Kim, J., & Mueller, C.W. (1978). *Introduction to factor analysis*. Beverly Hills, CA: Sage.
Lissitz, R.W., & Green, S.B. (1975). Effect of the number of scale points on reliability. *Journal of Applied Psychology, 60*(1), 10–13.
Mauro, J. (1992). *Statistical deception at work*. Hillsdale, NJ: Lawrence Erlbaum.
McDermott, S.T. (1999). Quantitative/sampling. In D.W. Stacks & J.E. Hocking (Eds.), *Communication research*, 2d ed. (pp. 209–232). New York: Longman.
Morse, R. (1998, June 2). Back to school on shootings. *San Francisco Examiner*, p. A2.
Murphy, D.J. (1992). Electronic communication in smaller organizations: Case analysis from a theoretical perspective. *Technical Communication*, 39(1), 24–32.
Murphy, D.J. (1994). *NASA/DoD aerospace knowledge diffusion research project*. Report Number 30. Washington, DC: National Technical Information Service.
Murphy, D.J. (1997). The influence of analyzability, equivocality, uncertainty, and variety on communication in small, medium, and large U.S. aerospace corporations. In T.E. Pinelli, R.O. Barclay, J.M. Kennedy, & A.P. Bishop (Eds.), *Knowledge diffusion in the U.S. aerospace industry* (Part B; pp. 581–610). Greenwich, CT: Ablex.
Neuman, W.L. (2000). *Social research methods: Qualitative and quantitative approaches* (4th ed.). Needham Heights, MA: Allyn & Bacon.
Nielsen Media Research. (1999). What TV ratings really mean ... and other frequently-asked questions. http://www.nielsenmedia.com/whatratingsmean/.
Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
Rummel, R.J. (1970). *Applied factor analysis*. Evanston, IL: Northwestern University Press.
Schutt, R.K. (1996). *Investigating the social world*. Thousand Oaks, CA: Pine Forge Press.
Stacks, D.W., & Hocking, J.E. (1999). *Essentials of communication research* (2d ed.). New York: Longman.
Sumser, J. (2001). *A guide to empirical research in communication*. Thousand Oaks, CA: Sage.
Young, R.K., & Veldman, D.J. (1981). *Introductory statistics for the behavioral sciences* (4th ed.). New York: Holt, Rinehart and Winston.

CHAPTER

Experimental and Quasi-Experimental Research

Davida Charney

ROLES FOR EXPERIMENTAL METHODS IN TECHNICAL COMMUNICATION RESEARCH

Our reasons for conducting research in technical communication are often practical and progressive. We seek to understand how communication works in technical and professional settings in order to make things better: to promote text designs that are easy for readers to use, to acculturate students into professional discourse communities, and to identify and promote effective and ethical communication practices in the workplace. All these goals may be furthered through experimental research methods. Technical communication researchers have used experiments to investigate such questions as these:

- Are structured abstracts for medical research articles easier to read than traditional abstracts? Do readers of medical journals prefer structured abstracts? (Hartley & Sydes, 1997).
- Are job recruiters willing to grant interviews to an applicant whose résumé contains grammatical errors, nominal sentence style, or irrelevant details? How do these features interact in forming a recruiter's judgment of a résumé? Do undergraduates assess these factors in the same ways as recruiters? (Charney, Rayman, & Ferreira-Buckley, 1992)
- As compared with working face-to-face, how does using electronic communication technologies from distant locations change the way a business team produces a written text, the quality of their report, or their satisfaction with the project? (Galegher & Kraut, 1994)

- Does oral feedback on a text differ from written feedback? When reviewers audiotape their feedback, do they provide comments on different topics or comments of different length, content, or style as compared with when they write out their comments? How do comments in oral or written media affect writers' revisions and attitudes toward the reviewers? (Neuwirth, Chandhok, Charney, Wojahn, & Kim, 1994)

Experimenters investigate the possible causes of a phenomenon. Experimenters work systematically to create situations in which different possible causes are present or absent. Then they observe how typical groups of people respond in each situation. In a "true" experiment, the situations are designed to be as similar as possible, differing only in the presence or absence of the causal factor under investigation. In the first example, some readers were given structured abstracts of a set of medical journal articles, while others read traditional abstracts of the same articles. Achieving the conditions of a "true" experiment often involves a high degree of control over the setting so that, for example, equal numbers of randomly chosen participants try out each causal factor. Quasi-experimental methods were designed for real-world settings where the controlled conditions necessary for "true" experiments are more difficult to arrange.

RESEARCH METHODS AND RESEARCH CLAIMS

To understand how experiments differ from other research methods, it is useful to think about the types of questions that research is used to address. A useful typology of research questions is the classical rhetorical system of stases, a sequence of questions that guides critical inquiry from the point when attention is drawn to a phenomenon to the point when we decide what to do about it. In Jeanne Fahnestock and Marie Secor's (1988, 1990; Fahnestock, 1986) formulation of the stases, five questions can be investigated:[1]

- Existence—whether a phenomenon exists or happened
- Definition—whether it belongs to some established (albeit fuzzy) category
- Cause—how it came about or what effects it has
- Value—whether it is to be considered good or bad
- Action—what should be done about it

The stases were originally formulated for use in courts of law, where lawyers argued over whether a crime took place, "whodunit," and so on. In recent years, rhetorical theorists have found the stases useful for analyzing and constructing arguments in a wide range of domains, private, public, and professional.[2]

The stases form a sequence; arguments at later stases build on consensus established from arguments at the earlier ones, even if agreement is provisional or accounts for only part of the audience.[3] The stases can be seen at play in a recent sequence of news stories about an astronomical phenomenon. Astronomers gained front-page newspaper headlines when they reported detecting a mysterious flash of light. They first sought grounds for agreeing that the flash in fact took place out in space and was not a local atmospheric effect. They worked to establish the timing, location, and intensity of the light, all claims at the stasis of *existence*. Then considerable debate ensued as astronomers attempted to determine what the flash was, to *define* it. A leading researcher eventually concluded that it was a special form of quasar, not a star and not something completely new. Finally, astronomers argued over what *caused* this quasar to behave differently from other more common quasars.

As a guide to inquiry, the sequence of stases is recursive and open ended rather than strictly linear. Progress in the sciences often involves revisiting earlier stases. For example, many years after agreeing that genes are *causal* agents in heredity and evolution, biologists launched the Human Genome Project to identify (or *define*) the sequence of genes in a human chromosome. The results of this project are expected to help researchers conduct further experiments about the causes of particular diseases and the effects of various medical treatments.

In some discourse contexts, such as a law court or a public policy debate, arguments are expected to range across the entire sequence of stases. But the entire sequence need not be addressed; a scholarly study often addresses an argument at a single stasis. Fahnestock and Secor (1988) analyzed the main argumentative focus of scholarly articles in literary and scientific journals. In literary criticism, they found that arguments often hinged on the stasis of value (such as the relative worth of an author or a text). In scientific articles, the argument focused on one of the three lower stases: existence, definition, or cause.[4]

Within the framework of the stases, research methods can be seen as standardized approaches that a discipline agrees on as suitable for supporting claims of certain types. Specific research methods are warranted for supporting arguments at specific stases. In most disciplines, descriptive and observational methods, whether qualitative or quantitative, including case studies, ethnographies, surveys, process-tracing, and textual analyses, are warranted for supporting claims of existence and definition. Experimental and quasi-experimental methods were developed specifically to warrant causal claims. Because the causes of a phenomenon often cannot be explored in a meaningful way until its nature is understood

somewhat, descriptive and observational methods often precede experimental research. But, as noted earlier, research is best understood as recursive, with results at a "higher" stasis often leading to further work at a "lower" stasis. Similarly, within composition studies, Bereiter and Scardamalia (1987) describe how different "levels" of writing research (including practitioner observation, experimentation, and simulation) interrelate cyclically.

It is important to emphasize that using a standard method does not guarantee that findings will be accepted as true or important, merely that the arguments will be treated as worthy of serious consideration. Using a standard method of research opens up the work to scrutiny by others who are familiar with either the methods or the phenomena (Charney, 1996). In his history of the research article, Charles Bazerman (1988) describes how specific methods evolved as ways to anticipate and respond to challenges to experiments that were initially conducted in very public arenas. In the course of extending and challenging each other's work, specialists in an area develop, apply, and refine a repertoire of methods that they consider appropriate for certain types of inquiries. Over time, some methods gain credibility on the basis of their elegance and their reliability across many applications. When scientists introduce new and unfamiliar methods, they argue at length for their reliability and productivity (Thompson, 1993). But even though researchers rarely have to justify using a standard method, they do have to argue that they applied it appropriately because almost every application involves creativity and hard choices. Over the course of time, even standard methods are subject to challenge; advances in knowledge, technology, or methodological standards frequently alter the credence that scientists invest in specific methods and their interpretations of studies that employed them.

Experiments as Causal Inquiry

The basic strategies for causal inquiry derive from John Stuart Mill (1843, 1930) who advocated experimentation, active manipulation of a situation to observe the effects of an intervention in controlled circumstances. Mill proposed four methods for causal inquiry:

- Agreement: searching for relevant factors (candidate causes) that are always present before the outcome (or effect) occurs
- Disagreement: searching for a single relevant difference between situations, such that the factor is present whenever the outcome occurs and absent whenever it doesn't

- Concomitant Variation: searching for relevant factors whose strength or frequency is positively or negatively correlated with the outcome[5]
- Residues or Elimination: identifying the roles of multiple factors by systematically removing known causes to see if the outcome continues to occur

The first three methods, agreement, disagreement, and concomitant variation, may be carried out by observing natural events. A researcher may seek out all possible cases in which an outcome occurred and analyze which factors were present and which were absent. The final method calls for experimentation. Experimenters try to create conditions in which the outcome occurs and then vary those conditions to test the contribution of individual factors.

To see how factors can be tested systematically, consider a series of experiments conducted by psychologist Lynne Reder. She started with a seemingly simple question: Do students learn more of the central ideas in a standard introductory college textbook by reading fully elaborated texts (including explanations, evidence, restatements, and so on) or by studying the main ideas in isolation? Reder took chapters from several widely used textbooks and prepared summaries that were one-fifth as long. In her first experiment, she gave half the students the original chapter to read and the other half, the summary. Then she tested the students' comprehension and recall of the main points. Reder was surprised to find that the students who had read the summaries performed better on the tests than those who had read the full chapters. Her colleagues suggested a wide number of accidental reasons why she might have gotten these results. She eventually completed ten studies checking out these factors (Allwood, Wikstrom, & Reder, 1982; Reder, 1982; Reder and Anderson, 1980). She varied how soon the test was administered (immediately after reading or after delays of up to one year); what type of test questions were asked (true/false, short answer, or free recall); and how the outcome was measured (accuracy or speed). She also varied the reading conditions. In one study, students were allowed to take the materials home to read at their leisure; in other studies, the duration of the students' exposure to each main idea was carefully equated. Consistently, students who read the summaries learned the main points better. One factor that Reder did not vary in these studies was the kind of learning expected of the students. However, remembering facts is only one goal of learning. Students also need to know how to use new knowledge to solve problems. In our research together, Reder and I investigated whether a full text would produce better results when readers needed to *apply* what they learned. We prepared full and summary versions

of a manual for a computer operating system and asked students to learn a set of basic commands. We found that students who had read manuals with certain types of elaboration performed better at a set of computer tasks than those who had read summary versions of the manuals (Charney, Reder, & Wells, 1988).

In their classic text *Quasi-Experimentation*, Thomas Cook and Donald Campbell (1979) explicitly relate experimental methods to causal argument and sensitively address recent philosophical concerns about causation (see also Cook & Shadish, 1994). Like many other researchers, they are cautious about making or accepting causal claims—and with good reason. As psychologists have repeatedly observed, most people are overly eager to infer causal connections between phenomena that simply occur together (perhaps by accident) and overly reluctant to say that some factor could not be a cause of some outcome, even when sufficient evidence for ruling it out is available (Kuhn, Amsel, & O'Loughlin, 1988). Because mistaken causal inferences can have serious social as well as scientific consequences, the standards for designing and reporting experiments are intended to encourage self-critical reflection and to maximize opportunities for public scrutiny of both methods and results.

Toward these ends, researchers have developed sophisticated protocols for conducting experiments. Cook and Campbell (1979) provide an excellent discussion of the general classes of experimental and quasi-experimental research designs. They also lay out some bases for judging an experiment by identifying a number of "threats" to validity that researchers routinely consider in designing their experiments and by explaining how specific research designs and practices can minimize these threats. Some of these strategies are sketched in the following sections (see also Shaughnessy, Zechmeister, & Zechmeister, 2000; Slavin, 1992).

PRINCIPLES FOR EXPERIMENTAL RESEARCH

An experiment is a comparison between a situation in which a putative causal factor is present and a situation in which it is absent. The aspects of the situation that are varied are called *independent variables*. Each independent variable has at least two levels, to reflect the presence or absence of the cause. For example, in Reder's (1982) study, the text variable had two levels: either many elaborations or no elaborations at all. Independent variables can also have more than two levels, allowing different degrees or kinds of the causal factor to be present. In our studies of computer manuals (Charney, Reder, & Wells, 1988), the text variable had several levels,

including no elaboration, explanations of the syntax of the computer commands, and elaboration of the reasons for using the commands.

Other aspects of the situation, called *control variables*, are held constant or equated. In Reder's experiments, students were asked to learn the same set of main points, regardless of whether they saw them in the full chapter or in the summary. Reder also took steps to ensure that the summary and full-text groups were equally typical of the student body as a whole and that they studied the texts under similar conditions. Ideally, all aspects of the situation except the independent variables are controlled. But control does not always mean conscious regulation. Paradoxically, a variable can be controlled by letting it vary as freely or randomly as in natural settings. For example, if enough participants are chosen randomly from among the students on campus, they will vary naturally in height, weight, religious affiliation, amount of sleep the previous night, and so on.

The sequence of events in which the participants are presented with the materials is called the intervention or treatment. Participants carry out tasks in which they perform the activities or form the attitudes or beliefs that the independent variables are hypothesized to affect. The tasks might include reading, writing, answering questions, solving problems, or using a device. The effects of the independent variables are measured by tests administered before, during, or after the treatment. The ways in which performance on the tests is evaluated are called *dependent measures* or *dependent variables*. These may include correctness of response, speed of response, strength of preference on an attitude scale, quality judgments by raters, or frequencies of occurrence within a participant's response (e.g., length in words or references to the audience in a written passage). In Reder's (Allwood, Wikstrom, & Reder, 1982; Reder, 1982; Reder & Anderson, 1980) studies, the students' learning was measured in various ways, with different types of questions (true/false, short answer, or free recall). She measured how long students spent reading the texts, how long they took to answer individual questions, how many of their answers were correct (on true/false tests), and how many of the main ideas they wrote down (on free recall tests).

In analyzing the results, experimenters do not usually produce profiles of individual participants. Instead, they try to characterize the central tendency of each group, its average or most representative behavior pattern, as well as its range and variation. Reder's (1982) report that readers of summaries learned more on average than readers of full-length chapters was based on consistent findings that the average correctness score for the summary group as a whole was higher and their average response time was shorter. Statistical analyses are used to compare the patterns of scores to

determine whether differences observed between the groups are robust enough to warrant a claim that they were not produced by chance, but instead were caused by the factor under study (Abelson, 1995). If so, the results are considered *reliable* or *statistically significant*. Statistical significance is expressed as a probabilistic confidence judgment rather than an assertion of fact. If none of the results of a study are statistically reliable, the researchers may modify the experiment to try to create conditions in which the results are reliable. Or they may ultimately conclude that the initial hypothesis was not tenable or that that particular experimental approach is not viable.

If the only systematic differences in the treatment of the groups are the levels of the independent variables and if there are reliable differences between the groups on the test scores (the dependent measures), then researchers may claim that changing the independent variable caused the difference in results. The results of a single experiment are rarely completely clear-cut. As described previously, Reder and her colleagues (Allwood, Wikstrom, & Reder, 1982; Reder, 1982; Reder & Anderson, 1980) conducted a series of ten experiments to check whether the superior performance of the summary group was due to specific aspects of how they had conducted the study. The consistent finding that the summary group performed better under a variety of conditions increased their confidence that isolating the main ideas in the summaries made them easier for students to understand and recall. At that point, experimenters weigh the pragmatic importance of the results for disciplinary or real-world issues. For example, if Reder had found that reading summaries reliably increased an average student's score by only 1 percent, this result might be considered real but unimportant. As Linda Flower (1989) has noted, statistical evidence has meaning only as part of a cumulative, communally constructed argument, in which "the special virtue of a claim that has earned the name 'result' is that it has been subjected to a given research community's more stringent rules of inference" (p. 300).

Experimental Designs

One common way to set up an experiment is to assign participants randomly to different treatments, in a *between-subjects* design. Then one (or more) group of participants receives *experimental* treatments, treatments that are hypothesized to cause a change. Reder's (1982) study employed a between-subjects design, because one set of students was assigned to read the summary and a different set to read the full-length chapter. A between-subjects design may also involve a *control* group. The control group

participates fully in the study and is treated as similarly as possible to the experimental group, but does not receive a treatment that is expected to cause a change. In medical research, for example, patients in a control group may be given placebos (sugar pills) instead of medication. The control and experimental groups may receive their pills on exactly the same schedule, with neither patients nor staff members aware of who is receiving medication and who is receiving a placebo.[6] The control group provides baseline information about how this medical condition might proceed without medication, but with the same experiences as the medicated group of other aspects of health care, confidence in their doctors, and so on. A challenge for between-subjects designs is forming equivalent groups of participants. In many cases, randomly assigning participants to groups is sufficient; however, many research designers also recommend giving pretests so that the groups' starting points on the dependent measures may be directly compared.

A second way to design a study is to see how the same individuals respond to a variety of situations, both when the putative cause is present and when it is absent. In a *within-subjects* design, every participant eventually receives all the treatments. Fewer participants are needed in studies with within-subjects designs because the participants, in effect, serve as their own controls; they bring in the same mix of preferences, experience, and physical characteristics when they act in one condition as when they act in the other. The simplest way to implement a within-subjects design is to have each participant go through the same experiment twice. For example, in a study of the effects of familiar and distant audiences on children's writing styles (Cohen & Riel, 1989), each child took part in two sessions, in one, writing an essay to their teacher for a grade, and in the other, writing a letter on a similar topic to be posted to a child in a foreign country. Cohen and Riel found that when students wrote to distant audiences their essays scored higher in content, organization, and language use than when the same children wrote to their teachers. Cohen and Riel ruled out the possibility that the effects were due to the order of completing the assignments, by having half the students write to the teacher first and the other half write to the peer first.

A within-subjects design can be implemented within a single session. This kind of design can be illustrated with a study I conducted with two colleagues to investigate how writing features affect job recruiters' judgments of student résumés (Charney, Rayman, & Ferreira-Buckley, 1992). We asked job recruiters visiting campus to rate a set of thirty-six student résumés on a four-point scale, to indicate their willingness to interview the students for a job in mechanical engineering. The résumés were fictitious,

but their content was drawn from real job application materials. The résumés were carefully constructed to vary four factors: sentence style (nominal or verbal), grammatical errors (present or absent), elaboration (no elaboration, object-based description, and function-based description), and relevance of previous work experience (low, medium, and high). The final set of thirty-six résumés comprised one of every possible combination of these four factors, so the same recruiters rated résumés in every condition. Other aspects of the résumés were held constant, such as format, grade point average, and degree program. Each recruiter gave only one overall score to each résumé. However, by sorting the scores for résumés representing the different factors (a task simplified by statistical computer applications), we could see the independent effect of each factor. For example, we calculated the average score for the eighteen résumés with grammatical errors and compared it with the average score for the eighteen correct résumés. We found that all the factors influenced the ratings. Our design allowed us to assess the strength of each factor and see how they interacted; for example, recruiters were sometimes harsher in penalizing mechanics errors from students whose résumés listed more relevant work experience.

Between-subjects and within-subjects factors can also be combined in a single study, for example, if different groups of participants perform the same tasks. In a later version of the résumé study, we compared job recruiters' ratings with those of technical writing students. In this comparison, participant status (student or recruiter) was a between-group factor, and the résumé variables (grammatical errors, elaboration, relevance) were within-subjects factors.

Strengthening Confidence in an Experiment's Validity

Because researchers try to control so many aspects of an experiment, there is often good reason to question whether the results are valid, to ask whether an experimental treatment really caused an apparent difference in the outcomes. Even if researchers find a big difference in the performance of two groups, it might not be due to the intervention. It might have been caused by some other factor or it might be an accident. Or researchers might see no difference in the outcome measures, even though the intervention actually caused a change—perhaps it was a change that the outcome measures were incapable of detecting. To help experimenters anticipate such challenges and design studies that avoid them as much as possible, Campbell and others developed a general list of threats to validity (and possible recourses) that researchers now routinely consider as they plan experiments

(Abelson, 1995; Cook & Campbell, 1979; Cook & Shadish, 1994; Shaughnessy, Zechmeister & Zechmeister, 2000; Slavin, 1992).

History and Maturation. Unanticipated events may occur during the study. Or participants may simply change over time. Unless they are equally likely to affect all participants, these events and changes may produce spurious differences in the outcomes for the two groups. Randomizing assignment of participants to treatments, randomizing the order of tasks, and conducting studies in a controlled quiet setting can reduce these threats. For example, suppose that between the time Reder's (1982) students read the passage and the time they took the test, a large freshman dormitory on campus was flooded and the students who lived there went without a good night's sleep. If many of the participants in the full-chapter group were recruited from this dormitory, their poor performance might not be due to the chapter, but due instead to their intervening history. But if students recruited across campus were randomly assigned to groups, then students from this dorm should be equally heavily represented in both groups, so the detrimental effects of sleeplessness should balance out.

Testing. A pretest may alter performance on a posttest, either because participants have a chance to practice or their attention is attracted to a key topic or strategy. Using a control group reduces the problem because the control and experimental groups should be affected equally by the pretest; differences detected on the posttest are then likelier to be due to the treatment.

Instrumentation. The quality of the test may obscure differences between groups in several ways. First, the test might vary during the study, if, for example, a study involves networked computers and there is great variation in the speed of response. Second, multiple tests might not be equivalent. If, for example, the posttest is harder or less interesting than the pretest, then participants' gains from the treatment might not be detected. Third, the tests might be too easy or too hard. If a test is too easy, then the scores all bunch at the top of the scale (a ceiling effect). If the test is too hard, the scores bunch at the bottom (a floor effect). The tests may spuriously show no difference between experimental and control treatments because there is no way for the improvement to register. Pilot testing can prevent many of these problems.

Selection Bias, Attrition, and Regression to the Mean. These threats reduce the chances that groups of participants are functionally equivalent, when the groups should start off on average at the same ability level, with the same overall mix of attributes. Assigning participants randomly to groups avoids the problem of *selection bias*, such as steering the most "promising" participants to the experimental group (consciously or accidentally), which would spuriously increase the chances of finding that the intervention "succeeded." *Attrition* is a problem if participants drop out of one group more than another. If, for example, many women dropped out of the experimental group, then at the end of the study, the control and experimental groups would no longer be equivalent. Any differences in the outcomes may be due to gender differences rather than the treatment. *Regression to the mean* describes the probability that people who score extremely high or extremely low on a test will score closer to the average if given another test. The problem arises if researchers try to compare treatment groups chosen from the extreme ends of a scale, because participants at the bottom may spuriously appear to improve and those at the top may spuriously appear to regress.

Nonrepresentativeness, Artificiality, Reactivity. Researchers are often concerned that their findings will only represent the behavior of the specific group of participants from the local setting in which the study was conducted. In planning their studies, they may take several steps to increase confidence that the findings generalize to a larger population. Any argument for generalization is an assessment of plausibility. To increase plausibility that the findings apply to a general group, researchers try to recruit participants who are representative, create conditions that are realistic, and use measures that are stable. Participants are *nonrepresentative* if they are, as a group, unlike the larger population of interest. It is not enough for all the members of the participant pool to be members of the larger population. They are nonrepresentative as a group if they do not have the same mix of attributes as any other sample chosen at random from the population. The costs of a nonrepresentative sample may be great. In 1987, the IRS user-tested its new W-4 tax form on nontechnical IRS clerical staff, who the IRS assumed knew no more about taxes than any typical taxpayer; however, these employees were far more capable of understanding the form than ordinary taxpayers (Gutfeld, 1987). The forms turned out to be unusable and were recalled and redesigned at great expense. Random selection is the best way to avoid the problem of nonrepresentativeness. *Artificiality* is the problem that the conditions of the study were so unusual that the same results

would not occur in other, more natural settings. Some artificiality is inevitable; researchers reduce the threat by conducting similar experiments under many kinds of conditions, as Reder (Allwood, Wikstrom, & Reder, 1982; Reder, 1982; Reder & Anderson, 1980) did. Some clearly artificial studies may be perfectly valid (Stanovich, 2001). For example, researchers studying the effects of visual feedback on writing style might legitimately ask students to compose on a computer with no monitor.

Tolerance for Threats to Validity. Beginning researchers who read experiments sometimes take validity as a single-elimination contest—as if finding any weakness whatever makes a study entirely invalid. But some flaws are more serious than others. No study can test every factor; the conditions can never be completely controlled. Robert Slavin emphasizes that readers must apply "educated common sense" to judge a study: the results of a seemingly perfect experiment are not guaranteed to be valid and the results of a seemingly flawed study may yet be strong enough to serve as a basis for further research (1992, p. 22). The indeterminacy of scientific methods does not mean that anything goes. Keith Stanovich (2001) notes that theories can be rigorously evaluated on the basis of a large number of partially flawed experiments, especially when the limitations of one are addressed in another (see also Charney, 1996). Meta-analysis is one way to see if an effect consistently occurs across experimental studies, such as George Hillocks's (1986) meta-analysis of various writing pedagogies.

Random Selection, Random Assignment, and Random Ordering

In contrast to everyday parlance, doing something "at random" in an experiment does not mean being careless. Instead, randomization means relying on the laws of probability to reproduce naturally occurring mixtures of attributes. Experimenters randomize for two main purposes: to select a sample of participants that is, in aggregate, representative of a larger population and to divide this pool into treatment groups that are, in aggregate, equivalent to each other. Experimenters also frequently randomize the order of tasks or materials to avoid the threat to validity of history. For all these purposes, randomization works on the principle of giving everyone an equal opportunity to be chosen.

The rationale behind randomization is that individuals are unpredictable. Members of a particular population, such as active NBA basketball players, share many salient characteristics because of selection criteria, training,

and acculturation. Even so, they vary in their personal traits, beliefs, politics, habits, moods, and current states of mind. Basketball players are obviously taller than male adults in the United States generally, but the heights of the individual players represent a range, with most players grouped around the average and a few who are quite a bit taller and a few quite a bit shorter than average. This kind of variation is what is "normal" about a "normal" bell curve distribution. To select a random sample of active NBA basketball players, we might take the official rosters of all the teams, start in an arbitrary spot, and select every sixth name. The resulting sample should mirror the range and variation of heights of the entire population, as well as their ages, fitness conditions, ethnicities, and a host of other attributes, none of which played any explicit role in the selection process. Random sampling does not ignore or suppress individual differences or commonalities; rather it treats them as too subtle and too complex to apportion and it gives them free play. If all members of a population have an equal chance to be selected, then common and rare attributes of the population should be represented as common or rare in the sample.

Some variations on random sampling can ensure that certain features of interest are included in the sample. Stratified sampling involves close analysis of some community in an effort to include some important constituency in representative proportions. So in constructing groups of taxpayers to test a new form, one might ensure that each group contained representatives of the various income brackets in the same proportions as the U.S. population. No matter how many categories are formed in a stratified sample, none may be reliably represented by only one person. The more individuals in the sample, the less any one participant may be mistaken as typical of the whole group.

Quasi-Experimental Designs

Quasi-experimental designs were developed for research in real-world settings where it is more difficult to randomly select participants or assign them randomly to conditions, such as schools, neighborhood literacy centers, nonprofit organizations, or workplaces. In a typical college classroom, for example, the selection of students is far from random; some take the course to fulfill a requirement, others out of interest, and others because it fits their schedule. The twenty-five students hardly represent a random sample of the entire student body, or of all seniors, or of all engineering majors. The considerations for avoiding threats to validity listed earlier apply even more strongly to quasi-experiments. As Cook and Campbell (1979) emphasize, because researchers using quasi-experimental designs

cannot rely on random selection to avoid some of these threats and cannot control the conduct of the study as closely, they must analyze the situation much more closely to make the possible threats explicit and find ways to reduce them. Slavin (1992) provides a useful discussion of strategies for reducing threats in research in classroom settings.

Control groups are frequently used in quasi-experiments. When researchers are unable to select participants randomly or assign them randomly to conditions, they usually collect more detailed information on the backgrounds of the participants and pretest their abilities, in order to check as much as possible on the equivalence of control and treatment groups. Some quasi-experimental designs use strategies similar to the within-subjects design, by providing a sequence of phases in which treatment is provided, withdrawn, and repeated, with performance tests at the end of each phase. An argument for causation can be made if the outcome measures reliably change when the treatment is present and reverts to control levels when it is withdrawn. Another strategy in quasi-experiments is to administer the outcome test repeatedly over a period of time before and after the treatment. Rising and falling test scores at the predicted times might then show the time-course of any effect of the treatment.

In and of itself, the setting does not determine whether a study should be designed as an experiment or quasi-experiment. In Cook's recent discussion of quasi-experimental designs (Cook & Shadish, 1994), he emphasizes that experimental methods are widely applicable in real-world settings. The choice of design is more likely to depend on the experimenter's degree of access, matters of timing and convenience, and other aspects of the situation.

ETHICAL TREATMENT OF PARTICIPANTS IN EXPERIMENTS

Some people object to experimental research because participants are taken as "objects" of study, which critics assume must be dehumanizing. These critics also object to the distant, impersonal stance that experimentalists adopt—as compared with ethnographers. Many of these criticisms essentialize experimental researchers too hastily as cold and uncaring (Charney, 1996, 1997). For most experimentalists, impersonality is intended as a form of ethical behavior that preserves participants' freedom of action. An impersonal stance minimizes the chances that a researcher will (even unknowingly) pressure participants to adapt to his or her predispositions, as in placebo effects. Experimentalists, like everyone else in the world, are prone to biases. Practices that promote objectivity cannot train researchers

out of their biases—rather, they reduce the effects of biases by limiting and systematizing interactions with participants and by making methods and results more available for scrutiny and replication by researchers with different sets of biases.

Experimental researchers plan out their interactions with participants ahead of time, and many even write out a "script" detailing all planned communications with participants—except, of course, for open-ended questions and comments. These plans and scripts avoid some threats to validity by ensuring that all participants are treated the same. They also prevent unethical exploitation of the participants. Written descriptions of how participants will be recruited and what they will be asked to do are submitted for external independent review to an Institutional Review Board (IRB), which at most universities is made up of faculty members and members of the public. Before any data are collected, the IRB checks the procedures for obtaining participants' informed consent, ensures that participants' right to privacy and right to withdraw are protected, and assesses the procedures' risks and benefits (and may require modifications). For a more detailed discussion of the IRB and its processes, see the Breuch, Olson, and Frantz chapter in this volume.

Although this planning and review process is the most visible arena for protecting the rights of participants, other important protection processes take place after the data are collected and an article is drafted or even published. At this point, the discussion moves from the institution in which the researchers practice to the discipline as a whole, where new standards of conduct might be developed. The method sections of an experimental research article opens the procedures to scrutiny by the research community at large, allowing problematic procedures to be challenged. For example, it is largely because of routine reporting of sampling procedures that feminists documented the unwarranted exclusion of women participants in some social science and medical studies, and it is because of such reports that ongoing reforms can be monitored.

Researchers must also take steps to protect participants when they describe their results. To protect participants' rights to privacy, researchers usually confine the data analysis to summaries of group tendencies, rather than to descriptions of individual participants. Because the scores are reported as averages rather than as individual scores, the participants in the study remain anonymous; their scores and their personal histories cannot influence future teachers or administrators. This approach limits the types of claims that researchers may make. Researchers may only make causal claims about how their factors influenced the tendencies of groups, not the

behaviors of individuals. Generalizations about the central tendency of a group are not distributive to the members; in other words, claims about the group as a whole are not assumed to hold of each member. For example, a study of the effects of a Head Start program may report that children from the most economically disadvantaged groups make the greatest gains. Some students in the most economically disadvantaged Head Start group probably make no gains at all, and the progress of some might seem to be held back. Predictions or judgments of outcomes for individual participants, such as the likelihood of academic success for any particular child in the Head Start program, are not warranted. Unfortunately, average results are sometimes taken as "normal," even though there is no basis for concluding that children who did not benefit from Head Start are "abnormal." Normative interpretations can be and are challenged in disciplinary as well as public arenas. In fact, statistical conventions for reporting average scores and variances are designed to help researchers and readers assess a group's heterogeneity. Using large numbers of participants and discussing group tendencies can thus be a way of respecting individual differences and a way of resisting totalizing or deterministic conclusions.

RESEARCH IN REAL-WORLD SETTINGS

The common assumption that experimental methods are not possible in real-world settings is incorrect. Even though Cook and Campbell are credited with developing quasi-experimental methods, they increasingly advocate conducting true experiments whenever possible (Cook & Campbell, 1986; Cook & Shadish, 1994). And they argue strongly that true experiments can be conducted in more settings than one might expect. They cite many successful studies that use random assignment of participants to treatments in field settings, including schools, housing developments, and clinics. Conducting experiments may take additional imagination, planning, and effort, but we should not choose methods or research questions primarily on the basis of convenience. All important research skills, from foreign language learning to statistics, require specialized training that may take years to master. Many interesting causal questions remain in technical communication and they are worth pursuing.

NOTES

1. For an alternative formulation of the stases for scientific discourse, see Prelli (1989). I have chosen to follow Fahnestock and Secor (1988, 1990) because

they treat cause as an independent stasis. This seems justified, especially in the current discussion, because of the substantial attention experimentalists devote to causal arguments. In most important respects, the two systems are compatible.

2. Notably, these strategies for inquiry are in no way restricted to scientific research. Similar techniques appear in popular argument textbooks for first-year composition (e.g. Fahnestock & Secor, 1990; Lunsford & Ruszkiewicz, 1999; Ramage, Bean, & Johnson, 2000).

3. Certainty about a claim at any stasis is not necessary in order to go on with an argument. In fact, research at a higher stasis is often conducted in order to address continuing disagreements about the nature of a phenomenon. As Fahnestock (1986) and many others have observed, claims in scientific texts tend to be probabilistic with qualifiers giving explicit signals of the degree of the writer's degree of confidence.

4. Fahnestock and Secor (1988, 1990) note that scholarly arguments almost always address value, at least implicitly. In order to persuade colleagues to read their work and take it seriously (by challenging it, replicating it, or building on it), scholars have to argue that the work is important and relevant to ongoing work in the discipline. (For an analysis of how they do so, see Swales, 1990.) Recently, some critical theorists have argued that scientists are remiss ethically for not addressing value arguments more explicitly—and some see this silence as acquiescence or complicity in inequity. I respond to these critiques at length elsewhere (Charney, 1996, 1997), arguing that scientists use other forums than research articles to address moral and ethical questions, that qualitative methods that seem to discuss ethical issues more explicitly do not necessarily avoid the problems, and that experimental methods should not be dismissed wholesale on ideological grounds.

5. Cook and Campbell point out that Mill drew a sharp distinction between correlation and causation—and so do experimentalists today. Correlational arguments describe the frequency of co-occurrence, such as the presence of fire engines at fires. Causal arguments require evidence of a direct connection between factors and outcomes, such as the presence of oxygen for fire to occur. For these reasons, in their discussion of Mill, Cook and Campbell combine concomitant variation and elimination.

6. Reder's (1982) studies compared two treatment groups, a summary group and a full-text group. In Reder's case, a control group might have been used in the early stages of the research to check on the difficulty of the passages and the test questions. A group of students might have been asked to read a passage on a different topic than the experimental passage but of equivalent length and difficulty and then asked to take the tests on the main ideas from the experimental passage that they had not read. If the control group's scores were high, then the test questions or the passages themselves might have been too easy. For a lively (if not brash) discussion of control groups and artificiality, see Stanovich (2001).

REFERENCES

Abelson, R. (1995). *Statistics as principled argument*. Mahwah, NJ: Lawrence Erlbaum Associates.

Allwood, C.M., Wikstrom, T., & Reder, L.M. (1982). Effects of presentation format on reading retention: Superiority of summaries in free recall. *Poetics, 11*, 145–153.

Bazerman, C. (1988). *Shaping Written Knowledge*. Madison: University of Wisconsin.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Charney, D. (1996). Empiricism is not a four-letter word. *College Composition and Communication, 47*, 567–593.

Charney, D. (1997). Paradigm and punish (response to Marilyn Cooper). *College Composition and Communication, 48*, 562–565.

Charney, D., Rayman, J., & Ferreira-Buckley, L. (1992). How writing quality influences readers' judgments of résumés in business and engineering. *Journal of Business and Technical Communication, 6*, 38–74.

Charney, D., Reder, L., & Wells, G. (1988). Studies in elaboration in instructional texts. In Steven Doheny-Farina (Ed.), *Effective documentation: What we have learned from research* (pp. 47–72). Cambridge, MA: MIT Press.

Cohen, M., & Riel, M. (1989). The effect of distant audiences on students' writing. *American Educational Research Journal, 26*, 143–159.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Cook, T., & Shadish, W. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology, 45*, 545–578.

Fahnestock, J. (1986). Accommodating science: The rhetorical life of scientific facts. *Written Communication, 3*, 275–296.

Fahnestock, J., & Secor, M. (1988). The stases in scientific and literary argument. *Written Communication, 5*, 427–443.

Fahnestock, J., & Secor, M. (1990). *A rhetoric of argument* (2d ed.). New York: McGraw Hill.

Flower, L. (1989). Cognition, context, and theory building. *College Composition and Communication, 40*, 282–311.

Galegher, J., & Kraut, R. (1994). Computer-mediated communication for intellectual teamwork: An experiment in group writing. *Information Systems Research, 5*, 110–138.

Gutfeld, R. (1987, February 17). Sad returns of W-4: One bureaucrat's "spirit" becomes another's "parody." *Wall Street Journal*, p. 26.

Hartley, J., & Sydes, M. (1997). Are structured abstracts easier to read than traditional ones? *Journal of Research in Reading, 20*, 122–136.

Hillocks, G. (1986). *Research on written composition: New directions for teaching*. Urbana, IL: National Council of Research in English.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills*. San Diego, CA: Academic Press.

Lunsford, A., & Ruszkiewicz, J. (1999). *Everything's an argument*. Boston: Bedford/St. Martin's.

Mill, J.S. (1843; 1930). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. London/New York: Longmans, Green.

Neuwirth, C., Chandhok, R., Charney, D., Wojahn, P., & Kim, L. (1994). Distributed collaborative writing: A comparison of spoken and written modalities for reviewing and revising documents. In *Proceedings of the Computer-Human Interaction '94 Conference*, April 24–28, 1994, Boston, MA (pp. 51–57). New York: Association for Computing Machinery.

Prelli, L. (1989). *A rhetoric of science: Inventing scientific discourse*. Columbia: University of South Carolina Press.

Ramage, J., Bean, J., & Johnson, J. (2000). *Writing arguments* (5th ed.). Boston: Allyn & Bacon.

Reder, L.M. (1982). Elaborations: When do they help and when do they hurt? *Text, 2*, 211–224.

Reder, L.M., & Anderson, J.R. (1980). A comparison of texts and their summaries: Memorial consequences. *Journal of Verbal Learning and Verbal Behavior, 19*, 121–134.

Reder, L.M., & Anderson, J.R. (1982). Effects of spacing and embellishment on memory for the main points of a text. *Memory & Cognition, 10*, 97–102.

Shaughnessy, J.J., Zechmeister, E.B., & Zechmeister, J.S. (2000). *Research methods in psychology* (5th ed.). Boston, MA: McGraw-Hill.

Slavin, R. (1992). *Research methods in education* (2d ed.). Boston: Allyn & Bacon.

Stanovich, K. (2001). *How to think straight about psychology* (6th ed.). New York: Longman.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.

Thompson, D. (1993). Arguing for experimental facts in science: A study of research article results sections in biochemistry. *Written Communication, 10*, 106–130.

CHAPTER 7

Identifying and Accommodating Audiences for Technical and Professional Communication Research

Jo Allen and Sherry Southard

Since the early 1980s, technical communication literature is replete with calls for more research (see, for instance, Moran & Journet, 1985; Rubens, 1982; and Flatley, 1994), for better research (see, e.g., Goubil-Gambrell, 1992), and for more applicable research (see, for example, Carliner, 1994). Other critics ask for a better clarification of the roles, sites, and purposes of research and its attending concepts: theory and practice (see Debs, 1993; Doheny-Farina, 1993; Sullivan & Porter, 1993; and, especially, Gross, 1994). And arguments over the methods and methodologies of technical communication research reveal our angst over definitions, clarifications, and applications—especially as they insinuate particular philosophical stances toward knowledge making and interpretation (see Blyler, 1995; Charney, 1996; Herndl, 1993; and Lay, 1991).

One source of confusion seems to be over terminology, especially regarding what qualifies as "research." We define *research* to mean an ideally systematic, though fluid process for uncovering or generating knowledge that should hold meaning for a particular audience. As such, investigations may constitute either established research or formal research, terms we have appropriated to differentiate between rather passive and active forms of information seeking and knowledge building. *Established research*, which is a prerequisite for formal research (as well as scholarship and theory-building[1]), is the search for information in already available, usually published or online, forms—the uncovering of existing knowledge.