2

Questions and Comparisons: Methods of Research in Social Psychology

PHOEBE C. ELLSWORTH
and RICHARD GONZALEZ

This chapter is based on two fundamental premises. The first premise is that the task of the researcher is to ask questions. It is said that in real estate there are three key features for success: location, location, and location. But in research the three key features are questions, questions, and questions. The researcher should constantly be asking questions. What is the phenomenon I want to capture; for example, is it marital satisfaction or marital dysfunction? What is the best way to test my idea? What is the best measure to use? How do I want to explain the findings? Does the research design provide a fair test of this explanation? What other explanations might account for the findings? How should I analyze the data? Are my analyses consistent with the design I used? Does my written description adequately explain the model, the design and procedure, the analysis?

Curiosity, excitement, and passion should drive the consideration of such questions. Simply consulting a checklist of standard questions for each new research project will not do. The researcher should play the role of the child at the Seder who asks the Four Questions (for example, 'On all other nights we eat herbs of every kind; on this night, why do we eat only bitter herbs?'). If the phenomenon really captures the researcher's interest, inquisitiveness comes naturally. Answers to questions lead to new questions, and one of the intellectual attractions of research is that often one cannot predict the next two or three questions down the line. The feeling that one does not know the answer to a research question (but knows how to design a study to find an answer) is exactly what makes empirical research interesting and stimulating. That feeling of uncertainty is sometimes an obstacle for new researchers, who must learn to channel their sense of confusion and vagueness into a workable plan. Indeed, the elimination of confusion is a key motivator for many successful researchers.

Of all the questions a researcher could ask, perhaps the most fundamental one is, '*What is the phenomenon I want to study*?' Am I interested in willing compliance or in the failure to resist pressure to conform? Do I think that attitudes can fundamentally change after a persuasive message, or do I think that attitudes remain stable and only the overt response changes? Before the researcher can begin to think clearly about which measures to use, which control variables to include, what the design should look like, how many subjects should be included, what should be counterbalanced, and so on, it is necessary to be clear about what the research phenomenon is. For any new research project, the best way to define the phenomenon under investigation or to frame questions about it may not be obvious, so it is useful to revisit the fundamental research question frequently. The basic message here is know thy research question.

There is a legend among cognitive psychologists that Endel Tulving stumped first-year graduate students by asking them the question, 'How do you measure a potato?' Thinking this question was an opportunity to demonstrate their creativity, the students generated multiple answers: you can weigh the potato, you can compute the potato's volume,

you can measure its luminosity, its water content, its chemical composition, and so on. The list would grow quite long, until someone finally realized the point of the exercise. How can you measure something without knowing what interests you about it? What is it that you want to know about the potato? Once you know what it is you want to know about a potato, you can figure out the best way to measure it.

This charming story makes a deep point. Students often ask us, 'What statistics should I use on my data?', 'What method should I use in my dissertation?', 'Which measure should I use in my next study?' These questions are difficult to answer without knowing the answer to the fundamental question: *what is the research question you want to answer*?

The second premise of this chapter is that *comparison* is essential in research and should be omnipresent. Our view of comparison is not limited to experimental designs, where, typically, one compares cell means to other cell means; the concept, as we use it, is more general. Predictions can be compared against a known or expected standard (such as Milgram's [1974] use of prediction by experts as a baseline comparison); hypotheses can be compared with each other (for example, Triplett's [1897] analysis of seven hypotheses that could account for social facilitation in bicycle racing); experimental conditions or interventions or subject groups can be compared; individuals can be compared across time; and measures of related/correlated concepts can be compared to each other. A single research design need not include all types of comparison: which comparisons matter depends on the question being asked, but some form of comparison should be present in every research project.

From the second premise (comparison is essential) follow two corollaries: (1) research should attempt to reduce the number of alternative explanations; (2) research programs should be based on multiple methods. We should *compare* possible explanations of research findings to see whether any of them can be ruled out. We should *compare* research methods to see whether analogous empirical findings emerge across different paradigms.

One goal of social psychological research is to offer explanations for phenomena under investigation; these explanations sometimes take the form of process models (for example, stimulus S triggers psychological process P, which elicits behavioral response B). A social psychological finding typically can be explained by more than one process. A dissonance finding can be interpreted with a motivational explanation (Festinger, 1957), a behavioral explanation (Bem, 1967), or a self-threat explanation (Aronson, 1969; Steele and Lin, 1983). The emotional reaction to a sad event can be interpreted in terms of primacy of emotion or primacy of cognition (Lazarus, 1982; Zajonc, 1980). Carefully designed studies allow one to compare different explanations and, with a little luck, to distinguish between them.

Studies that directly compare theoretical explanations by pitting one prediction against another are important for the advancement of theory in the field (Platt, 1964). Greenwald (in press) argues that comparing theoretical explanations has not managed to resolve many theoretical disputes, and suggests that the field should focus instead on establishing the boundary (limiting) conditions of phenomena. We agree that boundary conditions (for example, how 'low' does a low reward have to be in a dissonance paradigm) are important to know, but we also believe that theory has a role and that empirical research in social psychology should contribute to the development of theory (Kruglanski, 2001).

If we ask what value social psychological research adds to our understanding and appreciation of phenomena that artists and novelists have grappled with for centuries, perhaps the best answer is that social psychology can offer theories that work. A theory should provide insight into a phenomenon. It should organize apparently disparate research findings. It should reveal why the observed boundary conditions are the way they are. Although it is rare that one theory emerges intact as the definitive winner, that does not really matter. What matters is that the confrontation or comparison of theories refines and redefines the questions, and generates fresh questions that we might otherwise have been unable to imagine. What matters is the process of creation.

In addition to alternative explanations of basic psychological processes, there are alternative explanations that arise merely because of limitations in a particular research design. Perhaps the reason that participants in the high-threat condition did not perform as well on a memory task as those in the low-threat condition was not because the two groups were differentially 'threatened', but because participants in the low-threat condition were bored and inattentive, or the high-threat participants were distracted. These types of alternative explanations are not as interesting as those that arise from alternative theories, but they can seriously limit the knowledge that emerges from a study. It is incumbent on the researcher to design a study that reduces the number of possible alternative explanations, both theoretical and methodological.

Multiple methods force a type of comparison that is often neglected in social psychological research. The researcher who uses only one method to tackle a problem introduces a confound because all conclusions are conditional on that particular research strategy. Our point is deeper than the issue of whether a finding is generalizable, or robust across research methods. Consider a physicist studying the effect of gravity on objects in free fall. The physicist chooses to study this problem in the context of a vacuum so that extraneous factors can be

controlled, and derives interesting mathematical relations between variables such as time and the distance traveled by the object in free fall. This strategy, however, does not allow for the discovery of the effects of friction on objects in free fall (because the study is performed in a vacuum, without friction). The point is not merely that conclusions will not generalize to settings outside the vacuum but that the understanding of the underlying physical laws that can emerge from a single research technique is limited. The physicist could not discover more general laws involving the conservation of energy.

One of the basic messages of social psychology is that the situation is a powerful determinant of behavior and that its influence usually is not salient to the perceiver. Like the ordinary perceiver who commits the fundamental attribution error (Ross, 1977) by not giving sufficient weight to situational factors, the researcher who holds the method constant (that is, the 'situation') will not be in a position to give sufficient weight to method. By using multiple methods, an investigator can detect these 'situational' factors and hopefully develop a deeper understanding of the phenomenon under investigation. Social psychologists tend to favor experiments because of their potential for high internal validity (that is, their capacity to determine cause–effect relations); however, as we argue below, internal validity is only one of many criteria that research should satisfy. Even if internal validity were the only desideratum of research, the example of the objects in free fall shows that the use of a single paradigm, no matter how 'clean' from an experimental view, precludes complete understanding of a phenomenon.

Regardless of the method chosen, there are certain issues that all researchers must deal with as they progress through the various stages of the research process from generating an idea to writing up the results. In the remainder of the chapter, we will discuss these issues, following the researcher step by step through the process, noting how the choices made at each step constrain other options, and describing how the process itself may differ depending on the type of question and the basic method chosen.

THE LIFE HISTORY OF A RESEARCH PROJECT IN SOCIAL PSYCHOLOGY

Generating questions

Courses and textbooks on methods rarely have much to say about finding good research questions, forming hypotheses, or, least of all, generating theory. William J. McGuire (1973, 1997, 1999) has been a tireless advocate of the importance of hypothesis generation, and has even provided a comprehensive list of tactics designed to stimulate would-be researchers who are looking for ideas. But these exhortations have had little impact on our discussions of research methods, probably because most people *do* have ideas about what they want to study. Turning a general idea into a researchable question, however, is not usually a simple, straightforward task, and here also our courses and textbooks provide little guidance.

Most students begin their research careers in one of two ways: 1) they work on a variation of a question that their adviser is studying, or 2) they discover a flaw in some study that they have read and design research to show that the conclusions of that study were wrong. Research usually leads to further research, and over the course of graduate school most students either find a topic that they care about or choose another career.

A topic is not identical to a researchable question, however. Different researchers tend to prefer topics at different levels of abstraction, and to favor different approaches to formulating research questions. Some naturally think in terms of conceptual variables and abstract constructs: inconsistency creates dissonance (Festinger, 1957); comparative judgments are more rational than absolute judgments (Hsee et al., 1999). Others go around noticing behaviors that seem surprising, irrational, or simply interesting: some people use a lot of hand gestures, others do not (Krauss et al., 1996); people get more upset when they miss a plane by two minutes than when they miss it by two hours (Kahneman and Tversky, 1982); I get mad at the inattentive driver who does not move when the light changes and honk my horn, but I also get mad at the guy who honks his horn at me when I am slow off the mark at the green light (Jones and Nisbett, 1972). Still others want to understand some general domain of behavior: what are the causes of aggression? (Berkowitz, 1993); what makes people happy? (Kahneman et al. (eds), 1999); what makes a decision good? (Hammond et al., 1999). None of these approaches is the right approach; none is the wrong approach. The examples we have given should make it clear that all can advance the field. But each one raises a somewhat different set of methodological challenges (Brinberg and McGrath, 1985).

Whatever one's approach, an important first step is to find out what other people have said about the topic, and whether a body of empirical research already exists. This will be relatively easy if one is interested in a topic that has already been defined by the field – stereotyping, aggression, or attitude change, for example – but rather more difficult if one thinks one is onto something new. Suppose a person wants to study 'betrayal'. A PsychInfo search reveals that there is next to nothing listed under that term, but it would be wrong to conclude that there is no pertinent research. It may be that the

person has merely coined a new term for an old concept, and is at risk of wasting a great deal of time rediscovering familiar ideas (Kruglanski, 2001; Miller and Pederson, 1999). It is important to explore related topics – trust, deception, honor, defection – maybe even anger or expectancy disconfirmation. Even at the literature search stage, it is important to consider alternatives. Doing so can help to refine or redefine the concept, and may often help to turn a vague idea into a set of concrete questions.

If it turns out that there really has not been much psychological research on a topic, the researcher in search of a question must try other strategies – broader reading (of philosophers, biologists, novelists, or news reporters, for example), observation, and thinking. Research on lay reasoning has been influenced by the work of Francis Bacon (Nisbett and Ross, 1980); research on bystander intervention, by news reports of 'urban apathy' (Latané and Darley, 1970). Anthropologists commonly go off to the field without a clear question in mind, allowing their questions to emerge from their observations on the scene. This is a strategy that makes sense at the initial stages of research about unfamiliar people or settings, since predetermined questions might turn out to be inappropriate in the new context. Focus groups and 'grounded theory' methods (Kreuger and Casey, 2000; Strauss and Corbin, 1998) are also strategies for deriving concepts from observation. Finally, sooner or later, most of us who are trying to come to grips with a new topic spend a lot of time in intense thinking, alone or in conversation with others, in the car, in the kitchen, in the shower.

In order to get on with the business of actually designing research, however, at some point our reading, observation, and thinking must coalesce into a manageable hypothesis or question. In some disciplines, a rich description is the end product; for social psychology, it is usually not; we are looking for meanings and ideas that can be tested with other methods and in other settings. A good question should be clear and comprehensible to ourselves and others. It should be neither intractable nor too easy: an answer should be *possible* but not self-evident. A good question should allow for several possible answers, whose relative correctness can be evaluated (Hastie, 2001; Hilbert/Newson, 1900/1902).

Pilot testing

The purpose of pilot testing is to capture the phenomenon embodied in your question – to measure what you intend to measure, and to find or create conditions that match your conceptual variables. The intricate business of pilot testing is not much emphasized in current discussions of research methods, which seem to devote more and more attention to the final stages of the research process – measurement and data analysis – and less and less to the initial planning stages, where the design and procedures are worked out (Aronson, 2002; Ellsworth and Gonzalez, 2001). This is unfortunate. If the design is missing crucial controls, if the treatments and measures do not capture the intended meaning of the conceptual variables, if the participants are bored, or confused, or suspicious, no amount of sophisticated post-hoc statistical repair work can rescue the study. In recent years we have read MA and PhD theses in which the treatments failed to have the intended effect – the subjects did not believe that the 'race-neutral' intelligence test was really race neutral; people given a positive mood induction felt no better than those in the control group; the researcher's idea of a highly credible communicator did not match the ideas of the population being studied. We have seen theses where there were floor or ceiling effects on the crucial measure – everyone's test performance was excellent, almost everyone in all conditions thought the defendant was guilty. With careful pilot testing, none of these problems should occur in a completed study: they should have been discovered and corrected before the study was run. Pilot testing, like all fine craftsmanship, can be frustrating and time-consuming, but surely running an entire study that gets null results because of flaws that could have been corrected is an even more frustrating waste of time.

Pilot testing is not a matter of running the whole experiment through from beginning to end to see if it 'works.' It is an opportunity to test the separate components of a study to see whether they have the intended meaning. Brown and Steele (2001) discovered that it requires a fairly elaborate presentation to get African-American students to believe that an intellectual ability test is really race neutral: simply calling it an unbiased test is not enough. A mood induction used by another researcher, studying a different question in a different context, may not work for your question and your context. A heavy-handed prime – for example, showing people a series of blonde bimbos right before testing their feminist attitudes – may be easy for the participants to figure out: the technique designed to prime sexist attitudes may actually prime defensive self-presentation strategies. During pilot testing, it is possible to stop the study immediately after the treatment has been administered and assess its effects: what did the person being studied think of the race-neutral test? Has her mood improved? What does he think was the point of the bimbo pictures? What stands out about the study so far?

It is often impossible to get information about questions like these during the actual experiment, and the information usually comes too late to be useful. Moreover, so-called 'manipulation checks' can sometimes cause more problems than they

solve. An immediate manipulation check can interfere with the psychological processes the researcher cares about; a delayed manipulation check may be so delayed that the effects it is designed to 'check' have changed or vanished. For example, a verbal manipulation check immediately after the treatment might draw people's attention to the treatment and alter their responses to it, so that the real independent variable is not the one intended, but a combination of the independent variable and the probe. If there are several independent variables, the participant's experience becomes cluttered with distractions. Waiting until the end of the study to check on the manipulations is also problematic: other events and measures have taken place, and people may not be able to disentangle their current feelings from their earlier responses.

Likewise, it is possible to pilot test possible measures, whether the research involves an experiment, a survey, or an open-ended interview study. Pilot testing measures can accomplish several purposes. First, it is possible to test people's psychological reactions to the measures themselves. Do they understand the questions? Do they find them offensive? What do they think you are interested in? Second, it is possible to find out about baseline levels of response – are there floor or ceiling effects? Is this a measure that is likely to vary with other variables in a correlational study, or to respond to changes in the situation? You may find that a task you thought was highly demanding is in fact very easy, that the criminal case you planned to use is overwhelmingly one-sided. You may find that people consider some of your questions to be too personal, and give neutral safe answers that do not reflect their actual beliefs (Visser et al., 2000). You may find that they think the questionnaire or interview is much too long and tedious, so that their earlier answers are much more careful than their later ones.

People are unlikely to give you honest evaluations of your incomprehensible, offensive, or boring questions during the actual study, but during pilot testing, if you tell them (honestly, as it happens) that you are still developing the questionnaire and that you want their help in designing a measure that will be acceptable and meaningful to people like them, they may be more forthcoming. Often, they are eager to be involved in the design of new research and to make suggestions. All empirical research, even the most qualitative, involves measures, and pilot testing is the time to develop measures that are involving, that mean what we want them to mean, and that people will answer honestly.

Especially for studies with manipulated independent variables, a few run-throughs of the whole experiment from beginning to end are useful – not so much to find out whether the experiment is going to 'work,' but to find out whether we have successfully translated our conceptual question into a coherent set of procedures that makes sense to people and holds their attention, and to find out whether the separate components work together. The pictures of blonde bimbos, for example, might be a perfectly good prime for sexist sentiments in some contexts, but not when followed by a dependent measure that is obviously a test of feminist attitudes. Either a subtler treatment or a subtler measure may be necessary to keep the subject from figuring out what you are trying to test.

Pretest–posttest designs can raise similar issues: asking people how they feel about affirmative action at the beginning of the hour, and then presenting them with a communication designed to change their attitudes about affirmative action may be quite different from simply presenting the communication. The pretest may sensitize the subject to the communication (Smith, 2000), making its effects stronger or weaker, but in any case *different* than they would have been otherwise. If a pretest is necessary, it is generally better to administer it much earlier, in a context apparently unrelated to the research. Often, it is not necessary to pretest at all, since random assignment can be used to ensure equivalence among groups (Greenwald, 1976).

Pilot testing is also the best opportunity to discover whether a study accords with ethical standards for research. One can find out whether the actual experience of participating in a particular study is upsetting, painful, or humiliating in a carefully monitored context where the experimenter is prepared to stop everything and talk to the participant at the first sign of distress. If some people find some of the emotionally arousing pictures too disturbing, or the task too embarrassing, or our questions too personal, or the deception unjustifiable, we must make changes. We can look for different stimuli or measures that still get at what we are interested in but that are less upsetting. Or we may learn that we need to screen out certain people, people whose life experiences might make them especially sensitive to some aspect of the research procedure. Occasionally, we may have to try a whole different approach to studying our question.

Most human-subjects review boards emphasize informed consent and debriefing as the primary ethical requirements. Important as these are, they are brackets around the person's actual experience, far less important to the participants than what actually happens to them during the study. In order to write an honest informed consent request, it is necessary to know how real people have actually reacted to the procedures – that is what participants need to know in order to be accurately informed. Of course, just as people may be reluctant to admit that they did not understand the questions or thought the experiment was stupid or transparent, they may be reluctant to admit that they were upset. Again, one way to elicit honest responses is to tell them that you are still developing the experiment, and ask

them whether they think *other* people like them might be disturbed about some aspect of the procedure.

Generating alternative hypotheses

It is important to develop multiple working hypotheses early in the research process, (Platt, 1964). Sometimes, if the topic is very new, the very first study may be designed to discover whether the phenomenon exists: whether people conform to group pressure when it contradicts the evidence of their senses, whether people from different cultures agree on the meaning of emotional facial expressions, or whether people are more concerned about losses than gains. At this stage, a simple demonstration can be an important contribution. Even if the phenomenon is brand-new, the researcher still has to consider and rule out artifactual alternatives, as we will discuss below. Moreover, it is important to think carefully about how novel one's phenomenon actually is – to ask, 'What are the most closely related phenomena that *have* been studied, and how can I argue that mine is really distinct?' Or, 'What are the most closely related treatments (or definitions of the independent variable) that have been used, and can I differentiate my conceptual variable from the ones they were designed to study?'

If the researcher has gone beyond the simple demonstration of a phenomenon and is interested in causes, processes, moderators, or mediators, it is important to consider other possible causes, processes, moderators, or mediators. One technique is to imagine that we have already found the results that we expected, and consider other possible interpretations. If the researcher is invested in his or her own favorite hypothesis, it is often very hard to generate plausible alternatives (e.g., Griffin and Ross, 1991; Ross et al., 1977), but it is important to make a serious effort while there is still time to modify the design or procedures. Discussing the research with other people is a good way to get beyond one's own biases; for example, if you describe your study as though it were already finished and your hypothesis was confirmed, your colleagues, anticipating the responses of your reviewers, will often suggest numerous alternative explanations. These may not, at first blush, seem *at all* plausible to you, but they are worth considering because they are plausible to *someone*. McGuire's (1997) detailed advice on how to generate initial hypotheses is at least as useful for generating alternative hypotheses.

Although only you yourself, aided by colleagues with whom you share your initial ideas and an open-minded literature search, can really judge the plausibility of alternative answers to the specific question you are posing, there are some generic, formal alternatives that should be considered, regardless of your question (cf. Brewer, 2000; Ellsworth, 1977). If your hypothesis is that X causes Y, a common alternative is that Y causes X. Ordinarily, when random assignment is possible, this will not be a plausible alternative (although in other settings, without random assignment, it might). In a correlational study of chronic attributes, reverse causality can be a serious rival. Does attractiveness lead to perceptions of competence, or vice versa? Or both? It is important to find a setting in which you can be sure that one precedes the other, or where you can introduce one independently of the other.

A second common alternative is the familiar third-variable correlation: X and Y occur together because some third variable, Z, is responsible for both. If a study finds that boys who are heavily involved in sports cry less often than boys who are not, it would be wrong to conclude that sports lead to stoic behavior or to happiness; one (of many) possibilities is that parents who have strong beliefs about appropriate sex-role behavior push their sons into sports and punish crying. Again, this alternative is generally ruled out in settings where people can be randomly assigned to the X treatment, but can be a serious problem in quasi-experimental and correlational designs.

A third possibility is that X alone is not enough – that X interacts with some other variable Z; that is, *Z moderates* the effects of X. Looking people in the eye makes them like you, but only in social contexts that are already positive; otherwise, it can be threatening (Ellsworth and Carlsmith, 1968). This kind of alternative cannot be ruled out by random assignment in a laboratory experiment. Typically, many features of the situation in a laboratory experiment are *held constant*, meaning that any one of them could potentially interact with the intended independent variable, but the interaction would not be discovered. If the experimenter studies only positive social interactions, she will conclude that eye contact is a positive signal; if only women are studied, if the communicator is always credible, or if the experimenter is extremely attractive, results that apparently confirm the hypothesis could actually be due to a combination of these constant variables with the variable the experimenter cares about.

These are the most common formal alternatives, but they are not the only ones. For a more extensive discussion, see Cook and Campbell (1979) or Ellsworth (1977). In designing a study, it is useful to consider these various alternatives, substituting one's own variables for the X's and Y's, thinking carefully about possible Z's, and building in controls (groups, measures, or occasions) to test any that seem to be plausible rivals. For any given question, some of the suggested rivals may be completely *implausible* – the investigator will not be able to think of any credible alternative that takes that particular form (West et al., 2000). If the hypothesis is that males are more assertive than

females, the reverse causality hypothesis that assertiveness causes gender can instantly be ruled out. Similarly, in a situation where random assignment is not possible, the addition of measurement occasions may make the alternative explanation of time implausible (for example, three observations before the naturally occurring 'treatment' followed by three observations afterwards help rule out alternative explanations if the only change seen in the dependent variable occurred between the third and fourth observations). Adding a second condition to a before/after design in which different participants respond to the pretest and posttest measures but do not experience the intervening intervention helps rule out alternative explanations such as the possibility that the effect is due to the same participants answering the same question twice.

In addition to alternative hypotheses about the relationship among variables, it is also important to consider alternative hypotheses about the *definition* of the variables. These are problems of *confounds* and *construct validity*. (Usually, the term *confound* is used for a correlated variable that is specific to the research environment and relatively trivial, while *construct validity* is implicated if important variables are correlated in a wide range of settings.) If males behave more assertively than females in a given setting, for example, it may be due not to their gender per se, but to their greater power. Females who have become accustomed to power may be as assertive as males. (If men were given more resources in the context of a particular experiment, this would be a confound; the more general problem of males' power in the wider society raises questions about the construct validity of the concept of gender in relation to assertiveness.) Both independent and dependent variables may be correlated with other variables besides the one we care about, and so efforts must be made to vary or measure them independently.

Dull but serious alternative explanations

Finally, there is a set of boring but fatal alternatives that are a risk of the research process itself: methodological artifacts such as demand characteristics and experimenter bias. *Demand characteristics* are 'subtle or not-so-subtle cues in the experimental setting that influence subjects' perceptions of what is expected of them, and that might systematically influence their behavior' (Aronson et al., 1990: 347).

Research participants usually know that they are being observed and studied, and they try to figure out what the research is about, and what is expected of them. They may not take our treatments and measures at face value, but may instead interpret them in the light of our imagined intentions, and adjust their behavior accordingly. We may *tell* participants that they are participating in two entirely different studies or that 'there are no right or wrong answers to these questions', but the fact that we have said it does not guarantee that they believe it. Research participants are interested, curious human beings. Most of them associate psychology with tests of mental health and mental abilities, and they may be apprehensive about appearing smart, or sensitive, or normal, or whatever they think we are concerned about ('evaluation apprehension'; Rosenberg, 1969). If we expect differences among different groups of participants (whether randomly assigned or 'natural' groups) and the demand characteristics are different across groups, we have no way of being sure that the results were due to our conceptual variables rather than to differences in demand characteristics. Demand characteristics can be discovered and remedied during pilot testing, when we can ask people what they think the questionnaire is getting at or what the treatment means, and how they think a normal, smart, successful person would respond.

Experimenter bias occurs when an experimenter unintentionally influences the participants to behave in a way that confirms the hypothesis. Neither the experimenter nor the subject has any awareness that this has taken place. Sometimes its causes can be easily identified – for example, videotapes of pilot testing showed that one of our research assistants, when presenting an array of faces for children to choose from, inadvertently (and consistently) pointed to the one we expected would be chosen. More often, such cues are subtle, and cannot be identified, even though the effects are strong (Rosenthal, 1969). Experimenter bias is a serious, pervasive alternative explanation (Rosenthal and Rubin, 1978) that can occur in any kind of research involving interaction with human beings – whether the research is qualitative or quantitative; observational, survey, or experimental – and has even been demonstrated in studies of rats (Rosenthal and Fode, 1963). In fact, the first famous demonstration of the phenomenon was Clever Hans, the arithmetically gifted horse who could solve numerical problems by tapping out the right answers with his hoof. It turned out that his gifts were psychological rather than mathematical – he was able to notice subtle changes in his trainer's posture and expression when he reached the right answer, and stopped tapping.

Pilot testing usually cannot diagnose experimenter bias. The design of the experiment must include controls to rule out this alternative explanation. A purely automated presentation of stimuli and measures can be effective, but it sacrifices the 'social' aspect of many social psychological questions (and in any case is not a guarantee, because differences in the printed or recorded words across conditions can also be a source of bias (Krauss and Chin, 1998).

Having experimenters or interviewers or observers who are unaware of which experimental

condition the participant is in provides complete protection, although it is sometimes hard to achieve. For example, if the study concerns obvious visible characteristics such as race or gender, experimenters may communicate different expectations even if the researcher does not tell them that the study is about race. Keeping the experimenters blind to one of the variables in a study protects against bias on that variable and on interactions with that variable. Keeping the experimenters unaware of the hypothesis may seem like an effective solution, but it is not. First, since the experimenters run all of the conditions and can observe the differences among them, and know what is being measured, the hypothesis may be pretty easy to figure out; second, even if the researcher tries to keep the experimenters in the dark, they will inevitably develop their own hypotheses, and communicate those to the participants. Like the people being studied, research assistants are sentient human beings who want to know what is going on. (Protection against experimenter bias, though difficult, is achievable: fuller discussion may be found in Aronson et al., 1990.)

The independent variable

The term 'independent variable' is often used to refer to the cause in a hypothesized cause–effect relationship, but it can also refer to any variable that *predicts* another. One might ask, for example, whether high levels of education predict better health; whether low popularity or high popularity is correlated with being a bully; whether moral standards vary with social class.

Independent variables differ on how much latitude their definitions offer. Abstract conceptual variables, such as power, mood, or group cohesiveness, can carry several concrete representations. We could give participants descriptions of high- and low-power people, or bring them into the laboratory and assign some the role of boss and the others the role of subordinate, or go to actual workplaces and interview real bosses and subordinates, or do an observational study in a classroom to see who influences whom, or simply ask people how powerful they think they are. Each of these methods will raise a different set of alternative hypotheses, which must be ruled out in the design of the study. Using more than one definition over a series of studies greatly increases the researcher's confidence that the important variable is really power, and not some correlated variable, because the correlated variables are likely to be different for different operationalizations of power. This provides convergent validity for the meaning of the independent variable.

Other independent variables offer less freedom in definition. A researcher interested in gender differences or other demographic variables, or interested in a particular educational method (such as small class size), legal reform (such as allowing jurors to take notes; providing legal aid), or other social policy (such as a tax cut) has far fewer choices. But these 'obvious' independent variables do not necessarily make the researcher's task easier. True, there is no question of how to operationalize the concept of 'woman' as opposed to 'man', but there is still work to be done before one can conclude that gender is really the variable that matters. If a psychologist had decided to use Harvard and Radcliffe undergraduates to study gender differences in the 1960s, aside from the obvious problem of generalizability to the population at large, there would be a serious selection problem affecting the comparability of men and women even within this elite population. Harvard was large; Radcliffe was small. Given the norms of the time and the size of the classes admitted, the Radcliffe students represented a much more highly selected group than the Harvard students. Differences that looked like gender differences could have been differences in qualifications. Whether this would be a *plausible* alternative, of course, depends on the hypothesis. A finding that women had superior verbal skills would be suspect; however, the women's higher qualification would not be a plausible alternative explanation for Matina Horner's (1972) finding that women demonstrated 'fear of success'.

Again, one goal in research is to reduce the number of alternative explanations that can be attributed to an independent variable. One technique to reduce the number of alternative explanations is to show convergent validity – show that you get qualitatively similar results under different definitions of the independent variable. Convergent validity is merely one criterion. The discussion of additional criteria requires a classification of predictor variables into those that can be manipulated, measured, or 'found'.

Independent variables that can be manipulated

Traditionally, social psychologists have favored the manipulated independent variable with random assignment of participants to two or more conditions, and there are good reasons for this preference, especially if the researcher has a causal hypothesis. Random assignment of participants to conditions is an enormously powerful technique because it allows the researcher to rule out whole categories of alternative explanations at once and guarantees that, on average, all of the groups are the same before the treatment is given. If the researcher finds differences between the groups, these differences can be attributed to the independent variable the experimenter cares about, and not to differences in the participants' backgrounds, personalities, abilities, motivation, or anything else

about themselves or their lives before the random assignment took place.

Random assignment does not solve all problems, however. First, any differences in the experiences of the experimental group and the control group(s) that occur *after* the participants have been randomly assigned and that are not an essential feature of the treatment are possible alternative explanations for the results. If the participants in the experimental group interact with different people, work on a more interesting task, or have experiences that make them more confused or more suspicious, any of these differences could account for the results, rather than the independent variable the experimenter has in mind. In the ideal random assignment experiment, the experimental and control participants are given the same information, spend the same amount of time in the study, interact with the same people, and engage in the same activities, except for the introduction of the treatment. If the manipulation varies on many dimensions, it is difficult to pin down what it is really manipulating.

Second, if participants are more likely to drop out of one condition than the other(s) after they have been randomly assigned, we can no longer be sure that the groups are equivalent. Suppose participants in the experimental group are to see a film of a rape trial and those in the control group a film of an assault trial involving two men, in a test of whether viewing violence against women affects feminist attitudes. If many more people withdraw from the study when they are told that the film involves a rape, one hypothesis is that these are the participants with the strongest attitudes about violence towards women. This means that even before the experimenter shows the films, the average feminist attitudes of the treatment group and the control group might already differ, and any differences on the dependent variable measure might not be due to the rape film at all but to prior group differences. Note that if a study involved showing *all* participants a film of a rape trial, a high dropout rate is a less serious problem: it can raise questions about the generalizability of the results to people who are unwilling to view rape films, but not about their validity among the people who are willing.

In a laboratory study, careful pilot testing can usually ensure that loss of participants is minimized and differential dropout is not a problem. In field studies, in those rare and precious instances where random assignment is possible, participant attrition or reassignment often poses a more serious threat. Parents may agitate to get their children moved out of racially integrated classrooms or into programs designed to improve school achievement, compromising the initial random assignment. Students or workers with weak skills may drop out of programs they find too challenging. The result is that the groups that are measured after the treatment are no longer the groups that were randomly assigned, and any differences found could be due to differences in the composition of the groups rather than to the intended independent variable (Cook and Campbell, 1979; West et al., 2000). Sometimes in field studies, random assignment may be undermined at the very outset, as when doctors surreptitiously assign their high-risk patients to a promising treatment group, perhaps furthering humanitarian goals, but invalidating the results of the study (Kopans, 1994). Unless the experimenter has full control over who *gets* into the various conditions of the experiment and who *stays* in them, random assignment may be an illusion, and it is important to keep track of the *actual* composition of the groups throughout the study.

The most common use of random assignment in social psychology is in the laboratory study. A consequence of this preference is that our independent variables are often weak, our dependent variables often inconsequential, and our effects inevitably short term. *This does not mean that our studies are invalid or unimportant.* The insights and theories tested in laboratory studies – about conformity, altruism, attitudes, expectancy effects, cognitive biases, and many other topics – have proven to be powerful and often generalizable to a wide range of nonlaboratory settings. Still, working within such narrow confines, it is almost impossible to test the boundary conditions of our findings. Consider the manipulation of letting one student boss another around for 45 minutes, giving some ego-bolstering praise. If we find that participants given this brief power manipulation are likely to take the credit for successes and deny blame for failures (Tiedens et al., 2000), can we conclude that this is also true of government officials or CEOs who have experienced power on a daily basis for years? What do a few moments of criticism have in common with chronic low self-esteem? If the only attitudes we study are the sort of trivial attitudes that can be changed within the course of an hour, what have we learned about deep-seated ideological convictions? In fact, we may have learned a great deal – it is as wrong to claim that laboratory results do not generalize to the real world as it is to claim that they do; as wrong to claim that short-term acute laboratory manipulations of variables are different from their real-world counterparts as it is to claim that they are the same. These are open questions, and only research that uses different methods in new settings can answer them.

Independent variables that are measured

When the independent variable is a measured variable, new problems arise. For example, when using self-report, the researcher must consider whether people can give an accurate assessment (for example, people may not be good judges of how powerful they are), as self-report measures can be woefully

inaccurate (Reis and Gable, 2000; Wentland, 1993). Self-report measures can also make the variable salient in people's minds and bias future responses: a person who has just described herself as powerful might be especially likely to feel a burst of high self-esteem. Reverse causality and third-variable causality are also problems with measured variables; for example, low self-esteem may cause people to see themselves as powerless (or is it that not having power causes low self-esteem?). Thus, the problems with *measured* variables are first, construct validity – are we really measuring the variable we care about and only that variable? – and second, reactivity – whenever people are aware that they are being measured and have any control over their responses, the measure can be affected by the image the participant wants to convey, rather than by the variable we care about. The techniques for dealing with construct validity and reactivity problems of measured predictor variables are the same techniques one uses with 'found' independent variables, so we will review the solutions together in the next subsection.

Independent variables that are 'found'

But what if the researcher is really interested in a variable that cannot be manipulated at all – gender, for example, or social class, or culture. These are, by definition, 'found' variables, although one's degree of gender, class, or cultural *identification* might be measured and might be relevant to some questions. In these cases, the strategy is to identify correlated variables, such as wealth or power, and attempt to rule them out, or to look for a variety of measures that might reflect the processes one cares about but that would not be affected by plausible third variables.

A psychologist interested in the effects of power on self-esteem, for example, might go to the field and study high- and low-power people in a hierarchically organized workplace ('found' power). Studying people who occupy real positions of high and low power has an appealing real-world relevance, but there are many possible variables that could cause differences in self-esteem between people in real-world positions of power and their subordinates. People could have attained powerful positions because they are older, more skilled, more educated, richer, whiter, or male – or even, perhaps, because they had higher self-esteem to begin with. Without random assignment, the researcher has to consider each of the plausible alternative hypotheses one by one and find ways to rule them out – by finding an all-black female group; by comparing jury forepersons on juries where the role is randomly assigned to those on juries where the members elect the foreperson; by statistically controlling for age, income, education, and so on. Rarely are these complete solutions. Thus, the problems with *found* variables are problems of nonrandom selection (self-selection or selection by others) and correlated variables – income is correlated with success, health, power, and SAT scores, so in comparing people on any of these dimensions we have to worry about whether we are really comparing them on income.

There are three common ways to rule out correlated variables. First, the researcher can use careful selection to make sure that the correlated variable does not vary. The third variable is 'held constant'. You may have the hypothesis that women are more prone to depression than men, but you also know that women earn less money. Thus, the depression that appears to be due to gender could actually be due to lower income. So you attempt to hold income constant by studying only people within a narrow economic range. You lose generality by this method (that is the gender effect you have demonstrated may be limited to people in that narrow income range), but you gain confidence that gender (or, alas, something else that is correlated with gender but not with income) plays a role in depression.

Second, the researcher can construct a model that includes both variables, not only the one he or she cares about, but also the troublesome correlated variable (or several hypothesized variables and several correlated variables). This makes it possible to examine the effects of both variables separately. In actually conducting the research, you would have to find people in each of the groups you want to compare; in this case, women and men who were characterized by different levels of the correlated variable – poor men and women, middle-income men and women, and rich men and women, for example. This is analogous to the technique of systematic variation in a random assignment study. If women are more depressed than men in all three income groups, you are more sure of your hypothesis; other patterns of results force you to consider new hypotheses. The main problem with this method is that it can be difficult to implement, especially if you want to rule out several correlated variables, as some combinations of variables may be quite rare (for example, very rich women). New statistical procedures based on 'propensity score' techniques help make this problem tractable (Rosenbaum and Rubin, 1984).

Third, the researcher can use statistical methods, such as analysis of covariance (ANCOVA), to control for correlated variables. These methods, as typically implemented, control for the linear association of the third variable or set of variables. In effect, a linear regression is computed where the linear effect of the third variable on the independent variable is subtracted from the independent variable, and the remainder – the residual – is used as the independent variable instead of the original independent variable. When using these techniques, the researcher must be careful not to to make

general pronouncements, such as 'we controlled for the effect of income'. More precisely, what typically was controlled for is the linear effect of the covariate. The data may still contain nonlinear effects of the correlated variable or interactions between the third variable and the independent variable.

We conclude this section on independent variables with this point: as long as a variable has been studied with only a single set of procedures, it is impossible to distinguish the role of the variable from the role of the procedures (Campbell and Fiske, 1959). The procedures, or the-variable-in-the-context-of-these-procedures, constitute an alternative explanation of the results that ordinarily cannot be ruled out in a single study. In order to make real progress, sooner or later it is important to study the same question by a different method – to compare the measured version of a variable with the manipulated version, to use an entirely different manipulation, or to find an instance of the variable rather than creating one. If the results are different, the researcher is confronted with a whole new set of questions about *why* they are different, questions that can stimulate real theoretical progress and understanding that would otherwise be unlikely. For example, laboratory experiments on social comparison showed strong evidence that people tend to compare themselves to others who are slightly better than they are on whatever dimension they are concerned about (upward social comparison). However, in field research, Taylor (1983) found that breast cancer patients tended to use downward social comparison, comparing themselves to patients who were not doing so well, with the result that almost all of the women thought that they were adjusting very well. These field data extended our understanding of social comparison processes in ways that were not suggested by the experimental research.

The dependent variable

Measured outcome variables raise the same issues as measured predictor variables. It is important to consider what *else* the measure might be tapping besides one's intended variable (construct validity). It is important to find out during pilot testing what sorts of motivations and interpretations participants experience when they encounter the measure (reactivity). And it is important to consider alternative explanations of the whole process, and include measures designed to assess other possible outcomes that might address these alternatives (tests of multiple working hypotheses) (John and Benet-Martinez, 2000). Sometimes the actual variable captured by the measure may be broader than the construct the researcher has in mind; for example, the researcher may be interested in favorable attitudes toward an outgroup, but the measure might actually reflect global good mood, in which case the predicted outcome is just a byproduct of a more general phenomenon. This possibility can be addressed by adding additional measures that are unaffected by mood, or additional measures that have nothing to do with the particular outgroup (known in the statistics literature as an *instrumental variable*). Sometimes the actual variable may be narrower than the researcher's concept. For example, the researcher may be interested in individualism as opposed to collectivism, but the scale may reflect *only* differences on the collectivism items while the groups are identical on the individualism items (Oyserman et al., 2002). This possibility can be assessed by looking for patterns and discrepancies in the individual items, for example, with confirmatory factor analysis. Sometimes the measure may tap a different variable altogether. For example, if one wants to measure knowledge or accuracy of perception, it is important to create a measure that is not affected by attitudes: if most of the 'correct' answers on a person perception measure involve negative qualities, what looks like accuracy could actually just be simple dislike.

Reliability and validity

The reliability of a measure refers to its consistency: consistency over time, consistency over observers, or consistency over components of an overall measure, such as items on a questionnaire. All three are essential if one is trying to measure a stable attribute such as a personality trait, as is often the case when one is interested in a measured predictor variable (Bakeman, 2000; John and Benet-Martinez, 2000). Consistency over observers and consistency over components or items are analogous, in that both involve multiple attempts to measure the same thing at a given point in time. A researcher may use two or more observers to score how aggressively a person is behaving, how often dispositional attributions occur in a narrative or a conversation, or any number of other variables. Or a researcher may ask several different questions all designed to tap aggression, the tendency to make dispositional attributions, or any number of other variables. If observers or items disagree, the measure is unreliable, and needs to be modified. Observers or coders may need further training (Bartholomew et al., 2000); items or coding categories may need to be revised or discarded. Generally, the more open-ended or unstructured the behavioral or verbal responses, and the more abstract and inferential the coding categories, the more difficult it is to develop a reliable measure. It is easier, for example, to measure competitiveness reliably in a game-like situation where there are only a few response alternatives, some competitive and some cooperative, than it is when observing

playground behavior. It is easier to achieve reliability if one is measuring concrete behaviors ('hits', 'kicks', or 'shoves') than more abstract categories ('shows aggression').

While consistency over observers and measures is always essential in social psychological research, consistency over time is often not relevant. Very often we are interested in people's responses to an immediate situational stimulus – a threat to self-esteem, a subliminal prime, or a persuasive argument. We do not expect lasting effects; in fact, we go out of our way to debrief the participants in order to make sure that the effects are undone.

A measure that is unreliable cannot be valid. If observers cannot agree on whether behaviors are aggressive or merely assertive, if the items on a test are uncorrelated, or if a person gets different scores from one day to the next on a measure of a supposedly stable trait such as IQ or extraversion, the measure is useless. (Of course, the fact that a construct we thought was coherent or stable turns out not to be so may lead us to new theoretical insights, but the measure is useless for its original purpose.) Thus, reliability must be established before a measure is used.

A measure is *valid* if it measures what it is supposed to measure and nothing else. A reliable measure is not necessarily valid. A blood test, for example, may be a highly reliable (and valid) measure of whether people are HIV positive, but an invalid measure of whether they are immoral or gay. Validity is not easy to establish in social psychology, because our conceptual variables – variables such as prejudice or anxiety – often represent families of related variables rather than pure states, so there is no gold standard by which to measure them. Certain cognitive biases may be rigorously demonstrated as departures from a statistically correct response (Kahneman et al., 1982), but interpersonal biases are not so easily verified. 'Criterion-related validity' often makes no sense for social psychological variables, at least at the current stage of development of the science, because there is no single criterion that definitively identifies most of our variables. A recent example is work attempting to develop a measure of attitude ambivalence (Breckler, 1994; Priester and Petty, 1996; Thompson et al., 1995). Researchers disagree about whether ambivalence should be measured from a structural point of view (that is, ambivalence as a combination of separately measured positive and negative attitudes) or from an experiential view (that is, the subjective phenomenology of attitude ambivalence). There is currently no clear criterion against which to assess the validity of the various proposed measures of ambivalence.

For many of our variables, validity must be established slowly, by triangulation. If we want to use frowning as a measure of anger, for example, we might look to see whether frowning occurs with other variables plausibly associated with anger: with independent variables such as being thwarted or insulted, with dependent variables such as yelling, threatening, and slamming doors. This is the process of establishing *convergent validity*: many other indicators of the conceptual variable 'anger' are associated with frowning. Just as important, we want to make sure that frowning is unique to anger, that it is not associated with other mental states. This is the process of establishing *discriminant validity*: demonstrating that a frown discriminates anger from other states such as fear or sorrow (Campbell and Fiske, 1959; Judd and McClelland, 1998). In fact, it does not; frowning is characteristic of various kinds of mental effort, uncertainty, and perceived obstacles. Thus, it would not be a very good measure of anger *unless* other supporting measures were included that were *not* related to mental effort, or *unless* the situation was structured so that none of the other mental states that go with frowning was plausible in context.

What have we just said? We have said that frowning lacks discriminant validity as a measure of anger, but that in a context that rules out other types of uncertainty or obstacles, it could be a valid measure. There is an important *general* message here: that in social psychology many of our measures are not valid or invalid per se, but are valid or invalid in a particular context. Personality psychologists generally look for measures that are stable across time and context, but this is far less true of social psychologists. We are generally interested in situational variables, we expect our measures to be responsive to the particular situation, and therefore we should not expect to find measures that are universally valid or applicable. Just as the answers to individual questions (for example, 'Overall, how satisfied are you with your life in general?') can have different meanings depending on the questions that preceded them (Schwarz et al., 1998; Tourangeau and Rasinski, 1988), so *any* measure might have different meanings in different contexts. Many social psychologists seem to have forgotten this important fact. Whenever a set of messages is sent out over the Social and Personality Psychology Listserver, for example, there are almost always some that ask whether anyone knows of a good off-the-shelf measure of some variable – regret, or mistrust of authority, or vengeance – as though any measure that someone used successfully in one context is a generally valid measure. Often, these are not intended to be used as measures of enduring traits, but as measures of responses to situational variables. Rarely do these questioners ask about the context in which the measure was used or describe the context in which they plan to use it. This is a serious mistake. First, the measure may not be appropriate in the new context; for example, questions about racial prejudice may elicit different answers in all-white groups than they do in mixed-race groups. Second, the researcher often has a wide range of measures to choose from, each appropriate to some

contexts but not to others, and looking for a generic measure may prevent the researcher from finding or creating a measure that fits the particular context. Racial prejudice, for example, may be measured by questions about affirmative action, welfare mothers, or the guilt of a particular criminal defendant; or by eye contact, conformity, helping, or any number of behaviors that in other contexts might have nothing to do with racial prejudice.

The analogous argument can be made about reliability. Social and personality psychologists often report the reliability of a scale (such as Cronbach's alpha) as though the value of the reliability measure is an inherent property of the scale. Our journal articles contain sentences such as 'Scale X has been shown to have acceptable reliability, alpha = 0.82', with a reference to another article. Typical measures of reliability are a function of error variance, so anything that changes the error structure of the data (change in subjects, change in experimenter, change in instructions, change in manipulation, change in task, change in length of the study, etc.) will change the reliability of the scale. Thus, the reliability of a scale should always be reported for that particular study; it is meaningless to claim that a scale is reliable in one context because it was found to be reliable in another.

Many of our measures are open to multiple interpretations. A direct gaze, for example, can imply liking, subordination, disapproval, or simple attention. This does not mean that gaze direction is a bad or invalid measure; it can be an excellent measure of any of these concepts provided that that is the only meaning that makes sense in the particular context, that precautions have been taken to rule out alternative explanations. Nonverbal, behavioral measures (and manipulations) often come in for criticism because their link to the intended concept is less transparent than that of verbal measures. A scale that asks people to rate their anxiety on a seven-point scale seems to be a more direct measure of anxiety than a measure of speech hesitations or fidgeting. But this advantage is often more apparent than real. Verbal measures almost always come with *built-in* alternative explanations such as reactivity, social desirability, and cultural stereotypes or folk theories. Nonverbal measures are relatively free of these problems, because they are usually under less conscious control than verbal reports, and because it is often possible to keep the participant unaware that a measure is even being taken. For nonverbal measures (and sometimes for verbal measures as well) alternative explanations usually have to be figured out on an ad hoc basis in each context.

Internal and external validity

If all of the procedures in this section are followed, an experiment should have high internal validity. *Internal validity* means that in this particular study, any differences observed between the participants in different conditions or groups are due to the treatment, not to any artifact or confounded variable: being given a high-status role caused participants to respond to failure with anger, and being given a low-status role caused participants to respond to failure with sadness. Of course, to make even this claim, we have to be sure that our status manipulation affected status and not some related construct, and our anger and sadness measures reflected anger and sadness, and not something else. If so, we know that our treatments were responsible for the outcomes.

External validity means that the results will generalize to other people and other settings (Brewer, 2000; Campbell, 1957). No single study can have external validity, since it is impossible to know whether the results will replicate in another context. The findings of a study using college students as participants may or may not generalize to senior citizens; the findings of a study using senior citizens may or may not generalize to college students. The results of a laboratory study of productivity may or may not generalize to telemarketers; the results of a study of telemarketers may or may not generalize to postal workers. External validity is always an empirical question, requiring further research. Thus, there is no 'trade-off' between internal and external validity. If a study lacks internal validity, nothing has been learned, so there is nothing to generalize. If a study has internal validity, its external validity is always an open question.

Social psychologists are sometimes criticized because they hardly ever bother to use truly representative samples in their research, and often just use whatever participants are most ready to hand – for many of us, this means college students who are taking a course in introductory psychology. There are serious costs to restricting our research to one small segment of the population, just as there are serious costs to relying on a single type of method. Any results we find could be peculiar to the college student population, or could represent an interaction between some feature of that population (youth, IQ, interest in psychology) and the variable we are interested in, rather than the variable itself (Sears, 1986). Ultimately, no result can be trusted as general – or even as real – until it has been tested on different kinds of people with different kinds of methods.

However, conducting research on a truly representative sample of almost any population is enormously expensive. For some kinds of question, a representative sample is necessary; for others, it is not. It is important to think carefully about the kinds of samples that are appropriate for your research question and the kinds that are not.

A representative sample – a sample in which every member of some population has an equal

chance of being included – is imperative if you want to make valid statements about the absolute frequencies of various responses in that population. For example, in predicting the outcome of a national election, you want to make accurate estimates of how many people favor each candidate and how many are undecided. In order to do this, you must draw a representative sample of voters. Likewise, if you want to know how blacks, whites, Hispanics, and Asians feel about affirmative action, or how often men are victims of violent crime compared to women, you need a representative sample.

But often in social psychology, our hypotheses are not about base rate differences among groups, and often we are not concerned with the absolute percentages or exact numerical levels of the variables we measure. We ask questions such as: 'Can information people learn after an event change their memory for the event?' (Loftus, 1979); 'Does sorrow lead to a perception that events in general are uncontrollable?' (Keltner et al., 1993); 'Is a person more likely to help another when alone or when there are other people around?' (Latané and Darley, 1970). We are interested in the effects of psychological variables on other psychological variables and behavior. We do not particularly want to make statements about the exact percentage of people whose memory will be distorted with and without new information, or the precise size of the decreases in perceived controllability caused by sorrow. To us, estimates like these do not even make sense – there *is* no exact number: it will vary depending on the type of event, the type of new information, and all sorts of other factors. Testing a large random sample of Americans in one particular experiment designed to ask one of these questions would be a huge waste of time and money. Vastly more could be learned by a judicious choice of small, nonrepresentative samples in a variety of experimental contexts.

This is not to say that college students are fine for all our questions. They are not. The examples described above were chosen partly because they were plausibly true of old and young people, rich and poor, male and female. For questions like these, there is no compelling reason *not* to start with college students, although later on in one's research program it is important to move on to other groups in order to test generality and boundary conditions.

But for other questions, any old sample will not do. The researcher needs to consider what kind of sample will most likely provide useful answers to the question. The sample need not be representative, but it must meet certain specifications. Rather than a sample of convenience, a *sample of forethought* is needed. Sometimes the sample specifications are obvious. In research on aging, college students can only aspire to be in the control group; in research on cultural differences, you need people of different cultural backgrounds. But, at least at the outset, when you are trying to establish the existence of a relation between variables, you do not need *representative* samples of old and young people, or members of the cultures you want to compare. You must make sure to choose samples that are uncontaminated by correlated variables that might be alternative explanations for your results (e.g., you would not go to a hospital if you want old people, because they would be not only old but also unhealthy), and eventually you must test your hypotheses on different samples, but you do not need a fully representative sample.

For these questions, the need for samples of forethought is obvious. For other questions, the need to seek out special samples may be important, but less obvious. College students have certain characteristics that make them a poor choice for some questions (Sears, 1986). Much social psychological research on racism and prejudice, for example, has shown surprisingly weak effects, at odds with what we know about pervasive racial segregation, poverty rates, and the racial populations of America's prisons. Some of this discrepancy is undoubtedly due to the fact that undergraduates in research universities are much less likely to express overt prejudice than are some other segments of the population (Sommers and Ellsworth, 2000). By sticking to the college student population, we have learned more about weak prejudice laced with liberal guilt than we have about the sort of strong prejudice that inspires hate crimes. Likewise, college students would not be a good population for a researcher interested in fundamentalist religious beliefs, or the joy, pain, and guilt that come with assuming a responsible adult position in society.

For other variables, college students may be a poor choice because there is so little variability among them: most college students are pretty high in self-esteem and pretty low in depression, for example, and show a highly restricted range on many other psychological variables that might interest us. A median split on a college student sample does not *really* yield high and low self-esteem groups, however the researcher labels them. Usually, the comparison is actually between a high self-esteem group and a moderately high self-esteem group.

The main reason we overuse introductory psychology students is convenience, a reason which is scientifically unsound. But although it is extremely difficult and expensive to use a truly representative sample, it is relatively easy to find alternative samples that lack the drawbacks of college students. Researchers have recruited participants in airport waiting areas, departments of motor vehicles, courthouses, malls, and science museums. Especially in contexts where they are just waiting, people are usually quite willing to participate. If the study can be administered by telephone, community members can be used instead of college students. Of course,

none of these are 'representative samples' of anything except themselves (for example, airline passengers flying out of Detroit), but they are likely to be *more* representative of the general population than college students are, and to be relatively free of the particular problems with college students (politically correct attitudes, lack of serious responsibility, bright future, and many more).

Data analysis

After formulating the research question; thinking of alternative explanations; designing the study; pilot testing the procedures and materials; thinking about reliability, internal validity, and external validity; and selecting an optimal sample, you proceed with data collection. Then comes the stage of analyzing the data and reporting the results. There are excellent books and chapters on data analysis, so we need not reiterate those techniques here (e.g., Cohen and Cohen, 1984; Judd, 2000; Maxwell and Delaney, 1999; McClelland, 2000). Instead of reviewing specific procedures in basic data analysis, we provide a few prescriptions for reporting results.

First, report descriptive statistics. The results of a study are not just a p-value. The most important purpose of data analysis is description. Simple summary scores such as measures of central tendency, measures of variability, measures of association, and plots are what should be highlighted in a results section. If a complicated statistical model is used, the parameters of that model should be emphasized and interpreted. Results sections should emphasize results, not statistical tests (the section is not called 'Statistical Tests'). Sentences should begin with the results themselves – 'Attitudes in the prime condition, $M = 5.2$, $sd = 1.1$, were more favorable than attitudes in the control condition, $M = 4.4$, $sd = 1.3$, $t(130) = 3.82$, $p < 0.05$' – rather than with the statistical test (for example, 'A two-sample t-test reveals that mean scores in the two conditions differed, $p < 0.05$'). Use the test statistic (the t, the F, the chi-square) and its corresponding p-value as punctuation marks at the end of the sentence, giving the conventional 'stamp of approval' on the pattern you observed.

Second, be aware of the statistical assumptions you make when conducting a test and check that your data are consistent with those assumptions. All statistical tests invoke a model that makes assumptions. Social psychologists appear to ignore this fact and act as though their hypotheses are tested in some absolute Platonic sense. A significant two-sample t-test does not show that one mean differs from another; instead, it provides a criterion by which to compare the means of two distributions under the assumption of equal variances, independence, and normality, leading to a particularly defined type I error rate. In other words, the researcher never tests a research hypothesis in isolation, but tests the *conjunction* of the research hypothesis and the set of assumptions required by the statistical test. A test may fail to reach statistical significance not because the research hypothesis failed (or there was not sufficient power), but because the assumptions were violated. For a discussion of how to check statistical assumptions, see McClelland (2000). Inform the reader that you checked the statistical assumptions and explain how you dealt with any violations.

Third, discuss a result in a manner consistent with the way you modeled it. An illustration of the violation of this prescription is seen in social psychologists' typical discussion of the Pearson correlation coefficient. Their usual language conveys an ordinal relation, in as 'the correlation shows that as anxiety increases, so does susceptibility to context effects'. As the reader knows from introductory statistics, the actual model underlying the correlation is a straight line (linear) relation between two variables. Therefore, the Pearson correlation assesses the degree of fit (defined in a particular way) between one variable and a linear transformation of the other variable (for example, 'The high correlation supports the model that anxiety and depression are linearly related'). If an ordinal relation is what the researcher wants to test, a different measure of association, one that measures the monotonic relation between two variables, should be used (e.g., Gonzalez and Nelson, 1996). It is possible for a Pearson correlation to be 0, and yet for the relation between those two variables to be systematic (that is, a Pearson correlation of 0 does not imply independence).

Fourth, do not describe an effect size as a measure of the underlying relation between constructs. The effect size is a normalized descriptive statistic. For example, the difference between two means is a descriptive statistic. The effect size measure normalizes that difference by dividing it by the standard deviation. The term 'effect size' tends to convey more than the computation implies. We have seen researchers discuss effect sizes in a manner that implies a deep, fundamental relation. For example, in an experiment examining the effects of reward on performance, a researcher can easily fall into the trap of claiming to demonstrate the 'effect size of reward on performance'. This language, which is at the level of constructs, suggests that the effect size has uncovered some underlying constant – reward influences performance (two abstract constructs) with an effect size of 0.2. Indeed, the use of meta-analysis connotes that multiple studies each provide estimates of this 'effect size' and that one can average over such studies to arrive at an even better estimate of effect size. In the physical sciences, there are examples of underlying constants that are invariant and can be estimated (for example, Planck's constant and the speed of light).

Are there such constants in social psychology? We doubt it. So do not fall into the trap of reading more into an effect size than is warranted by the ingredients – the descriptive statistics – that created it.

Data analysis should stay as close to the data and as close to the research hypothesis as possible. Present data and test the hypotheses that you have made (that is, if you made an ordinal prediction, use a test designed for ordinal hypotheses). Students frequently ask us to evaluate the 'proposed analyses' section of their dissertation proposal to check whether they will be 'analyzing the data correctly'. Such an evaluation is impossible for us to make out of context – we need to see the introduction, the hypotheses, the materials, and the procedure before we can make an evaluation of the 'correctness' of the analysis section. For us, a data analysis plan is correct if it addresses the research question being asked and is consistent with the research design. All too often, researchers focus on only one of the two (for example, my design is within-subjects so I need to run a repeated-measures ANOVA).

The great contribution of social psychology has been to illuminate the ways in which people's beliefs, values, emotions, and behaviors are affected by their social context. Statistical tests, on the other hand, are designed to be relatively context-free, widely applicable, and sensitive only to crude psychological differences (is the variable one of frequency in a population or degree within individuals? is it dichotomous or continuous?) or to peculiarities in the underlying distribution of variables in a sample (for example, various departures from normality). From a statistical point of view, a person's response – *any* response – is a data point, and the challenges of statistical analysis involve problematic data points, not problematic people.

Advances in statistical and computer methodology have benefitted our field enormously, but they have seriously skewed our recent writings on research methodology. We all know the old slogan of the computer analysts, 'Garbage in, garbage out', but, lately, we have said very little about what goes in. We seem to be impressed more with what we can now churn out of a fancy statistical package than in choosing our ingredients carefully.

Conclusion

The purpose of this chapter has been to rectify the dominance of analysis over design and procedure in methodological discussions; to remind ourselves and our students of the importance of the stages before the data are analyzed, indeed, even of the stages before the study is actually run. The most important phases of research are formulating a research question, creating a design that includes the comparisons required to answer it fairly and the comparisons required to test alternative possibilities, and devising a procedure that will represent that question in a way that is meaningful and involving for the people we are studying. Social psychology demands not just one talent, but many: cold logic, the free-ranging ability to see a problem from multiple points of view, and sympathetic human understanding. It demands them anew and in a different form, every time we plan a new study. Hackneyed research makes for dry social psychology. Intuition is not enough; we have to try out our methodological ideas on real people like the ones we plan to study before we can be sure that the ideas make sense. Often we have to revise them. Our questions are deep and difficult, and we have to sneak up on them through triangulation and intelligent compromise. Always we have to consider what else our results might mean, and design our next study to figure out which explanation is best. It is this combination of skills that has made our research a part of Western culture (Milgram's work on obedience, Asch's on conformity) and our technical terms a part of everyday discourse (dissonance, self-fulfilling prophecy), and it is the challenge of using all these skills together that makes our research so exciting.

Acknowledgments

We are grateful to Wendy Treynor and Alexandra Gross, who helped us to make the writing clearer, and to Barbara Zezulka Brown, who instantly incorporated our revisions, and made it possible to come close to meeting the deadline.

References

Aronson, E. (1969) 'A Theory of Cognitive Dissonance: A Current Perspective', in L. Berkowitz (ed.), *Advances in Experimental Social Psychology* (vol. 4). New York: Academic Press. pp. 1–34.

Aronson, E. (2002) 'Drifting My Own Way: Following My Nose and My Heart', in R. Sternberg (ed.), *Psychologists Defying the Crowd: Eminent Psychologists Describe How They Battled the Establishment and Won.* Washington, DC: APA Books. pp. 2–31.

Aronson, E., Ellsworth, P.C., Carlsmith, J.M., and Gonzales, M.H. (1990) *Methods of Research in Social Psychology*, 2nd edn. New York: McGraw-Hill.

Bakeman, R. (2000) 'Behavioral Observation and Coding', in H.T. Reis and C.M. Judd (eds), *Handbook of Research Methods in Social and Personality Psychology.* Cambridge: Cambridge University Press. pp. 138–59.

Bartholomew, K., Henderson, A.J.Z., and Marcia, J.E. (2000) 'Coding Semistructured Interviews in Social Psychological Research', in H.T. Reis and C.M. Judd (eds), *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press. pp. 286–312.

Bem, D.J. (1967) 'Self-Perception: An Alternative Explanation of Cognitive Dissonance Phenomena', *Psychological Review*, 74: 183–200.

Berkowitz, L. (1993) 'Aggression: Its Causes, Consequences, and Control', New York: McGraw-Hill.

Breckler, S.J. (1994) 'A Comparison of Numerical Indices for Measuring Attitude Ambivalence', *Educational and Psychological Measurement*, 54: 350–65.

Brewer, M. (2000) 'Research Design and Issues of Validity', in H.T. Reis and C.M. Judd (eds), *Handbook of Research Methods in Social and Personality Psychology*, Cambridge: Cambridge University Press. pp. 3–16.

Brinberg, D. and McGrath, J. (1985) *Validity in the Research Process*. Beverly Hills, CA: Sage.

Brown, J.L. and Steele, C.M. (2001) 'Performance Expectations Are Not a Necessary Mediator of Stereotype Threat in African American Verbal Test Performance.' Unpublished manuscript, Stanford University.

Campbell, D.T. (1957) 'Factors Relevant to the Validity of Experiments in Social Settings', *Psychological Bulletin*, 54: 297–312.

Campbell, D.T. and Fiske, D.W. (1959) 'Convergent and Discriminant Validation by the Multitrait–Multimethod Matrix', *Psychological Bulletin*, 56: 81–105.

Cohen, J. and Cohen, P. (1984) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 2nd edn. Mahwah, NJ: Lawrence Erlbaum.

Cook, T.D. and Campbell, D.T. (1979) *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston, MA: Houghton Mifflin.

Ellsworth, P.C. (1977) 'From Abstract Ideas to Concrete Instances: Some Guidelines for Choosing Natural Research Settings', *American Psychologist*, 32: 604–15.

Ellsworth, P.C. and Carlsmith, J.M. (1968) Effects of Eye Contact and Verbal Content on Affective Response to a Dyadic Interaction', *Journal of Personality and Social Psychology*, 10: 15–20.

Ellsworth, P.C. and Gonzalez, R. (2001) '"*The Handbook of Research Methods in Social and Personality Psychology*": A Tool for Serious Researchers', *Psychological Science*, 12: 266–8.

Festinger, L. (1957) *A Theory of Cognitive Dissonance*. Stanford, CA: Stanford University Press.

Gonzalez, R. and Nelson, T. (1996) 'Measuring Ordinal Association in Situations That Contain Tied Scores', *Psychological Bulletin*, 119: 159–65.

Greenwald, A. (1976). 'Within-Subjects Designs: To Use or Not To Use?', *Psychological Bulletin*, 83: 314–20.

Greenwald, A. (in press) 'The Resting Parrot, the Dessert Stomach, and Other Perfectly Defensible Theories', in J.T. Jost, M.R. Banaji, and D. Prentice (eds), *The Yin and Yang of Progress in Social Psychology*. Washington, DC: APA.

Griffin, D. and Ross, L. (1991) 'Subject Construal, Social Inference, and Human Misunderstanding', in M. Zanna (ed.), *Advances in Experimental Social Psychology* (vol. 24). San Diego, CA: Academic Press. pp. 319–59.

Hammond, J.S., Keeney, R.L., and Raiffa, H. (1999) *Smart Choices: A Practical Guide to Making Better Decisions*. Boston, MA: Harvard Business School Press.

Hastie, R. (2001) 'Problems for Judgment and Decision Making', *Annual Review of Psychology*, 52: 653–83.

Hilbert, D. (1900) 'Mathematische Probleme', *Goettinger Nachrichten*, 24: 253–97. (M.W. Newson [trans.] [1902] 'Mathematical Problems', *Bulletin of the American Mathematical Society*, 8: 437–79.)

Horner, M. (1972) 'Toward an Understanding of Achievement-Related Conflicts in Women', *Journal of Social Issues*, 28: 157–75.

Hsee, C., Bloundt, S., Loewenstein, G., and Bazerman, M. (1999) 'Preference Reversals Between Joint and Separate Evaluations of Options: A Review and Theoretical Analysis', *Psychological Bulletin*, 125: 576–90.

John, O.R. and Benet-Martinez, V. (2000) 'Measurement: Reliability, Construct Validation and Scale Construction', in H.T. Reis and C.M. Judd (eds), *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press. pp. 339–69.

Jones, E.E. and Nisbett, R.E. (1972) 'The Actor and the Observer: Divergent Perceptions of the Causes of Behavior', in E.E. Jones, D. Kanouse, H.H. Kelley, R.E. Nisbett, S. Valins, and B. Weiner (eds), *Attribution: Perceiving the Causes of Behavior*. Morristown, NJ: General Learning Press. pp. 79–94.

Judd, C.M. (2000) 'Everyday Data Analysis in Social Psychology: Comparisons of Linear Models', in H.T. Reis and C.M. Judd (eds), *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press. pp. 370–92.

Judd, C.M. and McClelland, G.H. (1998) 'Measurement', in D. Gilbert, S. Fiske, and G. Lindzey (eds), *Handbook of Social Psychology*, 4th edn. New York: McGraw-Hill. pp. 180–232.

Kahneman, D. and Tversky, A. (1982) 'The Simulation Heuristic', in D. Kahneman, P. Slovic, and A. Tversky (eds), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press. pp. 201–8.

Kahneman, D., Diener, E., and Schwarz, N. (eds) (1999) *Well-Being: The Foundations of Hedonic Psychology*. New York: Russell Sage Foundation.

Kahneman, D., Slovic, P., and Tversky, A. (eds) (1982) *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Keltner, D., Ellsworth, P.C., and Edwards, K. (1993) 'Beyond Simple Pessimism: Effects of Sadness and Anger on Social Perception', *Journal of Personality and Social Psychology*, 64: 740–52.

Kopans, D.B. (1994) 'Screening for Breast Cancer and Mortality Reduction Among Women 40–49 Years of Age', *Cancer*, 74 (Suppl.): 311–22.

Krauss, R.M., Chen, Y., and Chawla, P. (1996) 'Nonverbal Behavior and Nonverbal Communication: What Do Conversational Hand Gestures Tell Us?', in M. Zanna (ed.), *Advances in Experimental Social Psychology*. San Diego, CA: Academic Press. pp. 389–450.

Krauss, R. and Chin, C. (1998) 'Language and Social Behavior', in D.T. Gilbert, S.T. Fiske, and G. Lindzey (eds), *The Handbook of Social Psychology*, 4th edn. Boston, MA: McGraw-Hill. pp. 41–88.

Krueger, R. and Casey, M.A. (2000) *Focus Groups: A Practical Guide for Applied Research*. Thousand Oaks, CA: Sage.

Kruglanski, A. (2001) 'That "Vision Thing": The State of Theory in Social and Personality Psychology at the Edge of the New Millennium', *Journal of Personality and Social Psychology*, 80: 871–5.

Latané, B. and Darley, J.M. (1970) *The Unresponsive Bystander: Why Doesn't He Help?* New York: Appleton-Century-Crofts.

Lazarus, R. (1982) 'Thoughts on the Relationship Between Emotion and Cognition', *American Psychologist*, 37: 1019–24.

Loftus, E.F. (1979). *Eyewitness Testimony*. Cambridge, MA: Harvard University Press.

Maxwell, S.E. and Delaney, H.D. (1999) *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Mahwah, NJ: Lawrence Erlbaum.

McClelland, G.H. (2000) 'Nasty Data: Unruly, Ill-Mannered Observations Can Ruin Your Analysis', in H.T. Reis and C.M. Judd (eds), *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press. pp. 393–411.

McGuire, W.J. (1973) 'The Yin and Yang of Progress in Social Psychology: Seven Koan', *Journal of Personality and Social Psychology*, 26: 446–56.

McGuire, W.J. (1997) 'Creative Hypothesis Generating in Psychology: Some Useful Heuristics', *Annual Review of Psychology*, 48: 1–30.

McGuire, W.J. (1999) *Constructing Social Psychology: Creative and Critical Processes*. Cambridge: Cambridge University Press.

Milgram, S. (1974) *Obedience to Authority*. New York: Harper and Row.

Miller, N. and Pedersen, W.C. (1999) 'Assessing Process Distinctiveness', *Psychological Inquiry*, 10: 150–5.

Nisbett, R. and Ross, L. (1980) *Human Inference: Strategies and Shortcomings of Social Judgment*. Englewood Cliffs, NJ: Prentice-Hall.

Oyserman, D., Coon, H.M., and Kemmelmeier, M. (2002) 'Rethinking Individualism and Collectivism: Evaluation of Theoretical Assumptions and Meta-Analyses. *Psychological Bulletin*, 128: 3–72.

Platt, J.R. (1964) 'Strong Inference', *Science*, 146: 347–53.

Priester, J. and Petty, R. (1996) 'The Gradual Threshold Model of Ambivalence: Relating the Positive and Negative Bases of Attitudes to Subjective Ambivalence', *Journal of Personality and Social Psychology*, 71: 431–49.

Reis, H.T. and Gable, S.T. (2000) 'Event Sampling and Other Methods for Studying Everyday Experience', in H.T. Reis and C.M. Judd (eds), *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press. pp. 190–222.

Rosenbaum, P. and Rubin, D. (1984) 'Reducing Bias in Observational Studies Using the Subclassification on the Propensity Score', *Journal of the American Statistical Association*, 79: 516–24.

Rosenberg, M.J. (1969), 'The Conditions and Consequences of Evaluation Apprehension', in R. Rosenthal and R. Rosnow (eds), *Artifact in Behavioral Research*. NY: Academic Press, pp. 279–349.

Rosenthal, R. (1969) 'Interpersonal Expectations: Effects of the Experimenter's Hypothesis', in R. Rosenthal and R. Rosnow (eds), *Artifact in Behavioral Reserach*. New York: Academic Press. pp. 181–277.

Rosenthal, R. and Fode, K.L. (1963) 'The Effect of Experimenter Bias on the Performance of the Albino Rat', *Behavioral Science*, 8: 183–9.

Rosenthal, R. and Rubin, D.B. (1978) 'Interpersonal Expectancy Effects: The First 345 Studies', *Behavioral and Brain Sciences*, 3: 148–57.

Ross, L. (1977) 'The Intuitive Psychologist and His Shortcomings', in L. Berkowitz (ed.), *Advances in Experimental Social Psychology* (vol. 10). New York: Academic Press. pp. 173–220.

Ross, L., Greene, D., and House, P. (1977) 'The False Consensus Effect: An Egocentric Bias in Social Perception and Attribution Process', *Journal of Experimental Social Psychology*, 13: 279–301.

Schwarz, N., Groves, R.M., and Schuman, H. (1998) 'Survey Methods', in D.T. Gilbert, S.T. Fiske, and G. Lindzey (eds), *The Handbook of Social Psychology*, 4th edn. Boston, MA: McGraw-Hill, pp. 143–79.

Sears, D.O. (1986) 'College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology's View of Human Nature', *Journal of Personality and Social Psychology*, 51: 515–30.

Smith, E.R. (2000) 'Research Design', in H.T. Reis and C.M. Judd (eds), *Handbook of Research Methods in Social and Personality Psychology*, Cambridge: Cambridge University Press. pp. 17–39.

Sommers, S. and Ellsworth, P.C. (2000) 'Race in the Courtroom: Perceptions of Guilt and Dispositional Attribution', *Personality and Social Psychology Bulletin*, 26: 1367–79.

Strauss, A.L. and Corbin, J.M. (1998) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 2nd edn. Thousand Oaks, CA: Sage.

Steele, C.M. and Lin, T.J. (1983) 'Dissonance Processes and Self-Affirmation', *Journal of Personality and Social Psychology*, 45: 5–19.

Taylor, S.E. (1983) 'Adjustment to Threatening Events: A Theory of Cognitive Adaptation', *American Psychologist*, 38: 1161–73.

Thompson, M., Zanna, M., and Griffin, D. (1995) 'Let's Not Be Indifferent About (Attitudinal) Ambivalence', in R. Petty and J. Krosnick (eds), *Attitude Strength:*

Antecedents and Consequences. Hillsdale, Lawrence Erlbaum. pp. 361–86.

Tiedens, L.Z., Ellsworth, P.C., and Mesquita, B. (2000) 'Sentimental Stereotypes: Emotional Expectancies for High and Low Status Group Members', *Personality and Social Psychology Bulletin*, 26: 560–74.

Tourangeau, R. and Rasinski, K.A. (1988) 'Cognitive Processes Underlying Context Effects in Attitude Measurement', *Psychological Bulletin*, 103: 299–314.

Triplett, N. (1897) 'The Dynamogenic Factors in Pacemaking and Competition', *American Journal of Psychology*, 9: 507–33.

Visser, P.S., Krosnick, J.A., and Lavrakas, P.J. (2000) 'Survey Research', in H.T. Reis and C.M. Judd (eds), *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press. pp. 223–52.

Wentland, E.J. (1993) *Survey Responses: An Evaluation of Their Validity*. New York: Academic Press.

West, S.G., Biesanz, J.C., and Pitts, S.C. (2000) 'Causal Inference and Generalization in Field Settings: Experimental and Quasi-Experimental Designs', in H.T. Reis and C.M. Judd (eds), *Handbook of Research Methods in Social and Personality Psychology*. Cambridge: Cambridge University Press. pp. 40–84.

Zajonc, R.B. (1980) 'Feeling and Thinking: Preferences Need No Inferences', *American Psychologist*, 35: 151–75.