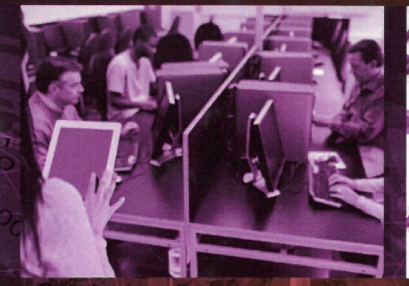
Laboratory Experiments in the Social Sciences

Second Edition


Edited by

Murray Webster, Jr. and Jane Sell

AP

granted by the university's IRB. The scheduler then uses the approved script to contact potential participants. In our experience, contacting participants during the weekend before the week they would be scheduled to participate is most effective. Phoning in the late afternoon is also effective. After the scheduler fills all possible experimental sessions for the upcoming week, making reminder calls to participants the day before they are scheduled to participate is good practice. We find that this strategy decreases the rate of "no-shows" for experimental sessions. Paid schedulers can also be offered a bonus for each scheduled participant who successfully completes participation in the study. Another useful strategy, albeit a more costly one, involves "overbooking" experimental sessions and arranging payments for those who come but do not take part in the experiment.

A second method of recruiting a pool of research participants in sociology is semester-commitment recruiting. As its name suggests, researchers using this technique secure a commitment from participants to actively participate in experiments for an entire semester. This method is only valuable in situations in which participants in the pool may take part in multiple experiments or in the same experiment multiple times. As such, this technique is not recommended for experiments in which deception is used early in the term.

Making initial contacts with potential participants for semester-commitment recruiting can be achieved by way of e-mails or the in-class technique. The primary way in which this method differs from others is that researchers obtain a commitment from participants to participate in as many experiments (or sessions of the same experiment) as they are able during a semester (i.e., as is reasonable given their schedule). If participants are paid for their participation, the pool of semester-commitment participants can be thought of as employees hired to contribute to the research being conducted. If participants receive course credit for their participation, these participants may be thought of as students enrolled in a semester-long course whose grade depends on the frequency of their participation. One way to manage course credit in lieu of payment is to create a research practicum course in which participants may register. Although participants in this framework may not receive instruction in the traditional format (i.e., with lecture and discussion), their experiences and observations of the research process justify college credits in the same way internship credits often count toward obtaining an undergraduate degree.[2]

D Online Experimentation and Virtual Worlds

Although the technological and organizational issues and "best practices" for carrying out controlled social science experiments on the Web have only recently started to receive due attention (Kohavi, Longbotham, Sommerfield, & Henne, 2009), programs such as TESS (Time-Sharing Experiments for the Social Sciences) provide researchers with a unique opportunity to participate in this new and exciting area of experimentation in sociology and related disciplines. Funded by the National Science Foundation (SES-0818839), the TESS program reviews brief, 5-page proposals from social scientists interested in conducting Web-based experiments with participants who are representative of the U.S. general population. Studies detailed in successful proposals are carried out at no cost to the principal investigator by the private research company, GfK (formerly Knowledge Networks). Population-based experiments such as those that can be conducted through TESS are particularly useful in cases in which the researcher seeks to combine the internal validity of experimentation with the external validity of probability sampling (Mutz, 2011).

In addition, popular multiplayer online role-playing games (MMORPGs) and other kinds of online virtual worlds represent another emerging frontier in social science laboratory research (Lovaglia & Willer, 2002). For example, the online virtual world "Second Life" has had more than 35 million "Residents" (i.e., players) since it was launched in 2003 (http://gridsurvey.com), many of whom participate on any given day to create a place for themselves in a virtual society in which social meaning and structures are created, negotiated, and modified as users of varying power and status interact through avatars, become involved in groups, and participate in an internal economy by exchanging a variety of goods and services with one another. The principal advantage of such Internet laboratories is that complex social situations can be followed by investigators and outcomes can be observed at various points as Residents play out their roles in an ongoing manner during the course of months or even years. For example, Bakshy, Karrer, and Adamic (2009) obtained data directly from the makers of Second Life (without personally identifying information) and analyzed complete Resident data over a 130-day period. They used these data to identify and model the influence dynamics underlying the diffusion of content (buildings, fashion, etc.) through evolving social networks. One drawback of this kind of research is that control over the characteristics and circumstances of participants is virtually absent. However, the drawback of participant heterogeneity is compensated by the huge number of participants and the amount of demographic information that can be collected to establish statistical control.

University undergraduates are not ideal for studying some research questions, as dictated by theory and sometimes practicality. Web-based experiments and online virtual worlds with simulated communities are a promising alternative for studying questions not amenable to analysis with recruits from university student populations.

IV PARTICIPANTS IN POLITICAL SCIENCE

Political science studies research questions for which undergraduates often make suboptimal research participants. Participation is reserved for adults and undergraduates who have little or no previous experience as research participants.

2. Although, as we discussed in the previous section, the *required* nature of such participation raises ethical concerns.

A variety of techniques are used in political science to recruit participants for experimental research that could be classified as laboratory based, whether those experiments occur in a university laboratory, in the field, are survey based, or online. Typically in political science, potential participants are told they will receive some form of reward, usually monetary pay, for participation.

A Laboratory Locations

Recruitment varies greatly even in laboratory experiments. As is often the case now in psychology and sociology, one method is to use a Web site where individuals can register and sign up for participation. One example is the Interdisciplinary Experimental Laboratory (IEL) at Indiana University (http://www.indiana.edu/~ielab), a joint endeavor of faculty and staff from political science, psychology, economics, and geography. Details on the variety of participant recruitment methods employed by the IEL are available on its Web site. Following the usual procedures, students at Indiana University first visit a Web site and register an account. They then sign up for an experiment by reviewing a dynamically updated calendar.

Another means of recruiting participants in political science involves visiting college dormitories to collect personal information from students who want to participate over the course of the year. This information can be entered into a database, and a selection of prospective participants may be generated from that list. The next step is to send an e-mail to each student directing him or her to a Web site to sign up for the experiment. Wilson and Eckel (2006) used this method to recruit participants to explore beauty and expectations in trust games.

Non-Internet methods to recruit participants have been used as well. In an interdisciplinary project (sociology and political science), Sell and Wilson (1999) recruited participants from introductory social science and humanities classes. Students were told they would be paid in cash for volunteering in "decision-making" experiments. Those who volunteered were scheduled at their convenience and randomly assigned to experimental conditions.

Another example of non-Internet recruiting is seen in the work of Bottom, Eavey, Miller, and Victor (2000). They recruited 240 participants from undergraduate and graduate classes in the school of business, the school of engineering, and the college of arts and sciences. They advertised an experiment in "collective decision-making" in classrooms, through an electronic bulletin board, and through sign-up sheets posted in the student union. All methods mentioned a minimum payment of $3 plus an opportunity to earn more based on group decisions.

B Laboratory Locations Using Nonstudent Participants

In laboratory experiments, when the student population is not desired, researchers may also draw from the general public. For example, Berinsky and Kinder (2006) enlisted participants through posting advertisements and also recruited from local businesses and voluntary organizations. Participants reflected great diversity (compared to the college student sample), although as discussed previously, for theory-testing purposes this is not desired. In addition, Ansolabehere, Iyengar, Simon, and Valentino (1994) examined the effects of negative campaign advertising on voter turnout. During an ongoing political campaign (therefore featuring actual candidates and voters), they recruited participants by placing advertisements in local papers, handing out flyers in shopping malls and other public venues, posting announcements in employer newsletters, and telephoning people from voter registration lists. All participants were promised payment of $15 for an hour-long study. Although the sample was not random, descriptive statistics suggested that it reflected the population from which it was drawn. Another study by Iyengar, Peters, and Kinder (1982) recruited participants from a specific city using classified advertisements that offered $20 to those who participated in "research on television." Interested citizens responded by phone and were randomly assigned to experimental conditions and scheduled at their convenience. Descriptive statistics suggest this method also produced a roughly representative sample of the city population. Redlawsk (2002) recruited participants in a large city by contacting different organizations (including the YMCA and a senior citizen center) and requesting that they invite their members to volunteer in experiments in return for a $20 donation to the organization per member who participated.

C Laboratory Experiments in the Field

In some field experiments, a community becomes the laboratory. For example, Eldersveld's (1956) often cited early work examined the effects of personalized versus impersonalized propaganda techniques on voting behavior. Eldersveld mailed out different forms of propaganda and followed up with post-experiment interviews. Local participants came out of a sampling frame of city clerk records. He selected all people living in four precincts of a central area and who had voted regularly in both state and national elections (but not in local elections, for reasons related to his research question). Although not perfectly representative, the sample size of 187 in two conditions allowed much statistical power for the use of statistical control variables.

Gerber and Green (2000) randomly selected households and exposed them to direct mailings, telephone calls, or personal appeals before a general election to determine which had the most impact on voter turnout. From a complete list of registered voters, they created a sampling frame of households. This technique generated a sample of 22,077 households. The effectiveness of randomization was checked using voter turnout data from an earlier election—a technique based on statistics and that showed there would be no significant difference between current and past voting behavior. The benefit of this technique is the large sample size that allows statistical control to overcome the loss of experimental control occurring with a heterogeneous sample.

Bahry and Wilson (2006) recruited participants for their field experiment using a sampling frame of individuals who had participated in an earlier interview pool in Russia. A total of 646 participants were included, with 252 from Tatarstan and 394 from Sakha. Experiments were conducted in small villages, medium-sized cities, and large urban areas within these Russian republics. Experiments were limited to villages and medium-sized cities where at least 20 individuals had been interviewed previously. Some medium-sized cities were skipped where travel was difficult or impossible. Payment for approximately 2 hours of participation reflected a week's wage or more for 62% of their participants.

Finally, Wantchekon (2003) conducted an experiment in the Republic of Benin in West Africa. Working with a team of consultants who helped him contact the leadership of selected parties, he communicated directly with them and campaign managers who then agreed to run an "experimental political campaign" in select districts. From his list of 84 districts, Wantchekon chose 8 districts and divided each into three subgroups. Each subgroup was exposed to either one of two experimental conditions or served as a control.

D Survey and Online Experiments

In survey-based experiments, investigators use secondary data while adding a manipulation. Gilens (1996) did this to examine whether white Americans' opposition to welfare is rooted in prejudice against African Americans or nonprejudice reasons. Using the National Race and Politics Study data set—a national telephone survey—he applied a manipulation in the survey in which half of the respondents were asked a specific attitudinal question about whites, and the other half were asked the same question about African Americans. Nelson and Kinder (1996) also used a secondary data source to recruit participants and create an experiment. In their work, participants were recruited from the sampling frame of respondents who completed the 1989 National Election Study (NES) and who also had provided their telephone numbers. Randomly drawn from this frame, the researchers created a representative sample of the American adult population. Advantages of survey experiments are large sample size, the ability to randomly assign the respondents to questions, and the ability to generalize results to a larger population if desired. They also have disadvantages. Gaines and Kuklinski (2007) review the typical uses of survey experiments in political science and identify problems and solutions specific to this methodology.

Online experiments in political science are also performed using a variety of recruiting techniques. OxLab at the Oxford Internet Institute (http://www.governmentontheweb.org) maintains a database of research participants including both University of Oxford students and nonstudents from the city of Oxford. Margetts, John, Escher, and Reissfelder (2011) studied how information on the Internet affects political participation by recruiting 668 individuals from the OxLab database and having them participate remotely in a Web-based

experiment using their own Internet connection. Using a less active approach to participant recruitment, investigators may also rely on "drop-ins," in which participants come across the experiment while surfing the Internet. Another method uses banner ads that offer some kind of incentive for participation. Finally, Iyengar (2002) has used a market research firm, Knowledge Networks, to reach a nationwide representative sample. Through standard telephone methods, Knowledge Networks recruits a continuous sample of individuals between the ages of 16 and 85 years who are provided free access to WebTV. In exchange, these individuals agree to participate on rotation in different studies. Iyengar examined online self-selection and found that drop-in Internet experiment participants reflect reasonably well the online user population, but participants still differ from the general population because non-Internet users are not reflected in the experiment sample. Iyengar also noted that among participants in online experiments, Republicans outnumbered Democrats and Independents compared to the broader online population. This is an important issue for political scientists and others who may prefer a "party-representative" sample for their research. In general, using the Internet as a platform for experiments offers many advantages (e.g., a worldwide geographic domain, the ability to reach diverse populations, and low cost). As with any format, however, there are drawbacks as well (e.g., sample selection bias, excluding the population with no Internet access, and lack of participant homogeneity for theory testing).

V CONCLUSION

In describing the methods used by laboratory researchers in several social science disciplines to recruit and work with human participants, we hope to have gone into enough detail to allow interested researchers to begin research with human participants in their own laboratories. As we have noted, recent technological advances require an expanded definition of laboratory experiments to include theory-driven fundamental research carried out in a variety of physical settings and using a variety of participant interface and data collection techniques.

REFERENCES

Ansolabehere, S., Iyengar, S., Simon, A., & Valentino, N. (1994). Does attack advertising demobilize the electorate? *American Political Science Review*, *88*(4), 829–838.

Aviv, A. L., Zelenski, J. M., Rallo, L., & Larsen, R. J. (2002). Who comes when: Personality differences in early and later participation in a university subject pool. *Personality and Individual Differences*, *33*, 487–496.

Bahry, D. L., & Wilson, R. K. (2006). Confusion or fairness in the field? Rejections in the ultimatum game under the strategy method. *Journal of Economic Behavior and Organization*, *60*(1), 37–54.

Bakshy, E., Karrer, B., & Adamic, L. A. (2009). Social influence and the diffusion of user-created content. In *Proceedings of the 10th ACM conference on electronic commerce, Stanford, CA*. New York: ACM.

Berger, J., M. Fisek, M. H., Norman, R. Z., & Zelditch, M., Jr. (1977). *Status characteristics and social interaction: An expectation states approach*. New York: Elsevier.

Berinsky, A. J., & Kinder, D. R. (2006). Making sense of issues through media frames: Understanding the Kosovo crisis. *Journal of Politics*, *68*(3), 640–656.

Bernard, L. C. (2000). Variations in subject pool as a function of earlier or later participation. *Psychological Reports*, *86*, 659–668.

Bottom, W. P., Eavey, C. L., Miller, G. J., & Victor, J. N. (2000). The institutional effect on majority rule instability: Bicameralism in spatial policy decisions. *American Journal of Political Science*, *44*(3), 523–540.

Bowman, L. L., & Waite, B. M. (2003). Volunteering in research: Student satisfaction and educational benefits. *Teaching of Psychology*, *30*, 102–106.

Britton, B. K. (1979). Ethical and educational aspects of participating as a subject in psychology experiments. *Teaching of Psychology*, *6*, 195–198.

Calder, B. J., Phillips, L. W., & Tybout, A. M. (1981). Designing research for application. *Journal of Consumer Research*, *8*, 197–207.

Coulter, X. (1986). Academic value of research participation by undergraduates. *American Psychologist*, *41*, 317.

Daniel, R. S. (1987). Academic value of research participation by undergraduates: Comment on Coulter. *American Psychologist*, *42*, 268.

Davis, D. D., & Holt, C. A. (1993). *Experimental economics*. Princeton, NJ: Princeton University Press.

DiMaggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, *48*, 147–160.

Eldersveld, S. J. (1956). Experimental propaganda techniques and voting behavior. *American Political Science Review*, *50*(1), 154–165.

Flagel, D. C., Best, L. A., & Hunter, A. C. (2007). Perceptions of stress among students participating in psychology research. *Journal of Empirical Research on Human Research Ethics*, *2*, 61–67.

Gaines, B. J., & Kuklinski, J. H. (2007). The logic of the survey experiment reexamined. *Political Analysis*, *15*, 1–20.

Gerber, A. S., & Green, D. P. (2000). The effects of canvassing, telephone calls, and direct mail on voter turnout: A field experiment. *American Political Science Review*, *94*(3), 653–663.

Gilens, M. (1996). "Race coding" and white opposition to welfare. *American Political Science Review*, *90*(3), 593–604.

Gil-Gomez de Liano, B., León, O. G., & Pascual-Ezama, D. (2012). Research participation improves students' exam performance. *Spanish Journal of Psychology*, *15*, 544–550.

Hertwig, R., & Ortmann, A. (2008). Deception in social psychological experiments: Two misconceptions and a research agenda. *Social Psychology Quarterly*, *71*, 222–223.

Iyengar, S. (2002). Experimental designs for political communication research: From shopping malls to the Internet. *Work presented at the Workshop in Mass Media Economics, Department of Political Science, London School of Economics*, June 25–26.

Iyengar, S., Peters, M. D., & Kinder, D. R. (1982). Experimental demonstrations of the "not-so-minimal" consequences of television news programs. *American Political Science Review*, *76*(4), 848–858.

Jackson, M. J., & Cox, D. R. (2013). The principles of experimental design and their application in sociology. *Annual Review of Sociology*, *39*, 27–49.

Jung, J. (1969). Current practices and problems in the use of college students for psychological research. *The Canadian Psychologist*, *10*, 280–290.

Kagel, J. H. & Roth, A. E. (Eds.), (1995). *The handbook of experimental economics*. Princeton, NJ: Princeton University Press.

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the Web: Survey and practical guide. *Data Mining and Knowledge Discovery*, *18*, 140–181.

Landrum, R. E., & Chastain, G. (1995). Experiment spot-checks: A method for assessing the educational value of undergraduate participation in research. *IRB: A Review of Human Subjects Research*, *17*, 4–6.

Landrum, R. E., & Chastain, G. (1999). Subject pool policies in undergraduate-only departments: Results from a nationwide survey. In G. Chastain & R. E. Landrum (Eds.), *Protecting human subjects: Departmental subject pools and institutional review boards* (pp. 25–42). Washington, DC: American Psychological Association.

Leak, G. K. (1981). Student perception of coercion and value from participation in psychological research. *Teaching of Psychology*, *8*, 147–149.

Lovaglia, M. J. (2003). From summer camps to glass ceilings: The power of experiments. *Contexts*, *2*(4), 42–49.

Lovaglia, M. J., & Willer, R. (2002). Theory, simulation, and research: The new synthesis. In J. Szmatka & K. Wysienska (Eds.), *The growth of social knowledge* (pp. 247–263). Westport, CT: Praeger.

Lucas, J. W. (2003). Theory-testing, generalization, and the problem of external validity. *Sociological Theory*, *21*, 236–253.

Margetts, H., John, P., Escher, T., & Reissfelder, S. (2011). Social information and political participation on the Internet: An experiment. *European Political Science Review*, *3*, 321–344.

McCord, D. M. (1991). Ethics-sensitive management of the university subject pool. *American Psychologist*, *46*, 151.

Miller, A. (1981). A survey of introductory psychology subject pool practices among leading universities. *Teaching of Psychology*, *8*, 211–213.

Mutz, D. C. (2011). *Population-based survey experiments*. Princeton, NJ: Princeton University Press.

Nelson, T. E., & Kinder, D. R. (1996). Issue frames and group-centrism in American public opinion. *Journal of Politics*, *58*(4), 1055–1078.

Nimmer, J. G., & Handelsman, M. M. (1992). Effects of subject pool policy on student attitudes toward psychology and psychological research. *Teaching of Psychology*, *19*, 141–144.

Padilla-Walker, L. M., Zamboanga, B. L., Thompson, R. A., & Schmersal, L. A. (2005). Extra credit as incentive for voluntary research participation. *Teaching of Psychology*, *32*, 150–154.

Redlawsk, D. P. (2002). Hot cognition or cool consideration? Testing the effects of motivated reasoning on political decision making. *Journal of Politics*, *64*(4), 1021–1044.

Roberts, L. D., & Allen, P. J. (2013). A brief measure of student perceptions of the educational value of research participation. *Australian Journal of Psychology*, *65*, 22–29.

Rosell, M. C., Beck, D. M., Luther, K. E., Goedert, K. M., Shore, W. J., & Anderson, D. D. (2005). The pedagogical value of experimental participation paired with course content. *Teaching of Psychology*, *32*, 95–99.

Rosenthal, R., & Rosnow, R. L. (1969). The volunteer subject. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 61–112). New York: Academic Press.

Sell, J., & Wilson, R. K. (1999). The maintenance of cooperation: Expectations of future interaction and the trigger of group punishment. *Social Forces*, *77*(4), 1551–1570.

Sieber, J. E., & Saks, M. J. (1989). A census of subject pool characteristics and policies. *American Psychologist*, *44*, 1053–1061.

Sona Systems. (2013). *Product: Web-based subject pool management*. Estonia: Sona Systems. Retrieved August 21, 2013, http://www.sona-systems.com/product.asp.

Stevens, C. D., & Ash, R. A. (2001). The conscientiousness of students in subject pools: Implications for "laboratory" research. *Journal of Research in Personality*, *35*, 91–97.

Szmatka, J., & Lovaglia, M. J. (1996). The significance of method. *Sociological Perspectives*, *39*, 393–415.

Thye, S. R. (2000). Reliability in experimental sociology. *Social Forces*, *74*, 1277–1309.

Trafimow, D., Madson, L., & Gwizdowski, I. (2006). Introductory psychology students' perceptions of alternatives to research participation. *Teaching of Psychology*, *33*, 247–249.

Turner, J. H. (2006). Explaining the social world: Historicism versus positivism. *Sociological Quarterly*, *47*, 451–463.

Wantchekon, L. (2003). Clientelism and voting behavior: Evidence from a field experiment in Benin. *World Politics*, *55*, 399–422.

Wilson, R., & Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly*, *59*(2), 189–202.

Witt, E. A., Donnellan, M. B., & Orlando, M. J. (2011). Timing and selection effects within a psychology subject pool: Personality and sex matter. *Personality and Individual Differences*, *50*, 355–359.

Wright, D. B. (1998). People, materials, and situations. In J. Nunn (Ed.), *Laboratory psychology: A beginner's guide* (pp. 97–116). Hove, UK: Psychology Press.

Chapter 6

Developing Your Experiment

Lisa Slattery Walker
University of North Carolina–Charlotte, Charlotte, North Carolina

I INTRODUCTION

For decades, at least, social scientists have discussed *whether* we should do experiments, *why* we do experiments, and even *when* we do experiments. But rarely do we discuss—particularly in print—*how* we do experiments. Reports of experimental research do not regularly describe the myriad minor and major decisions that went into the project. However, a social science experiment is made up of many details, and most of the details have to be addressed competently or the outcomes of the experiment can be useless for the purposes intended or, worse, misleading. With this chapter, I hope to remove some of the mystery from conducting social scientific experiments by presenting a few particulars about how one should design, conduct, and analyze results from an experiment in order to maximize the usefulness of its outcomes.

Elsewhere in this book, you can read about certain elements of how to conduct an experiment (e.g., technological issues, ethical concerns, training those who will be conducting your experiment, recruiting participants, and maintenance of records), but I focus on the initial design of an experiment and on the development of that design. Especially for new investigators, translating abstract considerations of good design into an actual, workable experiment can seem a daunting task. I hope this chapter can help with the process of doing a real experiment. I hope also to make clear some elements of experimental design that are often overlooked in more philosophical or abstract discussions.

For convenience, I present three steps in creating and conducting an experiment: (1) *designing the experiment*; (2) *pretesting* the operations and *pilot testing* the experiment; and (3) *analyzing and interpreting the data* it produces. Each stage in the execution presents challenges and requires decisions on the part of the experimenters. As an overview, it is helpful to keep in mind that the essence of experimental design is to create a situation (or possibly multiple situations) that includes *all* the factors described in a theory, and *only* those factors, in order to test ideas from the theory. In most cases, an experiment will contrast

Laboratory Experiments in the Social Sciences. http://dx.doi.org/10.1016/B978-0-12-404681-8.00006-6

multiple conditions, and those will ideally be identical to each other except for differences required by contrasting hypotheses.

II DESIGNING THE EXPERIMENT

Good experiments begin with an explicit theory, which has the structure to permit predictions of derived consequences. Theoretically derived consequences are sentences (hypotheses) detailing outcomes that a theory predicts, given a specified kind of situation. However, derived consequences usually contain abstract theoretical terms, not concrete terms that are immediately observable.

For instance, a derived consequence of David Willer's network exchange theory (NET) (e.g., Willer et al., 2002) might be the following: "A person occupying a central node in a network will have more negotiating power than someone occupying an isolated node." Although the sentence's meaning may be clear, it does not tell us in terms of operations just what being "central" means or how to observe "power." On the other hand, "A person with two potential exchange partners will gain more points in negotiation than someone with only one partner" translates the theoretical terms into observable facts in an experimental situation. The first sentence is a derived consequence of a theory; the second is a testable hypothesis. In designing an experimental test of NET, it is necessary to create such testable hypotheses.

No experiment can test all of the derivations of a theory; one must choose some of those derived consequences for hypotheses, preferably a set with some range of theoretical assumptions. For instance, if the theoretical foundation of the experiment is a theory having five general propositions, it is wise to examine which propositions are used in the derivation of each derived consequence. Usually, any two, three, or four propositions will yield many derivations. The experimenter must choose to test a few derivations from among a large set. Although the choice is somewhat dependent on personal preference and empirical simplicity, it is usually wise to be sure that the experiment tests as many of the propositions as possible. Thus, testing two derivations that both are implied by propositions 2 and 3 is redundant, and if no derivation that uses propositions 4 and 5 is tested, the experiment will provide only a partial test of the theory.

The design task is then to translate the conceptual terms in which the theory is couched into a realistic, although not usually real-world, situation in which the experimenter controls many of the elements. By "realistic," I mean that the situation must be understandable to the participants, and it cannot be so bizarre that they feel they have entered the "Twilight Zone." On the other hand, an experiment is not a natural setting. If a natural setting existed that provided a good test of theoretical derivations, there would be no need to design the experiment. Thus, most experiments seem a little strange to participants, but as long as they understand the important aspects of it and their behavioral options, realism is neither needed nor desirable. The more "realistic" an experiment seems, the more likely that some (but probably not all) of the participants will fall into

familiar role behaviors in it. If an experiment reminds some participants of a high school classroom, for instance, they may activate ways they typically behave in classes: some will be attentive, some defiant, some bored, etc. Those role behaviors and the variability across participants, of course, are not what an experimenter wants. What she wants is for the situation to present all and only the previously set initial conditions of her design. The situation should seem real to the participants in that it is understandable and it has consequence, although it may be unlike anything they have ever encountered before.

The important practical consideration is how to make the experimental situation understandable and relevant to the participants, with thought to their culture and background, while staying true to the theory. A number of abstract design elements thus come into play, particularly variables and conditions, manipulations, and manipulation checks.

A Standard Protocols

Many theoretical programs develop standardized experimental protocols. If you happen to be working in one of these areas, such a protocol is an excellent tool for developing your own experiment. In addition to providing you with what amounts to a shortcut to a good experiment, the use of such standardized protocols increases the likelihood that your results will be meaningful in an ongoing scientific conversation.

Changes made to standardized protocols should be driven by theoretical questions. That is, you should only alter the elements of the design needed to test your hypotheses. Making other changes to the protocol for reasons such as "it seems better" or "it will be easier this way" often leads to unintended, and generally unmeasured, consequences.

When you do find it necessary to change an established design, be sure to carefully pretest any altered elements. Later, I discuss in detail the importance of pretesting, but in the case of standard protocols it is particularly important. When you use a standard protocol, you increase the comparability of your results to others that use the same protocol. When you alter the design, you may just decrease that comparability. However, others will certainly make comparisons anyway, and it is your responsibility to ensure that they are valid.

B Variables and Conditions

I discuss, in turn, a number of abstract design considerations with which researchers must deal when conducting a social scientific experiment. Primary among these considerations are manipulations, where the researcher puts into motion the initial conditions and independent variables as specified by his or her theory. It is important that the researcher is clear from the outset just what are the independent and dependent variables in the hypotheses, and which ideas are being tested in the experiment, in order to create the experimental conditions.

In other words, it is important to be clear just which ideas from the theory are to be tested and what sorts of situations are appropriate for that purpose. What sorts of situations the theory describes are often couched as scope conditions or, sometimes, limiting conditions.

Hypotheses (like derived consequences) can usually be stated in the form "If X, then Y." More completely, they state, "Given a situation of a specified sort, if X then Y." The first part of that sentence, a situation of a specified sort, describes the *initial conditions* of the situation. This governs the kind of experimental situation the researcher will create. (Of course, the experimental situation must also instantiate the scope conditions of the theory, as discussed by Foschi in Chapter 11.) The "X" represents the *independent variable(s)*. This is the element that will be introduced in some experimental conditions and not in others, or introduced at different levels in different conditions. The "Y" is the *dependent variable(s)*. This is what the researcher will measure once he or she has devised a suitable measurement instrument.

Once the variables are clear, one can determine how many experimental conditions are required. This requires a good understanding of the variables and the relevant number of levels each has. An incomplete number of conditions can be the downfall of an otherwise well-conducted experiment if the important comparisons cannot be made with the data. This includes the problem of not having baseline conditions when needed. However, not every experiment requires a full crossing of all of the levels of the independent variables in order to have a set of conditions that is complete for the purposes of testing the theoretical hypotheses under consideration. Again, it is important to be clear about exactly what one is testing when designing the experiment.

Knowing the predicted relationships among the values on the dependent variables is also important. Again, the theoretical concerns allow one to determine just what the relevant comparisons are across conditions. The design of the experiment should allow for, and lead inexorably to, making comparisons among conditions that will give a true and meaningful test of the hypotheses and therefore of the theoretical concepts.

C Manipulations

Manipulations are the process by which an experimenter creates the independent variables operationally within the experimental setting. Manipulations in social science experiments frequently fall into the category of information that is given to the participants about themselves, anyone with whom they might be interacting, the situation, the task, or the social world. Other types of manipulations include the behavior of others in the situation (often computer programmed or performed by a confederate) or an imposed social network or structure.

The process of manipulations often includes a cover story or process of setting the stage as the researcher creates the scope and initial conditions and independent variables in an experimental condition. This cover story may or

may not include deception. Often, it is through this cover story that the experimental manipulations are made. It often comes in the guise of the instructions that participants receive regarding their participation in the study—what they are to do, when, how, and with whom. Manipulations can come in the form of commission—what is said in the given condition—or omission—what is not said in one particular condition that is said in others.

It is best to make sure that participants hear all of the relevant pieces of information at least three times during the cover story. As a rule, experimental participants are not especially attentive, and they often miss crucial pieces of information if it is only said once or even twice, so three times is required. They are also not usually very suspicious. Thus, although they might find the repetition of hearing something three times slightly tedious (if they notice it at all), it rarely causes them to disbelieve the cover story. It is better to err of the side of saying things too often, even with a risk of irritating the participants, than to err on the side of not saying things often enough and failing to properly create the conditions needed to create useful data.

(Please notice that I wrote "three times" three times in the preceding paragraph. If you even noticed the repetition, did it bother you? Probably not—and you are attending to the topic right now. Experimental participants certainly are not bothered by this kind of repetition, especially when it does not take place in just one paragraph!)

Here is an important rule about creating experimental manipulations: SUBTLETY IS OUT OF PLACE IN EXPERIMENTAL DESIGN. I trust I do not need to repeat that point. Sometimes investigators try to create subtle manipulations in a misguided attempt to preserve "naturalness" or, they think, to avoid drawing participants' attention to hypotheses under test. The problem with subtlety is that it goes against the goal of creating a situation that instantiates conditions and variables of the hypotheses. Subtle elements of a situation are missed by some people and can be interpreted differently by different people. That means that if the manipulations are subtle, some participants will fail to notice them (and thus will not be in the situation the researcher thinks they are in), and some will interpret them differently. Both those effects will introduce variance in the data because people will be responding to different sorts of situations.

For instance, I once heard about an experimental design in which the researcher was interested in whether white participants would play a competitive game differently when they thought their opponent was white than when they thought the opponent was black. Because participants would never see the opponent, who in fact existed only as a computer program, the researchers intended to identify the opponent's skin color by giving him what they thought was a "typically black" or a "typically white" name. These researchers wrote that they did not want to explicitly identify the opponent's ethnicity for fear that it would activate either stereotypes or concerns about appearing egalitarian.

The problem is that the researchers do not know whether participants in the experiment will code the names they chose as revealing ethnicity; probably

some would and others would not. Worse, many participants may not even attend to the name of the opponent; who cares about his name if we are never going to meet? If it is important that participants classify their opponents in the game, then good experimental design makes that element unmistakable.

Generally, the more full a picture the researcher can paint of an element, the better the design. For instance, in the preceding design, the researchers could identify partner's skin color with an instruction such as "Your partner today is named ___. He is, like you, a white student here at State University." That, at least, is clear and unambiguous. However, to really activate any behavioral tendencies participants may have so that they may be seen in this situation, it would be even better to show a photograph, or a videotape with action and speaking cues, to instantiate this variable. The clearer and the more complete the instantiation of important design elements, the better.

Knowing who the participants are, in terms of their background and culture, is also useful in creating the manipulations of an experiment. Making the situation presented in the cover story relevant to the participants creates a more believable situation and one that they are more likely to take seriously. Students at elite universities, for example, might be more motivated by studies presented as furthering basic science, whereas those at less elite schools may be more focused when the study purports to help them learn something about themselves. Participants who are not used to the laboratory setting may need more friendly and repetitious instructions.

D Manipulation Checks

One of the greatest strengths of laboratory experiments is the control the researcher has over the independent variables. However, researchers often fail to fully realize the potential of their experiments because they do not create the situation they intended to create. Careful experimental manipulation is important but not difficult. One tool all researchers should employ is the manipulation check.

Manipulation checks can take several forms. During pretesting (discussed later), experimenters should be sure to discuss with participants what they heard, how they interpreted it, and how it affected their behavior. In addition, a part of all experiments should be a questionnaire or interview (or both) in which the participants are asked about the experiment. The experimenter should verify that the information given to participants during the cover story was heard correctly and believed. This check should include any embedded information about partners, the task, the situation, or any other manipulations.

III "THE GENDER EXPERIMENT": A PRACTICAL EXAMPLE OF ABSTRACT CONSIDERATIONS

It may be easiest to understand how these abstract considerations look in an experiment by examining an actual experiment that has been conducted. I discuss

an experiment I designed and conducted (Rashotte, 2006). I use an example of my own work not out of ego but, rather, because it is only for an experiment of my own that I will know fully the considerations and decisions that went into its design, pretesting, and conduct.

I designed an experiment to examine how to control status beliefs associated with gender. This experiment was intended to test several related hypotheses from the status characteristics branch of the expectations states theoretical research program within sociology. I was not concerned with *whether* gender was associated with status for my participants; rather, I wished to demonstrate that *when* status beliefs were present, they could be controlled through certain mechanisms described in the theory. Thus, I made sure that status beliefs—favoring men or favoring women—were present in every condition of the experiment.

The theory posits a number of mechanisms that might allow general status beliefs to be overcome. I tested two in this study: (1) by presenting status information about a task that contradicts the generally held status beliefs (e.g., saying that women are generally better at the task at hand); and (2) by providing specific evidence that the generally held status beliefs do not hold for these individuals (e.g., saying that although men are generally better at the task, in this case the particular male is not very good at it and the woman is exceptionally good at it).

A Standard Protocols

To design this experiment, I began with a design that has been used in dozens of previous experiments and thus had a number of known properties, a variant of the standard experimental situation described by Berger in Chapter 12. The task at hand, the delivery of experimental instructions, and the cover story have been well-established over decades of research. Technological advances have allowed for recent improvements as well.

B Variables and Conditions

My independent variables were gender of participant, status information regarding gender and performance on the task, and performance feedback. Participants always interacted with purported partners of the opposite gender. In some conditions, participants were told that males would do better at the task to be completed; in others, they were told women would do better. In certain conditions, participants were given (fictional) feedback on a pretest.[1] I was interested in comparing the effect on my dependent variable of the general information that

1. All participants completed the short trial version of the task in order to maximize comparability among the conditions. Only in the "feedback present" conditions were participants given (fictional) scores for the trial version.

women did better versus the specific feedback that the female partner did better (but not how the two combined). I thus needed six conditions:

- Male participants, told males generally do better, with no feedback
- Male participants, told females generally do better, with no feedback
- Male participants, told males generally do better, but with feedback that the female partner did better
- Female participants, told males generally do better, with no feedback
- Female participants, told females generally do better, with no feedback
- Female participants, told males generally do better, but with feedback that the female participant did better than her male partner.

My dependent variable was how often the participants deferred to their partners in making decisions on the task when the partner disagreed with the participant.

C Manipulations

Participants were brought into an isolated room containing a computer monitor, a television, and a video camera. They were told that the study would begin when everyone was settled into the various rooms (leading them to believe there are real other participants nearby)[2] and that, when the time came, they would need to look into the camera to introduce themselves.

The participants then saw a videotape of instructions (said to be live via closed circuit television). The instructions were presented by a "Dr. Gordon" who claimed to be an expert in the task at hand and the ability underlying good performance at that task. The tape included a "live" introduction from their "partner" and a chance for the participant to introduce him- or herself, at which time the participant appeared on the television in the room. The introductions included information about the school attended (always our institution, to equate on that status variable)[3] and hobbies, to make the partners seem more real to the participants.

The instructions delivered all three of the independent variables. The gender of the partner was first introduced when Dr. Gordon said, "I see we have two people working together today, a man and a woman," and reinforced by seeing the partner on screen and by the partner reporting gender stereotypic hobbies. The status information ("Previous studies have shown that men/women are generally better at this task") was repeated three different times, including once just before the data collection phase began. The feedback information was provided

2. In fact, I usually did conduct several participants at the same time in order to support the belief that they were interacting with a real other, even though in reality each was interacting with the same fictional partner.

3. Participants were also similar to their partners on race and age in order to eliminate other status effects.

in those two conditions by a colleague of Dr. Gordon's, "Ms. Mason," who was an expert at scoring performance at this task. Ms. Mason repeated the scores, and their meanings (unusually high or unusually low), three times. Ms. Mason was also videotaped but said to be just down the hall.

D Manipulation Checks

I conducted several kinds of manipulation checks. The interviews mentioned previously were conducted in order to verify that the participants heard all of the relevant information regarding the independent variables and that they understood the task they completed. The interviews were also used to determine if the scope conditions of the theory were in place. During a pilot testing phase of the study (more on this later), these interviews were even longer and covered other topics, such as whether "Dr. Gordon" was pleasant yet scholarly, if the session was an appropriate length, and how much of the instructional detail the participant could recall (beyond the basics related to the independent variables).

Participants also completed a questionnaire just prior to the interview. The questionnaire served as a double check to the interview and also provided some guidance for the experimenter in terms of where there might be some problems with a particular participant. The questionnaire covered factual information as well as impressions and affective responses. Extreme emotional responses can indicate a participant for whom the study was problematic and not properly prepared, and they can be competing processes to the ones of theoretical interest in the study.

IV PRETESTING AND PILOT TESTING

In addition to the abstract considerations described previously, pretesting and pilot testing are important elements of good experimental design and conduct that are, unfortunately, sometimes overlooked by researchers. Pretesting involves examining certain elements of the experiment in isolation; pilot testing involves conducting complete experimental sessions with an eye to what is and is not working as expected.

Both pretests and pilot tests are different from actual experimental sessions in that the participants are required to act as informants. They let the researchers know what works and what does not work in the cover story, the task and/or interaction situation, the data collection, and all other parts of the experiment. This information can be gathered through questionnaires, interviews, or free-response surveys; ideally, it is obtained from several methods. They allow the researcher to fix unanticipated problems in the design.

The main elements of the design that need to be examined during both pretesting and pilot testing are scope conditions of the theory, the initial conditions of the experiment, and the instantiation of the independent variables. Researchers need to find out from the participants if the scope conditions are

holding in order for the theory's predictions to have any validity. The initial conditions, the cover story, must be understandable and believable for the data to have any value. Thus, for example, some groups of participants may require that information be repeated more than three times to be comprehended. The independent variable(s) must be clear and reasonable to have an effect on the dependent variable(s).

In addition, researchers must ascertain that the measures are working as expected. Participants must be paying attention, so the task must be somewhat interesting to them. Usually, experimenters wish to have a task that challenges participants without being so difficult as to cause undue frustration. If participants become distracted or emotional, their behavior may reflect those effects rather than effects of the theoretical factors as expected by the experimenter. The measures—especially key measures of the dependent variables—must be both valid and reliable.

Participants must buy into the cover story, the situation, and the task. They must also believe that the experiment has some importance—if not to them personally, then to the researchers and to science generally. Cultural factors come into play here. Experimenters must determine, through pretests and pilot tests, how to frame the situation in order to get their participants to believe it and want to take it seriously. Different populations of participants will require different frames for the cover story, and often it is not possible to know how participants perceive the encounter until pretests and pilot tests are conducted.

It is also in the process of pretesting and pilot testing that experimenter effects (discussed more fully later) can be identified. By having multiple experimenters conduct pretests and pilot tests, with thorough double checking, experimenter expectancy and observer effects can be identified early, before they can contaminate the data collected. Technology can be introduced when needed to reduce experimenter–participant interaction. Double-check systems can be implemented, and training can be increased if necessary. (For more on training, see Shelly, Chapter 4, this volume.) New measures, less participant to observer effects, can also be introduced if other steps are not effective at minimizing experimenter effects.

When pretests or pilot tests show things are not working as expected, experimenters must determine what to fix and how to fix it. It is much easier to determine *that* things are going wrong than it is to determine just *how* things are going wrong. Pretesting various elements of the design, prior to starting pilot tests or after, can allow experimenters to isolate where the problems are occurring. Thorough interviews with participants, with specific questions related to important elements of the experiment such as scope and initial conditions, can also help pinpoint the issues.

Once identified through pretests or pilot tests, problems must be fixed by the researcher. Issues with scope and initial conditions can usually be corrected by altering the cover story to resonate more with the participants. Issues of hearing, comprehending, or believing the independent variables can be corrected by rewording—and/or reiterating—the information. If the measures of the dependent variable are not working as expected, new measures can be added to or used to replace existing measures. If participants are reacting in an emotional way that is distracting them from the experiment, additional information must be provided to the participants to help them contextualize what they are experiencing.

A Pretesting

Pretesting most frequently is used for experimental instructions, but it can also be used for tasks, confederates, and instruments. These elements are isolated from the rest of the experimental setting, and participants are asked to evaluate them independently. The important considerations are often the means of conveying the situational definitions and all of the interaction elements. Pretesting is essential to ensure that the abstract and theoretical concerns are translated into a practical reality for the participants.

Again, I use an example from my research to illustrate details about pretesting. In a different but related experiment from the one described previously, new actors were used to portray "Dr. Gordon" and "Ms. Mason." To be sure that this new portrayal created the right situation, and created the desire on the part of the participants to be focused and serious, pretests were conducted just on the videotape of instructions (Rashotte, Webster, & Whitmeyer, 2005). In fact, only a short 10-minute segment of the tape was tested (this is long enough to get a sense of the situation but not so long as to require students to be brought into the lab to view it).

Dr. Gordon and Ms. Mason were intended to be authoritative and pleasant and to hold the attention of the participants. The short segment of tape was shown to students in classes at the same university where the experiment was to be conducted. Students also saw a 10-minute segment from another experiment previously conducted in which the "host" experimenter on camera had been shown to be effective at creating the right situation. The previous segment was sort of a control condition. By that, I mean it had been used successfully in prior research, and our question was whether the new segment was at least as good as the established one.

Students rated each person they viewed on 40 seven-point semantic differential items. An ideal answer for each item was determined by the researchers (though not shared with the students doing the rating). A final open-ended question was also presented to allow the students to raise any issues that might not have already been covered. The 40 items were classified into four general categories: authority and competence; absence of distractions; clarity; and serious manner. Comparisons between the mean ratings for each individual to the ideal rating showed that the new Dr. Gordon did not perform as well as the previous experimental host, but the new Ms. Mason performed satisfactorily.

The particular failings of the new Dr. Gordon indicated that the problem was with the demeanor of the actor and not with the instruction script. Thus, a new

tape was produced with a different actor (in fact, it was the same actor who was in the tape from the previous experiment). This tape was also pretested, and the ratings were then compared with those from the previous experiment and the original actor. Things were then satisfactory, and the experiment could proceed. Since then, we have used the pretest technique on additional actors playing the experimenters' roles and have found good results; these new actors can be used for future experiments using this design.

B Pilot Testing

Once the various elements of the experiment have been pretested, pilot tests can begin. Pilot tests are complete experimental sessions, designated "test groups" or "test sessions," in which the researcher spends additional time questioning the participants about their participation. Pilot tests can identify any problems that did not arise in pretests because pretests often only examine segments of the experiment in different settings. Pilot tests are like dress rehearsals in the theater: everything is done together, in sequence, to judge how it looks as a whole. Also like rehearsals, sometimes results show that minor changes are needed, and other times everything is fine and the design can be used for the rest of the run of the show or of the experiment.

It is not until pilot tests that competing processes are usually discovered. Competing processes include fatigue, hostility, and withdrawal. Experiments that are too long—from the participants' points of view—or take place at the wrong time of day can lead to fatigue, which can cause the participants to be less focused and less serious about the session. When part of the cover story involves providing the participants with information about themselves or others in terms of abilities or other such characteristics, emotional responses can occur. Sometimes this leads to anger and hostility; sometimes it leads to sadness or other negative affect and withdrawal on the part of the participants. If these emotional reactions are not a relevant part of the experiment (i.e., if they have nothing to do with the theoretical derivations under test), they can be distracting and lead to corrupted data. Thorough questioning of pilot test participants can lead researchers to detect competing processes.

After pilot tests are completed and all identified problems addressed, the actual run of the experiment can begin. Once problems are fixed, sessions may be called "experiment" rather than "test groups." If no problems arise and no changes are made to the procedures of the experiment, the "test groups" can be reclassified "experiment groups" retroactively.

V ANALYZING AND INTERPRETING DATA

The final stage of an experiment is the analysis and interpretation of the data it produces. Here, I do not address statistical methods for experimental data generally because those have been well covered elsewhere and most social scientists

are well versed in them. However, there are two elements of data interpretation that I believe are frequently overlooked by researchers. First, power analyses are often skipped altogether, which may lead to researchers missing evidence that their hypotheses are supported, even in an otherwise excellently designed experiment. Second, experimenter effects must be considered during data interpretation in order to rule out competing explanations for one's findings.

A Power Analyses

Statistical power analyses are easy calculations that allow one to determine the number of participants that will be required in an experiment in order to reliably detect meaningful differences in the dependent variable. Calculators to determine statistical power are readily available online.

Statistical power is best thought of as the likelihood of not making a Type II error (failing to reject a null hypothesis that is not true). As you reduce the chance of making a Type II error, you increase the statistical power and thus the test is more sensitive (Keppel, 1991). The likelihood of a Type II error can be lessened by having a sufficient number of participants in the experiment and reducing the variability within conditions.

Statistical power depends on three factors: the significance level α (representing the probability of making a Type I error, or rejecting a null hypothesis that is true); the magnitude of the differences across conditions on the dependent variable(s); and the sample size n (Keppel, 1991). Most often, researchers are only concerned with the sample size because the effect sizes are predicted by the theoretical constructs and α is set by convention.[4] Thus, many of the statistical tools that have been developed, including those online, are geared toward determining needed sample sizes.

Researchers should conduct power analyses to ensure that the data will be useful. If one does not have enough participants in each condition in order to detect the differences between the conditions on the dependent variable, then all will have been for naught. The calculation of "how many is enough" requires knowing the expected differences between conditions, the variability within conditions, and the desired level of significance.

For example, let us think about a simple experiment. This experiment has only two conditions, and the dependent variable is measured as a proportion. The value of the dependent variable for each condition will be compared to a fixed value, 0.60. The predicted mean of the dependent variable is 0.65 for condition 1 and 0.54 for condition 2, with a standard deviation in each condition of

4. The significance level that is used varies across disciplines and also according to the kinds of data available. Most experiments in sociology and economics currently use 0.05; psychologists also use 0.05, and sometimes 0.01, as do researchers in education. In political science experiments, 0.05 is often used, although for nonexperimental work in which samples may be smaller, 0.10 or 0.15 is sometimes used.