

Chapter 2 Reading Research Reports

The purpose of this chapter is to provide you with some guidance that we hope will help you to read and evaluate research reports more easily. It consists of four sections. In Section A, we discuss how the parts of the research report are organized and provide suggestions about what is most important to attend to in each part. In Section B, we discuss the major classes of empirical studies and the sorts of conclusions that can and cannot be drawn from each. In Section C, we present some of the most useful critical strategies for evaluating empirical research studies. In Section D, the final section, we introduce a few of the measures which are very widely used in empirical studies. This section is purposely short because we prefer to introduce methods and concepts associated with empirical research in the context of actual studies. Thus, most of the concepts relevant to critical reading are introduced in our commentaries on the articles we have included in Part 2 of this volume. Our objective is not to give you extensive knowledge of subject matter or of statistical methods. Rather, we will try to give you enough information so that you can understand and evaluate the author's main line of argument. For more detailed understanding of subject matter or statistical method, you will have to consult other sources, such as those listed in the "Additional Readings" section at the end of this book.

Section A: The Format of the Research Report

As the articles in Part 2 illustrate, research reports vary in structure. Heath's article has a narrative structure, and Hunt's article is structured to help the reader empathize with the reading experience of the study's participants. Usually, though, research reports have five parts:

1. an abstract that provides a brief summary of the report,
2. an introduction that identifies the questions to be addressed in the report and provides a brief review of the most relevant literature,
3. a methods section that describes the participants, the materials, and the procedures for collecting data,
4. a results section that reports what was observed together with statistical analyses of the observations, and finally,

5. a discussion section that presents the author's claims about the meaning of the results.

Many are inclined to interpret the research report as a narrative of the research process. The structure of the report might lead one to believe that the researcher first read all of the relevant literature, then conceived a research idea, designed a study, carried it out as designed, and found the results that were later reported. What probably happened, though, was that the researcher, who was broadly familiar with the area, conceived an idea, designed a study that didn't work as expected, changed the idea, designed a new study, et cetera, and finally, after getting an interesting result, checked the relevant literature to be sure that nothing had been overlooked or misinterpreted. A research report differs, often substantially, from the events that actually took place because the author omits all of the false starts and blind alleys that were very much a part of the research process but are not needed to make the author's point. We have included authors' comments along with the articles in Part 2 to make the point that the precise, well-ordered argument in the research report may arise from a much less tidy research process, a process in which the author may have been far from clear about just where the research was going when it was carried out. Such contrasts lead to a more adequate idea of what is entailed in the process of research and the relation of research to reports about it. It also, we hope, makes empirical research on problems of literacy seem more accessible to the reader trained in the humanities.

The Abstract

The purpose of the abstract is to provide readers with just enough information so that they can decide whether or not they want to read the article. Generally, the abstract will include a statement of the issue investigated, the population in which it was investigated, and a brief summary of what was found. The abstract below, from the Wallace and Hayes article (p. 352, this volume) exemplifies the form.

This study investigates the impact of task definition of students' revising strategies. Our primary aim was to determine if freshmen students could revise globally if instructed to do so and if those global revisions would result in improved texts. We asked two groups of freshmen to revise a text provided by the experimenters; one group was given eight minutes of instruction on how to revise globally, and the other was simply asked to make the text better. The texts written by students who received the instruction were judged both to be of significantly better quality and to have

included significantly more global revision. Further, the improvement appears to affect the treated population generally rather than just a small part of that population. Thus, at least for these college students, the change in task definition allowed them to tap revision skills that they already had available.

The Introduction

The purpose of the introduction is to explain to the reader what issues were investigated and why they were worth investigating. It is in this section that the author identifies the problem, or problems, that the study is designed to solve. Often, the problems to be solved are presented as clearly marked questions or hypotheses. In the Hillocks article (p. 413, this volume), for example, the hypotheses are labeled and numbered for easy recognition and reference.

Swales and Najjar (1987), who studied research reports in physics and educational psychology, found that the introduction to a research report typically involves the following four moves: First, reasons are given for being interested in an area of investigation. Then, the relevant previous research is summarized briefly. Next, a gap is identified in the previous research, and, finally, the gap is used as a rationale for the current investigation.

As the reader will find by examining the introductions of the articles in Part 2, there is considerable variation from the pattern that Swales and Najjar have described. The introduction by O'Donnell et al. (p. 374-375, this volume) fits the Swales and Najjar pattern pretty well, but, in their first move, they simply state the topic rather than arguing for its importance. On the other hand, Enos's introduction varies widely from the Swales and Najjar pattern. Enos identifies a gap with his first sentence. He elaborates on this gap and then identifies a second gap. Finally, he states the purpose of his study as contributing to filling these gaps.

While readers should expect variation in the form and content of introductions, it will still be useful to look for the moves that Swales and Najjar describe. Perhaps the most revealing move is the identification of gaps in current research because it shows us the author's representation of the problem to be solved.

The Methods Section

Methods sections are usually divided into three subsections: a "Participants" subsection, which describes the participants and how they were selected; a

"Materials" subsection, which describes any physical artifacts employed in the study such as reading material, computer software, pictures or diagrams to be examined, etc.; and a "Procedures" subsection, which describes what the participants were asked to do and how the researchers made their observations.

The primary purpose of the Methods section is to describe the study in sufficient detail so that readers are convinced that the study was conducted competently. Ideally, the Methods section should be presented in enough detail so that other researchers could repeat the study in all its essential features if they chose to do so. By describing their methods in detail, researchers offer the disciplinary community the opportunity to find fault with the study, to reinterpret its results, or, indeed, to repeat it. Researchers must do this to be credible. Critical readers would be unlikely to take a study seriously if its methods were not described clearly and in detail.

What should readers attend to while reading the methods section? Unfortunately, the answer is "many things." Critical readers would certainly want answers to the following:

1. How were the participants selected? Were they a representative sample or merely a convenient group, e.g., friends of the researcher who happened to be available?
2. What were the instructions? Could they have biased the outcome by suggesting the result that the researcher was looking for?
3. How were the results scored? Might the judges have been biased by knowing beforehand which participants were in which condition?

The last section of this chapter, "Rival Hypotheses," provides some additional guidance concerning the critical reading of both the methods and discussion sections. However, while many things are important to attend to, one is probably more important than any of the others — identifying the type of the study. Differentiating between descriptive, correlational, and experimental studies is crucial because you cannot interpret the results without knowing the type of the study. This issue is discussed in detail below in the section entitled "Types of Empirical Studies."

The Results Section

In research reports, the results section is kept separate from the discussion section for the same reason that newspapers keep the editorial section separate from the news. In both cases, the intention is to keep "observations"

(those matters about which people can be expected to agree) separate from "interpretations" (those matters about which people may well disagree). The results section should present a summary of the observations of the study, e.g., "The average scores for the groups were . . ." together with a summary of the statistical analysis of the observations, e.g., "The differences among groups were significant." Interpretation of the observations should be reserved for the discussion section.

In reading the results section, you should attend carefully to information which bears on the major hypotheses of the study. Notice especially the *size* of the effects being reported. Sometimes this information is presented in sentence form, but often it is presented in a graph or in a table. (If you have trouble reading graphs, the "Visual Communication" references in the "Additional Readings" section of this book may be useful.) For example, in a study of three methods for teaching spelling, you might find a table such as this:

Table 1
Teaching Methods Studied

Teaching Method			
	A	B	C
Words Spelled Correctly	125.3	127.4	143.1

Table 1. Average number of words spelled correctly in three teaching conditions.

This table shows that students in conditions A and B do about equally well (within a percent or two of each other) but that students in condition C do considerably better.

Now that you know roughly how big the differences between conditions are, you should look for information about the **reliability** of the differences. (Throughout the text, we present certain terms in boldface that we believe are both important for understanding empirical studies and perhaps unfamiliar to the reader. Each of these terms is defined in the glossary.) Information about reliability is provided by the results of statistical tests which indicate the **statistical significance** of the result.

Statistical significance and **practical significance** are very different concepts. A difference is statistically significant if it is unlikely to have occurred by chance. A difference is practically significant if it is important to

someone. If a difference is statistically significant at the five percent level (written $p < .05$), then the chances are five in a hundred or less that it occurred by accident. If the difference is significant at the one percent level (written $p < .01$), then the chances are less than one in a hundred that the difference would have occurred by accident. The smaller the p-value, the more significant the result. Any difference with a p-value of less than .05 is usually referred to as being statistically significant.

Looking at Table 1, we can believe that a teacher might consider the differences between condition C and the other two conditions practically significant, but not the difference between conditions A and B. However, the differences in practical significance don't necessarily correspond to differences in statistical significance. For example, if the averages in Table 1 are based on a very small number of observations, then none of the differences may be statistically significant. On the other hand, if the averages in Table 1 are based on a very large number of observations, then all of the differences may be statistically significant. The important point to remember is that the concepts of practical significance and statistical significance are different. Table 2 suggests in simplified form how you might interpret results with various combinations of practical and statistical significance.

Table 2
The Relationship Between Practical and Statistical Significance

		Practically significant?	
		Yes	No
Statistically Significant?	Yes	An important and reliable difference	A reliable difference but not an important one
	No	A potentially important difference but not definitely so because it is not reliable. It is probably worth collecting more data to establish reliability	A difference that is neither important nor reliable

Table 2: Combinations of practical and statistical significance

The Discussion Section

The discussion section and the methods section are the two parts of the research report that are most difficult to learn to read. The problem is that in these sections, authors must defend themselves against sophisticated sorts of counterarguments that may be unfamiliar to the general reader. Thus, readers may fail to see the point of certain features of the study design and of certain sorts of arguments presented in the discussion. In the final section of this chapter, "Modes of Thought and Argument in Empirical Research," we discuss some of the most important of these counterarguments or **rival hypotheses**. Familiarity with these counterarguments can help you to understand methods and discussion sections more fully, to become a more perceptive critic of research studies, and to think about studies as a researcher would.

Section B: Types of Empirical Studies

Suppose that a researcher observes that college students who have taken an argument course are more critical readers than other students. Now, one might think that an observation is an observation is an observation. But, alas, what we can conclude from an observation depends not only on the observation itself but also on the type of study the researcher was conducting when the observation was made. To evaluate research findings effectively, readers must make two important distinctions among types of empirical studies: 1. they must distinguish between **hypothesis-testing studies** and **descriptive studies**, and 2. they must distinguish, within hypothesis-testing studies, between **correlational studies** and **experimental studies**. Hypothesis-testing studies are designed to evaluate one or more hypotheses that the researcher has proposed *before* conducting the study. A hypothesis is a question for which the researcher wants to find a positive or a negative answer, e.g., "Does this training improve performance?" or "Do good writers typically use topic sentences?" Hypothesis-testing studies rarely provide an absolute "yes" or "no" answer to the question asked, but often they do provide evidence which helps us to decide whether "yes" or "no" is the more plausible answer.

In contrast to hypothesis-testing studies, descriptive studies are not designed to provide an answer to a "yes" or "no" question. Rather, they are designed to answer more open-ended, exploratory questions, such as "What happens during the process of revision?" or "How do students of various ethnic groups prepare for exams?" Descriptive studies employ various research methods including **case study analysis**, the careful and detailed

description of one or a few individuals or events; **participant observation**, the description of a group by a participating member of that group; and **protocol analysis**, the moment-by-moment observation of individuals as they perform tasks. In this volume, Nelson uses the case study method to investigate students' practices in completing writing assignments. Haas and Funk use the participant observation method to study communication in a Japanese firm, and Schumacher uses protocol analysis to explore processes in journalistic writing. We should note that although case studies, participant observation, and protocol analysis are often used in descriptive studies their use is not limited to descriptive studies. A case study, for example, can be used for hypothesis testing, since a single counter instance is sufficient to disconfirm categorical claims that something is always true or never true.

Correlational and experimental studies are both classed as hypothesis-testing studies, but they differ in what researchers do in order to make their observations. In experimental studies, researchers take an active role in setting up their observations. In particular, experimental conditions are created by manipulating some aspects of the situation, called **independent variables**, to observe the effects the manipulation may have on other aspects of the situation, called the **dependent variables**. In studies of teaching, the manipulation often involves providing one kind of instruction to one group of participants and another kind of instruction to a second group. In such studies, the independent variable is the kind of instruction provided; the dependent variable might be changes in knowledge or skill resulting from instruction.

In a well-designed experimental study, participants would be assigned to experimental conditions at random. For example, in the teaching studies just mentioned, the kind of instruction that an individual would receive would be determined by a random procedure, such as drawing names from a hat or selecting numbers from a table of random numbers. The primary virtue of **random assignment** is that it is fair and unbiased. The selection is not influenced by preferences of the researchers or the participants. With random assignment, every individual has the same chance as every other of being included in a particular experimental condition. (We will have more to say about randomization later in this chapter.) Experimental studies, then, are characterized by 1. The purposeful manipulation of independent variables to create the study conditions, and 2. The selection of participants to conditions.

To illustrate the difference between experimental and correlational studies, suppose that a researcher is interested in the effect of note taking on learning from text. In an experimental study, the researcher might randomly assign,

say, half of the students to a group required to take notes during reading (the "notes" condition) and the other half to a group required not to take notes (the "no notes" condition). The researcher would then compare the two groups to see what effect notetaking had had on learning. However, in a correlational study the participants are not assigned to conditions; the researcher would take the world the way he finds it. For example, to study the impact of notetaking on learning from text, the researcher doing a correlational study would identify two groups of students. One of the groups would consist of students who spontaneously take notes while studying, and the other would consist of students who do not take notes. He would then compare the two groups to see if they differed in how much they learned from the reading task. The difference in procedure between these two types of studies is that in the experimental study, the researcher manipulates the conditions and assigns individuals at random to treatments; and in the correlational study, the researcher does not. As we will see below, this difference makes an important difference in what we can conclude from the two types of study.

Let's return to the observation that we discussed at the beginning of this section, namely, that college students who have taken an argument course are more critical readers than other students, and let's interpret it as if it had been made in the context of each of the three classes of study. First, suppose that the observation had been made in a descriptive study. Imagine a descriptive study in which the researcher was interested in the effects which various educational experiences in the first semester of college might have on students. To conduct this study, he administered a battery of tests to students after their first semester at college. These included tests of reasoning, memory, attitude, personality, reading and writing. The researcher then examined the data to see if he could find interesting patterns or relationships among scores and the courses that the students had taken. The relation between critical reading and enrollment in the argument course was found during this search process. That is, the observation that students who have taken argument are more critical readers was an outcome of the study.

What has the researcher learned from observing this relation? Maybe a lot and maybe nothing. Because this is a descriptive study, he can't conclude with any certainty that there really is a relation between the argument course and critical reading. Why? Because when one examines any complicated body of data, there are very likely to be some patterns in it just by accident. Students who have taken a particular physics course may happen to be taller than average. There may be more red heads than one would expect among the modern language students. Students who are taking a psychology course

may have especially healthy gums. The apparent relation between critical reading and taking the argument course seems to make sense, but then it may be just one of those accidents. If the researcher repeated his measurements in subsequent years, the relation might never reappear. However, the fact that we need to be cautious in drawing definite conclusions from descriptive studies doesn't mean that they are useless. The real strength of descriptive studies lies not in demonstrating the existence of such patterns but rather in identifying possibilities and in suggesting hypotheses for further investigation.

Now let's suppose that the observation had been made in a correlational study rather than in a descriptive one. As we noted above, researchers in correlational and experimental studies have sharply focused hypotheses. They are looking for particular patterns in the data. Imagine a correlational study in which the researcher hypothesized, before he collected any data, that an argument course would improve students' abilities to read critically. To test this hypothesis, he administered a critical reading test to all of the students in a college after their first semester. (Notice the use of a single test of critical reading rather than a large battery of tests reflects the narrower focus of the researcher's hypothesis.) The researcher then compared the critical reading scores of students who had taken the argument course to those who had not. He found, just as the researcher in the descriptive study had found, that the students who had taken the argument course were much more critical readers than were other students. However, while the researchers who conducted the descriptive study could only hypothesize that there might be a relation between the course and the test scores, the researcher who conducted the correlational study can argue much more convincingly that the relation is real rather than accidental. To understand why this is so, consider this analogy. Imagine that you are playing roulette. The croupier spins the wheel and the ball lands on number 27. The guy next to you says, "I knew that was going to happen!" You are not much impressed because you know lots of people who can predict things with extreme accuracy after they have happened. You would have been much more impressed if he told you before the event what the outcome was going to be. This is exactly the difference between descriptive studies, on the one hand, and correlational and experimental studies, on the other. It is much easier to believe that a result is not just an accident if it was predicted beforehand.

While correlational studies provide us with good evidence that the observed relations are not just accidents, they don't necessarily give us good evidence about the cause of the relation. We know that students who took the course scored high in critical reading, but we don't know why. It might

be that the course actually does improve the students' reasoning skills. However, it also might be that the course has no effect at all but that students who are naturally good at critical reading tend to sign up for such courses. A third possibility is that students who have already had argument courses in high school tend to sign up for them again in college perhaps because they like such courses. These last two possibilities make it clear that we can't conclude from this study that taking the argument course was definitely the cause of the students' superior performance. We can only conclude that there is some sort of relationship between the two.

Now, imagine that the observation had been made in an experimental study, that is, a study which involves not only one or more hypotheses to be tested, but deliberate manipulation of the experimental situation and random assignment of participants. Suppose that one year twice as many students signed up to take the argument course as could be accommodated in a single semester. Thus, half of the students would have to take the course in the Fall and the other half, in the Spring. This circumstance meant that at the end of the Fall semester there would be an opportunity to compare the critical reading abilities of the students who had already completed the course with the abilities of those who were waiting to take the course in the Spring. However, allowing the students to decide for themselves whether they would take the course in the Fall or the Spring would create the sorts of problems in interpreting the results that we discussed earlier. For example, the Fall class might attract an inordinate number of "eager beavers," or it might attract few engineers because of a conflict in the Fall schedule with a required science course. To avoid the possibility of such **self-selection biases**, the researcher assigned students at random to the Fall and Spring classes. He then compared the critical reading scores of students who had taken the argument course to those who had not yet taken it, and found, just as the other researchers had, that the students who had taken the course were much more critical readers than those who had not.

The experimental study allows researchers to argue that the course actually *caused* the improvement in student scores. Because students were randomly assigned to conditions, the only systematically imposed difference between the two groups was whether or not they had taken the argument course. Of course, we have to recognize the possibility that the random assignment process might accidentally have placed all, or most, of the smart students in the Fall class. Statistical tests are designed specifically for the purpose of estimating the probability that such accidents might have occurred. If, for example, statistical testing indicated that the results of the experiment above

were significant at the 5% level, that would mean either that the course had improved critical reading skills or that the assignment had accidentally been biased in a way that would happen only 5% of the time. The experiment, then, doesn't prove that the course caused the improvement, but it does provide evidence that makes that conclusion much more plausible.

As we indicated in the beginning of this section, the interpretation of an observation depends on the kind of study in which it was made. In the descriptive study, the observation provided only a suggestion that the argument course might be related to students' higher critical thinking scores. In the correlational study, the observation provided evidence that the argument course was related to the students' higher critical reading scores, but it did not provide evidence about what the relation was. In the experimental study, the observation provided evidence that the argument course caused students to score higher in critical reading.

The three types of study clearly differ in the persuasiveness of the evidence they provide. However, the reader should not conclude that one type of study is, without qualification, better than another. Each type of study should be regarded as a tool with a specific purpose. Each type has advantages and disadvantages which depend on the situation. If we are investigating an area new to our experience, it makes good sense to do a descriptive study first in order to get oriented. Doing experiments first might lead us to narrow our focus before we really know what the focus ought to be. Correlational studies can also have an important role to play in an extended investigation of a new area. Some variables, such as age and gender, simply can't be studied experimentally, because they can't be assigned at random. We can't say "Hey, you on the left side of the room! You're going to be the four-year-old girls in this study." For this reason, developmental psychology is built largely on correlational studies. But correlational studies can also play an important role even when the factors being studied can be randomly assigned. In some cases, we may suspect that the process of assignment itself may have an important impact on what is being observed. Assigning students to take the course in the Fall semester might make them feel coerced, or it might make them feel privileged. Either effect could bias the result of the experiment. In such cases, we may get a better understanding of the situation by using both types of study. The experimental study would convince us that the effect wasn't due to extraneous factors such as a self-selection bias. The correlational study would convince us that the effect wasn't due to the process of assignment. In many situations, the increased cost involved in using multiple methods to approach

the same question is more than offset by the greater confidence we can place in the results.

Table 3 summarizes some of the circumstances relevant in deciding what type of study is most appropriate for approaching a particular research problem.

Table 3
Uses of the Three Major Types of Research

Type	Appropriate Use	Characteristic Outcome
Descriptive	The domain is not yet well explored. The researcher does not have specific hypotheses to test.	Formulate hypotheses
Correlational	The researcher has specific hypotheses to test, but they involve variables which cannot be manipulated, e.g., age, gender.	Establish relations
Experimental	The researcher has specific hypotheses which involve variables which can be manipulated.	Establish causal relations

Table 3. Types of research, appropriate use, and characteristic outcomes

Section C: Modes of Thought and Argument in Empirical Research

Perhaps the most characteristic feature of the empirical research community is its insistence that researchers exercise special care in making and interpreting observations. Researchers are expected to be scrupulous in collecting data. Evidence of carelessness or fraud will seriously tarnish the researcher's reputation. In addition, researchers are expected to describe their procedures carefully enough so that other researchers can repeat them if they want to check the observations for themselves.

Equally important, researchers are expected to be very careful in drawing conclusions from their data. They should draw definite conclusions from their observations only after they have tried to rule out all of the plausible alternative interpretations. For example, suppose that a researcher finds, by conducting a survey, that schools providing their students with computer training have fewer disciplinary problems than other schools, and the researcher then claims on the basis of this observation that computers promote student discipline.

Other researchers might question this claim by noting that schools with computers also tend to be the wealthier schools. In this case, the alternate interpretation that the observed difference in disciplinary problems may have to do with economics rather than with computer use is quite plausible.

The researcher in this example would be embarrassed to have overlooked such an obvious alternative interpretation. Researchers are expected to be good at anticipating such rival hypotheses. They are expected either to provide evidence which casts doubt on such alternatives or to hedge their conclusions—that is, to acknowledge the possibility of alternative interpretations explicitly. Researchers who fail to recognize alternative interpretations which readily occur to their colleagues are considered unsophisticated. Indeed, training in recognizing alternative interpretations of results is an essential part of the education of empirical researchers. Texts such as Campbell and Stanley's *Experimental and Quasi-experimental Designs* and Huck and Sandler's *Rival Hypotheses* distill years of experience in the critical analysis of data-based claims by identifying a number of the most common sources of alternative interpretations. In the following section, we list the kinds of alternative interpretations which must most frequently be considered in reading and conducting empirical research in literacy. We have grouped them into five families: Stability, Bias, Reactivity, Confounding Factors, and Faulty Inferences.

Common Sources of Alternative Interpretations

1. **Stability:** Do the observations warrant the conclusion that some stable event or thing has been observed?

When we look at clouds, we can sometimes see fairly definite patterns such as faces or animals. However, we don't attribute stability to these patterns. We interpret them as accidental arrangements of light and dark patches without permanence or significance. Similar sorts of accidental patterns can arise when we examine data collected in research. If we look long and hard enough at complex data, we are bound to see some patterns. If we do see a pattern in our data, we need to ask if that pattern has stability, in the sense that it can be observed on other occasions and by other people. The next three questions bear on this issue of stability.

- a. **Observer reliability:** Do independent observers agree about the objects or events being observed?

Imagine that a school is planning to institute the following placement policy: all incoming students are to be placed into basic, average, or honors classes on the basis of their performance in writing an impromptu theme. For reasons of fairness, all of the students' essays are to be graded wholistically by the same English instructor, Mr. Brown. At a faculty meeting, another member of the English department, Ms. Jones, voices concern about the placement policy. "Suppose," she says, "that these essays were graded independently by two English instructors and their judgments didn't agree at all. Wouldn't that mean that by the proposed policy we would be assigning students to basic and honors courses pretty much arbitrarily. Why not just draw lots?" Clearly, Ms. Jones has raised the question of **observer reliability**.

The issue here is not Mr. Brown's qualifications. Even if all faculty members were equally qualified as writing critics, they still might not see the same things or value the same things in the students' essays. Thus they may not agree in placing students into basic, average, and honors sections. (See Chapters 12 and 13 of this volume for causes of agreement and disagreement among readers of students' essays.) To estimate how well instructors agree in placing students, it would be appropriate to ask two or more instructors to judge the essays independently. Their judgments could then be compared. The usual way to do this would be to calculate the **correlation** between the judgments of pairs of instructors. A correlation of +0.80 or better is generally taken as indicating a satisfactory level of observer reliability.

- b. **Test-retest reliability and replication:** Are the observations repeatable?

Suppose that a team of reading researchers report that reading scores for primary school children show a surprising increase in March. They have named this effect "The March Surge" and are giving newspaper and magazine interviews about it all over the country. A critical minded researcher suspects that "The March Surge" may be just another pattern in the clouds, a one-time "glitch" in the data that won't be found in repeated observations. To check out her suspicion, she initiates a replication study. That is, she tries to repeat the original study with a new group of primary school students. To do this, she selects the new group in the same way as those in the original study were chosen and collects data from them by methods as close as possible to those

described in the original report. Whether her study confirms or fails to confirm "the March Surge," she will seek to publish it in the same journal that published the original report. This will help to insure that the community that was audience to the original report will consider the new results.

- c. **Statistical reliability:** Is the number of observations sufficient to warrant the conclusion?

Imagine that you have designed a new method for teaching paragraph writing. You choose two children at random from a friend's class and teach one by your new method and the other by the standard method. Suppose further that on a subsequent test, the child who was taught by your method writes a better paragraph than the other child. Assuming that your teaching and the test were fair to both methods, is it time for celebration? Of course not! Even if the two teaching methods were equally good, one of the children would likely have done better on the test than the other. The chances that the child who did better would happen to be the child trained by your method would be 50%. You would need to repeat your observations with many children before you could be confident that your method was an improvement over the standard one.

2. **Bias:** Are the observations fair and accurate?

Bias in reporting the facts is something we learn to take into account everyday. We expect salespeople to tell us the good things about their products but not the bad. We know that people often play up their successes and play down their failures. We know that people's judgments are biased by "**halo**" effects, e.g., teachers may grade students who have done well in the past more leniently than they grade other students. It is an important part of the tradition in empirical research to try to eliminate or at least control bias in observations. We will talk about three sources of bias: Sampling Bias, Self-reporting Bias, and Researcher Bias.

- a. **Sampling Bias:** Is everyone represented equally?

In many cases, considerations of cost and time make it impractical to study all the members of a population we are interested in, e.g., all of the students in a large college or all the children in an urban school system. Instead, we often have to study a **sample** of the population that we hope will represent the whole. How that sample is chosen is

important because some ways of choosing samples may systematically bias the results of the study. For example, suppose that a researcher wanted to know what teachers believe about the role of word processing in the teaching of writing. One very natural thing for the researcher to do would be to interview friends who teach since they are easy to access. Unfortunately, one's friends are usually not a representative sample of the population. Friends tend to resemble one another in age, gender, race, and professional specialization. If attitude toward word processing varies with age, as seems likely, or any of the other factors in which friends tend to resemble each other, then samples of friends are likely to provide biased results.

Another procedure would be to post a bulletin asking teachers to volunteer to be interviewed. Unfortunately, samples of volunteers are usually not representative of the population either. Teachers who are hostile to word processing or uninterested in it are less likely to volunteer for interviews than are those who are very enthusiastic about it. Thus, results based on a volunteer sample are likely to be biased.

The best way to pick a sample is through randomization. A **random sample** is one in which every member of the population has an equal chance of being included. Whether a given individual is included in the sample or not depends on a random choice, such as one made by picking a number out of a hat or rolling a die. For example, suppose that we wanted to pick a sample of 20 teachers from the 400 teachers employed at a college. First, we would assign a number from 1 to 400 to each teacher. Then, we could write the numbers on slips of paper, put the slips in a bowl, and stir them up. The sample of teachers would be those corresponding to the first 20 numbers chosen from the bowl. Using a table of random numbers to make the selection or even the last three digits from a page of listings in a phone book might be easier, but the effect would be the same. Choosing samples by randomization is the surest and simplest way to avoid bias.

Sampling bias can apply to materials and contexts as well as to participants. For example, a researcher who was interested in demonstrating that students have trouble understanding their textbooks might be inclined to study the worst of the available textbooks. In fairness, the results of such a study would have to be qualified to indicate how the texts were chosen (e.g., "Students have trouble understanding bad

textbooks"). If the results are to be applied to textbooks generally, then the texts studied should be a random sample of those available.

Similarly, if we are to draw reasonable conclusions about how students write in college, we should not base that conclusion on observation of students in a single context, for example, English classes. Rather, any such conclusion should be based on observation of students in a random sample of the various and often very different contexts in which college students write, e.g., biology labs, school newspaper offices, philosophy exams, etc.

- b. **Self-reporting bias:** Are people reporting about themselves fairly and accurately?

Suppose that a college committee designing a humanities course wants to know what students read on their own outside of class. One way to try to get that information would be to ask students to fill out a questionnaire in which they are asked to report how frequently they read novels, poetry, non-fiction, plays, biography, science fiction, etc. The problem with such self-reporting is that respondents may tend to over-report high prestige reading, e.g., novels and poetry, and under-report low prestige reading, e.g., *Playboy* and *The National Enquirer*. Researchers for the National Assessment of Educational Progress attempted to reduce this kind of bias by asking students to name the books they had read. The assumption here is that students' claims to have read novels are more believable if they are actually able to name some.

Another way to deal with bias in this situation would be to make use of multiple sources of information about student's reading habits. For example, one might get information about what sells at student bookstores, what books are checked out of college libraries, and what reading materials students carry around with them. Using many sorts of measures to investigate the same question from different angles is called **triangulation**. In this case, triangulation can help to control for self-reporting bias because it would allow the researchers to compare survey data which may be biased with bookstore sales which are less likely to be biased.

- c. **Researcher bias:** Is the researcher evaluating the results fairly?

Being a researcher doesn't make a person unbiased. Researchers

usually care a great deal about how their research comes out. They spend much of their time thinking up clever theories and hypotheses, and they are, quite naturally, eager to see these theories confirmed in their data. The result is that researchers can be quite biased observers. Knowing this, good researchers take special care to reduce and control their own biases in evaluating data.

Suppose that a researcher has designed a new method for teaching argument. He identifies a number of instructors who agree either to teach his method or to act as control teachers. Knowing that he might be tempted to choose the better teachers to teach his method, he decides to assign the teachers at random. To assess the effect of the instruction in argument, all of the students are asked to write an argument on a specified topic. Again, because he is aware of his own bias, he does not grade the essays himself. Rather, he has the essays evaluated by other instructors who are not permitted to know which of the students have received the experimental instruction. This last procedure is called **blind grading**; like the other tactics just mentioned, it is a common and useful technique in controlling researcher bias.

3. **Reactivity:** Is what is observed changed by the act of observing?

The reactivity of a method of observation is the tendency of that method to change what is being observed. Methods of observation differ in how reactive they are. A method in which a person of one race interviews a person of another race about racial attitudes is likely to be highly reactive. In contrast, an anonymous questionnaire about political attitudes will probably be less reactive. Some observational techniques are not reactive at all. For example, assessing the popularity of various types of reading by examining library circulation records probably has no effect on the public's reading behaviors. Historical studies, of course, are completely non-reactive. There are two major loci of reactivity in empirical studies: researcher effects and participant effects.

- a. **Researcher effects:** Is the researcher influencing the outcome?

The potential for researcher effects is especially high in participant observation studies. Participant observers are typically visiting members of the group being studied and are, therefore, not "insiders." Because group members may be uncertain of the goals and loyalties of

the observer, especially if the observer is socially or ethnically different, they may not behave naturally in the observer's presence. For example, the presence of a white observer in a black social group might well change the nature of conversation in the group. The more obviously different the observer is from the group being observed, the more one must be concerned with researcher effects.

Researcher effects are important in experimental studies as well. In these studies, researchers often know the condition to which each participant has been assigned. As a result, they may, without intending to, act slightly differently toward participants in the experimental and control conditions. For example, experimental participants might receive somewhat friendlier, more encouraging treatment than controls. Those differences may be conveyed in facial gestures or body language of which the researcher may be completely unaware. It is now well established (Rosenthal, 1976) that researchers can influence a study's outcome through such subtle cues.

Researchers can and should reduce such effects on the outcomes of studies. For example, they can restrict the opportunities they have for treating differently participants in different conditions. Instructions might be scripted or, in some cases, tape recorded or videotaped to insure uniformity of presentation across conditions. Responses to questions might be scripted as well. In many studies, participants' questions are answered simply by referring them to the appropriate section of the instructions. Researchers can try to fade into the background. In a writing study, for example, the researcher might try to stay out of the participant's line of sight and intervene as little as possible during data collection. And some studies can be designed so that the person running the study doesn't know the conditions the participants have been assigned to. Such care in the design of studies can greatly reduce the possibility of researcher effects.

- b. **Participant effects:** Is the experience of being observed influencing how people behave?

People who participate in research studies respond in many different ways to being observed. They may feel honored to be singled out for study, they may feel anxious to perform well, or they may feel spied on. In any case, they probably feel and act somewhat differently than they would if they weren't being studied. The probable extent of these

differences must be taken into account in interpreting the results of studies. Case studies are especially subject to participant effects because individuals participating in them are singled out for exceptional attention.

Participant effects can be reduced in a number of different ways. For example, it may be possible to design an experimental study so that all of the participants undergo both the experimental and control treatments. In a study of the impact of computers on writing, each participant might write some essays with a computer (the experimental treatment) and other essays with pen and paper (the control treatment). Another strategy is to make it appear that all participants are in the same condition even though they are not. Suppose that a study is to be conducted by distributing packets of experimental and control materials to students in the same classroom. If the researcher does not call attention to the differences in the packets being distributed, the participants may not notice them. Participant effects can often be greatly reduced by incorporating such strategies in the design of the study.

4. **Confounding factors:** Might the effect be due to something else?

When researchers conduct correlational or experimental studies, they naturally focus their attention on certain variables or on certain differences among groups in which they are particularly interested. But this exceptional selectivity of focus may produce a kind of intellectual myopia; the researchers may fail to attend to other variables which change along with the ones they are interested in. For example, a recent article, which one of us reviewed, focused on the relation between writing ability and personality traits. The author had found a strong relation between writing ability and the presence or absence of a particular personality trait. What is relevant for us here is that the personality trait was strongly related to age. Older children had it, and younger children didn't. Thus, what the author of the article had interpreted as a rather surprising effect of personality on writing ability may well have been a not very surprising effect of age. In this study, age was a confounding factor because it varied along with the variable that the researcher was highlighting. As such, it provided opportunities for alternative interpretations of the author's claims about the meaning of the data.

In the following section, we describe five common sources of confounding:

a. **Effects of drop-outs:** Has the population changed during the study?

In a study designed to evaluate the effectiveness of a remedial reading course, researchers compared the average score on a reading test achieved by students at the beginning of the course with the average achieved at the end of the course and found a substantial increase. On the basis of this result, the researchers concluded that the course was very effective. In deciding whether or not to accept this conclusion, it is important to ask how many students' scores were involved in each average. If there were many more students at the beginning of the course than at the end, it is possible that drop-outs may contribute heavily to the result. Suppose for example, that the students who dropped out were the ones who had greatest difficulty with the course. These may well also be students with relatively low reading scores. If this were the case, it might be that the reported gains occurred just because these students' scores were not included in the end-of-course average.

A good way to protect against drop-out effects would be to base the evaluation only on those students who completed the course, eliminating from consideration the pre-test scores of the students who dropped out. If big gains were found for these students, then we would have evidence that the course actually helped some students.

b. **Maturation.** Could the effect be due simply to the passage of time?

People, especially young ones, get better at a lot of things as they get older. For example, we would expect kindergarten students' vocabularies to increase over the course of time whether we gave them special vocabulary lessons or not. Thus, if a researcher reported that kindergarten students increased their vocabularies by 15% when they were exposed to an enrichment program for a year, we wouldn't know whether to be impressed or not. After all, between ages four and five, children increase their *age* by 25%. Perhaps vocabulary could be expected to increase by a similar amount. To assess the effectiveness of the enrichment program, we would have to compare the reported improvement to the amount students would have improved their vocabulary over the same period of time without the enrichment program. That is, we would need to measure vocabulary gains over the same period in a **control group** of students who *did not* have the enrichment program.

c. **Concurrent events:** What else was going on?

Even though researchers are studying them, people continue their lives in the usual complex way beyond the researcher's view. Sometimes events in the lives of the people being studied are confounded with what the researchers are interested in studying. Such confounding, if unnoticed, can lead researchers to draw unwarranted conclusions from their observations.

Suppose, for example, that researchers interested in the development of writing skills ask students in grade 4, 5, and 6 to write on the topic, "making new friends." They find modest gains in essay quality between grades 4 and 5 but very dramatic gains between grades 5 and 6. One explanation of this result that the researchers consider is that writing development is accelerating during these years. However, one of the researchers notes that there is a possible explanation of the results based on concurrent events in the students' lives. In the school district being studied, students move from primary to middle school between the fifth and sixth grades. For this reason, the sixth grade students have just had a lot of experience in making new friends—experience which may make it especially easy for them to write on the assigned topic. To test whether writing development really does accelerate between grades 4 and 6, the researchers should, at least, have students write on several different topics.

d. **Testing effects:** Could the improvement be due to test practice?

Years of experience in mental testing has turned up a rather odd finding. When individuals are tested twice, they usually do better on the second test than on the first when the two tests are parallel forms, that is, when the tests are on the same topic but with completely different questions. By better, we mean getting higher scores on tests of knowledge or skill or getting more "normal" scores on tests of personality. This effect occurs even when participants are given no feedback or opportunity to study between testings.

The implication of the testing effect for research involving pretests and posttests is clear. A comparison or control group may be necessary if one wants to sort out the effect of the variable being investigated from the effect of testing. For example, suppose that we want to examine the impact of writing on students' topic knowledge. To do this, we give the

students a knowledge pretest. We then ask them to write an analytic essay on some aspect of the topic. Finally, we give them a knowledge posttest. Suppose that the students score better on the posttest than on the pretest. What can we conclude? Not much unless we know how big the testing effect is. A good way to assess the testing effect is to use a control group of students who are pretested and posttested but who do not write the analytic essay. A positive effect of writing on knowledge would be indicated by greater **pre-post gains** in the essay group than in the control group.

- e. **Order effects:** Could the order of testing have influenced the result?

Suppose that a researcher was interested in determining whether children wrote better on a topic they chose themselves or a topic that the teacher chose. To answer this question, the researcher might have the children write first on the teacher's topic and then on their own topic. If the researcher did this, though, the study would be open to the criticism that any difference found might be due to the order in which the essays were written. Perhaps the second essay would benefit from the students' practice or suffer because of their fatigue or boredom. To deal with this problem, the researcher can require half of the children to write on the teacher's topic first and the other half of the children to write on their own topic first, and then reverse the assignments for the second task. In this way, each topic will, on average, be equally advantaged or disadvantaged by the order in which the essays are written. This process of balancing the testing order is called **counterbalancing**.

5. **Faulty inferences:** Can we really draw that conclusion from the observations?

- a. **Causation and correlation:** One of the most common problems to be found in research studies in literacy is a tendency to draw causal conclusions from correlational data. If two factors are correlated with each other, they may be causally related, but they also may not be; all that we can say for sure is that they vary with each other. Both may be caused by some third factor. For example, writing skill and shoe size are strongly correlated with each other. Why? Not because having big feet makes it easier to write, but rather because both writing skill and foot size are strongly correlated with age. Two year olds' writing skills are weak, *and* they have very small feet.

A frequent form in which this sort of faulty inference appears in the literature is the drawing of pedagogical conclusions from correlational data. For example, if it is observed that writers who plan well produce higher quality essays than those who don't plan well, it is tempting to draw the conclusion that teaching planning would help students improve their writing. However, teacher's time and student's time is too valuable to justify drawing such hasty conclusions. It may be that planning and writing are not directly related but rather are connected through a third factor such as motivation or fluency. Perhaps students who are motivated to do well in school put extra effort both into planning and writing. If the underlying causal factor were really motivation, then teachers might be wasting their time in teaching planning. The simplest and surest way to establish a causal connection is through an experimental study. If we teach planning to one randomly selected group and not to another, and the group taught planning writes better, then we can conclude with much greater confidence that teaching planning skills can improve writing.

- b. **Drawing inferences from non-independent samples.** It is important to know if samples are **independent** or not because most statistical tests are appropriate only for independent samples. If the samples are not independent, the test may be invalid and its results misleading. Before discussing independence and non-independence, though, we need to discuss units of observation and to define the term "sample".

Studies differ from one another in the units of observation they employ. In some studies, the unit of observation is the individual. In others, it may be the classroom, the city block, the school, the county, or any of innumerable other units. If the unit of observation is the individual, then the data of the study will consist of measurements of individuals, e.g., individual's scores or individual's ratings on essays. If the unit of observation is the classroom, then the data of the study will consist of measurements of whole classes, e.g., average achievement scores or average attendance records for the class as a whole.

The samples in an experimental study are the groups of units that are assigned to the various conditions of the study. For example, if the units of observation were students, the samples would be the collections of individual students assigned to the various experimental conditions. If the units are assigned to conditions at random, then the

samples are said to be **independent**. They are independent in the sense that the chances that a unit is assigned to a particular condition doesn't depend on which of the other units are assigned to that condition. For example, if children in a kindergarten class are assigned at random to experimental and control conditions, then the chances that a particular child will be assigned to the control group doesn't depend on who the other children in the control group are. The samples would *not* be independent if the researcher tried to create congenial groups by placing best friends in the same group. Such samples are not independent because the chances that a child is included depends on whether that child's friend is included.

The following example shows some of the implications that independence (or non-independence) of samples has for the analysis of study results. Suppose that a high school has 300 ninth grade students in 12 classes of 25 students each. A researcher wants to divide the students into an experimental and a control group of 150 students each and is considering two methods for accomplishing the division. The first method, which might be called "the individual method," involves random assignment of individual students to the experimental and control groups. To do this, students' names would be put into a drum and mixed thoroughly. Then 150 names would be drawn. The students whose names were drawn would constitute the experimental group and the remaining students, the control group. These two groups would be independent samples because each individual was selected for inclusion in the experimental or control group independently of every other individual.

The second method, "the class method," involves random assignment of whole classes to the study conditions. To do this, classes would be assigned numbers from 1 to 12. Six numbers would then be selected at random to identify the classes to be assigned to the experimental group. The remaining six classes would be assigned to the control group. The class method has some practical advantages. Many studies can't be run in classes if members of both experimental and control groups are present. Arranging for times when students can participate in the study out of class may be difficult or impossible. These difficulties can be reduced if whole classes are assigned to conditions. However, the class method also has some serious limitations. If the class is taken as the unit of measurement, then the six classes in the experimental group and

the six classes in the control group constitute independent samples. What is important to note here is that, if it is the classes that are selected randomly for assignment to the study conditions, the individuals in the classes cannot be treated as independent units; all the individual students in a single class must be treated as constituents of a larger independent unit, i.e., the class. If one student in a class is in the control group, then *all* the students in the class are in the control group. Statistical tests can appropriately be applied only if the whole class is viewed as the unit of measurement. In this example, appropriate statistical tests would compare *six* experimental measures with *six* control measures — one measure for each *class* in each group. Usually the measure for a class would be the average performance of all of the students in the class. In contrast, if experimental and control groups were selected by the individual method, statistical tests could appropriately compare *150* experimental measures with *150* control measures — one measure for each *student* in each group.

It is not unusual to find studies in which whole classes are assigned to the experimental and control groups but which are never the less analyzed as if students were independent units of measurement. The effect of this error is to make results appear to be more statistically significant than they should. Let's consider an extreme case to emphasize the point. Suppose that a researcher wants to evaluate a new procedure for teaching spelling. To do this, he identifies two classes of 30 students each and randomly assigns one to be the experimental group and the other, the control group. If the researcher analyzes the data as if the student was the unit of measurement and as if there were 30 independent measurements in each group, he might believe he had found a significant advantage (or disadvantage) for the new procedure. However, an appropriate analysis could claim only two independent measurements, one for each group. This more conservative approach to the analysis makes sense because all of the students within a class may be strongly influenced (either positively or negatively) by their teacher. The difference between two classes might simply reflect a difference between two teachers. Two measurements, in any case, are not enough to warrant a claim of statistical significance.

Section D: Some Commonly Used Empirical Measures

In this section, we will discuss three sorts of empirical measures:

1. measures of central tendency, 2. measures of variability, and 3. measures of association.

1. **Measures of central tendency:** Measures of central tendency are used to identify a central or representative value for a set of observations. The most familiar of these is the **arithmetic average** or **mean**. The mean age of a class of 25 children is the sum of their ages divided by 25. The mean works well for characterizing factors such as age and reading score, because about as many observations can be expected to be above the mean as below it. However, for some factors, such as the number of books that students read for pleasure, the mean may not produce a representative value. Most children will read fewer than the mean number of books, because the mean is inflated by a few children who read lots of books. In cases such as this, two other measures, the **median** and the **mode**, may yield measures more representative of the group. The value of the median is the value of the middle observation. For example, imagine that the 25 children in a class are arranged in order of age: the median age is the age of the middle or 13th child. The median is representative in the sense that there are as many observations above the median as below it. The mode is the most frequent observation. Imagine now that there are more children who read exactly two books a year for pleasure than any other particular number of books: two is the mode. As the most frequent observation, the mode may be viewed as representative of the group of observations.

2. **Measures of variability:** Groups of observations differ in how variable or "spread out" they are. One measure of variability is the **range** or the difference between the largest and the smallest observation. For example, one class of 30 college freshmen might range in age from 17 years 6 months to 18 years 6 months, and another from 16 years 2 months to 52 years 8 months. The range in ages in the first class is just one year, but in the second class, it is 36 years 6 months.

Another commonly used measure of variability is the **standard deviation** or **S. D.** The standard deviation increases as the variability or spread of the observations increases. Generally, about two-thirds of the observations can be expected to fall in the interval extending from one standard deviation

below the mean to one standard deviation above the mean. Standard deviations are included in many data displays so that the reader can judge the degree of precision of the observations being reported (See O'Donnell et al., Tables 1 and 2; Spivey, Tables 3 and 5; and Palmquist and Young, Figures 1, 2 and 3, all this volume.)

3. **Measures of correlation:** measures assess the extent to which two sets of observations vary together. For example, correlations may be used to determine the extent to which reading scores change with age or the extent to which independent raters agree with one another in scoring a set of essays. There are many different measures of correlation. The **Pearson product moment correlation**, which is the one most frequently found in the research literature, is used with interval data. **Spearman's Rho** is used for comparing ranked data, such as judgments of essays ranked by quality. **Cohen's Kappa** is used for comparing categorical data. For example, it could be used to measure the strength of the relation between readers' guesses of the writer's gender and the writer's actual gender. All correlations vary from plus one through zero to minus one. Plus one represents a perfect positive relation, e.g., every paper rated high (or low) by Rater 1 was also rated high (or low) by Rater 2. Minus one represents a perfect negative relation, e.g., every paper rated high (or low) by Rater 1 was rated low (or high) by Rater 2. A zero correlation indicates no systematic relation, e.g., the raters' grades are unrelated to each other.

A Final Word

Reading a research report is a skilled activity. Doing it well requires acquaintance with the conventions of thought and language of the empirical research community. We have tried to introduce you to many of these conventions in this chapter. Further development of your critical skills requires the use of the conventions in the careful analysis and evaluation of representative research studies. We hope that you will be able to read the studies in Section 2 with understanding and enjoyment. In addition, we hope that you will be able to sense some of the excitement of the research enterprise.

References

- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.
- Huck, S. W., & Sandler, H. M. (1979). *Rival hypotheses*. New York: Harper & Row.

Rosenthal, R. (1976). *Experimenter effects in behavioral research (enlarged ed.)*. New York: Irvington Press.

Swales, J., & Najjar, H. (1987). The writing of research article introductions. *Written Communication*, 4 (2), 172-191.

Webb, E.J., Campbell, D.T., Schwartz, R.D., Sechrest, L., & Grove, J.B. (1966). *Nonreactive measures in the social sciences*. Boston, MA: Houghton Mifflin.

Part 2