Editor's Introduction

W. Paul Vogt

The field of research methods in the social sciences is richly endowed with excellent texts and reference works. That makes it relatively easy for a researcher to look up *how* to employ a particular method – surveys or interviews or regression analysis or grounded theory, and so on. We are less richly endowed with works that help us decide *which* methods to use, be they cluster sampling, cognitive interviews, multi-level modeling, participant observation, and so on. *Which* method to use is arguably a more important question than *how to* use the method. "*Which method?*" is, at least, a necessary prior question. One cannot look up how to do something until one has decided what that something is.

Planning a research project requires the researcher to make many choices. These choices fall into four broad categories: (1) *design* or the methods for collecting the evidence, such as surveys, experiments, interviews, observations; (2) *sampling* or the methods of selecting cases to study; (3) *coding and measurement* or the methods of classifying and recording the evidence; and (4) *analysis* or the methods for interpreting the evidence.

Design, sampling, measurement, and analysis are closely related. One cannot discuss any one of them fully without discussing the others. But it is hard to think about everything simultaneously, and the volumes in this collection on 'Selecting Research Methods' surely cannot cover them simultaneously. While any ordering of these connected elements is somewhat arbitrary, there is a natural sequence, illustrated in Figure 1. This sequence of methodological choices will be used to order the content of these four volumes.

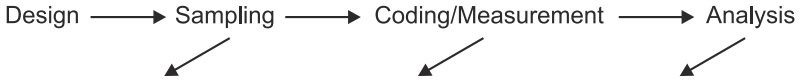

Figure 1: Logical sequence of methods choices

The logic of the order in Figure 1 is fairly clear. A design is a plan for collecting evidence. Once one determines the design, then one turns to sampling, which is the process of selecting sources of evidence. When the design and sources of evidence (cases) are determined, one needs to address coding and measurement, which are methods for handling evidence. Finally, after having determined what to collect from whom and how to record it, analysis, or methods for interpreting evidence, becomes central. It is generally most effective to proceed according to

this sequence, that is, to select one's design before making decisions about sampling, coding/measurement, and analysis. However, the feedback arrows in Figure 1 indicate that choices at each step in the sequence can have an influence in both directions: not only ahead to the next link but also back to a reconsideration of the previous ones.

To help researchers with the choices they must make, *Selecting Research Methods* provides a total of 79 previously published articles addressing the themes of design, sampling, coding, and analysis. The articles have been selected from a wide range of sources (28 different journals) in the social and behavioral sciences. The articles discuss a broad array of methods; they draw from all major research traditions and from all major disciplines in social research – sociology, political science, psychology, economics, anthropology, as well as applied fields such as education and demography. All the articles are recent and represent up-to-date thinking about methodological issues. This introductory essay is a review of the literature on the subject of selecting research methods. Unlike most reviews of the literature, however, this one concludes by presenting the reader with the studies on which the review was based: the 79 chapters of *Selecting Research Methods*.

Volume 1: Selecting Designs to Gather Evidence

A research design is a plan for collecting evidence that will be used to answer a research question. Selecting the design is the most important choice in planning a research project, because it determines all subsequent choices of methods. The design should be selected principally on the basis of how well it addresses the research question and enables the researcher to resolve a research problem.

A good research *design* is justifiable in terms of the research question. A good research *question* is researchable, and it is substantively important. Beyond those two obvious criteria (researchable and important), the issue of where we get good research questions becomes more intricate, perhaps even mysterious, since creativity is involved. Problems, which the research questions are meant to resolve, typically come from the external world, or from a researcher's personal interests, or from a discipline-based theory. Of the three, the last is by far the most prestigious, especially among discipline-based basic science researchers, but practical problems more often motivate applied researchers, and personal interests surely motivate us all. Regardless of where research questions come from, design choices will be influenced by whether or not one's research question aims to discover causal relations, to understand how causal influences work, to generalize from a small group to a larger one, to study change over time, or to understand a small number of cases in great depth.

The articles included in Volume 1 contain unusually cogent explanations of how designs can be effective for the envisioned research tasks, that is, how they can be appropriate to answering one's research questions. The specific topics addressed are: the theoretical underpinnings of design choice; qualitative and quantitative approaches to design; designs to collect qualitative data; varieties of experimental research design; and design models for quantitative data.

Philosophical foundations of design choice

We begin with the broad theoretical and philosophical underpinnings of design choice. Researchers in the social and behavioral sciences are, and have long been, divided on the question of how important understanding epistemological issues is to conducting intelligent, self-aware research. Epistemology is, briefly, the study of knowledge: what it is and how we obtain it. Major epistemological positions include realism, idealism, positivism, pragmatism, and constructivism. Some researchers say that epistemological principles are the foundation of all serious research; others conclude that epistemological controversies are a distraction from the real work of scientists. Researchers probably ought to make an effort to decide where they stand on such issues. This is an important decision if for no other reason than because one is expected to have a reasoned position about epistemological matters, even if it is to contend that they are unimportant. More substantively, epistemological sophistication is important because interpreting the results of research entails addressing uncertainty about knowledge claims. Uncertainty is inevitable when knowledge claims are based on evidence rather than faith. One's epistemological position ultimately describes how one understands and copes with uncertainly.

Two chapters help guide the reader through what is an old and complex theoretical thicket and one that is continuously sprouting new tendrils. Each, though it argues for distinct conclusions, discusses the range of other positions in a way that provides a good overview. These two chapters are especially important because the epistemological discussions in them are tied to practical questions of conducting research. Specifically, Siegel in Chapter 1 reviews what is meant by 'epistemological diversity' in education research including differences in cultural beliefs, methodological techniques, and research questions. He says that while important, these are not, properly speaking, *epistemological* differences. Arguing vigorously against epistemological skepticism and relativism, Siegel contends that the scope of diversity that is truly epistemological is rather narrower and less important than is usually supposed.

Traditions of inductivism and deductivism in sociology and political science, particularly as they relate to drawing conclusions about causation, are addressed by Gorski in Chapter 2. The best known model for thinking about knowledge claims in the social sciences is probably Karl Popper's 'falsificationism,' which basically contends that, while we cannot prove a theory, we can disprove one. Science makes progress by *conjectures* (thinking up theories) and *refutations* (testing theories). A good theory is one that many researchers have tried unsuccessfully to refute. (Readers may notice parallels to null hypothesis testing.) Gorski argues that this model is inadequate for sociology and political science; he then proposes an alternative, which he calls a constructive realist model.

Quantitative and Qualitative Approaches to Design

The next question sure to confront the practicing researcher making design choices is whether one's approach is quantitative or qualitative. As with the questions surrounding epistemology, researchers preferring to ignore this issue will often find

that other researchers are not inclined to let them do so. I believe that, properly understood, the terms quantitative and qualitative refer to ways of coding data, not to types of research design. This belief is easy to support empirically. All major research designs – interviews, surveys, experiments, and observations – can be and in fact have been used to collect either qualitative or qualitative data, or both. For instance, observational researchers not infrequently collect quantitative data, such as the number of interactions of particular types, and survey researchers often collect qualitative data by asking and analyzing open-ended questions.

While I think that the most important work involving the distinction between qualitative and quantitative should come at the third and fourth stages of the research sequence (coding and analysis), not at the first (design) stage, this may be a minority view. It is certain that practicing researchers will often be asked to label themselves as either a Quant or a Qual at the design stage. Those who resist these labels may call themselves mixed-method researchers. Note, however, that what is usually being mixed is quant and qual. Thus, intentionally or not, the mixed methods position ratifies the distinction between quant and qual. Without the division between quant and qual, there would be nothing to mix.

Several aspects of the qualitative-quantitative debate are examined in the next three chapters. Mahoney and Goertz (Chapter 3) argue that the difference between the two approaches to research is fundamental. Ercikan & Roth (Chapter 4) claim that the polarization is an inaccurate description of the nature of knowledge and thus impedes research progress. While my sentiments lie with Ercikan & Roth, who dispute the validity of the distinction, Mahoney & Goertz, are very persuasive about the differences between qual and quant, which they describe in rich detail. They make an excellent argument for the importance, especially in discussions of causation, of the distinction between what is sometimes called case-based (qualitative) research and variable-based (quantitative) research. Ercikan & Roth see the differences between the two research traditions as a matter of degree, while Mahoney & Goertz see them as distinct types: in measurement terms, this is a difference between continuous and categorical coding. But both pairs of authors suggest ways that the two, whether distinct types or degrees on a continuum, can work together.

Very practical discussions of quant and qual working together are illustrated in Chapters 5 and 6. In a study of the influence of religion on childbearing practices in Nepal, Pearce (Chapter 5) combines methods in a kind of dialogue between survey research and ethnographic interviews. The survey results are used to identify ethnographic informants for in-depth interviews, and information from these interviews is then used to recode some of the survey responses.

The problems of combining quantitative and qualitative approaches in large-scale evaluation research are reviewed by Plewis & Mason in Chapter 6. Evaluation researchers, motivated by pragmatic concerns, have generally been less likely to be methodological purists and, instead, to look for what works, whether quantitative or qualitative. This is the theme of Plewis & Mason's chapter. It also nicely illustrates some of the special travails of evaluation research, an area of study where researchers often find themselves 'cooking in someone else's kitchen.' Evaluation researchers often have little if any control over variables, the way they are defined, and the way they have been delivered. Control is usually in the hands of those

who have designed and implemented the programs being evaluated. Combining qualitative and quantitative approaches can enable evaluation researchers to work around some of the difficulties that this lack of control generates.

Designs for Collecting Qualitative Data

While all research designs can be used to collect either qualitative or quantitative evidence, some designs are most closely associated with gathering qualitative data. The actual distinction referenced when people talk of quantitative and qualitative designs is not principally one of numbers versus words. Rather the distinction is between designs good for studying a few cases intensively, such as interviews, as compared to designs better at studying many cases less intensively, such as surveys. This tradeoff between breadth and depth is one researchers routinely need to make. The most important grounds for making the choices stem from the nature of the research problem. If one's research problem involves determining *whether* a relationship exists, breadth is often more useful, but if the problem is to explain *how* the relationship works, depth may be more important. For example, to establish that a causal connection exists, methods emphasizing breadth and the relation of variables may be needed, but to understand the how's and why's of a process, in-depth methods for understanding cases are required.

One feature all of the approaches in this section have in common is a tendency to focus on a smaller number of cases but to study them in more detail than would be possible if a larger number of cases were to be examined. Yet, the desire to have both in-depth knowledge and general knowledge remains a big theme, and all our authors in this section struggle with it. Five designs for studying small numbers of cases intensively and in depth are reviewed: fieldwork, participant observation, grounded theory, focus groups, and narrative research.

Economists do not often use fieldwork, but Udry in Chapter 7 shows how, for some research questions in economics, it is a crucial method. His research questions pertained to landholding in Ghana; they could not have been adequately addressed without intensive fieldwork that linked observational methods to survey data and statistical analyses. After having decided on intensive fieldwork, one often also has to decide about which approach or approaches to take. In the context of workplace ethnographies, the three most common methods have been interviews, participant observation, and non-participant observation. Tope and colleagues (Chapter 8) examine these three in terms of their 'information yield.' They argue that participant observation is usually the superior method, but there is no one best method for all circumstances. For example, for some research questions, such as those requiring comparative data from several organizations, participant observation can be impractical or inappropriate.

One popular method of intensive analysis, based on repeated in-depth interviews, is grounded theory. This is described by Harry and coauthors in Chapter 9. Their research question is how and why students who are members of racial and ethnic minorities are more likely to be assigned to special education classes. Broad quantitative studies have routinely shown *that* they are. Intensive naturalistic methods are required to see how and why. Harry and colleagues provide a good example and excellent summary of the main tenets of grounded theory. Grounded

theory is especially appropriate for investigating what respondents believe, but unless events are a direct consequence of beliefs, it may not be sufficient for identifying causal mechanisms behind events.

A refinement of individual interviews is focus groups, which is short for focused group interviews. A group of similar informants are interviewed together and discussion among them is encouraged. The idea is that a group interview will yield different kinds of information than will individual interviews. As Munday points out in Chapter 10, this technique is especially appropriate for studying topics in which group interaction plays an important role, such as the formation of collective identities. Munday reviews major debates among methodologists specializing in collective identity and, like other authors in this section, emphasizes the importance of studying processes, not only content – not only what interviewees believe but how they came to believe it. She demonstrates the effectiveness of the focus-group method for uncovering such in-depth knowledge.

Narrative analysis is another important method used to collect in-depth qualitative data. Pedriana, in Chapter 11, employs it to study a problem in historical sociology. Narrative analysis is usually applied to psychologically-based individual narratives, but as this chapter shows, it can also be applied to historical processes. Even more unusual for narrative analyses, Pedriana shows how this versatile method can be used in conjunction with rational choice theory, specifically in this case to understand the development of the Equal Employment Opportunity Commission in the United States. Narrative research has in common with the other methods discussed in this section – fieldwork, participant observation, grounded theory, and focus groups – the desire to pursue a subject in more depth than breadth or, to change the metaphor, through a microscope rather than a wide-angle lens. Experiments also often aim for detail, but in ways quite distinct from the methods just reviewed. In experiments the emphasis tends to be on variables rather than on cases.

Varieties of Experimental Research Design

In discussions of research design, experiments are often referred to as the 'gold standard.' Ironically, this label for research designs emerged only after the original gold standard, in monetary policy, was completely defunct. Be that as it may, experiments have certain undoubted advantages, especially in terms of internal validity. Internal validity can be briefly defined as studying what you intend to study. You know you are studying what you intend to study in an experiment because you *control* it. You manipulate the variables and assign subjects to treatments so that you can see the effects of treatments on subjects. In many circumstances, of course, such assignment and manipulation are impossible, or if possible, unethical. Political scientists studying the relation of the development of democratic institutions to war, economists studying the relation of changes of unemployment rates to the emergence of recessions, or sociologists studying the relation of racial attitudes to crime rates can hardly manipulate their variables. However, researchers from these fields have sometimes been able to find inventive ways to use experiments to study aspects such problems.

Political science has traditionally been a field where experiments have been uncommon and often not considered highly important. In education research, by contrast, experiments have generally been considered extremely important but have been almost equally rare. This contrast between the two disciplines sets the stage for the chapters in this section. Druckman and colleagues have written an exceptionally persuasive account (Chapter 12) of the growth of laboratory, field, and survey experiments in political science. Not only have these increased in number, the studies reporting the results of experiments have been highly influential. Druckman et al. conclude that, like sociology and economics, 'political science may not be an experimental discipline, but with creativity it can become a discipline whose contributions are deepened and strengthened by experimental research' (p.634).

Survey experiments have been one of the most popular research designs employing experimental methods especially in sociology and political science. Gaines and colleagues, who are among the pioneers of survey experiments in political science, provide a critical examination of the method in Chapter 13. Survey experiments are often fairly simple in design: investigators randomly assign members of the sample to different survey questions or question orders to test the effects of these. A comparison of the results is especially interesting when the respondents are a representative sample, for then the experiment combines the internal validity of an experiment with the external validity, or generalizability, of a survey. As with many experiments, so too with survey experiments, a serious concern is 'ecological validity': how closely does the experimental manipulation match something in the real world? If it does not, there is not much point in studying it. If it does, then 'survey respondents interrupt real life to undergo an experimental simulation of real life' (p. 13). This paradox suggests why the pure conditions of the laboratory can rarely be met in survey experiments.

My impression is that when researchers in the social sciences say we should adhere to the experimental gold standard, what they most often have in mind is the randomized field trial (RFT). This is especially the case for the study of social and educational programs. These are not simulations like most laboratory experiments and survey experiments, but are studies of actual social interventions. Probably the most extensive of such randomized field trials in the United States have been studies of the effects of welfare policies. Three decades of this research is reviewed in Chapter 14 by Moffitt, a social scientist long involved in it. He provides a realistic look at the pros, cons, and complications of the method. His basic conclusion is that the limitations of RFTs mean that they have to be supplemented by other, non-experimental, designs.

The external validity of experiments in economics research is addressed by Harrison and colleagues in Chapter 15. External validity refers to the ability to generalize results from the study to a broader arena. The specific question in this chapter is: does 'behavior in a laboratory setting provide a reliable indicator of behavior in a naturally occurring setting?' The answer is clear: it depends. It depends specifically on the kind of behavior and the quality of the simulation. The main point is that the question is an empirical question, one that can be answered by reference to data rather than methodological posturing. By comparing results from a genuine market at a coin-show and experiments with some of the individuals

participating in that market, the authors draw important conclusions about which experimental manipulations are effective as simulations of market behavior.

Design Models for Quantitative Data

The word 'model,' like the word 'theory,' means many things. In this context it refers to the way researchers imagine that the variables they are studying will be related. Are the variables likely to interact, that is, are they likely to have joint effects that go beyond their separate effects, such as the effect of income and age on political beliefs? Or, are the variables likely to be multi-level or 'nested,' that is, does one variable form a context for another, as the influence of a teacher might be contextually shaped by, or nested within, the influence of a school? If the answer to either of these questions is yes, and empirically it probably should be yes more often than not, the researcher must plan for interacting and multi-level nested variables. Planning to collect the right kinds of information on the relevant variables needs to be built into a study's design from the outset. Such planning is central to any design, but is doubly important for complicated models where initial missteps can be hard to correct at later stages of the research.

The importance of interactions among variables is discussed by Brambor and colleagues in Chapter 16. When variables interact their joint effect is multiplicative, not only additive. This means that effect of the two variables together is greater than the sum of their individual effects. Such multiplier effects are likely to be relatively common, and researchers do check for them *fairly* routinely in the social sciences. However, while including and interpreting interaction effects may be thought of as routine, it requires more care than is often shown. The authors of Chapter 16 systematically review the use of interaction models in three top political science journals and find that the models are often misused. Mistakes have led to doubtful interpretations and sometimes to conclusions that were quite simply wrong.

The statistical modeling of causal effects is discussed in the next two chapters, where the particular focus is 'value-added models,' as these are used to assess educational outcomes. Rubin and colleagues in Chapter 17 and Raudenbush in Chapter 18 talk about these models as they are used to estimate the effects of teachers and schools on student learning. The basic idea of the value-added model is simple: one compares gains in learning of students who have experienced an educational intervention with the gains in learning of students who have not. More frequently, one compares gains of students receiving the intervention with what those gains *would have been* had they not received it. As these two chapters make clear, value-added models, while ostensibly simple, have many conceptual, inferential, and statistical difficulties associated with them. Here again, we see that careful research design and planning of what evidence to collect about which variables is crucial – if faulty conclusions are to be avoided.

Volume 2: Methods to Sample, Recruit, and Assign Cases

A key decision facing researchers is *which* cases to study; also important is *how many* of them to study. The cases may be individuals, organizations, events, or any other

distinct units of analysis. Thus, while a good design is crucial, so too is using that design to study relevant types and numbers of cases. Which cases and how many are particularly important questions if the researcher wishes to generalize from the cases studied to a broader population. After deciding which cases to study, then come the tasks of persuading, recruiting, or gaining access to those cases. In addition to basic issues of sampling, such as probability and non-probability sampling, researchers also must decide how to recruit informants for interviews, select sites for participant observation, and assign subjects to control and experimental groups. In the coming chapters, advice will be found for finding and recruiting cases to study, determining sample size, bias when selecting cases, sampling from qualitative data archives, and the use of a range of new technologies developed for interviews and surveys.

Research ethics are also a recurring theme in the chapters of this volume. Questions of ethics are most effectively addressed at the intersection of design and sampling. Once the researcher has considered the basic methods of gathering the data and decided from whom they will be gathered, then ethical issues are concrete enough that they can be addressed. Ethical concerns are best discussed in terms of specific designs and sampling plans. One can hardly consider, for example, the rights of subjects until one knows who those subjects will be and what will be asked of them.

Finding and Recruiting Cases to Sample

Most discussions of sampling focus on deciding which units of analysis to select and how many of them to include. The emphasis tends to be on selecting a sample that is representative of a broader population. If one has an accurate list of a population, then drawing a representative sample from it is a matter of using well-known probability techniques. Probability samples are those in which each unit selected has a known probability of being selected. Very often, however, selecting is not enough. There is also the task of *persuading* those selected to participate. An even more difficult situation is studying populations for which there is no workable list, or sampling frame, from which to choose. This problem occurs much more frequently than one might expect, given textbook discussions of sampling. Researchers have had to be inventive when studying rare or hidden populations for which no list of members is readily available. Their efforts are illustrated in the chapters of this section.

A frank account of difficulties encountered is provided by Wells and colleagues in Chapter 19. Their research was on Australian workers eligible to retire who subsequently did so. Despite having impressive resources, the researchers could not employ a probability sample, mostly because of the absence of a sampling frame. With no list of the population, they tried recruiting through employers, but they found that employers had little incentive to participate. Unions were more cooperative, but union members hardly flocked to be surveyed. Workers were often worried about their anonymity even though researchers quite specifically were not breaching it. Organizations gave workers a phone number they could call if they were willing to participate. The researchers had no list of names, but this was not clear to respondents who were frequently suspicious. Some 34 *thousand* surveys

were distributed through organizations to obtain just over 5 *hundred* participants, a response rate of less than 2%. Perhaps the difficulties encountered in this study were unusually great, but the researchers' problems, and how they dealt with them, yield important lessons. The clearest lesson is that the costs (time, effort, money) of recruiting a sample can often be *much* higher than a researcher might estimate.

The absence of a sampling frame (list of the population) was also a problem confronted by Lee and colleagues in Chapter 20. Because they wished to study a rare type of a public organization, secondary schools in the United States divided into smaller subunits called 'schools-within-schools,' their issue was more one of *finding* than recruiting. Their goal was not to use a sample to generalize to a population. Rather their purpose was to study all members of the population they could find. When conducting such a search, how do you know when it no longer makes sense to keep looking? The answer is often referred to as reaching a point of 'saturation.' You stop when you stop learning new things. Studying an entire population rather than a sample is fairly common. The best-known example is probably meta-analysis. Here too researchers usually seek to include all instances of a population (of research reports) rather than a sample of it.

In experimental studies the focus is usually on *recruiting* subjects. The goal is to study the effects that different treatments have on subjects; the emphasis is more on the treatment than on the representativeness of the subjects receiving it. Experiments are strong in terms of internal validity, but not external validity or generalizability. Henrich and colleagues in Chapter 21 discuss experimental research in economics, which is often based on games that simulate market situations. One basic assumption about markets made in classical and neo-classical economics is that individual behavior is motivated entirely by self-interest. Experiments have shown that this assumption does not always hold and that many experimental subjects show a preference for fairness and reciprocity over personal gain. But these experiments, like many in the social sciences, have been conducted mostly with university students in developed countries. To broaden the findings Henrich et al. conducted the same experiments among 15 small-scale traditional cultures in nations as diverse as Indonesia, Paraguay, and Tanzania. They discovered much variety among these cultural groups. This variation allowed very powerful explanations of behaviors, explanations that would have been impossible had they not recruited from a wide range of societies.

Populations that are hard to contact, rare, and/or whose membership is unknown pose special problems for researchers. These are addressed in the next three chapters. Bloor in Chapter 22 examines the use of 'contact-recontact' methods to study drug use, problem drinking, and 'rooflessness' in Scotland and Wales. The basic idea of contact-recontact methods was developed in the study of animal and fish populations, where it is usually known as capture-recapture. For instance, to study the fish population of a lake, researchers capture, mark, and release a sample of fish. By doing so repeatedly and noting how many of subsequent captures had been previously captured, it is possible to estimate the size and composition of the lake's population of fish. Adapting this method to human populations has special problems, some technical and some ethical, all of which are reviewed in Bloor's chapter. His conclusion is that while the method has serious limitations, it is often the only feasible one for populations that are rare or hidden or both.

Farquharson in Chapter 23 uses a 'reputational snowball' technique to identify members of a network of policy makers and implementers in Australia. Snowball methods essentially involve asking initial contacts for further contacts. The key assumption of snowball sampling is that members of the target population know one another. In this case Farquharson asked a small group of policy makers to identify others that they believed to be influential. After five rounds of having those who were named in one round nominate 'influentials' for the next round, the researchers on Farquharson's team reached the point of 'saturation,' where few new individuals were being named. The reputational snowball allowed not only identification of members of the group but estimation of their influence. With this as with other interview-based methods, where anonymity respondents and informants is impossible, ethical issues, especially of confidentiality, are important. Because these individuals were public officials, concealing their identity was probably less necessary than it would be for populations whose identification could put them at risk.

Respondent-driven sampling (RDS), pioneered by Heckathorn, is probably the best known and definitely the most technically well-developed of the methods for studying hidden and/or rare populations. Respondent-driven sampling can be used to derive unbiased, probability-based estimates about the populations studied. This is true even though RDS is based on snowball sampling, which is in no sense a method of probability sampling. As an example of the method, visitors to a drug rehabilitation clinic who are willing to participate in a study are interviewed and paid for their time. They are given three coupons that they can give to acquaintances who might be eligible for the study. Interviewees recruited from among the coupon holders are also given coupons to distribute. This technique enables researchers to learn about the network; they use this knowledge to draw inferences about the population. In addition to the attractiveness of a method that uses non-probability methods to make unbiased estimates, respondent-driven sampling has proven to be comparatively easy and inexpensive to implement. In Chapter 24 Wejnert and Heckathorn provide details and show how the method can be applied to sampling an Internet-based network.

Sample Size and Nested Models

What methods can one use to determine adequate sample size? The researchers writing in the previous section determined sample size as they went along; they had to do so because the populations were unknown, which made it impossible to calculate sample size in advance. By contrast, this section focuses on calculating sample size before one starts recruiting. All else equal, the bigger the sample size the better, but all else is seldom equal. For example, if a sample of 500 is sufficient for all purposes in your survey, sampling 1,000 is a waste of your time and, more important, that of 500 respondents. Can one reach a point of diminishing returns where resources would be better spent on other aspects of the research than on adding more cases? If there is such a point, where is it?

These questions are too often handled by plugging guesses into an equation or computer program. Parker & Berman in Chapter 25 provide thoughtful alternatives. The dubious approach to sample size calculation has become more frequent as

journals and granting agencies increasingly demand power estimates. The authors suggest an alternative paradigm, and a method of calculation, that does not focus on null hypotheses, as power calculations do. Kelley & Rausch in Chapter 26 also contend that determining sample sizes by emphasizing the rejecting of false null hypotheses (i.e., power calculations) can be time ill-spent. Rather, researchers should focus on effect size measures and especially on the confidence intervals of those effect sizes. When using effect sizes, the emphasis is on estimating whether the treatment has an effect large enough to be important, either for science or for practice. The confidence interval provides the margins of error around that estimate. For example, if the estimated mean difference between those receiving and not receiving the treatment is 6 points with a margin of error of plus or minus 5, then the true difference in the population is likely to be somewhere between 1 and 11. That range may be too large to be meaningful. If so, one can narrow the range by increasing the sample size. Kelley & Rausch provide tables to consult to learn how much the sample size needs to be increased for a given improvement in the accuracy of the estimate. They also provide links to software that can be used with the freeware statistical package 'R' to determine sample size when the goal is the accuracy of the estimate.

Many research questions require multilevel, hierarchical, or nested-variable approaches (the three terms are used almost interchangeably), but calculating sample size in multilevel modeling is complicated by the fact that each level will have a different sample size. At the higher levels, the samples can be very small, sometimes too small for statistical analysis. For example, if students are nested in classrooms and classrooms are nested in schools, the number of schools in a study could be quite small even if the number of students is substantial. Smaller sample sizes open the possibility of biased estimates at the higher levels, which is one of the reasons that most multilevel studies have only two levels. Maas & Joop in Chapter 27 use simulation studies to address the question of what constitutes a sufficient sample size for accurate estimation and provide tables and formulae researchers may use for this purpose.

Hierarchical models are also Achen's focus in Chapter 28. In most such studies, level 1 is the point of interest. For example, one might be interested in student learning; the other levels, such as classrooms and schools, would be introduced primarily to put student learning in context. In political science, especially in comparative international studies, the focus is often on level 2. Then the goal is to combine statistical models at stage 1, where the Ns are usually large, with localized case knowledge at stage or level 2. For example, researchers may study thousands of voters at level 1 but only dozens of countries at level 2. Achen provides a simplified approach to such research that many readers are likely to find helpful. Lieberman in Chapter 29 also attempts to unite the levels in a nested or hierarchical research problem. Specifically he tries to link large-N analyses, which are studied using quantitative techniques, with small-N analyses, which are studied using qualitative techniques. Lieberman sees these techniques as distinct but complimentary rather than as opposed.

Bias When Selecting and Analyzing Extreme Cases

A generalization rarely applies to the data so well that there are no exceptions, and sometimes the exceptions to the rule are extreme. Extreme outliers and negative cases raise special problems for researchers. Negative cases are as important for research on qualitative data as extreme outliers are for research on quantitative data. Exceptions test the rule – or, in the original saying, *prove* it. And counterfactual questions are essential for causal inferences: what would happen to the birth rate if the education level increased, or what would the economy be like if oil sold for $10 per barrel? King & Zeng (Chapter 30) tackle the problem as it occurs in quantitative data, while Mahoney & Goertz (Chapter 31) do so when the data are categorical or qualitative.

To take the quantitative case first: What if we want to study the influence of national income on whether a nation is likely to be democratic or autocratic? If there are no democracies with national incomes as low as those of most autocracies, then there are no comparable data with which to judge the effect of income on type of government. In a simple case like that one the problem is easy to see, but it is much more intricate for multivariate research questions where controlling for the effects of several variables is required. Prior to King & Zeng's work, there had been no readily available method to examine the question of whether the data were sufficient to test one's research questions in cases of potential extreme outliers. The authors provide free software for resolving a technical design and sampling problem that had hitherto been too computationally complex to be addressed in most circumstances.

Similar problems arise when the variables (especially the dependent variables) are qualitative. For Mahoney & Goertz the issue is one of distinguishing between cases that are relevant to testing a hypothesis and those that are not. This is a version of the same problem with qualitatively coded variables as the one studied by King & Zeng for quantitatively coded variables. In both chapters the problem is cases that are so far outside the boundaries of the possible that they are irrelevant to the research question. One needs negative cases to test hypotheses, but only negative cases that are possible. The authors provide guidelines for choosing a set of negative cases to test causal generalizations in research using qualitative data.

Some researchers focus on extremes rather than try to work around them. The extreme-groups approach is a widespread but dubious practice. Preacher and coauthors examine it in Chapter 32. It involves reducing a continuum of individual differences to a simple binary difference, such as high and low, and then comparing these two. The authors review the advantages and disadvantages of using extreme groups. They find hardly any advantages, and they point out several drawbacks. Using extreme groups has become a routine practice, perhaps even a tradition, in many subfields, but it is one that should be rethought. It is clear that the burden of proof is on those who would use the extreme-groups approach, since it inflates standardized effect size measures, risks model misspecification, and reduces measurement reliability.

A similar problem occurs in studies of social and psychological interventions. These are often directed at extreme groups, since those with graver problems are often more likely to be selected to receive an intervention. As Larzelere and

colleagues demonstrate (Chapter 33) this type of extreme-groups approach can also lead to misinterpretations. Sometimes it is easy to see how misinterpretations can occur. The sicker a person is the more likely s/he is to be hospitalized. Hence there will be an association between hospitalization and increased death rates, but few would conclude that hospitalization *causes* most of the increase. Here the confounding variable is obvious. But in other cases, it might not be. For instance, many studies of parents' help with their children's homework have found that this intervention is associated with lower achievement. However, this surprising result probably arises because parents are more likely to intervene when children are doing poorly. Intervention bias is a very widespread problem in many fields in large part because random assignment to interventions would be difficult and often unethical. In such cases, research has to be based on statistical associations. When statistical associations are computed without controls for the intervention effects, serious errors can result. Faulty conclusions about the effectiveness of treatments can lead to the misallocation of resources and they can impede the development of new treatments. The authors conclude that 'the most important step toward minimizing selection bias is to recognize its pervasiveness and potential magnitude in intervention research' (p. 296). This is a crucial problem and much is at stake in public policy. Both extreme-groups bias and intervention selection bias arise from unwise sampling practices. The big difference between them is that extreme-groups bias is self inflicted, while intervention selection bias often cannot easily be controlled by researchers studying a social policy.

Using Archives of Qualitative Data

Researchers analyzing quantitative data often make extensive use of archives and perform 'secondary analyses' on the data sets they contain. Indeed, in many subfields in economics and sociology it is almost unheard of for researchers to collect their own data. This reliance on previously gathered evidence is less common for researchers interested in problems more typically addressed with qualitative data. Some of the reason for the difference is that qualitative archives are much less extensively developed. There is also probably some reluctance of qualitatively oriented researchers to use archived data. The use of qualitative data archives is discussed in the next two chapters. Fielding (Chapter 34), while reviewing obstacles to more extensive use of qualitative data archives, is generally optimistic about their potential. Parry & Mauthner (Chapter 35) see these data sources as fraught with problems, mostly of an ethical nature. My guess is that most current researchers using qualitative data will find the worries and concerns expressed by Parry & Mauthner to be more salient than the opportunities highlighted by Fielding. Be that as it may, the two chapters provide a nice point-counterpoint on an issue that is likely to grow in prominence as qualitative data archives become more available and researchers become more aware of their potential.

Technologies for Interviews and Surveys

Research methods in surveys and interviews have in part been driven by the availability of technology for finding respondents and for recording what they say.

The next set of chapters reviews a range of technical innovations and their potential uses, from the tape recorder, through the internet, to the global positioning system. Lee in Chapter 36 provides a fascinating history that traces the evolution of techniques for recording what interviewees say: from shorthand to on-line interviewing. The tape (or digital) recorder has become the standard means of quickly gathering large amounts of text (although transcription is a serious bottleneck). The text can then be analyzed using techniques that range from traditional impressionistic methods to computer-based data crunching (see chapters 60-63).

Random-digit dialing to find participants for telephone survey interviews has become common in recent decades, but there has been a sharp decline in the response rates using this technology. One method to increase the response rate, or stem the decline, is to send a letter or a postcard in advance of the attempted telephone contact. The Internet is potentially even cheaper than the telephone, and is certainly less expensive than letters alone. But two main concerns limit its use: first, response rates to Internet surveys are often even lower than to random-digit dialing; second, bias results from the unequal availability of the Internet. To test the validity of results with new technologies, researchers may compare responses to their survey to those of well-known surveys with very high response rates. One theme that has emerged in such research about new technologies is that they often seem to perform as well as the tried and true methods (see chapters 37 and 38), but it is sometimes not clear why or how. A method that regularly generates a lower response rate, for example, 'should', a priori, perform less well than one that generates a higher rate.

Any interview method introduces a possible bias due to interviewer effect. Respondents may answer questions differently depending on the interviewers' gender, accent, and so on. Web-based surveying eliminates this potential problem because it is self-administered. The same advantage is true of telephone-based approaches called interactive voice response (IVR) methods. These use the recorded voice of a survey interviewer. The IVR is one of the technologies Blumenthal discusses in Chapter 37. There are many limitations to IVR methods, for example, what if a ten-year-old child answers the phone and responds to an election survey? Yet, as with all of these technologies, the proof is in the testing. In the U.S. in 2004, IVR surveys performed as well as or better than other statewide surveys in predicting outcomes of elections.

Regardless of the technique employed, there will be differences in how well they are implemented. Differences in the quality of implementation may be equally or more important than the technique itself. One of the important features of Chapter 38 by Sanders et al. is that it reports the results of an experiment in which a very well-designed Internet survey was compared to a very well-designed in-person survey. Again, the results are encouraging for the new method: the 'in-person and Internet surveys yield remarkably similar results' (p. 279). While Sanders and colleagues are cautious about generalizing these results to other contexts, there is little doubt that in the 2005 British elections, the new methods of Internet surveying performed exceptionally well.

A new technology to improve on an old idea is discussed by Landry & Shen in Chapter 39. When one does not have a reliable sampling frame (list of the population members) from which to draw a sample, an alternative is to select a sample

of geographic areas. Originally the idea was to, for example, draw a grid over a map of a city and then take a random sample of the areas formed by the grid. One problem with this approach was that it was hard to identify units that were small and accurate enough to be useful. Global positioning system (GPS) technology has solved this problem since the GPS has the 'ability to identify small units with considerable precision almost anywhere on the planet' (p. 7). Area sampling methods are especially helpful when the populations are unknown and/or highly mobile. Does this method of GPS area sampling yield good results? It did in the test case (Chinese cities), and the authors conclude that the method will work well in other areas – as long as all respondents in the sampled geographic areas are surveyed. Landry & Shen conclude that using their method results in an equal probability sample; this is a remarkable achievement. GPS technology has a lot of promise, as much perhaps as the Internet, but so far it has been used much less.

Volume 3: Coding and Measurement

Selecting methods to record observations is the next step after choosing one's design and selecting one's sample. When the data are quantitative, deciding how to record observations is usually called *measurement*. When the data are qualitative, it is often called *coding*. The concepts and problems of assigning codes to phenomena and using those codes to analyze phenomena are quite similar whether the codes are names, ranks, or numbers. Validity and reliability are the most widely used generic labels for the conceptual issues and problems involved in coding and measurement. Researchers whose work does not involve numerical data sometimes use alternative terms for these concepts, such as *trustworthiness* and *dependability* to denote, respectively, validity and reliability. Techniques for assessing validity or trustworthiness and reliability or dependability differ depending on the data (names, ranks, or numbers), but the conceptual issues are broadly similar.

The basic theme of Volume 3 is how to study concepts in ways that are reliable and valid. There are several types of reliability and validity, and each has multiple meanings. This reflects the complexity of the issues involved. For example, the validity of a conclusion might be thought of as its correctness, but validity of a measurement is better thought of as its appropriateness. Reliability and validity are related but distinct. They are terms for aspects of the process of trying to investigate concepts. In measurement and coding, reliability refers to consistency while validity refers to relevance. Reliability and validity are linked in one fundamental way: an unreliable (completely inconsistent) measure, it is generally believed, cannot be valid. However, reliability alone does not guarantee validity; a consistent measure can be consistently wrong, inappropriate, or irrelevant. The ultimate goal in Volume 3 is to give readers guidelines to improve the relevance and consistency of their investigations of concepts.

Approaches to Reliability

We begin with an exchange between two scholars on the meanings of reliability and the implications for validity of various understandings of reliability. The authors

take different approaches: Mislevy's (Chapter 40) is based on quantitative measurement theory, while Moss (Chapter 41) builds on a hermeneutic approach. Despite their disagreements, the two chapters provide a gentle introduction to this conceptual thicket.

Unquestionably the most widely used measure of reliability is Cronbach's alpha. Cronbach's alpha is a measure of the internal consistency of the items in a scale. More technically, it is the squared correlation of the observed score on the items with the 'true score,' which is the average score persons would get were they measured an infinite number of times. The measure has been extremely popular with researchers, and the 1951 article in which Cronbach introduced it has been cited a huge number of times. But Cronbach in Chapter 42 concludes that his own measure has been overused. It has been used when it is inappropriate and when better alternatives exist. Researchers tend to calculate and report alpha and think that they have 'covered' reliability, but Cronbach concludes that the alpha coefficient needs to be seen 'within a much larger system of reliability analysis' (p. 416).

Some aspects of that larger system of reliability methods are reviewed by Dimitrov in Chapter 43. He discusses, in more or less depth, classical test theory, Cronbach's alpha, generalizability theory, and item response theory, but he is most especially concerned with reliability generalization (RG). This was introduced by Vacha-Hasse in the late 1990s and has been quite influential. Reliability generalization uses the tools of meta-analysis to compute presumably more reliable estimates of reliability. These meta-analytic estimates also allow researchers to control for variables that may influence an instrument's reliability. Dimitrov believes that the RG method is open to several sorts of error and that it is less useful than some of its enthusiasts have claimed. Still, it remains a choice in a field with a growing number of increasingly sophisticated options for estimating the reliability of measurements. Debates among proponents of various approaches matter because reliability is of fundamental importance. Accuracy of measurement is directly determined by measurement reliability. A totally unreliable measure cannot be accurate: the lower the reliability, the lower the accuracy.

A Rasch modeling approach to reliability is reviewed and recommended by Schumacker & Smith in Chapter 44. Originally developed by Georg Rasch in the 1950s and 60s, this family of models, which today includes item response theory (IRT), improves upon reliability estimates by using non-linear regression, specifically logit (or logistic) regression. This form of regression is necessary because raw scores on an instrument measuring a concept, such as mathematics ability, are not linear. Rasch and IRT methods have largely replaced traditional ones in standardized testing and will probably eventually do so for computing reliability in other realms as well.

We conclude this section with a substantive article rather than a methodological one. The authors demonstrate several methods by using them to address a specific problem. The reader can see these methods in action. Newmann and colleagues (Chapter 45) sought to define and measure instructional program coherence (IPC) and then to estimate how much IPC contributed to fostering student achievement, and finally to investigate how it does so. After a review of the literature they constructed a survey instrument to measure IPC in Chicago 222 elementary schools

by surveying some 5 thousand teachers. They used Rausch methods in these measurements. Then they related IPC to academic achievement. Academic achievement was measured using IRT methods, which allowed scores to be compared across years and over grade levels. Then they used multi-level modeling, specifically a 3-level hierarchical linear model (HLM) to find a strong positive relationship between improving IPC and improving student achievement scores. Finally, to inquire into *how* IPC fostered learning, they conducted intensive fieldwork in 11 field sites. I know of no other article that gives practical illustrations of how to employ as many different methods, changing method as necessary to be reliable and valid at the current stage of the project. Even if you have little interest in school program coherence, it is well worth studying this chapter for its insights into survey research, reliability measurement, multi-level models, and fieldwork. It is particularly important for illustrating how these can be integrated.

Validity in Coding and Measurement

Validity is more complicated than reliability. At minimum, the literature on validity is more intricate. For example, the dozens of widely discussed 'threats to validity' come in four categories, threats to: *external* validity or problems with generalizability, *construct* validity or poor links between concepts and measurements, *statistical conclusion* validity or inappropriate selection of techniques for analysis, and *internal* validity or errors that impede reasoning about cause and effect. Internal and external validity are mostly matters of design and sampling. Statistical conclusion validity has to do with analysis, which will be discussed in Volume 4. Here we are interested mainly in measurement and coding, which means that our focus is on construct validity. The basic question for construct validity is: Are you collecting evidence about what interests you? In other words, are you measuring what you intend to measure? The criterion '*intend to measure*,' in turn, requires the researcher to know what is intended clearly enough that judging the quality of coding and measurement is possible. That necessitates careful reasoning about concepts, not just technical work.

The proliferation of 'types' of validity is the context in which Adcock & Collier work in Chapter 46. In a review of the literature they found 37 different adjectives attached to the term validity. The spawning of terms, often rather obscure terms, to discuss validity is a good indicator of the complexity of the topic or, perhaps, of confusion in the field. Various taxonomies, dating back to the 1950s, attempting to sort out the complexity (most notably by the American Psychological Association) may have aggravated the problem. Bafflement and annoyance are common when students first encounter, for example, construct, content, concurrent, convergent, and criterion validity ('who thought this up!?'). Adcock & Collier provide a comprehensive and clear analysis of it all by providing a template for thinking through issues pertaining to measurement validity. Especially impressive is the fact that their template, though simpler than most, is equally applicable to work using either qualitative or quantitative data.

Finding methods to bridge the qualitative and quantitative traditions is also the focus of Shaffer & Serlin in Chapter 47. They are looking for ways that quant and qual can be combined so that both are valid ways to approach research questions.

They look for true *integration*. Most mixed-method studies involve juxtaposition more than integration – desegregation, one might say, rather than true integration. Shaffer & Serlin are especially interested in using quantitative measures and analyses in support of qualitative inferences. They attempt this through what they call intra-sample statistical analysis (ISSA). The key to understanding ISSA is realizing that many qualitative studies, although they are often built upon a comparatively small number of cases, often contain a very large number of observations. Rather than trying to use statistical methods to generalize to a broader population of cases, ISSA aims at a broader population of observations. Generalizations about observations can be used to support qualitative inferences about cases.

Two of the biggest problems with measurement validity in surveys are: first, writing questions that accurately measure complicated concepts, such as democracy or social tolerance; and, second, writing clear questions when respondents understand key terms differently. The second problem may be especially likely to occur in cross-national surveys. The nature of the problem is thoroughly examined, and an ingenious method for addressing it is discussed by King and colleagues in Chapter 48. The problem is called differential item functioning (DIF). DIF means that an item (survey or test question) functions differently for different groups. To assess and adjust for DIF, King et al. use vignettes. They give an example comparing levels of political efficacy. Political efficacy is the belief that one has some influence on or 'say in' what one's government does. Chinese respondents had lower levels of political efficacy than Mexican respondents. But the Chinese have lower thresholds for defining political efficacy. These lower thresholds lead Chinese respondents to *declare* that they have higher levels of efficacy than do respondents from Mexico. Without adjustments for the lower thresholds used by the Chinese respondents, researchers would wrongly conclude that the Chinese levels were higher than the Mexican, when in fact just the opposite was true. The method of using vignettes to determine thresholds is comparatively simple, and the authors provide free software for applying it. The method is widely used to conduct research in many nations today, for example in the World Health Survey.

Valid measurement is a key issue with vital statistics (e.g., birth and death rates). Measurement is especially difficult in less developed countries because these nations often do not have comprehensive systems for collecting the information. One method for obtaining data in the absence of such records is to estimate death rates by using surveys of siblings. Examining data from the World Health Survey, Gakidou & King in Chapter 49 show that the most common techniques employed in such surveys are subject to serious bias because they are based on an untenable assumption – that death rates do not vary by family (sibship) size. The authors introduce a new method to work around that bias. The method is very important because it improves the accuracy of measurement for nations that do not have complete systems for registering vital statistics, and in fact most nations in the world do not.

Cross-national study raises important issues of measurement validity in the study of values. These are examined by Ovadia in Chapter 50. How do individuals conceptualize values, and are the questions used to tap those conceptualizations validly structured? Specifically, do people think about values by ranking or rating? Survey questions that use ranking ask respondents to place a list of values in order

of importance. A rating system has respondents tell how important a value is to them (highly, somewhat, etc.) but without reference to other values. The choice of ranking or rating is less a measurement issue and more one of what the researcher assumes about how respondents understand values. Is the importance of values independent, as a rating system assumes, or are values potentially in conflict or in a hierarchy of importance, as a ranking system assumes? The World Values Survey makes use of both systems (ranking and rating) to investigate their effects. When the same concept is approached in these two ways, sometimes one can get quite different results, results that are an artifact of the method of question writing. Ovadia argues that the choice of valid question form depends upon context and that the two formats are not mutually exclusive in all circumstances. Valid surveys will generally use both.

The valid measurement of change over time is very important in the social sciences, but except in economics, the data are seldom been rich enough to apply quantitative techniques. Either there are too few cases or the cases are measured too infrequently, or both. Wilson & Butler (Chapter 51) examine this situation in political science. There the most commonly used method is one suggested by Beck & Katz in 1995. It has become orthodox. As with most orthodoxies, however, this one tends to discourage rather than encourage reflection and conscious choice. Wilson & Butler reviewed 195 articles in political science journals and found that most authors using the method did not even consider problems with its application, problems that were the main point of the original article by Beck & Katz. Wilson & Butler conclude that in fields such as political science with small Ns, much data gathering will need to remain qualitative, and researchers will often have to foreswear using fancy techniques of quantitative analysis. There is a tendency on the part of researchers to turn guidelines into routines followed without a great deal of thought. In a world where much is complicated it is understandable that researchers should want to be able to use tried-and-true methods. But there is no escaping the need to think. Routinization can be highly detrimental to selecting research methods rationally.

Methods to Improve Survey Questions and Experimental Instructions

Obtaining reliable and valid information from survey respondents and experimental participants requires that they are asked clear questions and given unambiguous instructions. Most textbooks routinely state that pretesting is the single most important step that researchers can take to improve survey questions. However, as Presser and colleagues demonstrate in Chapter 52, pretesting is not often done, and there are few reliable guidelines for doing it. When it is done, the details are seldom given in the kind of depth that would allow for the accumulation of knowledge. Chapter 52 is the definitive review of the state of our knowledge about improving survey questions. There is no better guide to what conscientious researchers ought to do before fielding a survey. The overall picture is not encouraging. Survey researchers have much better methods for finding problems than for fixing them. Presser et al. provide a checklist of problems researchers may need to address and up-to-date knowledge about how to do so.

The theme of refreshing honesty about how much more we have to learn is continued in Chapter 53, by Beatty & Willis. Their topic is cognitive interviewing.

Cognitive interviewing focuses on the validity of questions, on whether an item on a survey 'is generating the information that its author intends' (p. 287). Data on validity are gathered by asking a small sample of respondents to tell survey interviewers what they mean when they answer questions. Cognitive interviewing is used increasingly, especially for large-scale survey projects, but research on the most effective methods of conducting cognitive interviews is not highly developed. Nonetheless, there is no better review of what we currently know about this increasingly important practice than Chapter 60.

Pretesting experimental instructions is surely as important as pretesting survey questions, but this is a field that is even less developed. Rashotte and colleagues in Chapter 54 provide a very systematic discussion of the topic. Their ability to make some progress on this question was facilitated by their use of an often-repeated experiment in social psychology. The effects of different instructions in the same experiment were substantial, substantial enough that the authors' call for continued work along these lines is clearly justified, if researchers are truly concerned about the validity of the data they gather. When dealing with human subjects, researchers must investigate whether communication is clear and not simply assume it is.

The most common statistical method for reviewing the meaning of survey questions (experimental instructions are much less studied in any manner) is factor analysis. In the context of survey research, factor analysis is essentially a means to study the patterns of correlations among scores on groups of questions. Hensen and Roberts in Chapter 55 review the way factor analysis has been used for this and related purposes. Again, the issues are reliability and validity. Are questions answered by respondents in consistent (reliable) ways? Do the questions validly represent the constructs the researchers are trying to study? How well are researchers employing factor analysis? The authors review the use of factor analysis in published articles in four psychology journals during the late 1990s. Conducting a factor analysis requires that a researcher make many decisions about subroutines, and experts disagree about which approach is appropriate in what circumstances. Hensen and Roberts review these controversies, but their main emphasis is on several undisputed procedures that researchers ought to follow. In brief, authors should explain what decisions they made and why. Hensen & Roberts find, alarmingly, that many published articles in respected journals do not meet minimum standards. They found some 'egregious errors concerning appropriate reporting practice' (p. 403). The majority of articles did not (1) report which matrix of association they used or (2) justify their rotation strategy or (3) list communality coefficients or (4) give the variance explained by each factor. The authors conclude with a series of recommendations all researchers would do well to follow.

Question Formats for Surveys and Interviews

We have reviewed several studies of the *mode* of survey administration (face-to-face, telephone, internet, etc.). And we have considered the issue of the reliable and valid *content* of questions. Here we turn to the *form* of questions, which is another area that beginning researchers might assume is unimportant, but as the following

studies demonstrate, they would be naïve to do so. The chapters in this section provide very careful comparative studies of different question forms. In a comparison of questions that ask respondents to check all that apply or questions that force them to make a choice, Smyth et al. in Chapter 56 show that the latter is always superior. I found the chapter so persuasive that I doubt I will ever use the check-all-that-apply format again and will always use the format in which respondents have to provide a distinct response to each item on a list. Christian et al. (Chapter 57) focus on the somewhat related issue of the visual design of questions. Their experiments demonstrate that small differences in the visual layout of questions can make a huge difference in the quality of the answers, sometimes a bigger difference than changes in wording.

The question of the best format for assessing how subjects allocate their time is important in several fields. Juster and colleagues in Chapter 58 compare the advantages and disadvantages of three methods for collecting data on time use: (1) *time diary* methods in which respondents are asked to recall what they did on a specific day; (2) *stylized* measures in which respondents tell how they spend their time on a 'typical' day; (3) *experiential sampling* in which respondents wear a pager and record what they are doing at the moment it goes off. The three methods differ in cost, and experts disagree about which is most effective in what contexts. The authors provide a useful comparative table of the measurement strengths of the three and good experimental data on the contexts in which each is likely to be preferable.

Similar comparisons are made by Belli and colleagues in Chapter 59, but their emphasis is on retrospective information. In some respects all data are from the past and all descriptions of time use are retrospective (e.g., what *were* you doing when the pager went off?), but Belli and colleagues focus on the long-term, not the recent, past. A retrospective approach has many possible uses and drastically reduces costs and time needed to collect longitudinal data. Belli et al. compare the standardized question list and the event history calendar. In the question list, interviewers read questions and respondents give a short answer either picking from a closed-ended list or providing an answer to a factual question such as: how old were you when you got married? The event history calendar presents respondents with a pencil and paper form to complete with the assistance of an interviewer. The idea is that listing events on a multi-year calendar helps respondents remember and sequence them better ('I got the new job just before we bought the house'). Belli et al. do an experimental comparison and find that while the costs of administering the two are about the same, the event history calendar is superior to the question list. It substantially improves respondents' ability to recall events.

Volume 4: Methods for Analyzing and Reporting Results

The distinction between quantitative and qualitative methods of analysis, tied as it is to quantitative and qualitative methods of measurement, becomes very important in the analysis stage. It is not absolutely determinative even here, because many kinds of data – textual, categorical, and ranked – can be studied with either quantitative or qualitative techniques. But types of data do importantly limit the range of appropriate techniques.

As was discussed in Volume 3, the two most frequent categories of ways data are coded and measured are verbal and numerical. Ordinal data is a middle type that can be coded and analyzed numerically or verbally. Other categories of data exist. The most important type discussed only marginally in these volumes is visual data. Visual data include a huge range of kinds of evidence, such as maps, satellite images, video tapes, and X-rays. Because these tend not to be used as extensively in most social sciences as verbal and numerical data (geography is an exception), coverage in this book is less extensive for visual data. Here we will focus on the kinds of data most used by social scientists: words and numbers.

Computer-Intensive Methods for Qualitative Data

The wide availability of computers and software has revolutionized data analysis. That statement is usually thought to apply mainly to numerical data, but it applies equally to the analysis of verbal data, perhaps more so. Computer programs may allow for more new kinds of qualitative data analyses. Most of what is done in quantitative analysis with computers is conceivable without computers and indeed was conceived before computers became ubiquitous. Quantitative techniques such as factor analysis might have been too time-consuming to actually execute very often, but the methods were well understood long before the computer became a mass market item.

Textual information forms vast data fields in the social and human sciences. Sometimes it is generated by researchers, but more often it is produced by people who do not have researchers in mind. The means of analyzing texts range from traditional read-and-think methods, through grounded theory, to the kinds of computerized techniques we will review in this section. The gigantic mounds of textual data that are available to researchers have several advantages. Unlike qualitative data gathered directly from human subjects, printed data generally raise no ethical problems, do not change over time or in context, and are infinitely available for testing, replication, and reanalysis by numerous researchers. And the very massiveness of the textual sources available, which made it intractable to traditional methods, is a benefit with computer software, since quantitative content analysis requires large amounts text. Add to all this that it is increasingly easy to find and access texts electronically and the result is a vast and largely untapped resource.

Some of the ways that researchers use computer software to help with the analysis of textual data are examined in this section's chapters. Analyses integrating qualitative and quantitative data using computer software are discussed by Bazeley in Chapter 60. Very few published research reports *integrate* methods at the *analysis* stage. Most mixed-method studies collect data with different methods (surveys and interviews are the most common mixture) and then conduct *separate* analyses of the data thus gathered. Integration of the analyses is usually informal and comes, if at all, at the final, speculative stages of interpretation. Bazeley thinks it is possible to do much better by using software (mostly text analysis software) to *combine* data and/or to *convert* one form of data into the other. Techniques for doing this are still being developed, but the theoretical foundation is not difficult to imagine, even if it is at present somewhat underdeveloped. Examples in the published literature

are also fairly rare, but Bazeley provides sufficient illustrations from her own work to be convincing about the promise of such methods.

The theme of computer analysis of texts continues in the next three chapters. The contents of the texts analyzed differ greatly: published journal articles in economics; answers to open-ended questions in a psychology experiment; and political party manifestos. Goldschmidt & Szmrecsanyi (Chapter 61) studied the discipline of economics by subjecting four decades of its published literature to systematic computer analyses. While they were interested in substantive conclusions about economics, the authors' main goal was to illustrate the potential of their method of textual analysis. A combination of word frequency counts and factor analysis enabled the authors to arrive at several important conclusions about the evolution of the discipline of economics and, through the analysis of journals in other fields, its distinctiveness as a discipline. The authors were able to review an astonishingly large amount of textual data and to reduce it to manageable proportions to draw credible conclusions. They thus provide an illustration of the possibilities for other researchers who confront massive amounts of text needing analysis.

Factor analysis of word frequencies was also used by Simon and Xenos (Chapter 62). They conducted an experiment with 150 participants. The output was answers to an open-ended question, which were then analyzed using quantitative techniques. They also analyzed the same texts using human coders and traditional methods. The two methods produced very similar results, which provides some (convergent) validity for the factor analytic method. If the two methods yield the same results, what is the advantage of the factor analysis? First, it was quicker and cheaper, an advantage that would grow in proportion to the length and complexity of the texts being analyzed. Furthermore, the exploratory factor analysis results could be incorporated into other analytic techniques, such as structural equation models, to allow for hypothesis testing. The authors concluded that factor analysis can be used to uncover the structure of participants' beliefs and values.

Computer methods are also compared with traditional hand-coding methods by Laver and colleagues in Chapter 63. Their research involved comparing test or 'virgin' texts with reference texts selected by the researchers. Selecting the reference texts is crucial since any attempt to assess a text must have some external context. To study the political manifestos that were the object of investigation, they compared them with the reference manifestos. Selecting these references 'involves crucial substantive and qualitative decisions by the researcher' (p. 314). After the selection, one can turn to computer software for word frequency counts to assess, for example, changes in the policy positions of political parties. The authors contend that the payoff can be great and the method can even be used on 'documents written in languages unknown to the analyst.' This is a remarkable claim, but one that, by the end of the chapter, to my surprise, I found persuasive. One can never dispense with crucial input from the researcher, much in the same way that interviewers have to ask good questions to generate useful answers. No one would deny the need for thoughtful expertise. The authors insist that they are not trying to have computers read text for meaning, for computer programs cannot 'understand meaning in context, something easily, if unreliably, performed by humans' (p. 329). But computer programs are good at raw counting and sorting, which for some problems can produce very credible results in the analysis of texts.

Qualitative Comparative Analysis

Qualitative comparative analysis (QCA) was developed in the 1980s by Charles Ragin, and it has become one of the more widely used, and discussed, methods in historical sociology and comparative politics. By employing Boolean logic, it combines some of the rigor of quantitative methods with the advantages of depth that come with case-specific knowledge. It works most effectively with a small to medium number of cases, perhaps 20 to 50, and a handful of dichotomous variables (or 'attributes' in the language of QCA), perhaps 3 to 6, although there are no set numbers. With more cases and variables QCA can become unwieldy; then quantitative methods for categorical variables, such as logistic regression, may be more effective. Since the number of cases and variables is limited, selecting appropriate ones is very important. To be effective, the choices have to be based on deep knowledge of the topic, the cases, and the variables or attributes. Often the case selection includes the entire 'population,' such as all modern industrialized economies or all parliamentary democracies or all members of the OECD or all states in the United States. In its original form only 'crisp' sets, composed of dichotomous variables, were analyzed in QCA. In a later development, methods for 'fuzzy' sets, composed of ordinal variables, were introduced. The three chapters in this section outline the method, illustrate it, and offer extensions and friendly amendments.

Crisp QCA uses sets that are Boolean or dichotomous – such as member of the European Union, yes/no; or engaged in a war, yes/no; has nuclear weapons, yes/no; has compulsory military service, yes/no. These can be coded 1 for membership in the set and 0 for non-membership. Often they are coded with capital and lowercase letters: WEAPONS/weapons, WAR/war, for membership and nonmembership in nations that have nuclear weapons and have engaged in a war. With such categorizations and codings, the researchers can use Boolean logic to discuss the causal relations among the sets/variables. Can the analyst find a causal link between membership in the European Union, membership in the nuclear club, and membership in the set of nations having compulsory military service on the one hand and membership in the set of nations that have engaged in war on the other? Caren and Panofsky, in Chapter 64, are advocates of this method, but think it has one serious shortcoming: it does not provide a way to include time or sequence as a variable. For example, in the hypothetical case above, did compulsory military service precede (yes/no) engagement in war? The main disadvantage of adding a temporal variable is that doing so can sharply increase the number of combinations of attributes to consider. But as the authors show with a worked out example, *temporal* qualitative comparative analysis (TQCA) can be worth the extra complications that result.

Schneider & Wagemann also add friendly amendment to Ragin's basic method. In Chapter 65 they introduce the distinction between proximate and remote causation. They claim that it is important to attend to the difference between causes that are more like background conditions (remote) and those that are more like triggers (proximate). This enables the analyst to proceed in two stages and in so doing keep the number of variables to a workable total. The problem of too many variables for the number of cases is especially likely to be particularly troublesome

in the case where membership in the sets is a matter of degree: fuzzy-set QCA (fs/QCA).

As Ragin & Pennings explain (Chapter 66), in fs/QCA cases do not have only two values (yes/no, 1/0) but usually have 5: 1 = completely in a set; 0 = completely out; .5 = wholly ambiguous; .8 = largely in; .2 = largely out. It is probably the case in the social sciences that sets with fuzzy boundaries or varying degrees of membership are more common than are sets with crisp boundaries. Assigning degrees of membership in sets to cases requires 'the use of theoretical and substantive knowledge' (p. 424); it is much more than a matter of simple classification and coding. The fuzzy-set approach was slow to be incorporated into the social sciences, but it has recently grown there thanks to the popularity of works by Ragin.

As we saw in Chapters 64 and 65, researchers are offering refinements to Ragin's basic model, which uses a logic based on set theory as an alternative to the quantitative logic so widespread in the social sciences. It is also possible to reunite the set theoretic approach with some aspects of traditional quantitative analysis, especially perhaps in the area of enabling researchers to determine when a fuzzy-set relationship is more effective for a given group of cases than is a crisp-set relationship. Ragin's work began in part as a middle way between quantitative and qualitative approaches to data analysis. Today, much work is being done around the fuzzy boundaries of the trichotomy: qualitative, ordered set, quantitative.

Choosing Analytic Models

The steps thus far in the research process have been: first select a design for gathering data; then decide from whom or what to gather it; then decide how to record and code it. Finally comes the fun part: analyzing the data to try to answer research questions. There are many options for how to proceed with this work. With analysis, as compared to the earlier choices, decisions are actually sometimes right or wrong. Mistakes can be made, as when one uses the technique that is not appropriate for one's data. In the chapters in this section, we examine seven examples of the wide range of choices that confront the researcher in selecting an analytic approach.

The theme of the previous section is continued by Grenstad (Chapter 67) who juxtaposes qualitative comparative analysis (QCA) with logistic regression. He compares the two by analyzing the same problem with the same data. Both methods are designed to deal with categorical variables, in the case of logistic regression, especially categorical dependent variables. Grenstad finds that the two yield similar results. Before concluding that the choice does not matter it is important to note that to conduct the comparison the author had to make several adjustments in the ways data were coded in each method, adjustments that an analyst using regular procedures with either one probably would not have made. Still, the chapter provides a very instructive review of two leading methods to analyze categorical data and an area of possible convergence between them.

Comparing methods for handling categorical data is also one of King's goals in Chapter 68. Logistic regression is rapidly becoming the method of choice in the social sciences for problems with a categorical dependent variable and several independent variables, whether categorical or continuous. Its advantage over QCA

is that it can handle large numbers of cases and numerous independent variables measured at any level. Logistic regression has also tended to replace other methods for multivariable quantitative analysis, such as discriminant analysis and log-linear analysis. That said, and convincingly, King turns to his more central concern: how to select the best predictors (independent variables) for a logistic regression. Any research project can only be as good as the variables included in the attempt to address it. The greatest error is sometimes called LOVE or left-out variable error. But it is also the case that including redundant or inappropriate variables can bedevil the analysis. To help with this problem, King advocates the 'best-subsets' method. This allows the researcher to generate and compare different models, which is a vast improvement over the widely used – but justly condemned by statisticians – stepwise routines. Stepwise software routines allow the software to do one's 'thinking' for one. Actually, they eliminate thought and replace it with an inappropriate algorithm. This chapter is likely to persuade researchers of the value of the best-subsets approach when they are working on a logistic regression problem.

The theme of how to choose a model or models among many possible alternatives is continued by Weakliem in Chapter 69. A model is, briefly, a collection of variables and a proposed pattern of relationships among them. Traditionally, models have been compared using the criteria of hypothesis or significance testing. This is a poor tool for the job, but until the 1980s good alternatives, which focused on *model* testing, not hypothesis testing, were not widely available. The two main alternatives now available for comparing and ranking all models, by providing measures of the degree to which one model is an improvement on another, are the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). These are closely related and also related to several other model selection criteria. The BIC is much more common in sociology publications, while the AIC is more often found in economics journals. Each yields a statistic that enables the researcher to determine whether it is useful to add a given variable to a model. Like the more familiar adjusted-R^2, BIC and AIC exact a 'penalty' for adding another variable. One main difference between the two criteria is the size of that penalty. The fundamental question when selecting models is 'how much additional information a parameter must add to justify the "cost" of its inclusion' (p. 170). The Weakliem chapter provides a highly accessible source to introduce the reader to this crucial aspect of choosing among complicated quantitative models.

In the three previous chapters, the focus was on analyzing large data sets that were not collected by the researcher. By contrast, with propensity score matching, described by Rudner & Peyton (Chapter 70), the issue is experimental studies where members of control and experimental groups are often determined by researchers. When membership in control and experimental groups cannot be assigned at random, an alternative is the matched pairs design, where subjects are assigned based on their scores on some covariate. For instance, if we wanted to compare two methods for teaching statistics, we might match subjects on their prior knowledge of statistics. In so doing we would ensure that the control and experimental group were matched on this important covariate. Propensity score matching, as developed by Donald Rubin, is an advance on this method; it employs multiple covariates to match subjects. Logistic regression or discriminant analysis is used to compute a propensity score for each subject and the subjects are then matched on

this multivariate score. The only problem that propensity score matching does not eliminate is left-out variable error (LOVE). This problem is minimized when random assignment is used. Propensity score matching is probably the best approach when random assignment is not possible.

The goal of propensity score assignment to control and experimental groups is to make causal inferences. Experimentation is very rare in demography and only somewhat less so in economics, but the desire to discuss causal relationships is no less widespread. Moffitt (Chapter 71) reviews recent developments and discussions in economics as they apply to demography. When one cannot conduct experiments one uses causal models instead. This raises the issue of model specification. How do we know we have included relevant variables? The answer to this question is much more important than knowing which methods of estimation to use. It is not much help to have a good estimate of an irrelevant variable. To select variables with which to build models there is no substitute for in-depth knowledge and solid reasoning based on it. Model building is difficult because the number of possible confounding variables is almost always large and unknowable – so large that causal conclusions are approximate at best. Confounding variables are in principle eliminated in controlled laboratory experiments, but most issues of interest to demographers are not susceptible to laboratory investigation. Some ingenious ways to conduct laboratory experiments in population research have been devised, and these experiments, like all experiments, are solid on internal validity, but very often they pay too high a price in external validity. In brief, Moffitt's message is rather pessimistic and the basic conclusion is that when it comes to causation 'modesty of claims for truth' (p. 106) should be a guiding principle. He concludes that progress in this area will most likely come through methodological pluralism and synthesizing research that takes different approaches. Formal theorizing will be required to stitch together the multi-method findings.

In any causal research, analytic methods and other tools to establish empirical patterns are crucial. It is essential to make sure that you have an association before you start generalizing about what its cause might be. The first empirical step in finding causes is finding patterns. Among the most difficult areas of research for finding patterns is international research. Moran (Chapter 72) uses studies of trends in income inequality in 17 nations to illustrate the use of bootstrap methods to facilitate generalizing across cases (nations) and over time. Bootstrapping is a computer-intensive method that uses randomized resampling techniques to estimate sample statistics, such as confidence intervals, without relying on unrealistic assumptions about the shapes of distributions. Specifically, Moran uses bootstrap methods to make national indicators more comparable. The techniques enable him to find three hitherto unrecognized patterns in the development of income inequality.

Establishing patterns and generalizations is also the subject of Payne & Williams's review of research using qualitative data (Chapter 73). They offer generalizations about the extent of generalization. Many researchers doing research using qualitative data assume that generalization is impossible or inappropriate. The drive to eschew generalization seems based on the notion that this is what the quants do. A reasonable insistence that one's cases are not a 'sample' from which it is appropriate to generalize to a 'population' does not mean that generalizations

are impossible or wrong-headed. Payne & Williams claim that not only are generalizations appropriate in qualitative research, *moderate* generalizations are unavoidable in sociology and by extension in the social sciences more generally. At minimum, articles published in a recent volume of the official journal of the British Sociological Association, *Sociology,* did not manage to avoid generalizations. Since no one manages to avoid generalizing, Payne & Williams argue, it should be done consciously and systematically rather than haphazardly. Some authors have apparently used a contention that one should avoid generalization as a pretext to lower standards. Generalizing using qualitative data is more demanding but potentially theoretically more rewarding. (This is one message of QCA.) Because qualitative generalization is more demanding it requires more effort, not less. Payne & Williams provide a thoughtful and easily applied set of guidelines for being more rigorous.

Data Handling and Interpretation Problems

Attention to detail is crucial in any form of research, whether ethnographic thick description or replication of computer-intensive findings. This is the theme of the next three chapters, which provide examples of some of the kinds of attention that needs to be paid. Errors in interpreting statistical significance are examined by Gelman & Stern in Chapter 74. As other chapters in these volumes stress, quantitative research has traditionally relied too heavily on measures of statistical significance, and its cousin statistical power, to assess the quality of research. This is bad practice. Statistical significance informs the investigator that, if a null hypothesis is true, a result of a given size (or larger) is unlikely to have been due to chance. This says little about the size of an effect except that it meets a threshold. And it says nothing at all about the substantive significance or practical importance of a finding. A related error, and one that has received rather less attention, is that it is inappropriate to compare the significance levels of the results of studies to judge them. For example, if Study A has a p-value of .3 while Study B has a p-value of .03, this does not mean that Study B is in any sense better, for example, that its treatment was more effective or that it was 10 times more significant. It could simply mean that a larger sample was used in Study B. Yet it is fairly common to find research reports comparing significance levels (p-values). This is an error that should be avoided because, as Gelman and Stern conclude, the difference between 'significant' and 'not significant' is not statistically significant.

Another ill-advised practice is to trust your data analysis software too much. Not only is it the case that the analyst has to decide among methods of analysis (e.g., logistic regression versus discriminant analysis), the software package used to conduct the analysis can affect the results, sometimes dramatically so. This is Altman & McDonald's point in Chapter 75. They show that problems are especially likely to occur when using computer-intensive procedures, such as maximum likelihood estimation. This later is the foundation for many important methods of analysis, such as structural equation modeling (SEM). At least 10 different programs are available for constructing and analyzing SEM, and they sometimes yield results that differ importantly. There are several steps that an analyst paying sufficient attention to detail ought to take, such as running the data on different programs

and comparing the results. At the very least, research reports ought to supply the reader with the name and version of the software used to do the analysis. Of course, trusting what goes on inside the 'black box' of the computer program makes no more sense in the analysis of qualitative data.

Attention to detail is required in many ways. When doing so, one often finds imperfections in one's data. Here is an example from some survey work I participated in.

Question 20. Do you have a part-time job? Yes or No? (If yes, please answer questions 21 – 24. If no, skip ahead to question 25.

What do you do when a substantial number of your respondents say, "No I do not have a part-job," and then answer questions 21 – 24 about the characteristics of the job that they do not have? Do you change the answer for question 20 from No to Yes reasoning that if they told you about their part-time job in questions 21 – 24, they must have had one? Or do you discard the answers to questions 21 – 24 as logically incoherent? Or do you discard their answers to question 20 *and* to questions 21 – 24? Do you do this even if the answers to those five questions are your main source of data about key variables in your study? You have to do *something*. If you do not change the answers to question 20, or if you do not discard the answers to questions 20 – 24, you will have to work with logically incoherent data, which could render any analysis useless. It probably matters less what you do than that you do something *and* tell the reader what that was. This is the main point in an ingenious study by Leahey and colleagues reported in Chapter 76. By presenting researchers with hypothetical data problems, they found that there was very little consistency about what researchers thought ought to be done. Messy data are universal, whether they are messy because they are missing, inconsistent, or impossible. But standards for dealing with the problems messy data cause are by no means standardized. Different researchers may offer radically different solutions to the same problem. The best advice is, perhaps, simple: Do something reasonable. Do it consistently. Explain what you have done.

Presentation of Findings

One of the final sets of choices in a research project is deciding how to present one's findings. Of course, thoughtful researchers will have anticipated this step and planned for it in advance. The final section of *Selecting Research Methods* presents suggestions about how to present one's findings. Structured abstracts are mooted and elaborated upon in Chapters 77 and 78. Mosteller and colleagues outline the basic idea. Make sure that each abstract covers the following essential items: Background, Purpose, Design, Conclusions; other sections would be included in some articles but not all, e.g., setting, population, intervention, and analysis. Readers of abstracts would know what every article would provide and in what order. Structured abstracts are longer than traditional abstracts but are still generally less than a page. They have been enormously popular in medicine where the format grew from none to widespread in a few years in the late 1980s and early 90s. In brief, structured abstracts facilitate communication among researchers and between researchers and practitioners. Kelly and Yin add to Mosteller's proposal by suggesting that the abstract should also contain information about the assumptions

and the nature of the arguments that lead to the article's knowledge claims – what they call the *warrant* for the conclusions. Kelly & Yin want to see 'greater clarity on the nature of research claims and the arguments and methodologies that underpin them' (p. 137). What Kelly & Yin ask for should indeed be included in every article, and it is important. But including it in the abstract might not facilitate communication if it made structured abstracts unwieldy.

When presenting quantitative findings the analyst almost always has several options. In part this is because various statistics for displaying results can be converted into one another and are sometimes fairly simple functions of one another. So how, in such circumstances, does one choose? One reason for choice has to do with the ability of a statistic to be used in further analyses, such as calculating power and statistical significance. Another sometimes related but distinct criterion has to do with clear communication of and comparison of results, either to other researchers and/or to a more general audience.

Effect size (ES) measures have become increasingly popular in recent decades. In some fields, especially non-experimental fields such as sociology, political science, and economics, effect size measures have always been used, particularly one form or another of the Pearson *r* correlation. In other fields, especially psychology and the many fields in which psychological methodology was the model, the ES was comparatively rare until the 1990s. In those experimental and psychology-dominated fields, the norm was to report the mean difference between the control and experimental group and the statistical significance of that difference. As seen above, measures of statistical significance (p-values) are inappropriate for comparisons across studies. Effect size measures are meant to remedy that weakness. Of course, a mean difference is a measure of the size of an effect. So what is usually meant by effect size is *standardized* effect size. Standardization (expressing size in standard deviation units) facilitates comparison across studies. In his history of standardized effect size indices (Chapter 79) Huberty shows that there are dozens of them. Despite their reputation in some circles as being a fad, standardized effect size measures have been used for a century. Huberty's historical account of their development provides the reader with a checklist of most of the measures available.

Conclusion

A vast range of methodological options is available to researchers. From those options an investigator selects a suite of methods to use on a research project. Choices are made in four broad categories: design, sampling, coding, and analysis. Those categories are also stages in the progress of a research project. In the design stage, one focuses on general strategies to gather evidence relevant to the research questions. These strategies differ according to one's philosophical orientations, whether one plans to gather quantitative and/or qualitative data, and on whether one will survey, interview, observe, or experiment. In the second or sampling stage, the researcher selects what or whom to study and decides how many cases are needed to answer the research question. Gaining access to the cases one wishes to study is as big an issue as deciding what those cases are. In the third or coding stage, the researcher selects a coding scheme, which amounts to deciding how to

record the evidence gathered and how to do so in ways that are both reliable and valid. In the fourth or analysis stage the researcher selects methods for interpreting and reporting the evidence gathered.

At each stage, choices abound. A good researcher does not avoid choices by replacing them with habits and traditions. Good research involves many things. Among them is using critical intelligence and rational means of decision making. Living up to that high standard, one cannot easily justify unthinking routines and traditions when making fundamental choices. *Selecting Research Methods* tries to assist researchers by providing resources for making decisions.