a quantitative approach can be used. The primary advantage of quantitative research is that statistical evidence can be used to enhance confidence in a knowledge claim. The second advantage is that many quantitative methodologies offer mechanisms to control nuisance variables, rule out rival explanations, and otherwise enhance confidence in knowledge claims.

More useful than the quantitative–qualitative distinction, however, is a broader distinction between social-scientific approaches and nonscience approaches such as rhetorical criticism, postmodern analysis, feminist scholarship, and critical scholarship. What makes an approach social scientific or not rests on issues other than whether or not numbers are involved. Science-based and nonscientific modes of research reflect very different understandings about the nature of knowledge, how knowledge is generated, what is useful to know, and how (or even if) we can have confidence in what we know. The quantitative–qualitative distinction need not involve these deeper philosophical differences about the nature of knowledge and knowledge generation. Nevertheless, most social scientists use quantitative methods at least some of the time, and the use of quantitative methods usually implies a social-scientific approach to knowledge generation, whereas qualitative research may or may not be social scientific in character. This said, a purely quantitative approach that shuns all qualitative inference will surely lack substantive feel. The best experimenters rely on personal experience for inspiration and talk to and observe their research participants to gain insight into their perspectives. Both qualitative and quantitative data have been extensively used in the works of master social scientists such as Musifer Sherif, Leon Festinger, Gerg Gigerenzer, and Paul Meehl, who happen to be my research heroes. My sympathies are for those not familiar with their brilliant work.

One hears from time to time that qualitative research is easier than quantitative work. After all, those wanting to do quantitative research need to take several statistics classes. Quantitative researchers need to do coding with multiple coders or spend hours in the lab running experiments. I suspect, however, that the opposite is more often true. Doing qualitative research well and effectively may be considerably more difficult than doing equally influential quantitative research. If one looks at the author studies that are published from time to time (e.g., Bunz, 2005; Hickson, Self, Johnson, Peacock, & Bodon, 2009), the most prolific authors in communication disproportionally seem to do quantitative research. The same trend is evidenced in the analysis of citation patterns (Levine, 2010b). The research in the field that is most cited is disproportionally quantitative. One (but certainly not the only) interpretation of these trends is that quantitative research may just be easier to do well.

Second, it seems fashionable in some intellectual circles to claim that one *does* both qualitative and quantitative research. Certainly, one can read and appreciate both and find both interesting, useful, and valuable. For the research consumer, being eclectic is a virtue. But for the aspiring young interpersonal researcher, perfecting one's craft in even one of these approaches is a lifelong pursuit, a goal that is unlikely to be ever fully attained even with considerable devotion. As for me, I know that I am still learning and am still perfecting my craft, even though I have been a full professor for a number of years now. There is so much to know, and life is so short. So for most young scholars, it is probably better to do one or the other reasonably well than to dabble in both. This does not mean that the two approaches cannot inform each other. But when actually doing research to create new knowledge, having only a passing familiarity with method is downright dangerous. It is so easy to be wrong, and when bogus information gets into the lore, the field suffers. Would-be researchers need to be sharper than the critics because criticizing others' knowledge claims is a whole lot easier than creating defensible new knowledge. All this suggests

that doing both qualitative and quantitative research well may be too ambitious for many mere mortals.

Philosophical Underpinnings

Most quantitative social-scientific research adopts the philosophical approach of scientific realism (Pavitt, 2001). The presumption is that there is a real world that exists beyond our perceptions and that the real world is potentially, at least partially, knowable. The goal of research is to get our understanding more closely aligned with this objective reality. The word *verisimilitude* describes this idea of closeness to reality. We want our theories and our findings to have verisimilitude, and the extent to which we can make a case that our theories and findings have verisimilitude is the bottom line in quantitative communication science.

Sometimes, quantitative social science approaches are mistakenly equated with logical positivism or operationalism, but these problematic philosophical perspectives have long (much before the current author was born) been out of favor (Meehl, 1986) and never held much sway in quantitative communication research anyway (Miller & Berger, 1999). Logical positivism was a philosophical view that held that the only meaningful knowledge is what we prove either by objective observation or by logical proof. Operationalism is a view of measurement that equates the attributes of things with their measures. For example, to an operationalist, communication apprehension (McCroskey, 1977) is a score on a communication apprehension scale (see Levine & McCroskey, 1990), not an estimate of an abstract construct. For the true operationalist, conceptual definitions are irrelevant.

The idea from Karl Popper (1959) that hypotheses and theories need to be falsifiable is both useful and widely accepted. This view holds that for ideas to be scientifically useful, they must be testable. Further still, disconfirming evidence has more logical force than supportive evidence. That is, evidence inconsistent with a hypothesis (i.e., not merely nonsignificant but significant in the wrong direction) provides stronger evidence that a hypothesis is false than findings consistent with a hypothesis provide evidence that a hypothesis is true.

I find that the ideas of Imre Lakatos (1978) expand on Popper's ideas and have much utility for understanding quantitative communication research. Very roughly, good research programs get out in front of the data. They allow the researcher to know what variables to study and lead the researcher to good new findings with verisimilitude. Degenerative research programs, in contrast, are always trying to catch up with the data. They are perpetually generating excuses for why the data were not supportive or require data spinning to create the illusion of support. Thomas Kuhn's (1996) philosophy of science, in contrast, is too relativistic and captures less well the aims of quantitative communication research.

Quantitative social-scientific research is usually *empirical*, meaning that knowledge claims are based on data and the data stem from observation. Quantitative social-scientific research also strives for *objectivity*. Complete objectivity is impossible to obtain, but methods are designed and evaluated by the extent to which the data are likely to be free from bias and the idiosyncrasies of the researcher. Finally, quantitative social-scientific research strives to be *authority free* and *self-correcting*. It is not the status of the researcher but the quality of the evidence that allows for confidence in a knowledge claim. Confidence in a finding or conclusion is enhanced though replication, and it is presumed that incorrect conclusions will ultimately be weeded out because they fail to replicate. If a finding has verisimilitude, other researchers should also be able to produce the same finding under conditions similar to the original research. For the social scientist, objective empirical observation, coupled with replication, provides the best path to verisimilitude over time. I do not myself place that much confidence

in my findings until I can replicate them or, better yet, until other researchers and other labs replicate my findings.

Research predictions and knowledge claims in quantitative communication research are usually probabilistic in nature, general within some specified conditions, and contextualized to within those specified conditions. Knowledge claims are usually probabilistic in that they are often based on statistical inferences that provide estimates of how likely or unlikely the data are given some set of assumptions. Findings and conclusions are general in that they tell us what is usual or typical within a situation or context. For example, deception research tells us that people tend to be truth biased and that people are more likely to believe other people regardless of the observed person's honesty (Levine, Kim, Park, & Hughes, 2006; Levine, Park, & McCornack, 1999). Such a finding does not imply that people always believe everything they hear or that they never think others are lying; it is just that this tends to be the case on an average. Finally, knowledge claims are contextualized in that they have boundary conditions; that is, all findings have conditions (specified or unknown) under which they apply and do not apply. For example, truth bias is diminished in situations where the person whose message is being judged is thought to have a strong motive to lie (Levine, Kim, & Blair, 2010).

Induction, Hypothetico-Deduction, and Abduction

Most quantitative social-scientific research involves the use of one or more of three types of reasoning: induction, hypothetico-deduction (HD), and abduction. With inductive reasoning, a researcher infers conclusions from data. Inductive conclusions have the basic form "Study *X* found such and such a result, therefore this is generally the case." A problem with this type of inference making is that the *data are always both finite and contextual*. Any given research finding is the result of a specific author or set of authors, involving some finite collection of research subjects, with the data collected at a specific point in time, with some specific method, and in some specific context. One always wonders if the conclusions would hold for different authors, with different subjects, at different times, with slightly different methods, and in different contexts. Unfortunately, there is no way to know this. Random selection and nonstudent samples do not help because one cannot random select across all time, researchers, methods, contexts, and so on. Sure, the findings can be replicated; replication is good, and replication enhances confidence. But replication does not fully solve the problem of induction either. Findings can never be replicated across all possible authors, subject populations, times, situations, and so on. And even if it were possible, surely researchers have better things to do with their time than pursuing near-infinite variations of silly replications. For these reasons, purely inductive conclusions are always logically unsatisfying. They leave the thoughtful research consumer with so many unanswerable "what if" and "why?" questions. Perhaps as a consequence, research involving purely inductive reasoning is often called *dust bowl empiricism*. The "dust bowl" charge is not flattering, and it refers to atheoretical research drawing purely inductive inferences. Induction, however, is not useless. Not by far. Inductive reasoning is a great way to get ideas and to document empirical regularities. Induction only becomes problematic when a line of research stops there and when it is the only inferential and logical tool in use.

While purely inductive research is often considered low rent and the stuff for intellectual bottom feeders, the *hypothetico-deductive* (HD) approach, in contrast, is often seen as the gold standard or the elite scientific ideal. HD research starts with theory. From theory, hypotheses are logically deduced. Research is designed specifically to test the hypotheses. If the data come out as predicted, support is inferred for the hypotheses

and for the theory that led to the hypotheses, otherwise the hypotheses and theory are considered unsupported and perhaps even falsified. Thus, the HD approach is all about formal a priori theory testing, and if one wants to formally test theory, the HD way is the way to go. Much has been written on the HD method, and Hemple's (1966) work is a classic reference that is worth reading.

The trouble is that formal theory testing and quantitative social science research are often considered to be one and the same. They are not. Yet most methods, texts, and classes (and graduate advisers, journal and grant reviewers, editors, etc.) tell students (and researchers) that the scientific method is about hypothesis and theory testing and that the only way science can legitimately be done is by creating and testing hypotheses, where all good hypotheses come from theory and produce $p < .05$ findings. But this cannot be so. Science is also about making new discoveries. Great discoveries such as the theory of evolution and DNA did not come from HD-style experiments. In fact, the HD method has little resemblance to most of the research published in the life sciences (Rozin, 2001). The HD method is fine and dandy for testing existing knowledge claims and theory, but it does not contain a good mechanism for generating new directions. And new findings are where the fun (and intellectual contribution) is. I do use HD when testing an existing theory, but this is not what I am trying to do in the vast majority of my published studies. A more flexible philosophical toolbox is needed.

When serving on graduate student thesis and dissertation committees, professors almost invariably ask the students questions such as "What is your theory?" and "What is you hypothesis?" but not "Do you have a hypothesis?" or "Does existing theory meaningfully apply to your question?" (Kerr, 1998). The presumption is that all good research needs hypotheses. After all, well-known and respected texts of scholars like Kerlinger and Lee (2000) say so explicitly. Recently, I was declined funding for a *National Science Foundation* grant application. The reviewer wrote that the project was really interesting and if the study were done, the results would be very important and the reviewer would cite such a study wherever possible.[3] But, the reviewer wrote, unfortunately the project did not merit funding because it is about finding out what people do, not about testing a specific hypothesis. This is ludicrous. My dissertation advisor has frequently told me, "If you always support your hypotheses, you will never learn anything new." The corollary to this is "If your research is confined to hypothesis testing, you are less likely to come up with an important new finding." Note that the argument here is not against theory or hypothesis. The argument is opposed to hypothesis testing as the *only* approach to science.

This is where abduction comes in. *Abductive reasoning* involves reasoning from the data to the best explanation of the data given the data. Think of abduction as a compromise between dust bowl induction and formal HD theory testing. A good example is my new "a few transparent liars" model (Levine, 2010a; Levine, Shaw, & Shulman, 2010). The model is an attempt to come up with an internally consistent logical framework that accounts for several reliable findings that do not make sense from existing theory. In short, abduction is a way to move from data toward theory. Because it is based on data, it has a better chance of being right than purely deductive theory building, but unlike purely inductive approaches, the goal is bounded explanation rather than mere generalization based on instances. Abduction is not formally valid as it exploits the affirming of the consequent fallacy, but it is nevertheless very useful in generating new theory. Abduction is not for the blind-rule follower, and the abductive-reasoning practitioner will be seen by the HD snobs as overly speculative (but not dust bowl). But if you look at what the really influential social scientists actually do rather than what the textbooks say they do, their theoretical ideas are typically generated through abduction rather

than pure induction, and only after they are formalized are the new ideas tested via HD.

A previous version of this chapter (by a different author) claimed that the HD approach characterized most published research on interpersonal communication. I have a different take and believe instead that most communication research is really induction masquerading as HD. It is a commonly held value that good research is theoretical (and I have much sympathy with such a belief so long as research working toward new theory is included as theoretical). The problem is that most research is not really theory driven. Instead, such research is really driven by what the author of the work finds personally interesting. Most interpersonal researchers do not start out by thinking "How can we give theory *XYZ* a really decisive test?" Instead, they think, "Such and such is interesting; how can I efficiently study it?" But authors can't say that if they want to get published. Convention disallows curiosity-based rationales. So the front end of the paper reviews "theory." Sometimes real theories are reviewed. Other times, the word *theory* is just put in, as if calling something a theory will make it so. But if a critical reader looks carefully at the hypotheses and the argument leading to the hypotheses, it will become evident that even though "theory" is given much lip service, the argument has this basic format: "Previous studies have found such and such, therefore we will too." Bingo, the argument is pure induction and is not HD at all! The twin tests to apply are (1) if the rationale for the hypotheses are based on what previous studies have found rather than on logical derivation from the theoretical proposition or (2) if you can take the theory out and the logic of the research still makes sense, without the guiding theory, it is not really HD research but instead pseudo-HD. I challenge the reader to apply these to tests to what they read in the published research. Inductive arguments dressed up as HD is probably the norm in interpersonal communication research; this unfortunate state of affairs stems from the misguided belief that HD is the only way to do good science. Ironically, if editors and reviewers were more tolerant of careful abduction, there might be more and better theory to test and then Professor Charles Berger would have less to complain about in his periodic theoretical rants (see Berger, 1991, 2005).

The Real Role of Hypotheses and Theory

The preceding should not be taken in any way as devaluating theory. It is often said that there is nothing more useful than a good theory. I wholeheartedly endorse this statement. But I would also advance a corollary that is sure to be more controversial. While a *good* theory is a truly wonderful thing, there are few things more intellectually damaging than a bad theory. This, of course, begs the question what makes a good theory good and what makes a bad theory bad? Good theories have lots of nice qualities. They are internally consistent, they have a large ratio of explanatory power to parsimony, they have good heuristic value in generating new predictions and reconciling old anomalies, they are testable and falsifiable, and so on. But the single most important criterion in distinguishing valuable theory from bad theory is verisimilitude. Good theories are, for the most part, consistent with the data, while bad theories get it mostly wrong. Good theories help us understand. They give us valuable insight into how the world works. Bad theories lead to misunderstanding and illusion. Good theories make us wiser. Bad theories make us not only more ignorant but also more pompous and arrogant in our ignorance. Good theories further science. Bad theories are mythologies that gather cultlike followings and seek to brainwash new converts into following false idols. You get the idea.

Good theories are valuable for a variety of reasons, but three of the most important are discussed here. The first and most obvious is that good theories provide explanation. They answer the "why" question. As interpersonal scholars know from uncertainty reduction theory (Berger & Calabrese, 1975), explanatory knowledge is deeper and richer than descriptive or predicative

knowledge. Good theories enable us to understand the principles, mechanisms, and processes behind the working of some interpersonal phenomena in a way that mere data cannot. Students of interpersonal communication want to know why things happen, and if for no other reason, this makes good theory invaluable.

Second, good theories point us in the right direction. They tell us what to study and how to study it. There are a near-infinite number of variables out there that we could include in our study. Theory not only helps us prioritize, it demands that we do so.

Besides the obvious importance of explanation and direction, a third indispensable function of theory and hypotheses is that they provide the best (and maybe only) path to external validity and generality. It has already been noted that all data are inherently finite and contextualized, which leads to the problem of induction. When we do a study, we want our findings to extend beyond the particular participants, time, and situation of our study. But our data are always limited in these ways. Some researchers try to overcome this problem by addressing what might be called the "surface features" of the research method. So they collect nonstudent data, they randomly select subjects from some population, they use multiple-message designs in conjunction with random-effect statistics, or they employ some similar ploy. All this is done to enhance the generality of the findings, but none of it ever solves the problem. The data are still finite. No statistical or methodological sleight of hand can change this. Imagine some large-scale random survey of the U.S. population. The findings may be general, within confidence limits, across the population, but they are still bound by time, the way the particular questions were worded, and so on. The problem simply cannot be overcome with a methodological fix. The only real solution is theory.

While data are always finite, theory is not constrained in the same way. Although theory is not infinite, good theory specifies boundary conditions that tell us when the theory does and does not apply. So theory, not data, allows bounded generalizations. What we can do is design our studies to test the boundary conditions of a theory. The purpose of data (and method), therefore, is very often not to make generalizations but to test generalizations (Mook, 1983). Well-designed studies can do that. Research, over time, allows for the development and testing of theory, and it is theory, in turn, that provides principles of human behavior with bounded generality. The theoretical knowledge of specifying under what conditions a proposition does and does not hold provides generality in a way that mere data cannot. This is why theory is important, and it is in this way, ironically, that basic theoretical research is often more useful, in application, than much so-called applied research.

Quantitative Basics

Constructs and Variables

Quantitative research involves variables. *Variables* are symbols to which numerals or numbers are assigned. Variables can also be defined as observable things that vary. That is, variables take on different values. Variables are contrasted with both constants and constructs. *Constants* are things whose values are fixed; they do not vary. *Constructs* are conceptual or theoretical entities that exist in the mind of researchers, whereas variables are observable. For example, the idea of the depth of self-disclosure is a construct, while scores on the depth dimension of a self-disclosure scale constitute a variable.

Quantitative researchers are interested in constructs and, usually, how two or more constructs are related to each other. Constructs are the ideas that are the topics of study. To research constructs, they are measured, and through measurement, values are assigned. The resulting values constitute a variable, and the relationships among variables can be tested, often with statistical analyses. When the variables are found to be

statistically related in some manner, then it is inferred that the constructs are likewise related in a similar manner. Thus, constructs are (albeit imperfectly) measured, and when values are assigned to represent these constructs, we call the resulting collection of values a variable. Variables are tested for statistical association or relationship, and inferences are made about how constructs are related based on the observed relationships among the corresponding variables.

When statistical analyses are used to test the relationships among variables, and some variables are conceived of as predictors or causes of other variables, the variables that are the predictor or cause variables are called *independent variables* while the variables that are predicted, the effects or outcomes, are called *dependent variables*. Dependent variables are so called because they are specified to depend on the independent variables. Often, the notation x is used to refer to the independent variable and y to the dependent variable. When graphing the relationship between x and y, x is plotted on the horizontal axis and y on the vertical axis.

There is nothing inherent in a variable that makes it independent or dependent. Instead, the identification of a variable as independent or dependent rests on theory, hypothesis, research design, and statistical analysis. For example, if we are looking at how biological sex affects the depth of self-disclosure, then depth of self-disclosure is a dependent variable (e.g., Dindia & Allen, 1992). In contrast, if we are considering how depth of self-disclosure reduces uncertainty in computer-mediated communication (e.g., Tidwell & Walther, 2002), then depth of self-disclosure is an independent variable.

Independent variables can be further classified as active (manipulated, induced) or measured (attribute) variables. The values of an active independent variable are set by the researcher. That is, they are induced or manipulated. This is not true for measured variables where the values are not under research control. Dependent variables are always measured and never active. As we will see later, the distinction between active and measured independent variables is the key to distinguishing experimental and quasi-experimental research from nonexperimental studies.

The Key Concepts of Variance and Control

The extent to which a variable varies is called *variance*. The more scores differ from one another, the more they vary, and hence the greater the variance. Statistically, variance has a more precise meaning. Variance refers to the average squared amount by which scores differ from the average score. It is helpful to remember that variance is in square units (Beatty, 2002). The square root ("un-square") of variance is the *standard deviation*. Both variance and standard deviation are metrics of dispersion, and the importance of these ideas is difficult to overstate.

Variance may be the single most important concept in quantitative research. Obviously, not all people are the same. Situations, too, differ from one another. So too do messages. Quantitative research helps us know why, when, how much, and to what effect things vary. This is often done by seeing if and how the variable we are interested in varies systematically with some other variable(s) of interest. That is, most quantitative communication research seeks to predict and/or explain how some variable of interest is related to another variable(s) of interest. This involves demonstrating that the variance in a variable is systematically related to the variance in another variable.

When variables are related, that is, one predicts, causes, or is associated with another, the variables are said to covary. So we will want to know if and how variables covary. This brings us to the idea of *function*. If x is an independent variable and y a dependent variable and x and y are systematically related to each other in some way, we can say that y is a function of x. Symbolically,

$y = f(x)$. So if we know what x is and we know the function, we can predict y. Then, we can do a study to see if our prediction is right. The trick, of course, is knowing the function. More about this later. Nevertheless, regardless of the specific function, it is a fundamental law of quantitative research that variance is required for covariance. That which does not vary cannot covary. If there is no variance, there can be no covariance and no meaningful function to test. If there is variance, however, then there may or may not be covariance, and if there is covariance, we will want to know the function that will let us predict y from our knowledge of x. In short, most quantitative research is about understanding variance, and understanding variance requires having variance to observe and the skills to figure out and test the function.

This can be done in two ways. First, we can try to measure naturally occurring variance and see if we find other variables that predict that variance. The second option is to try to create or induce variance. If you are not sure how this might be done, go find a working light switch. Turn it off and on. You just created variance in how much light there is. This is how experiments work, but that is getting too far ahead for the moment.

The flip side of variance is constancy. Constants are also extremely important in quantitative research because they are central to the idea of *control*. Because that which does not vary cannot covary, constants cannot be related to anything. Hence, holding something constant is a surefire way to control (rule out) its impact. What researchers try to do is to induce or assess the variance and covariance of the variables of interest while holding constant as much else as possible. Because constants never affect other things, they provide the best mechanism for the "control" of nuisance variance in research. An example from the research on the "probing effect" will be discussed later.

These principles of constancy and covariance provide the conceptual basis for experimentation. If some variable x has a causal impact on some other variable y, then changes in x will systematically produce changes in y. In an experiment, the researcher systematically alters the values of x and observes the values of y. If values of y systematically change when x is changed but y stays constant when x is constant, then evidence that x leads to y is obtained. Other potential causes of y are held constant so that they cannot have an impact on y and so that the impact of x can be isolated. The tighter the control over nuisance variables, the stronger the inference that is obtained from observing y vary as a function of inducing variance in x.

So just as the light switch controls the light, inducing variance in the switch induces variance in the lighting. Conceptually, I do the same thing in my deception experiments. I think that I know some situations in which people are more or less accurate at detecting deception. So I bring the research participants into my lab and throw the switch (my independent variable) and turn it on and off to find out how accurate they are. When I can turn my dependent measure on and off (so to speak) at will, I have evidence to enable me to understand what is going on.

A fun example is the Levine et al. (2000) norms and expectations experiment. The study was actually a class project in an MA class on research design, and my coauthors were the students in that class. The previous literature (Bond et al., 1992) had shown that people who observe unexpected, odd behavior find the target person less believable than a target person who was behaving in a more expected, typical way. Bond et al. (1992) would have us believe that this is due to expectancy violations. They claim that unexpected behavior is seen as less honest. It might be, however, that we are simply less likely to believe weird-acting people and that expectations have nothing to do with it. That is, if observation of unexpected, odd behavior makes it less likely that a person is believed, it might be due to expectancy violations or due to mere oddity. So we did an experiment to ferret this out. Research participants came into our lab and had a get-to-know-the-other-person interview with a

research confederate. The confederates either acted weird (e.g., obsessively picking their teeth with their fingers or following an invisible insect around the room with their eyes—this took lots of practice) or not, and the participants were forewarned ahead of time to either expect weirdness or not. We found that flipping the expectation switch made little difference. Flipping the weird behavior on and off, however, systematically raised and lowered perceived honesty. Bond's reasoning and the predictions of Burgoon's expectance violations theory (Burgoon & Hale, 1988) were wrong.

Constancy, variance, and covariance are also central in nonexperimental quantitative research. In nonexperimental research, variance is observed rather than created, and statistical analyses are used to document differences or association. Again, variance is essential because it is required for covariance.

Relationships Among Variables

Variables can be related to each other in different ways. This was discussed briefly earlier by introducing the idea of a function. Given that the goal of quantitative communication research is usually to document and explain how variables are related, knowing conceptually about the different types of relationships between variables (i.e., the different functions) is essential.

One possibility is that no relationship exists. That is, the variables are completely unrelated, and there is no covariance. Statistics, however, cannot be used to prove a complete lack of relationship, but statistical techniques such as confidence intervals (CIs), meta-analysis, or equivalence tests can show that a relationship is neither strong nor substantial (Levine, Weber, Park, & Hullett, 2008; for an example of equivalence testing in communication research, see Muthuswamy, Levine, & Weber's 2009 study of fear appeals in Africa). Nevertheless, it is unusual for communication researchers to purposely study weak relationships.

If variables are related, the simplest possibility is that the variance in one variable causes variance in the other. This situation is called a *direct causal relationship*. Documenting a direct causal relationship requires showing that (a) the variables covary, (b) the cause variable precedes the effect variable in time, and (c) the effect is not explainable by some other variable called a *spurious* cause. If some other variable causes both the independent and the dependent variables, then it will look like there is a direct relationship when actually there is not. The relationship is said to be spurious. A well-known example of a spurious relationship is that towns with more churches tend to have more bars. It would be a mistake, however, to conclude that church going and alcohol consumption are related based on such an association. Obviously, both are related to population size. Larger towns tend to have more of everything. Cook and Campbell (1979) offer an excellent discussion on the concept of causation.

Sometimes direct causal relationships are stringed together. So variable x may lead to variable y, and y, in turn, leads to z. This is called a *mediated* relationship, and y is said to mediate the relationship between and x and z. Mediated relationships are sometimes confused with *moderated* relationships. A moderated relationship exists when the effect of an independent variable on a dependent variable varies as a function of a third variable. That is, the focal relationship of interest is variable. For example, if the relationship between self-disclosure and liking is stronger for women than for men, then sex moderates the effects of self-disclosure and liking. Evidence for moderators is reflected by a statistical interaction effect. In the previous example, there is a (hypothetical) two-way interaction between self-disclosure and sex on liking. Baron and Kenny (1986) is the most cited reference on mediated and moderated relationships, but most communication researchers test mediation hypotheses with path analysis or SEM rather than with the Baron and Kenny approach. Most moderation hypotheses are tested with the interaction term in the analysis of variance (ANOVA), although

moderated multiple regression (see Cohen, Cohen, West, & Aiken, 2010) has been gaining popularity.

Research Design

Experimental Design

For a research study to technically be "an experiment," three necessary and jointly sufficient criteria must all be met: (1) at least one independent variable must be active, (2) at least one comparison or control group must be used, and (3) participants must be randomly assigned to the experimental conditions. Studies meeting the first two criteria but not the third are called quasi experiments. Studies lacking the first two criteria are nonexperimental. Nonexperimental research will be discussed later.

Random assignment means that each participant has an equal probability of being in each experimental condition. The primary purpose of random assignment is to guard against *selection effects* (Campbell & Stanley, 1963). A selection effect occurs when the participants in one condition differ systematically from the participants in another condition (a) in a way that is other than for the intended active independent variable and (b) in a way that affects the dependent measure. Random assignment should not be confused with haphazard assignment, which may have subtle systematic biases. If the first person who comes to the lab is in Condition 1, the second in Condition 2, and so forth, this is haphazard, not truly random. Who knows what kind of sneaky artifacts might slip into the study. Online random number generations that can be used to randomly assign participants to conditions can be obtained at http://www.randon.org.

Quasi experiments (i.e., studies with active independent variables but lacking random assignment) are not inherently flawed; they just lack one protection offered by a true experiment, and for this reason, researchers would not use a quasi-experimental design if a true experiment were a viable option. Furthermore, random assignment does not preclude selection effects; it just makes them less probable. Chance is lumpy (Abelson, 1995), and randomization is useful but imperfect.

Also, random assignment should not be confused with *random selection*. Random selection refers to how participants get into the study in the first place, and it means that every participant in the population from which the results will apply has an equal probability of participation. Random selection is an issue of generality or external validity, and studies of interpersonal communication seldom employ random selection.

Experiments and quasi experiments have a number of advantages over nonexperimental research. One main advantage is that better evidence for time ordering and causal order is obtained from experimental work than from nonexperimental research. Most theories and explanations specify some order among constructs/variables. Their logic says that this leads to that (and not vice versa). So let's say our theory states that x leads to y (i.e., y is a function of x). Let's make x an active variable, under the experimenter's control. If x leads to y, when we flip the x switch (i.e., turn it off and on), we should observe systematic variance in y. And if this is the case under nicely controlled circumstances, we will gain a real insight into the $y = f(x)$ relationship. Alternatively, if x leads to y and we turn the y switch, x is unaffected because "the effects" or "the cause" flows from x to y, not from y to x. The results of the experiments and quasi experiments give the researcher this type of evidence. This is the beauty of active independent variables. The experimenter can flip the switch at will. Causal evidence in nonexperimental work is more tenuous.

A second advantage of experimental research is greater control through research design. What one does in an experiment is flip the switch to your independent variable while holding all other potential causal variables constant. If, when we flip the x switch, y changes and the only difference is x being on or off, then we know that the variance in y is a function of x and nothing

else. The trick, of course, is holding everything else constant. To the extent we can do that, we have high-quality inference because we know that constants are inert. Control in real experiments, however, is never perfect, because it is impossible to be certain that everything else is really constant. Nevertheless, the better the control, the better the evidence obtained for the results. Nonexperimental work lacks control and comparison groups and has to rely on statistical rather than design-based controls, and thus concern over the impact of unknown spurious nuisance variables is typically greater in nonexperimental studies than in well-designed experiments. Furthermore, statistical control rests on the quality of the measurement, which is always imperfect.

Textbooks often say that experiments have disadvantages such as lacking realism and generality. Such assertions are not persuasive. First, such claims often conflate lab work with fieldwork. Both lab studies and field studies can be either experimental or nonexperimental. Second, while some lab studies clearly lack realism (e.g., Bond et al., 1992), so too do many nonexperimental studies. Questionnaire and interview research often asks people questions that they would not have thought about were they not in the research study (Schwarz, 1999). Third, I do not see the research lab as somehow divorced from the real world. On the contrary, people actually talk to other people in many lab studies. It is scandalous how few published studies of interpersonal communication involve people actually communicating with other people. Finally, as noted before, I see theory as the path to generality and external validity, not the surface features of the research. So if theory can be better tested with an experimental design, I would think that experimental works offers more, not less, generalizable knowledge.

I do both experimental and nonexperimental research. In deciding which way to go for a particular study, I ask myself two key questions. First, do the independent variables lend themselves to experimental induction? So, for example, if I want to find out the dimensionality of the communication apprehension scale (see Levine & McCroskey, 1990), all I need to do is collect responses to the PRCA-24 (Personal Report of Communication Apprehension-24). No experimental induction is needed. But if I want to test the effectiveness of the Joe Ayres communication apprehension treatment program (Ayres & Hopf, 1985) against a credible placebo control (Duff, Levine, Beatty, Woobright, & Park, 2007), then I need to do an experiment. The second question I ask myself is how important is causal order and control. If the variables can be experimentally induced and if I am interested in testing one process against another, then I think of doing an experiment. If I am just interested in description or looking at statistical association or if the independent variables do not lend themselves to experimental variation, then I take the nonexperimental path. The research question and theory behind it determine the method, not some fixed preference for one method over another.

A good example of what can be done with experiment design is the probing experiments I did with Steve McCornack (see Levine & McCornack, 2001). The probing effect relates to the presence of mere question asking in deception detection. The finding is that senders who are questioned are more likely to be believed than senders who are not questioned, regardless of actual honesty (Levine & McCornack, 2001). One explanation for the probing effect is offered by interpersonal deception theory (IDT; Buller & Burgoon, 1996). According to IDT, senders who are questioned with suspicious probing questions perceive suspicion. Once they recognize that the listener is suspicious, they strategically adapt their behavior to appear more honest. Listeners pick up on the honest behavior and are more likely to judge them as honest. Thus, probing questions lead to belief, because they prompt honest-appearing behavior. Steve and I suspected that the IDT explanation was bogus (see Levine & McCornack, 1996a, 1996b, for a detailed account of our reasoning). So we did a simple experiment. We videotaped senders being questioned with neural questions. Then, with a video editor, we altered the tapes. In a no-probe condition, we just

spliced out the question, leaving the answer intact. In supportive and suspicious probe conditions, we replaced the neural questioning with supportive or suspicion questions. But in all the conditions, the answers and the sender behaviors were constant. What we found was that the probed sources tended to be believed more when the listener heard the probes, regardless of probe type, thus replicating the probing effect. But since sender behavior was held constant in the research design, we could show that strategic behavioral adaptation was not responsible for the effect, at least in our data. In our more recent studies, we showed that suspicion-implying probing questions can lead to lower or higher accuracy depending on the wording of the questions (Levine & Blair, 2010; Levine, Shaw, et al., 2010). Experiments such as these provide convincing evidence for and against the predictions of different theories by inducing variance in some causal variables while holding other variables constant.

Design Basics

The simplest experiments involve a dichotomous on/off independent variable and a single dependent (outcome) variable. In a *posttest-only control group experiment*, participants are randomly assigned to one of two groups, an experimental group that gets the experimental induction or a control group that does not. Everything else is held as constant as possible. If the two groups differ on the dependent variable, that difference is attributed to the experimental induction. An alternative is a *pretest/posttest control group experiment*. Here, all participants are first measured on the dependent variable, then they are randomly assigned to one of two groups, the experimental group that gets the experimental induction or a control group that does not. Again, everything else is held as constant as possible. Finally, participants are tested a second time. If there is more change in the experimental group than in the control group, the difference is attributed to the independent variable (the induction). Posttest-only designs are sometimes also called *independent groups designs* or *between-groups designs*, while pretest–posttest designs are often called *repeated designs* or *within-subjects designs*.

There are pros and cons in going with a posttest only as opposed to having a pretest. There are three big advantages in pretest–posttest designs. First, every participant is his or her own control, and thus individual differences are held constant. Second, these designs are efficient in that they have more statistical power (see later in the chapter) and may require smaller samples. Third, they let the researcher look at change, which can be critical in some research areas such as persuasion. For example, if you wanted to study boomerang effects or psychological reactance in persuasion (e.g., see Dillard & Shen, 2005), you would need a pretest to show that the participants actually changed.

There are two main disadvantages in having a pretest. First, they increase the probability of nuisance variables and artifacts known as *testing effects*, *history effects*, and *maturation effects*. A testing effect occurs when the act of pretesting affects the posttest. With a pretest, we have to worry about order effects, priming, learning, and the like. A history effect occurs when an event other than induction occurs between the pretest and posttest that might affect the dependent variable. For example, my friend and colleague Joe Walther was doing a study of Israelites and Palestinians working together in online cooperative work tasks when Israel made a military incursion into the West Bank. Such events could affect his results, perhaps further polarizing his participants. Maturation is when participants change naturally over the course of the study. And in addition to all these concerns, change scores are more difficult to deal with statistically. For these reasons, posttest-only designs are much more common in interpersonal communication research.

Designs Researchers Really Use

Describing simple experiments is a useful teaching tool for conveying the basic ideas central to social-scientific experimentation, but most

experimental work in interpersonal communication involves designs that are more complex, with more than one independent variable, multiple dependent variables, multiple induction instantiations, or some combination of these. So we need to move from the simple "building-block" designs discussed so far to the designs that researchers actually use. Let us start with multiple independent variables.

Most experiments with more than one independent variable use *factorial* designs. In a factorial design, all active independent variables are *crossed* with each other. Crossing variables involves having all possible combinations of the levels of the variables. It is important not to confuse the levels of a specific independent variable with the existence of two or more different independent variables. The levels of a variable are the different experimental values that are induced. If there are just two levels, the levels might be on/off, treatment/control, high/low, male/female, and so on. Three levels could be low, moderate, or high suspicion, as in McCornack and Levine (1990). An example of four levels of a variable is systematic desensitization, the Ayres combination treatment, no treatment control, and placebo control—as in Duff et al.'s (2007) study of communication apprehension reduction treatments. Lee, Levine, and Cambra (1997) used five different grade levels of children (fourth through eighth graders) as a variable in their study of compliance resistance in children; thus, grade had five levels in their study. So each independent variable will have two or more levels; a variable with only one level would not vary at all. It would be a constant.

The simplest possible crossed design is a 2×2. In a 2×2 crossed design, there are two independent variables with two levels each. Imagine that we have two independent variables, $X1$ and $X2$, each of which can be on or off. In a 2×2 independent-groups factorial experiment, participants are randomly assigned to one of four conditions. The four conditions are $X1$ off and $X2$ off, $X1$ on with $X2$ off, $X1$ off and $X2$ on, and $X1$ on and $X2$ on. In a 2×2 repeated design, all participants are in all four conditions in some order, but the four conditions are the same as those just mentioned. In a $2 \times 2 \times 2$, there are three independent variables, each with two levels, yielding eight different experimental combinations or cells. In a $3 \times 2 \times 2 \times 3$, there are four independent variables, the first and the last with three levels each, the other two with two levels each. See the pattern? There is a number for each independent variable included in the design, the value of the number tells the number of levels, and the product of all the numbers identifies the number of cells.

There are a number of advantages in crossed designs, and these explain their popularity. Foremost among these is the ability to test for moderation (interaction effects) in addition to main effects. Main effects are the effects of an independent variable averaged across the other independent variables. With a crossed design, research can test for the main effect for each independent variable and for the interactions among the independent variables.

When independent variables are not crossed, they might be either *nested* or *fully confounded*. Nesting is when an independent variable only occurs within some levels of another independent variable. This is not desirable, but sometimes it is unavoidable. Confounding is when the effects of one variable are indistinguishable from those of another. For examples of nesting and confounding, consider my norms and expectations experiment mentioned previously (Levine et al., 2000). Our objection to a previous experiment by Bond et al. (1992) was that it had confounded norms and expectations. Bond et al. assigned their participants to one of two conditions: unexpected weird behaviors or expected normal behaviors. Participants, for example, either saw a person talking with one arm raised vertically over the head or a person in a more normal posture. The people in odd poses were judged to be less honest than the people in normal poses, and Bond et al. inferred from this difference that behaviors that are unexpected are seen as lies because they violate expectations. However, any differences between the two groups might have been because

the behavior was just weird. Maybe people evaluate norm breakers less positively. So we unconfounded the norms and expectations in our experiment by the crossing norms and expectations to create four experimental conditions: normal–expected, normal–unexpected, weird–expected, and weird–unexpected. This let us sort things out, and we discovered that it was being weird and not violating expectations that mattered. Not only did this resolve the confound in Bond et al., but it also provided a strong test of nonverbal expectancy violation theory (Burgoon & Hale, 1988).

To vary the weirdness, we decided to use four different weird behaviors. In the weird-behavior conditions, our participants interacted with someone who did one of four things: (1) followed an invisible insect flying around the room with the eyes, (2) picked his or her teeth with the fingers obsessively, (3) dropped down on the floor and stretched during the conversations, or (4) sporadically modulated his or her speech volume. Each of these odd behaviors was only enacted in the weird-behavior conditions. Such things do not happen when acting normally. So, in our design, weird behaviors were nested within the weird-behavior condition.

In addition to multiple independent variables, it is not unusual for studies to involve multiple dependent variables. Researchers with multiple dependent variables have a number of options. One option is to do separate statistical analyses for each dependent measure. The problem with this kind of analysis is that standard statistical inference presumes a single test. Multiple tests run the risk of error inflation (see Weber, 2007). There are corrections for multiple tests such as the Bonferroni test, but such corrections just trade off one type of error for another (Smith et al., 2002). Sometimes multivariate significance tests such as MANOVA are used, but this usually makes the situation worse, not better. When interpersonal communication researchers use MANOVA, they usually just report the multivariate test and then report the separate analyses for each variable (the so-called univariate tests), which is what is interpreted. So they end up doing more instead of fewer tests, making the problem worse, not better. Another option is that researchers could use some type of factor analysis to try to reduce the number of dependent variables. Finally, researchers can use path analysis or SEM to model the relationships among the dependent variables. In my own research, I try to keep my experiments as simple as possible and avoid variable inflation in the first place. When I have multiple dependent variables, I ask myself if I have hypotheses about if and how they are interrelated. If I think they are interrelated, I do path analysis to model those relations, otherwise I do separate analyses. I stay away from MANOVA. Simple researcher designs are usually better because there is less to go wrong and the results tend to be more interpretable (Abelson, 1995; Cohen, 1990; Smith et al., 2002).

A third type of complexity is the inclusion of multiple exemplars or instantiations nested within an induction, and if these exist, the decision to use *fixed-* or *random-effects* statistical models becomes an issue. In a basic experiment, participants are randomly assigned to a treatment (active independent variable "on") or control (active independent variable "off"). In such an experiment, the independent variable is considered a "fixed" effect, and the research participants are considered a "random" effect in the analysis. For a fixed effect, the values of the variable are set by the experimenter, and the findings only apply to those levels. We would not know, for example, what would have happened had a different treatment been used; the findings only apply to the treatment–control conditions actually tested. If you think about it though, the participants in the study are also a variable. Each and every one of them is different. In this example, however, the variability in the participants is not of interest. We are just going to average across them to get an average score in the treatment and control conditions. That is, we are interested in whether or not the average participant who got the treatment is different from the average control participant. Within-condition variability is

just error. It's a nuisance. So participants are treated as a random effect. We want to generalize across them. But note that there is a difference between "generalizing across" and "generalizing beyond." Statistically treating a variable as a random effect lets you generalize across nuisance variance. It does not let us generalize beyond. That requires not only appropriate statistical analysis but also random (or some other kind of representative) sampling, and even then, we still have the problem of induction. The fact that standard *t* tests and ANOVA treat the participants as a "random factor" in the analysis does not mean that the findings necessarily apply to all people everywhere who were not in the study. No statistical analysis ever magically grants generality; there are always other considerations.

Next, consider my norms and expectations experiment (Levine et al., 2000) discussed previously. We had confederates act normal or weird. In designing that study, does it make sense to use just one confederate, or might it make more sense to use more than one? What are the pros and cons? The same logic applies to weird behaviors. Is just one weird behavior best, or might several be better? In our study, we chose to have four different confederates who each enacted each of four different weird behaviors. The obvious cost is the added complexity and variable inflation. Nevertheless, we wanted to show that our findings would hold across different confederates and different behaviors. So we did multiple instantiations of our weird behavior. It added 16 cells to our design (four confederates crossed with four behaviors), but it gave us the logical equivalent of 16 mini replications, which made our claims stronger because the findings held across confederates and weird behaviors. When we did the analysis, however, we made confederates and type of weird behaviors fixed effects, not random effects. The decision to do this was easy and correct.

In interpersonal research, researchers often want to use multiple instantiations of variables to avoid the single-example/instance problem. This applies not just to confederates and behavioral studies but also to situations and messages (e.g., see McCornack, Levine, Torres, Solowczuk, & Campbell, 1992). In a very well-known and unfortunately influential article, Jackson and Jacobs (1983) advocated the use of random-effects analyses in these types of situations:

> One serious design flaw, which involves the use of a single message to represent a category of messages, occurs in nearly all of the experimental research on communication effects. The problem with such a design is that an observed difference between categories may reflect only differences between individual, idiosyncratic cases. A related error, the language-as-fixed-effect fallacy involves use of several replications of each category, but analysis of the cases as fixed effects. The consequence is that findings cannot be *generalized beyond* the sample used. *Future research should use multiple cases within each message category studied and treat cases as nested random effects* [italics added]. (p. 169)

This may sound reasonable, but it can be very bad advice. First, random effects do not allow for generality beyond the sample used any more than generalizing findings beyond other types of nonrepresentative samples. Second, random-effects analyses have a very nasty side effect. Random-effects analyses are often very low powered (Abelson, 1995; Hunter, Hamilton, & Allen, 1989). Statistical power will be discussed later, but basically the number of levels of the random factor becomes the sample size. So in my norms study, if weird-behavior example were a random effect, it would be like doing an experiment with a sample size of $N = 4$. It makes it highly probable that if your hypothesis is really right, the statistical test will yield the wrong answer. A better solution is replication and eventual meta-analysis (Abelson, 1995; Hunter et al., 1989). Nesting should only be done with care since crossing has definite advantages and random-effects analyses,

in most applications, have unacceptably high error rates. Besides, the point of much theory-guided experimentation is not to make generations but to test generalization. Theory provides a better path to generality than complex statistical models that are not well understood by the researcher using them.

Design Validity, Artifacts, and Confounds

In evaluating research, we can talk about design validity. Design validity is the extent to which the research design permits confidence in the research conclusions. In a totally invalid design, the findings have no meaning whatsoever. More typically, however, validity can be thought of as falling on a continuum from more to less confidence depending on the degree of control of rival explanations. A perfect study has never been done, but some findings are more likely to replicate than others and some designs are much "tighter" than others.

Design validity has traditionally been split into internal validity, external validity, and construct validity (Campbell & Stanley, 1963). Internal validity is the extent to which the intended independent variable, and not something else, produces the effects on the dependent variable. External validity refers to the extent to which the findings are generalized beyond the particular circumstance of the study. Construct validity refers to how well the inductions or measures capture the constructs they are supposed to induce or measure.

As Campbell and Stanley (1963) note, internal validity is the sine qua non (without which, nothing) of research. If researchers (and readers) cannot attribute the variation in the dependent variable to the independent variable(s), then the findings are meaningless, and there is no point in doing the research. Some textbooks talk about internal validity and external validity being a trade-off, but this is silly and makes no sense. If the findings are meaningless, then generality is moot.

Campbell and Stanley (1963) classify confounds as threats to construct validity, but when an induction (i.e., active independent variable) is confounded with some other variable, I find it useful to think of confounding as an issue of internal validity. A confounded induction is when more than one variable is (usually inadvertently) manipulated in such a way that the effects of the variables cannot be parsed. For example, Booth-Butterfield and Jordan (1989) were interesting in comparing communication adaptation patterns in same- and mixed-race groups. They recruited groups of college students who knew each other. Some of the groups were all white, and some were all black. In the same-race conditions, intact groups of five women interacted, and the interactions were videotaped and coded. In the mixed-race interactions, two intact groups of five were combined. So the same-race groups were all of the same race and consisted of five members, with all members known to each other previously, while the mixed-race groups had 10 members, who were mixed in race, and contained both members who were previously known to each other and others who were not known to each other. So in Booth-Butterfield and Jordan's design, the racial composition of groups (the intended independent variable) is fully confounded with group size and friends-only versus friends-and-stranger composition.

Most experimental studies are confounded in some minor ways, but not all confounds present meaningful threats to internal validity. The key question is whether the confounding variable(s) involves something that might actually change the results. In Booth-Butterfield and Jordan (1989), are group size and friend–stranger group composition likely to affect communication patterns? Because the answer would seem to be yes, the internal validity is highly suspect. There are other confounds in the Booth-Butterfield and Jordan study that are of less concern, such as interaction order (same-race groups interacted first), the number of experimenters present, and the topic of discussion. So the study was really at least a six-way confound. But not all these confounds are equally problematic.

The solution to confounds is first to identify them. When the confound is recognized, then the confound can either be held constant and thus neutralized or built into the design and thus tested. Confounds can be unconfounded and thus built in by crossing them with the other independent variables. The cost of building in variables is increased complexity, so it is often best just to hold them constant. For example, Booth-Butterfield and Jordan (1989) could have used strangers only in the same-race groups and could have had equal group sizes.

Three additional strategies for dealing with unwanted variables are randomization, double blinds, and placebo and other special controls. Random assignment was discussed previously. The idea is that by randomly assigning people to conditions, or messages to people, or randomly assigning experimental orders, and so on, the nuisance variables will average out. Another strategy is double-blind procedures. In double-blind experiments, both the researcher collecting the data and the participant are kept unaware of which experimental condition subjects are in so that this knowledge cannot inadvertently affect the results. A placebo is a believable but inert treatment used to rule out psychotherapeutic effects. Examples of communication research using these strategies effectively include the Duff et al. (2007) study of communication apprehension treatment effectiveness and the Levine, Feeley, McCornack, Hughes, and Harms (2005) study of nonverbal training in lie detection.

Quantifying Constructs

Nonexperimental Research

Most nonexperimental quantitative research on interpersonal communication involves asking people about something (e.g., survey, questionnaire, and interview methods), observation (e.g., behavioral observation, content analysis, textual analysis, physiological measurement), or both. Experiments, too, use these methods of measurement. The key difference is that in experiments and quasi experiments, at least one independent variable is active. In nonexperimental research, none of the variables are active. But all quantitative research has some variables that are measured, and the trick is in measuring the intended construct well.

As a general rule, if the construct of interest is overtly behavioral in nature, something that is directly observable, or both, then observation is often the preferred method. Alternatively, if the construct is internal-psychological in nature, such as a memory, opinion, attitude, emotional state, and so on, that is not directly observable, then it often makes sense to use measurement that involves asking questions of the participants. Furthermore, self-report measurement, including questionnaire and interview methods, are most useful when the participants know the answer and are willing to give the researchers the answer. Unfortunately, self-report questionnaires often ask questions that fail to meet these necessary conditions. Researchers often overestimate peoples' self-awareness or their ability to recall their own behavior (Schwarz, 1999).

Essentials of Valid Measurement

The new knowledge generated through empirical research can be no more valid than the measures used to make the observations. Thus, measurement validity is an absolute prerequisite for obtaining valid research results and for the defensible interpretation of findings. In short, the path to verisimilitude in quantitative research always goes through measurement.

Unfortunately, past and current trends in graduate education in most of the social sciences focus more on the statistical analysis of data than on the methods of observation and measurement used to produce data. Consequently, the typical quantitative researcher publishing in the social sciences probably took several advanced graduate-level statistics classes but likely had little training in psychometrics. Perhaps as a result, measurement is often the weakest link in our empirical knowledge claims.

It is my experience that most published scales and measures are never put to rigorous test, and of those that are put to the test, most do not fare very well. Fiske and Campbell (1992) made a similar observation. Published measures that have serious validity problems may even be more the norm than the exception. Examples that readily come to mind include measures of Machiavellianism (Christie & Geis, 1970), self-construal (i.e., that aspect of self-concept in which the self is differentiated from or converges with others; Gudykunst & Lee, 2003; Singelis, 1994), verbal aggressiveness (Infante & Wigley, 1986), and argumentativeness (Infante & Rancer, 1982). Research has shown substantial measurement problems in the scales measuring each of these constructs (Bresnahan et al., 2005; Hunter, Gerbing, & Boster, 1982; Kotowski, Levine, Baker, & Bolt, J., 2009; Levine, Bresnahan, Park, Lapinski, Lee, et al. 2003; Levine, Bresnahan, Park, Lapinski, Wittenbaum, et al. 2003; Levine et al., 2004).

To review, constructs and their interrelationships are the things researchers are interested in knowing about. The meaning attached to a given construct is specified in a conceptual definition. Measurement, on the other hand, is the act of assigning numbers or numerals to represent attributes of people, objects, or events (Nunnally & Bernstein, 1994). That is, measurement is the process of systematically converting abstract ideas into empirical data. Measurement allows for the representation of abstractions (i.e., constructs) with observable values or scores so that speculation, hypotheses, and theories about how various constructs are related can be put to empirical test.

Obviously, we want to maximize the degree of correspondence between the conceptual definition of a construct and the construct's measure. When the correspondence between construct and measurement is low, true relationships between constructs can appear to be false and false relationships between constructs can appear to be true. The degree of correspondence between a construct and its measurement is at the heart of measurement validity.

Measurement validity generally refers to how closely the values produced by a measure reflect the construct being measured. That is, a measure is valid to the extent that there is fidelity between scores and what the scores are meant to represent. Put differently, measurement validity is the extent to which a measure assesses what it is purported to measure, and nothing else.

As with design validity, it is important to point out that validity is not binary and measures need not be either invalid or valid. Usually, validity is considered as a continuum reflecting the degree of confidence that researchers have in a measure given the specific use for which the measure is being used. Even the best measures in the social sciences are not perfect. But some measures, such as IQ scores, for example, have considerable bodies of evidence suggesting substantial correlation between the construct and observed scores with only a small amount of systematic error (see Lubinski, 2004). For other measures, such as self-construal scales, the evidence is much more consistent with major problems: a weak correlation between the construct and observed score is attributable to substantial confounding. Consequently, researchers ought not to place much confidence in the meaning of the scores on self-construal scales (Levine, Bresnahan, Park, Lapinski, Lee, et al., 2003; Levine, Bresnahan, Park, Lapinski, Wittenbaum, et al., 2003).

Just as experimental inductions can be confounded, so too can measured variables. Measurement confounding exits when scores reflect more than one construct. A good example is self-construal scales. Several of Gudykunst's self-construal items ask people about consulting others (Gudykunst & Lee, 2003). This confounds self-concept and communication style (Levine, Bresnahan, Park, Lapinski, Lee, et al., 2003).

Another important point about measurement validity is that it is an empirical issue, requiring empirical evidence of different kinds and from a variety of sources to achieve. As evidence consistent with validity amasses over time, it is possible to have more confidence that scores on the measure are indeed indicators of the construct. In the absence of a substantial amount of evidence, arguments for a scale's validity cannot be

considered defensible. Lack of evidence does not mean that a measure is invalid but rather that validity is indeterminate. Further still, a scale is never proven valid because new data might arise in the future that tips the scales back toward invalidity. The definition of the construct may change over time, or responses elicited by the measure may change over time or application.

Because it makes little sense to measure something unless you know precisely what it is that you want to measure, the most reasonable place to start when thinking about a measure is with the conceptual definition of the construct that is to be measured. Once a good conceptual definition is adopted, then the measure can be created and the validation process can begin.

Measurement Validity and Validation Strategies

There are many different aspects of measurement validity. One sees reference to face validity, content validity, construct validity, structural validity, convergent validity, divergent/discriminant validity, and so forth. Each of these is a different aspect of measurement validity—that is, how closely the scores on the measure reflect the construct that we are seeking to quantify.

Face validity is about appearance and involves critical thinking and informed judgment. Try this test. A certain self-report scale item states, "Speaking up in a work/task group is not a problem for me." What construct is this item measuring? Maybe some aspect of leadership? Communication apprehension? If you guessed independent self-construal, you would be correct.[4] If you think it a problem that such an item would be used to measure self-concept, this too would be correct, and you get the idea of face validity.

Content validity is the extent to which a measure covers the breadth of the intended construct. A good measure would cover all the different facets of the construct and not just some narrow aspect. For example, a good measure of communication apprehension would not be limited to public-speaking contexts.

Structural validity usually refers to the "factor structure" or "dimensionality" of a scale. There should be a one-to-one correspondence between the number of constructs that a researcher wants to measure and the number of factors or dimensions of the scales used (factors and dimensions are used synonymously here). This is usually tested by some kind of factor analysis. Exploratory factor analysis (which is different from principal components analysis; see Park, Dailey, & Lemus, 2002) is used when a researcher does not know how many factors there might be or which items comprise which factor. Confirmatory factor analysis, in contrast, is used to test if a scale is factoring as intended. In either case, factor analysis, when used well, tells us how many constructs are measured but not about the substantive nature of the construct.

The principle of convergent validity (Campbell & Fiske, 1959; see Fiske & Campbell, 1992, for a useful retrospective on their idea) is that measures of the same construct should converge; they should be highly intercorrelated and should function in parallel manner with different constructs. It can be thought of as validation by triangulation; that is, if you measure a construct in different ways, all the alternative methods should triangulate or converge, all pointing to the same conclusions. When this is the case, we can have more confidence in each of the individual measures.

Divergence discriminance (Campbell & Fiske, 1959) is the flip side of convergence. The principle of divergence is that measures of different constructs act differently; they are not so highly intercorrelated as to be alternative measures of the same thing, and they act differently with respect to outside constructs. Convergent and divergent validity can be established with confirmatory factor analysis (by showing that different items measuring a single construct converge while items on different scales fall on different factors) or, better yet, multitrait–multimethod (MTMM) validation (Campbell & Fiske, 1959; for communication examples, see Bresnahan et al., 2005; Kotowski et al., 2009). MTMM offers the advantage of controlling method variance by crossing constructs and methods.

Construct validity is a more general term, referring to how well scores correspond to the construct

given the theory. It encompasses structural validity, convergent validity, discriminant validity, and nomological network validation (Cronbach & Meehl, 1955). Nomological networks are theoretically specified patterns of relationships among constructs–measures. Nomological network validation involves testing if a scale is associated with other measures as theoretically specified. Nomological network validation is often conflated with construct validity, but it is better seen as one (very limited) type of evidence for construct validity (Levine, Bresnahan, Park, Lapinski, Lee, et al., 2003; Nunnally & Bernstein, 1994). Full MTMM validation is usually preferred to rule out method variance confounding.

Reliability

Obviously, we want reliable measures. The reliability of a measure is the extent to which it is free from random response errors. Reliability is important, but it can be misleading.

Reliability is important because to the extent measures are unreliable, there is more random error; and random error obscures the relationships between variables, resulting in attenuated effect sizes and reduced statistical power. Reliability can be misleading because high reliability does not mean that a measure is valid. In fact, certain types of measurement confounds can inflate reliability (Shevlin, Miles, Davies, & Walker, 2000). This is why reliability was addressed after validity. From a research consumer perspective, we will want to know about validity first and interpret reliability in light of the validity evidence.

The Statistical Analysis of Data

The uses of statistics in research can be roughly divided into descriptive and inferential approaches. The goal of descriptive statistics is to give a simplified account of some data at hand, making the data understandable, while inferential statistics involves using the existing data to make inferences that go beyond those data. Inferential statistics can be used in two different ways: (1) to make inferences about populations based on samples and (2) for hypothesis testing using tests of statistical significance.

Descriptive Data Analysis

Descriptive statistics are just what the label implies. The goal of descriptive data analysis is to give an understandable summary of data. For example, in one of my most recent deception detection experiments (Levine, Shaw, et al., 2010), we had $N = 128$ students watch and judge 44 videotaped interviews. Of the students interviewed, half had cheated on a task and the other half were noncheaters. All the interviewees denied having cheated. We edited the interview tapes and showed the judges either the first few questions and answers or the last few. All judges were asked if they thought that each of the interviewees were cheaters or noncheaters, and these judgments were scored for accuracy. This resulted in 5,632 accuracy scores ($N = 128$ judges' judgments of 44 message senders). So once we got these numbers, we needed to figure out what the scores were telling us. Of course, just looking at the 5,000-plus individual accuracy scores is not very informative and is more than a little overwhelming. So we needed to summarize the results and convert the mass of numbers into sensible results. Therefore, we tallied the scores in a number of different ways (see Levine, Shaw, et al., 2010, Tables 1–3). Importantly, we looked at accuracy scores for both judges and senders. That is, we looked at how accurate each judge was averaged across the 44 senders, and we also looked at the average number of times each sender was judged correctly across the 128 judges.

The first thing we did was to look at the distribution of scores (Levine, Shaw, et al., 2010, Table 3). We also looked at the means (average scores), variances, and standard deviations because the distribution was such that the means were informative. The distribution of scores can be seen by arranging all scores in order from the smallest to the largest and looking at the spread and the clumping. This

is often done with a frequency distribution or graphed with a histogram or stem-and-leaf plot. When we did this with a frequency distribution, we saw immediately that judge accuracy scores were all tightly grouped; most judge scores fell within a 20% to 25% range depending on the condition. Sender scores, however, were all over the place, ranging from 7% to 100%—a 93-point spread. This difference in spread, of course, showed up as a difference in variance. The variance for the senders was 10 times as much as the variance for the judges! The other thing we noticed right away was that the judges who watched the last three interview questions obtained more accurate scores (68% on average) than those who saw the first three questions and answers (44%). In fact, there was not much overlap in the two distributions of judges. Finally, there was even more sender variance when the senders were assessed by the judges who saw the last three questions than by those who saw the first three.

So by looking at the central tendency (means), dispersion (standard deviation and variance), and distributions, we were able to make sense out of those 5,000-plus accuracy scores. We learned that judges tend to see things the same way but some senders tend to be seen much more differently than other senders. This told us that senders are where the action is. We also learned that questioning senders makes a big difference in judge accuracy by increasing sender variance. It also helped set up our next studies, with some looking at the effects of questioning and others taking a closer look at sender variance.

While means (i.e., average scores) are often informative, and certainly the most commonly reported descriptive statistic, they can be misleading when the distribution of scores around the mean is not symmetrical. Here is an example. An important question in my main area of research, deception, is how often people lie. Deception researchers often presume that people lie frequently, but if you think about this for a moment, you will realize that this is a difficult research question to answer well. Just how do you random sample lying?

The best known study on the prevalence of lying is a diary study by DePaulo, Kashy, Kirkendol, Wyer, and Epstein. (1996). They found and reported that people lie between once and twice a day. Recently, we tried to replicate this finding using survey methodology (Serota, Levine, & Boster, 2010). Just like DePaulo et al. (1996), we found an average of between one and two lies per day (mean = 1.65, standard deviation = 4.45, median = 0, mode = 0, maximum value = 53). But the average here does not tell the whole story. Sixty percent of our sample reported telling no lies at all in the past 24 hours, and 75% reported telling one or fewer lies. In contrast, about half of all reported lies were told by just 5% of the sample, and almost 25% of the lies were told by the top 1%, comprising most frequent liars. We called this finding "a few prolific liars." What's more, we requested data from previous authors and collected additional data to replicate our findings. All the data had the same pattern. Most people don't lie very often, but a few people lie with astounding frequency. The average does not reflect what the average person does! Looking at the distribution of data is crucial.

Another class of descriptive statistics consists of measures of *effect size* or strength of association. These statistics tell us how large a difference is relative to variability or how strongly variables covary. Basically, measures of effect size tell us about the magnitude of effect that one variable (or set of variables) has on another. Interpersonal communication researchers have long realized the importance of reporting and interpreting effect sizes in research, and interpersonal communication is well ahead of many other social sciences in this regard.

Common measures of effect size include d (the standardized mean difference), r (the correlation coefficient), R^2 (the multiple correlation), β (the standardized regression coefficient), and η^2 (eta squared, the ratio of the effect of sums of squares to the total). There are also others. Table 2.1 shows conversions between d, r, and η^2 for eight values of d. Each of these statistics tells the same story by just using a different metric.

Table 2.1 Examples of Conversions Between *d, r,* and η^2

d	*r*	$\eta 2$
0.10	.05	<.01
0.25	.12	.02
0.50	.24	.06
0.75	.35	.12
1.00	.44	.20
1.50	.66	.36
2.00	.70	.50
2.50	.78	.61

What the reader should take from this is the idea that descriptive statistics are really useful and informative. Compared with inferential statistics, they are simple and easy to understand. They are also essential. It is very easy to get lost in raw data. It is also sometimes tempting to focus exclusively on the inferential statistics and forget that it is really the descriptive statistics that help us see the patterns and the trends in the data. But we can't just look at averages. Dispersion, distributions, and effect sizes are important too.

Inferential Statistics Basics

The inferential statistics used in interpersonal communication research can be divided into two types: (1) CIs and (2) statistical significance testing. The goal of a CI is to estimate population values with sample data, while statistical significance is used to test statistical hypotheses. Unfortunately, in interpersonal research, significance testing is much more common than the use of CIs, even though it would be more rational if the reverse were true.

Readers are likely to be familiar with the idea of CIs, as they are frequently mentioned in media coverage of scientific opinion polling. We often hear that some opinion poll found that some percentage of people favor (or disfavor) some sentiment with a certain margin of error. The margin of error is the CI, which, by tradition, is usually set at 95%. What this means is that, presuming the sampling was done correctly, there is 95% chance that the CI includes the population value. Furthermore, CIs are informative about the precision of statistical estimates. The smaller the CI, the more precise is the estimate.

CIs can be calculated around means, but they can also be calculated around effect sizes. When provided for effect sizes, they provide the same information as a significance test. A result that is statistically significant is one where an effect size of zero is outside the 95% CI. But CIs also provide additional information, telling us how far above zero the lower limit and also the plausible upper bound of the effect is. A useful overview of CIs is provided by Cumming and Finch (2005), and Levine, Weber, Park, et al. (2008) provide a primer on how to calculate CIs around some common effect sizes.

The (Il)Logic of Significance Testing and the Tyranny of the *p* Value

Statistical significance testing (or, more precisely, *null hypothesis significance testing*, NHST) is a commonly used type of inferential statistics. When you read about a finding not being significant, note that a finding is $p < .05$, or see p values

reported in researcher articles, it is almost certain that NHST is being done. NHST is used in conjunction with popular statistical analyses such as chi-square (χ^2), *t* tests, ANOVA, correlation, and regression. Popular statistical software packages such as SPSS (now an IBM company—SPSS was acquired by IBM in October 2009) or SAS report NHST as the default when using these statistics.

Most NHST in interpersonal research is concerned either with testing for differences (between two or more means with *t* tests or ANOVA or between frequencies using chi-square) or with looking for the linear association between variables (correlation or regression). Each of these statistical techniques involves the calculation of a test statistic (e.g., *t*, *F*, or χ^2) from the data. Each of these test statistics has a theoretical probability density function, presuming that certain assumptions are met, and predicated on the null hypothesis being true. In NHST, the *p* value tells us the probability of the results given the null. That is, the *p* value from NHST tells us the probability of an outcome in data assuming a true null hypothesis. If the outcome of the test is sufficiently unlikely (where sufficiently unlikely means, by convention, a probability of 5% or less, i.e., $p < .05$, the "alpha level"), then it is inferred that the data are unlikely given the null hypothesis, the null hypothesis is rejected on this basis, and support is inferred for the researcher's hypothesis.

Confused? My guess is that (a) this is not what you learned in your basic statistics class and (b) this sounds like tortured logic to you.

What is taught in most statistics classes in the social sciences is that there is a null hypothesis, H_0, and an alternative hypothesis, H_1, reflecting the researchers' hypothesis, and that H_0 and H_1 are mutually exclusive. It is presented as follows:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2,$$

where μ_1 and μ_2 are the estimated population means from two independent samples.

Then, we calculate a *t* test, and if the test is $p < .05$, we reject H_0 and accept H_1. This is presented as "statistics" without reference to the mathematicians who developed this method. Readers actually interested in the origins of these ideas are directed to Chapter 1 in Gigerenzer and Murray (1987) and to Sir Ronald Fisher's (1990) three books for seminal work on NHST.

The logic of NHST typically confuses those who have not been fully indoctrinated. I suspect this is so because it does not make valid logical sense. The logical problems with NHST have been known for at least 50 years (see Rozenboom, 1960) and have been repeated by many authors (e.g., Kline, 2004; Levine, Weber, Hullett, Park, & Lindsey, 2008). Trafimow (2003) and Nickerson (2000) provide some of the most insightful analyses on the meaning of *p* values I have seen, but they are not the easiest reads. For those interested, I would suggest starting with Gigerenzer and Murray (1987), Boster (2002), and Cohen (1990) in that order, before moving on to Levine, Weber, Hullett, et al. (2008), Nickerson (2000), and Trafimow (2003) in that order.

Nevertheless, NHST is the accepted convention in social science, and there is little evidence that a more rational approach will take hold any time soon. The brainwashing of would-be quantitative communication researchers will likely continue, and continued use of NHST will surely be expected by journal editors and reviewers in the foreseeable future. My advice to researchers is to self-educate (the citations provided in the previous two paragraphs are good places to start) and to try and slip in as many descriptive statistics and CIs into their papers as they can. For research consumers, I suggest relying more on the descriptive statistics that are reported than on *p* values and findings with asterisks.

To this point, I have been pretty hard on NHST. The criticism is very well deserved. But let me make two points of qualification to this criticism. First, I report NHST in the majority of my own papers (although I try not to rely on it too much). Second, there is a difference between having a lot of problems and being worthless. Small

p values have some evidentiary value, and there is a need to rule out chance as a plausible explanation for a finding. An example in the next section demonstrates this point well.

The Importance and the Confusions of Statistical Power

For the past few years, when I teach my graduate methods class, I have a short research article evaluation assignment due at the end of the first week of class. The assignment is to evaluate the contribution and validity of Cohen, Nisbett, Bowdle, and Schwarz (1996) series of experiments testing the culture of honor. What I don't tell the students in my class beforehand is that it is one of my favorite articles. Students typically want to impress me by finding a lot of flaws and write papers that are highly critical of the target research. Then, they are shocked when I hand the papers back and tell them how much I admire the work they have just shredded.

The main point of the assignment is to be judicious in criticism, saving it for when it is well deserved and not dismissing interesting findings too quickly. But there is another lesson in the assignment that is directly related to the topic of this section—an important lesson about statistical power and what *p* values do and do not tell us. One of the really good aspects of the Cohen et al. (1996) experiments is the effort to triangulate through a multiple-method approach. They use self-report measures, psychological measures, and behavioral observation to test the same focused hypothesis in more than a dozen different ways. This allows for a strong evaluation of convergent validity in a way that is very unusual in social science.

Cohen et al. (1996) use a 2×2 independent-groups design, crossing the presence or absence of an insult with whether or not the participant is Northern or Southern. Their hypothesis is that one cell will be different from the other three: Southerners will respond with aggression to insult. They report something like 14 different tests of the hypothesis, with sample sizes of $N =$ 83, 146, and 173 in three studies. Although most of the findings have the same basic pattern, by my count, only 7 of the 14 tests come out with $p < .05$.

The fact that the results are mixed is typically an avenue of attack for the overly critical graduate student. The typical student argument goes like this:

(a) Cohen et al. claim support in their discussion section.

(b) Their actual findings are mixed. There is just as much nonsupport as there is support.

(c) Therefore, Cohen et al. argue far beyond their data and are guilty of overselling their conclusions. Cohen et al.'s findings fail to provide consistent evidence for their claim, and this lack of consistency in findings seriously limits the contribution of their work.

Here are my twin questions for the reader to test his or her knowledge of NHST:

(a) If the null was in fact true and there are no real differences to be found what outcomes are anticipated from the 14 NHSTs?

(b) Alternatively, if the null is false and Cohen et al. are right, what outcomes are anticipated?

If the null is true and we do 14 tests at $p \leq .05$, then we can use the binomial distribution to calculate the probability of various outcomes. There is about a 49% chance that all 14 tests will be nonsignificant; there is another 36% chance of getting one hit by chance, a 12% chance of getting two hits by chance, and a 3% chance of three Type I errors. That adds up, after rounding, to 100% and therefore exhausts the plausible outcomes. The chance of getting seven or more significant findings under the null hypothesis is $p < .0000001$. Thus, Cohen et al.'s findings suggest that the null is highly implausible at much less then $p = .05$.

Now, what if the null is false? This is where statistical power comes into play. Statistical power is the chance of finding a significant result if the null is false. One minus power is the chance of a Type II error. If we presume for the sake of demonstration an effect size of $d = 0.5$, then the power of their tests are in the .36 to .62 range depending on the sample size in each study. So what this means is that if their hypothesis is in fact right, they will likely get "support" on somewhere between one third and two thirds of their tests. So finding that half of the tests are significant and half are not is well within the range of what is to be expected if their hypothesis is correct and tested 14 times. Furthermore, consider this case. If power was say .5, then the chances of all 14 tests being significant is $p < .0001$. So had Cohen et al.'s findings all turned out to be significant, their findings would not have been plausible at all—they would have been much too good to be true.

This is why statistical power is so important. If power is not understood, research results will be misunderstood. And, if you are doing research, and using NHST, you need all the power you can get. For brief primers on power, see O'Keefe (2007) and Levine, Weber, Park, et al. (2008). The bottom line is that statistical power combined with publication bias (the tendency for nonsignificant results to be published less often than significant results) pretty much guarantees that all literatures that rely on NHST will show mixed results, making it very difficult to make sense of the findings (Meehl, 1986). This is one of many reasons why CIs are preferred.

Beyond ANOVA and Correlation: Multilevel Modeling, SEM, and Meta-Analysis

If you follow published original communication research, you will know that the once omnipresent ANOVA and correlation matrices are increasingly being replaced with articles reporting multilevel modeling (MLM), SEM, and meta-analysis. These techniques are gaining popularity for good reasons. Each solves problems and satisfies important statistical needs for interpersonal researchers. However, these techniques create additional problems when misunderstood or misused.

The advent of MLM is an important advance because it allows interpersonal researchers to deal with nonindependent data. Traditional inferential statistics all share a strong assumption of *independence of observation*. To introduce this idea, consider the probabilities associated with a string of coin flips. If we flip a coin, we have a 50–50 chance of getting a heads. If we flip the coin three times in a row, there is a .125 chance of getting three heads in a row ($.5^3 = .5 \times .5 \times .5 = .125$), assuming that the outcome of the first flip did not affect the probability of subsequent flips. This is what is meant by independence of observations. What if, however, getting a heads on the first flip meant that each subsequent flip was now more or less likely than .5 to be a heads? That would change the expected probabilities, and understanding the probabilities would have to take the initial outcome and the nature of the contingency into account. Subsequent flips would be nonindependent.

The degree of nonindependence can be quantified with a statistic called the *intraclass correlation* (ICC). A positive ICC means that data points in the same "class" tend to be similar while a negative ICC means that data points in the same class tend to be dissimilar relative to data points in other classes. With the coin flip example, a negative ICC would mean that a heads is more likely to be followed by a tails. More relevant to interpersonal communication, imagine that we are interested in communication traits in dating couples. If people tend to form relationships with dispositionally similar others, it would create a positive ICC, while opposites attracting would create a negative ICC. Only if the partner's traits were uncorrelated would the ICC be zero.

Nonindependence of data becomes a potential concern when data points are nested in larger categories. Individuals might be nested within couples, families, or friendship circles. Students might be nested within classes, classes within

majors, majors within colleges, and so forth. Fortunately, the existence of nonindependence can be tested with the ICC, and if the ICC is meaningfully large, then the nonindependence can be modeled with MLM. The two most popular MLM techniques in interpersonal research are David Kenny's social relations model, which is more ANOVA based, and hierarchical linear modeling (HLM), which is more regression based.

Currently, SEM is more popular than MLM. SEM combines measurement modeling (confirmatory factor analysis, CFA) and path analysis (structural causal modeling), although it is advisable to do CFA and path analysis separately. Commonly used SEM software programs include AMOS, LISREL, EQS, and M-Plus.

CFA is generally superior to EFA when it is known which constructs specific items are supposed to measure, because CFA is better able to deal with correlated factors and better able to identify measurement confounds. Path analysis, on the other hand, is useful whenever one wishes to test a causal model with at least one mediated relationship. Both CFA and path analysis require researchers to specify causal models in advance. The data are tested for how well they fit the model, and models are then accepted or rejected. The biggest problem with SEM, as it is used in interpersonal research, is that it is more often used for model fitting than for model testing. When model fitting, replication with new data is advisable to avoid capitalization of chance. This replication is called *cross-validation*, and unfortunately, it is not common practice.

Meta-analysis is a set of procedures that allow researchers to cumulate findings across studies to quantitatively summarize literatures (Hunter, Schmidt, & Jackson, 1982). This is quite useful because, as we have seen, the nature of NHST makes most literatures appear to have mixed findings whether or not consistent patterns exist. To the extent to which the findings from a literature converge on a single conclusion, the findings are said to be homogeneous. When findings are homogeneous, they can be averaged as a meaningful summary of the findings. When findings are heterogeneous, however, researchers need to try and find out why. Sometimes there exist some subsets of the findings that tend to cluster around one outcome while other subsets cluster around another point. This suggests an important moderator.

While a well-done meta-analysis can be extremely informative, a few words of caution are advisable. First, average effects can only be safely interpreted when the findings of individual studies are homogeneous. Second, the results of meta-analysis depend on the quality of the original research. If there is some artifact or bias that runs through an entire literature, meta-analysis will not catch it. Third, meta-analysis is susceptible to publication bias, and the existence of publication bias tends to inflate the average effects reported in meta-analysis (Levine, Asada, & Carpenter, 2009).

Conclusion

This chapter offers a relatively unconventional take on the topic of quantitative methods. As students of social science, we know that nonconformity is often not appreciated. At the same time, it is hoped that this chapter gives readers new to social-scientific thinking a glimpse at what is beyond the conventional takes on the topic that populate the overpriced but underinformed texts on the topic.

The bottom line is that the validity of research findings rests on their conceptual/theoretical foundations and the qualities of the researcher design, measurement, and statistical analysis. The findings are no better than the weakest of these crucial links. So good research requires getting theory, research design, measurement, and analysis into harmony. It is hoped that this chapter helps the reader see how this might be accomplished and highlights what the research consumer might profitably look for when making sense of research reports. While method is only a means to an end, valid and thoughtful method is necessary for a fuller, richer, and more accurate understanding on interpersonal

communication. Too much methodological rule following or ever more complex statistical analysis just won't get us there.

Notes

1. "Bullshit" is an important communication construct that is highly relevant to interpersonal communication. It was formally explicated, however, not by a communication theorist but by the Princeton analytic philosopher, Professor Harry G. Frankfurt (2005).

2. Throughout this chapter, I offer numerous examples from my own research. Although it is good practice to avoid excessive self-citation, using examples from my own research makes sense when writing about methods. I know why I made the methodological decisions I did, and thus I have insights into my works that I don't have about others' studies. Furthermore, this chapter reflects my research priorities and commitments. I gained these views, to a large extent, through the practice of being a researcher, and these views are reflected in my research. Thus, it is natural that my work exemplifies well the points I am trying to make.

3. We went ahead and did the study anyway. The funding came out of our own pockets, and the result was the lead article in Volume 36 of *Human Communication Research*.

4. Items such as these are at the heart of the self-construal debate between Levine et al. (2003a, 2003b), Gudykunst and Lee (2003), and Kim and Raja (2003) and explain why self-construal scales do not factor as intended. Factor analysis is helpful in showing validity problems, but a good critical eye for face validity is an important first step.

References

Abelson, R. P. (1995). *Statistics as principled argument*. Hillsdale, NJ: LEA.

Asch, S. E. (1956). Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological Monographs, 70,* 1–70.

Ayres, J., & Hopf, T. S. (1985). Visualization: A means of reducing speech anxiety. *Communication Education, 34,* 318–323.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Beatty, M. J. (2002). Do we know a vector from a scalar? Why measures of association (and not their squares) are appropriate measures of effect. *Human Communication Research, 28,* 605–625.

Berger, C. R. (1991). Communication theories and other curios. *Communication Monographs, 58,* 101–113.

Berger, C. R. (2005). Interpersonal communication: Theoretical perspectives, future prospects. *Journal of Communication, 55,* 415–447.

Berger, C. R., & Calabrese, R. J. (1975). Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research, 1,* 99–112.

Bond, C. F., Omar, A., Pitre, U., Lashley, B. R., Skaggs, L. M., & Kirk, C. T. (1992). Fishy-looking liars: Deception judgments from expectancy violation. *Journal of Personality and Social Psychology, 63,* 969–977.

Booth-Butterfield, M., & Jordan, F. (1989). Communication adaptation among racially homogeneous and heterogeneous groups. *Southern Communication Journal, 54,* 253–272.

Boster, F. J. (2002). On making progress in communication science. *Human Communication Research, 28,* 473–490.

Bresnahan, B. J., Levine, T. R., Shearman, S., Lee, S. Y., Park, C. Y., & Kiyomiya, T. (2005). A multimethod-multitrait validity assessment of self-construal in Japan, Korea, and the U.S. *Human Communication Research, 31,* 33–59.

Buller, D. B., & Burgoon, J. K. (1996). Interpersonal deception theory. *Communication Theory, 6,* 203–242.

Bunz, U. (2005). Publish or perish: A limited author analysis of ICA and NCA journals. *Journal of Communication, 55,* 703–720.

Burgoon, J. K., & Hale, J. L. (1988). Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communication Monographs, 55,* 58–79.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multi-trait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

Christie, R., & Geis, F. L. (1970). *Studies in Machiavellianism*. New York: Academic Press.

Cialdini, R. B. (1980). Full-cycle social psychology. *Applied Psychology Annual, 1,* 21–47.

Cialdini, R. B., Borden, R. J., Thorne, A., Walker, M. R., Freeman, S., & Sloan, L. R. (1976). Basking in reflected glory: Three (football) field studies. *Journal of Personality and Social Psychology, 34,* 366–375.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in Experimental Social Psychology, 24,* 210–234.

Cialdini, R. B., Vincent, J. E., Lewis, S. K., Catalan, J., Wheeler, D., & Darby, B. E. (1975). Reciprocal concessions procedure for inducing compliance: The door-in-the-face technique. *Journal of Personality and Social Psychology, 31,* 206–215.

Cohen, D., Nisbett, R. E., Bowdle, B. F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An "experimental ethnography." *Journal of Personality and Social Psychology, 70,* 945–960.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2010). *Applied multiple regression/correlation analysis for the behavioral sciences.* New York: Taylor & Francis.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation.* Boston: Houghton.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–301.

Cumming, G., & Finch, S. (2005). Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70,* 979–995.

Dillard, J. P., & Shen, L. (2005). On the nature of reactance and its role in persuasive health communication. *Communication Monographs, 72,* 144–168.

Dindia, K., & Allen, M. (1992). Sex differences in self-disclosure: A meta-analysis. *Psychological Bulletin, 112,* 106–124.

Duff, D. C., Levine, T. R., Beatty, M. J., Woobright, J., & Park, H. S. (2007). Testing public anxiety treatments against a credible placebo control. *Communication Education, 56,* 72–88.

Festinger, L., Riecken, H. W., & Schachter, S. (1956). *When prophecy fails: A social psychological study of a modern group that predicted the destruction of the world.* Minneapolis: University of Minnesota Press.

Fisher, R. A. (1990). *Statistical methods, experimental design, and scientific inference: A re-issue of statistical methods for research workers, the design of experiments, and statistical methods and scientific inference.* Oxford, UK: Oxford University Press.

Fiske, D. W., & Campbell, D. T. (1992). Citations do not solve problems. *Psychological Bulletin, 112,* 393–395.

Frankfurt, H. G. (2005). *On bullshit.* Princeton, NJ: Princeton University Press.

Gigerenzer, G., & Murray, D. J. (1987). *Cognition as intuitive statistics.* Hillsdale, NJ: Lawrence Erlbaum.

Gudykunst, W. B., & Lee, C. M. (2003). Assessing the validity of self-construal scales: A response to Levine et al. *Human Communication Research, 29,* 253–274.

Hemple, C. G. (1966). *Philosophy of natural science.* Englewood Cliffs, NJ: Prentice Hall.

Hickson, M., Self, W. R., Johnson, J. R., Peacock, C., & Bodon, J. (2009). Prolific research in communication studies: Retrospective and prospective views. *Communication Research Reports, 26,* 337–346.

Hunter, J., Gerbing, D., & Boster, F. (1982). Machiavellian beliefs and personality: Construct invalidity of the Machiavellianism dimension. *Journal of Personality and Social Psychology, 43,* 1293–1305.

Hunter, J. E., Hamilton, M. A., & Allen, M. (1989). The design and analysis of language experiments in communication. *Communication Monographs, 56,* 341–363.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies.* Beverly Hills, CA: Sage.

Infante, D. A., & Rancer, A. S. (1982). A conceptualization and measure of argumentativeness. *Journal of Personality Assessment, 46,* 72–80.

Infante, D. A., & Wigley, C. J. (1986). Verbal aggressiveness: An interpersonal model and measure. *Communication Monographs, 53,* 61–69.

Jackson, S., & Jacobs, S. (1983). Generalizing about messages: Suggestions for the design and analysis

of experiments. *Human Communication Research, 9,* 169–191.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of behavioral research*. New York: McGraw-Hill.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217.

Kim, M. S., & Raja, N. S. (2003). When validity testing lacks validity: Comments on Levine et al. (2002). *Human Communication Research, 29,* 275–290.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Kotowski, M. R., Levine, T. R., Baker, C., & Bolt, J. (2009). A multi-trait multi-method validity assessment of the verbal aggressiveness and argumentativeness scales. *Communication Monographs, 76,* 443–462.

Kuhn, T. S. (1996). *The structure of scientific revolutions* (3rd ed.). Chicago: University of Chicago Press.

Lakatos, I. (1978). *The methodology of scientific research programmes*. Cambridge, UK: Cambridge University Press.

Lee, C. R., Levine, T. R., & Cambra, R. (1997). Resisting compliance in the multi-cultural classroom. *Communication Education, 46,* 10–28.

Levine, T. R. (2010a). A few transparent liars: Explaining 54% accuracy in deception detection experiments. In C. Salmon (Ed.), *Communication yearbook 34* (pp. 40–61). Thousand Oaks, CA: Sage.

Levine, T. R. (2010b). Ranking and trends in citation patterns of communication journals. *Communication Education, 59,* 41–51.

Levine, T. R., Anders, L. N., Banas, J., Baum, K. L., Endo, K., Hu, A. D. S., et al. (2000). Norms, expectations, and deception: A norm violation model of veracity judgments. *Communication Monographs, 67,* 123–137.

Levine, T. R., Asada, K. J., & Carpenter, C. (2009). Sample size and effect size are negatively correlated in meta-analysis: Evidence and implications of a publication bias against non-significant findings. *Communication Monographs, 76,* 286–302.

Levine, T. R., Beatty, M. J., Limon, S., Hamilton, M. A., Buck, R., & Chory-Asada, R. M. (2004). The dimensionality of the verbal aggressiveness scale. *Communication Monographs, 71,* 245–268.

Levine, T. R., & Blair, J. P. (2010). *Questioning strategies, diagnostic utility, and expertise interactions in deception detection.* Manuscript submitted for publication.

Levine, T. R., Bresnahan, M., Park, H. S., Lapinski, M. K., Lee, T. S., & Lee, D. W. (2003). The (in)validity of self-construal scales revisited. *Human Communication Research, 29,* 291–308.

Levine, T. R., Bresnahan, M., Park, H. S., Lapinski, M. K., Wittenbaum, G., Shearman, S., et al. (2003). Self report measures of self-construals lack validity. *Human Communication Research, 29,* 210–252.

Levine, T. R., Feeley, T. H., McCornack, S. A., Hughes, M., & Harms, C. M. (2005). Testing the effects of nonverbal training on deception detection accuracy with the inclusion of a bogus train control group. *Western Journal of Communication, 69,* 203–217.

Levine, T. R., Kim, R. K., & Blair, J. P. (2010). (In)accuracy at detecting true and false confessions and denials: An initial test of a projected motive model of veracity judgments. *Human Communication Research, 36,* 81–101.

Levine, T. R., Kim, R. K., Park, H. S., & Hughes, M. (2006). Deception detection accuracy is a predictable linear function of message veracity base-rate: A formal test of Park and Levine's probability model. *Communication Monographs, 73,* 243–260.

Levine, T. R., & McCornack, S. A. (1996a). Can behavioral adaption explain the probing effect? *Human Communication Research, 22,* 603–612.

Levine, T. R., & McCornack, S. A. (1996b). A critical analysis of the behavioral adaptation explanation of the probing effect. *Human Communication Research, 22,* 575–589.

Levine, T. R., & McCornack, S. A. (2001). Behavioral adaption, confidence, and heuristic-based explanations of the probing effect. *Human Communication Research, 27,* 471–502.

Levine, T. R., & McCroskey, J. C. (1990). Measuring trait communication apprehension: A test of rival measurement models of the PRCA-24. *Communication Monographs, 57,* 62–72.

Levine, T. R., Park, H. S., & McCornack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the "veracity effect." *Communication Monographs, 66,* 125–144.

Levine, T. R., Shaw, A., & Shulman, H. (2010). Increasing deception detection accuracy with

strategic questioning. *Human Communication Research, 36,* 216–231.

Levine, T. R., Weber, R., Park, H. S., & Hullett, C. R. (2008). A communication researchers guide to null hypothesis significance testing and alternatives. *Human Communication Research, 34,* 188–209.

Levine, T. R., Weber, R., Hullett, C. R., Park, H. S., & Lindsey, L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research, 34,* 171–187.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'General Intelligence,' Objectively Determined and Measured." *Journal of Personality and Social Psychology, 86,* 96–111.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin, 132,* 895–919.

McCornack, S. A., & Levine, T. R. (1990). When lovers become leery: The relationship between suspicion and accuracy in detecting deception. *Communication Monographs, 57,* 219–230.

McCornack, S. A., Levine, T. R., Torres, H. I., Solowczuk, K. A., & Campbell, D. M. (1992). When the alteration of information is viewed as deception: An empirical test of information manipulation theory. *Communication Monographs, 59,* 17–29.

McCroskey, J. C. (1977). Oral communication apprehension: A summary of recent theory and research. *Human Communication Research, 4,* 78–96.

Meehl, P. E. (1986). What social scientists don't understand. In D. W. Fiske & R. A. Shweder (Eds.), *Meta-theory in social science* (pp. 315–338). Chicago: University of Chicago Press.

Miller, G. R., & Berger, C. R. (1999). On keeping the faith in matters scientific. *Communication Studies, 50,* 221–231.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–387.

Muthuswamy, N., Levine, T. R., & Weber, R. (2009). Scaring the already scared: Some problems with HIV/AIDS fear appeals in Africa. *Journal of Communication, 59,* 317–344.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York: McGraw-Hill.

O'Keefe, D. J. (2007). Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: Sorting out appropriate use of statistical power analysis. *Communication Methods and Measures, 1,* 291–299.

Park, H. S., Dailey, R., & Lemus, D. (2002). The use of exploratory factor analysis and principal components analysis in communication research. *Human Communication Research, 28,* 562–577.

Pavitt, C. (2001). *The philosophy of science and communication theory*. Huntington, NY: Nova.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Routledge.

Rozenboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review, 5,* 2–14.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 93–105.

Serota, K. B., Levine, T. R., & Boster, F. J. (2010). The prevalence of lying in America: Three studies of reported deception. *Human Communication Research, 36,* 1–24.

Shevlin, M., Miles, J. N. V., Davies, M. N. O., & Walker, S. (2000). Coefficient alpha: A useful indicator of reliability? *Personality and Individual Differences, 28,* 229–237.

Singelis, T. M. (1994). The measurement of independent and interdependent self-construals. *Personality and Social Psychological Bulletin, 20,* 580–591.

Smith, R., Levine, T. R., Lachlan, K. A., & Fediuk, T. A. (2002). The high cost of complexity in experimental design and data analysis: Type I and Type II error rates in multiway ANOVA. *Human Communication Research, 28,* 515–530.

Tidwell, L. C., & Walther, J. B. (2002). Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations. *Human Communication Research, 28,* 317–348.

Trafimow, D. (2003). Hypothesis testing and theory evaluation at the boundaries: Surprising insights from Bayes's theorem. *Psychological Review, 110,* 526–535.

Weber, R. (2007). To adjust or not to adjust in multiple testing. *Communication Methods and Measures, 1,* 281–289.