THINKING CRITICALLY ABOUT RESEARCH ON SEX AND GENDER

PAULA J. CAPLAN
University of Toronto

JEREMY B. CAPLAN

HarperCollins*CollegePublishers*

THINKING CRITICALLY ABOUT RESEARCH ON SEX AND GENDER

Acquisitions Editor: Catherine Woods
Project Editor: Brigitte Pelner
Text Design: Heather A. Ziegler
Cover Design: LaToya Wigfall
Cover Design Concept: Jeremy B. Caplan
Production Assistant: Hilda Koparanian
Compositor: Express 93
Printer and Binder: R. R. Donnelley & Sons Company
Cover Printer: The Lehigh Press, Inc.

Thinking Critically About Research on Sex and Gender

Library of Congress Cataloging-in-Publication Data

Caplan, Paula J.
Thinking critically about research on sex and gender / Paula J. Caplan, Jeremy B. Caplan
p. cm
Includes index
ISBN 0-06-501621-1
1. Sexology–Research. 2. Sex differences—Research. 3. Sex differences (Psychology)—Research. I. Caplan, Jeremy B. II. Title
HQ60.C37 1994 93-22718
305.3' 072—dc20 CIP

94 95 96 9 8 7 6 5 4 3

For Emily, for everything

CONTENTS

. . . science is the creative product of an engagement between the scientist's psyche and the events to which [the scientist] is attentive . . . .

Henry A. Murray, "The Case of Murr" in *A History of Psychology in Autobiography*, vol. 5, New York: Appleton-Century-Crofts, 1967, pp. 285-310.

PREFACE

Walk into a classroom filled with seven-year-olds, ask them to think of a claim they have heard made about sex differences, and then ask them how they might try to find out whether or not that claim is true. Then, ask them to think critically about the strengths and weaknesses of their research plan. When we did that, the children reacted enthusiastically, and their comments made it clear that no one had to teach them how to think critically. They only had to be *encouraged* or given permission to do so.

People are naturally able to think critically. How poignant that this is sometimes called "childlike curiosity," as though it were unseemly in adults. How interesting that formal systems of education and social pressures to accept what one is told so often suppress that keen, questioning attitude. The most powerful force in that suppression may be the sentence "That is the correct answer," which encourages us to believe that there is a correct answer for everything and that an answer once considered correct should never be questioned. But if we are honest with ourselves, we realize that what is considered "common knowledge" changes over time, and the fact that some assertion has been accepted as truth does not mean that it is inevitably and forever right. The only way we can make efficient progress is by always questioning the truth of those claims and evaluating both the good and the harm that they might do.

The exercise with seven-year-olds was done because, years ago, when Jeremy was in second grade, he heard Paula talking about her research in psychology and asked why they didn't learn that kind of science in school. He asked Paula to visit his class and discuss some of that work with them. She had been teaching the fundamentals of critical thinking about research on sex and gender to undergraduates since 1980 and later to graduate students. When Paula approached his teacher about Jeremy's suggestion, the teacher naively replied, "Well, you can try, but kids this age can't think abstractly." When Paula asked the students to think of claims they had heard about sex differences, one they mentioned was, "Boys are ruder than girls." Paula then said, "Today, we are going to do what scientists do. We are going to try to find out whether or not that is true." She asked them how they might explore the topic of sex differences in rudeness, and they decided, "We could go to a house where there is a brother and a sister, and while they are eating dinner, we could make a mark every time the boy is rude and every time the girl is rude." Paula then said, "Let's suppose we did that and found that the boy was rude 8 times, and the girl was rude 5 times. Would we have proven that boys are ruder than girls?"

Many of the children began to say "yes" but then stopped, looked quizzical, and said, "Not really." All Paula had to do was ask, "Why not?" and "Anything else?" repeatedly, and the children came up with a remarkably sophisticated critique of their research proposal. They pointed out that it might be that only in that particular family or only at dinner was the boy the ruder of the two children. They wondered whether one child or both might be behaving differently because there was someone else there taking notes about their actions. They even debated whether or not 8 was "all that much" bigger than 5, getting into statistical questions in that way.

As a result of the students' eagerness and performance, Paula worked with Margaret Secord on writing up the directions for this exercise in the form of a curriculum unit for grade-school children (Caplan, P. J.; Secord-Gilbert, M., Staton, P. [1990]. *Teaching children to think critically about sexism and other forms of bias.* Toronto: Green Dragon Press) and then testing its effectiveness with many other children and teachers. When Secord (Secord, M. J. [1987]. Teaching children critical thinking. Unpublished master's thesis, Ontario Institute for Studies in Education, University of Toronto) asked grade-school teachers to do the exercise in a single class period, she found that the children who had been exposed to that exercise were more likely than other children to think critically about claims made about other group differences (ageist, racist, etc.) as well as sexist ones.

We both have always felt a sense of excitement and discovery when engaging in critical thinking about research and about theories. Those people who have been most instrumental in encouraging such thinking in us have enriched our lives immeasurably: for Paula—my parents, Tac and Jerry Caplan; my uncle, Bill Karchmer; and my teachers, Jack Bush, Donal Stanton, and Bruce Baker; and for Jeremy—my great-uncle, Bill Karchmer, and many others. We offer this book in the hope that it will do the same for you.

Acknowledgments

The authors wish to thank Jennifer Chambers for her assistance with portions of the manuscript; Emily Julia Caplan and June Larkin for their thoughtful comments and suggestions throughout the entire manuscript; Gina Feldberg for help with some concepts and references; and our editor Catherine Woods for her help and support. Thanks also to the following reviewers for their suggestions: Claire Etaugh, Bradley University; Terri D. Fisher, Ph.D., Ohio State University at Mansfield; Karen G. Howe, Trenton State College; Lois Muir, University of Wisconsin; Rose Preciado, Mt. San Antonio College;

Paula J. Caplan
Jeremy B. Caplan

THINKING CRITICALLY ABOUT RESEARCH ON SEX AND GENDER

CHAPTER ONE

INTRODUCTION

❖ THE CYCLE OF BIAS

It is virtually impossible to grow up without believing that girls and boys, men and women, differ—not only physically but also in important aspects of behavior, attitudes, and abilities. Whether or not we realize it, many of these beliefs come to us directly or indirectly from scientific research on sex and gender. Because the most prevalent twentieth-century attitude toward scientists is that they discover and describe the Truth, it may not occur to us to question what we think are scientific truths. This unquestioning acceptance of scientists' pronouncements about sex and gender differences affects every aspect of our private and public lives, since—consciously or unconsciously—every time we interact with another person, we are making assumptions about what is true and natural for people depending on their sex. Countless people of both sexes invest great amounts of energy worrying about whether they themselves are doing, feeling, and believing what members of their sex are "supposed" to do, are naturally meant to do, are destined by their genes and hormones to do. When we believe that these sex-related patterns have been proven by researchers to be pervasive and inevitable, it can be disturbing to find that we, or people with whom we live or work, do not fit the patterns.

However, scientists do not simply discover and describe the Truth. Like everyone else, scientists who study sex and gender grew up learning what women and men are "supposed" to be like. They might have heard, for example, that boys don't want to play with dolls (unless the dolls have guns), and that girls cannot play hockey. These beliefs about how people are or should be influence how scientists do their research, how they see and describe the world. A girl and a boy could be doing exactly the same thing, but because one is a girl and one is a boy the activity may be described differently. For example, a girl playing with fire may be said to be demonstrating her inborn desire to cook, while if a boy plays with fire he would probably be called a natural fireman or naturally daring. It is often mistakenly assumed that scientists are able to be free from such bias—to be "objective" and able to see the world without being influenced by their own thoughts or feelings about it. Yet many psychologists do things such as appreciatively labeling as *assertiveness* such behavior in men as interrupting other people, which others might call *rudeness.* Whichever label one chooses in such a case reflects one's experiences and perspective. The truth is that no one is free from bias, but scientists often present their interpretations of their research as though they are absolutely and objectively true. Then, people hear researchers' claims about sex differences, assume they are true, and raise their children accordingly; and some of those children become scientists who investigate sex differences, and thus the cycle of bias continues.

This book is about how scientists have looked at women and men. Scientific research is intended to be a way of trying to find out the truth about the world. It is a way of asking questions and seeking answers. The thoughts and feelings of scientists influence what questions they ask and how they are answered. For instance, the research question, "Do women's cognitive abilities decline when they are premenstrual?" is likely to yield answers that may cast women in a bad light. By contrast, the question, "Do females' and males' cognitive abilities show cyclical patterns over time?" may yield information from which one might draw rather different conclusions. The answers we get always depend partly on the way we ask the questions.

In the midst of the twentieth century's information explosion, it has become impossible for any one person to stay informed of the results and the strengths and weaknesses of all the research that is important in our lives. Therefore, we often accept some scientists' claims as facts, not knowing that their approaches were narrow, biased, or otherwise limited. That means that our view of reality has become distorted. The purpose of this book is to assist those who wish to expand their vision by questioning some of the "facts" most of us have heard about males and females. Practicing a questioning, thoughtful approach to issues of sex and gender, and learning some of the common pitfalls in that area, is also helpful in developing the capacity to do careful thinking about other issues that are replete with bias, such as the work on race, class, age, sexual orientation, and so on. The critical thinking skills presented in this book can help us not just in knowing what to ask about research reports in scholarly journals but also in thinking about claims that are made in the popular media, by our co-workers, and by our friends and family and that can affect our feelings, our personal lives, and our experiences at school and work.

❖ TWO DANGEROUS ASSUMPTIONS

As you read and think about the research on sex differences, you will need to be aware of two major but *wrong* assumptions that have muddied our understanding of this work. They are:

1. The assumption that if we find a "sex difference" in some ability or kind of behavior, that means that all males do a particular thing and all females do some quite different thing (e.g., all males are aggressive, and all females are passive and peace-loving). If asked directly, most researchers would probably acknowledge that in every realm of psychological research, females' and males' test scores or behavior overlap a great deal. Finding a "sex difference" does not mean finding that all women are one way and all men are another way (Hochschild, 1973). For instance, even when a research team reports that men are more aggressive than women, that does *not* mean that *no* women are aggressive and *all* men are aggressive, nor does it mean that all men are aggressive to the same degree. But we need to remind ourselves repeatedly that few researchers or laypeople remember how much of males' and females' behavior is similar—or, in other words, how much overlap there is—when we hear the term *sex differences.* Hearing about a study that "proves" there is a sex difference in math ability, for instance, we often come to expect most or all females to perform worse than most or all males on math tests, although in fact the overlap in their scores is extensive. What is commonly called a *sex difference* is the difference between the *average* score of the women who were studied and the *average* score of the men who were studied. An average score is reached by adding up all the individual scores and dividing by the number of individuals. Most individual women and men do not have scores (or behavior) exactly like the average score for their sex. This means that even when a "sex difference" is found in a study, we can't predict how any individual will behave if all we know is the person's sex.

Another reason that most sex differences seem more extreme and dramatic than they really are is a result of the way most research is done. Researchers are more likely to predict that they will find *differences* than *similarities* between groups. They tend to look for differences because, if we give boys and girls a test and find that the sexes perform differently on it, we will probably get little argument if we claim to have found a sex difference. But if we give them a test and find that they do *not* perform differently, then it is harder to claim that there is no sex difference in the ability or behavior that that test is supposed to measure. It is very hard to prove convincingly that a difference between any two groups does *not* exist, since it is possible that there is a difference but you missed it. People can always make such claims as, "You didn't test *enough* children to get a difference; there probably is a sex difference, but it is small" or "Maybe that just wasn't a very good test for measuring skill or behavior X" or "Maybe the children you tested are not typical of most children." The term that is used to refer to trying to prove that there is *no* difference is *trying to prove the null hypothesis.*

2. The assumption that sex differences are biologically based and, therefore, inevitable and unchangeable. This is an unfounded assumption. Many differences result from the different ways girls and boys are raised, and even differences that *may* have *some* biological basis—such as differences in height—have been shown to be fairly easy to modify (Hubbard, 1990; Hubbard et al., 1982). In fact, although we tend to think that nothing changes our genes, biologists now know that genes can be changed by the chemical processes in genes near to them. This means that what seems to be a simple, straightforward question—"Is a particular sex difference caused by biology or by the environment?"—is not really so simple. As biologist Margaret Thompson has said, "The environment of genes is other genes."

It is important to understand that it is not so easy to distinguish the contributions of nature or genes from those of socialization, experience, or other environmental factors, because some people are quick to claim that what is biologically caused is not only natural and inevitable but even morally right.

For instance, some people have claimed that it is *natural* for women to stay at home fulltime to raise children while men leave home all day to work for pay. Some people believe that what leads to this sex difference (which isn't even very common these days) is the fact that women carry the fetus and then nurse the newborn. These same people believe that pregnant women and nursing mothers have no business being in the workplace; they tend to think that we would be tampering with Mother Nature if we tried to change that pattern, and even that that would be morally wrong. In contrast, others believe that the woman-at-home and man-in-the-workplace division came about because employers have historically paid men higher salaries than women, and so families with young children have found that it makes sense economically for the man to do the paid work; the people who believe that do not tend to feel so uncomfortable about deviations from that pattern, they don't tend to think that a deviation is morally wrong or unnatural, and they sometimes even think that what is morally wrong is paying men more than women for doing the same kind of work.

❖ WHAT YOU WILL LEARN

In this section we explain the details of the main goal of this book: to teach you to think critically about sex and gender. Subsequent chapters are directed at different aspects of achieving this goal. The chapters constitute a cumulative learning experience, but each stands on its own as well. To be able to deal adequately with the science of men and women, it is important to be aware of the variety of factors that are involved in the scientific process. In this way, you will learn the essential skills for making critical judgments of your own.

What do we mean by *sex* and *gender?* We shall use *sex* to refer to the biological sex of the individual—whether a person is born physically female or male. Sex is determined by the genes. In most, but not all, cultures, people assume that there are only two sexes: male and female. (In North America, most people think that females have two X chromosomes and males have one X and one Y chromosome, although in fact more than two genetic sex types do occur more often than most laypeople realize.) We shall use *gender* to refer to the social role of being a woman or being a man. Gender means "being feminine" or "being masculine," standards that look different in different societies. Gender is composed of the whole list of features that the society in

question labels as appropriate for, or typical of, one sex (but not the other, or more than the other), including feelings, attitudes, behavior, interests, clothing, and so on. The issue of how biological sex and social gender interact—how much our "masculine" or "feminine" behavior is unavoidably determined by our physical sex—underlies most of the controversies in the science of sex and gender.

In this book, as you learn to think critically about the scientific study of sex and gender, our specific goals are for you to:

1. learn how science is conducted—both actually and ideally;
2. become increasingly able to evaluate scientists' work (for example, to recognize that no scientist—and, therefore, no science—is completely free of prior expectations);
3. develop the conceptual tools you need in order to think critically about research;
4. question people's (especially scientists') expectations of and perspectives on women and men;
5. develop an awareness of the limitations of any individual perspective and use this awareness to analyze the limitations in perspective of all sources of information: TV, newspapers, scientific journals, and other types of media. This means ascertaining and questioning every author's frame of reference;
6. come to treat your own and others' expectations of men and women as theories, not facts, which may be confirmed or challenged by the use of logic and evidence;
7. become able to evaluate these hypotheses, or working assumptions, for their usefulness in discovering some aspect of truth;
8. learn to discuss all the different ways any given hypothesis could be tested;
9. increase your awareness of the scope or limitations of the methods used to test any particular theory (i.e., If a different test had been used, how might the results have been different?);
10. want to strive to generate as many different interpretations of the evidence as possible;
11. apply your critical analysis to statements made by scientists, or by people in everyday life, about the nature of women and men;
12. examine how language influences our perceptions of men and women, as well as how our perceptions of women and men influence our use of language about them;
13. explore the impact our beliefs about women and men have had on scientific theory and practice; and

14. consider how scientific theories of sex and gender have affected, and continue to affect, our lives.

Everyone who develops the ability to do these kinds of thinking—be they researchers or laypeople who hear the claims of researchers—acquires the power to help stop the cycle of bias.

❖ OUTLINE OF THE TEXT

Thinking Critically About Research on Sex and Gender begins with a summary of this book and what you can expect to learn as you use it.

The second chapter is a history of the science of sex differences. Understanding the science of the nineteenth century helps us to see clearly that science can be used to prove a point that has profound social and political consequences. For instance, when we read about scientists' intensive efforts in the nineteenth century to prove that men's brains were bigger than women's—and that, therefore, men were smarter than women—we may think it is a quaint example of unsophisticated research from the olden days; but once we are familiar with that research effort, we can more easily see the bias and the political or social consequences of research in our own era on sex differences in the size of *one particular segment of one part* of the brain. We now recognize that the nineteenth-century science that was used to "prove" that men were smarter than women was filled with or based on faulty logic and primitive research methods. Analyzing the biased science of the nineteenth century sharpens our ability to question modern science. Learning about this history also reveals the roots of today's scientists' attitudes toward women and men.

The third chapter is a description of *scientific method*—the way of doing scientific research—and some of the most common mistakes scientists make in their research on sex and gender.

The rest of the book deals with some of the most important current issues in the science of sex and gender. Some chapters are about sex differences; others are about women. Some include a focus on men, but the reason that we focus here somewhat more on females than on males is that the focus of most research has actually been on males. The major exception to the male-oriented focus of traditional researchers has been research that has been focused on women's and girls' supposed inferiority or pathology. In some chapters, we choose a single piece of research and examine it in great detail. In those chapters, the studies were chosen because they have been extremely influential, because they illustrate many of the research errors made in the field, or both. The topics covered in the following chapters include:

Math ability. Males are generally said to be superior to females in math ability. Some scientists say this is determined by a variety of social influences, including how girls *and* boys are encouraged in their math studies, whether they are taught math by teachers of their own sex, and whether being good at math is valued equally for boys and for girls. Other scientists consider biology to be the major factor determining males' superiority to females on some tests of mathematical ability.

Spatial ability. The ability to read maps or find one's way through mazes and other similar abilities are said by some to be another arena of males' innate superiority. Other researchers say that spatial abilities have been tested in ways that favor males and that when other tests are used the results are different. From this point of view, sex differences that are found in spatial abilities can be explained by the different experiences of girls and boys.

Women and masochism. Some scientists have suggested that women enjoy being hurt. This theory has been challenged by those who say that women stay in harmful or upsetting situations for various reasons, not one of which is enjoyment of misery.

Males and aggression. Many researchers conclude that males are "naturally" more aggressive than females, but the research results vary with the definition of aggression.

Mother-blame. The majority of explanations for the emotional problems of both adults and young people are based on mother-blame. Why *is* that, and what other factors could be involved?

Women and hormones. Many people, including many scientists, believe that women are mentally and emotionally unbalanced because of their menstrual cycle. Some researchers challenge this idea, suggesting that social influences, not hormones, cause the symptoms, also noting that men's hormonal cycles do not get the same attention as do women's.

Verbal ability. One of the few realms in which females have been considered superior is that of verbal ability. But even that supposed *advantage* has been used against women, and the research is riddled with problems.

Dependence of females. Girls and women have long been regarded as the more dependent sex, but recent work suggests that much of the relevant research actually shows them to be *skilled in forming and maintaining relationships* rather than dependent.

❖ KEEPING SEX DIFFERENCES IN PERSPECTIVE

When we study the research on sex differences, we can get so absorbed in thinking about the details of the studies that we lose sight of the larger perspective. Part of the larger perspective that we need to keep in mind is that, since each scientist will be able to explore only a limited number of research questions, there must be a reason that some choose to spend their lives trying to find sex and gender differences.

Since most "proof" of differences between groups is usually used to "prove" that one group is *better than* the other, and scientists are aware of this, we need to ask what motivates them to pursue such research. A few hope to prove that there are fewer sex differences than people thought there were; many, however, seem to be intent on justifying the treatment of females as inferior in terms of being, for instance, less intelligent, "overly emotional," or more dependent than males. Scientists who try to prove that there are important differences between members of different *races* are usually recognized, these days, as racist, but those who try to prove that there are important sex differences are not usually recognized as sexist.

Another part of the perspective that we need to maintain involves a clear view of *which* research questions become the focus of the greatest amount of research. For example, although early sex-difference researchers reported that females were superior to males in various verbal abilities (e.g., learning to speak at younger ages, developing greater vocabularies) and that males were superior to females in spatial abilities, *most* of the research effort has gone into work on spatial abilities. So has most of the attention from the media. The research effort has included trying to document how great the male superiority is and developing theories to explain why males are so superior in this regard. If we become caught up in exploring the details of the spatial abilities research, we fall into the trap of *assuming* that males are superior, forgetting that there is also evidence that females are superior to males. The goal should not be to reverse the pattern and focus on areas where females outperform males; rather, it should be to take care not to let our beliefs be shaped by the research topics that receive the most attention from scientists and from the media reports about them.

We do not claim that there are definitely no sex differences in humans' behavior. What we do believe is that, since so much of the research is deeply flawed, and since males and females have nowhere been treated identically from birth, it is virtually impossible to know what inevitable sex differences there might be. And if it seems to you, as you read this book, that most sex-difference research is riddled with problems, you are right. This is partly because of the difficulty of studying human behavior, which is so variable and complex, and partly because of researchers' biases and failures to plan their studies as carefully as they might.

We do not believe that most or all sex-difference researchers have consciously and purposely set out to do research that is harmful or demeaning to one sex or the other. We do believe, however, that it is hard, if not impossible, for any of us to be aware of all our biases and unquestioned assumptions, and those of us who do research will bring those factors into our research, like it or not. We are all products of our time and culture. No doubt the same applies to us in the writing of this book. So, after reading through the following chapters and honing your critical thinking skills, you may wish to apply them to the arguments and reasoning we use in this book.

A word of warning is in order. Some students have been accustomed to believing that scientists and teachers are always right, and they sometimes find it upsetting to be shown that the so-called experts have often made significant mistakes, unintentional as they may have been. It can feel like the rug of certainty pulled from under us when we start to question what we read or what

we are taught. To be sure, we cannot promise to give you new, absolute truths to replace some of the certainty you may lose, but we believe that it is important to know when what we thought was absolute truth is only partial or even nonexistent. Better, we feel, to know the limits of our knowledge than to believe we know more than we actually do.

Furthermore, developing critical thinking skills does *not* leave you with nothing. Instead, it leaves you with a wealth of important abilities which enable you to grapple with research in an active way. And by approaching research with an active mind, you will be in a good position to see which research *is* reasonably done and which researchers try to identify and freely acknowledge their own biases.

When we encounter experimental errors, we should not be surprised; after all, we cannot know anything in this world with absolute certainty. Naturally, error should be minimized, but we can go only so far. The important thing for researchers is to be as accurate as possible but to make sure the conclusions that they draw do not go beyond what the study's method and results, combined with the experimenters' biases, really show.

In the next chapter we shall look briefly at the ways that scientists in an earlier era made claims for their research that went well beyond their methods and results and failed to deal with their biases.

References

Hochschild, A. (1973). A review of sex role research. In Joan Huber (Ed.), *Changing women in a changing society* (pp. 249–267). Chicago: University of Chicago Press.

Hubbard, R. (1990). *The politics of women's biology.* New Brunswick, NJ: Rutgers University Press.

Hubbard, R., et al. (Eds.). (1982). *Biological woman: The convenient myth.* Cambridge, MA: Schenkman.

CHAPTER TWO

A BRIEF HISTORICAL PERSPECTIVE ON SEX-DIFFERENCE RESEARCH[1]

Research on sex differences in behavior has a long history and has aroused intense scientific and public interest in the nineteenth and twentieth centuries. In studying current sex-difference research, knowing something of its historical background is important for several reasons. First, we can learn from the mistakes of those who have gone before us. If we understand the errors researchers have made in the past, it can help us to try to avoid them in our own new research and to progress more efficiently. Second, knowing the history of sex-difference research helps us understand people's attitudes today. Experiments from the last century may sound quaint, naive, and obviously biased to us. We might assume that no researcher today could ever do that kind of research. However, becoming familiar with the faults in those early studies makes it easier to spot the (sometimes more subtle) descendants of such experiments in our own day. Often, we find that studies from a hundred years ago—and the attitudes and prejudices that characterized their experimenters—are really not very different from those of today.

In this chapter, then, we examine some of the influential early research on sex differences, pointing out some common themes, identifying some of the significant and common errors researchers have made, and studying the various assumptions and attitudes displayed by the people who have done the studies and interpreted the results. Then we shall note some parallels in the research of our own era.

1. This chapter was extensively based on Catherine Gildiner's 1977 paper, "Science as a political weapon: A study of the nineteenth century sex differences literature." York University, Downsview, Ontario.

❖ WHOSE HISTORY IS IT, ANYWAY?

Our culture's modern science was born and reared primarily during the Victorian Age in Europe. At that time and in that place, the people who were the most influential in scientific research were white, middle class or wealthy, and male. The people of that era had been heavily influenced by the Judeo-Christian tradition, which was filled with stories that were used to "prove" women's inferiority to men. At the very beginning of the Bible, for instance, Eve—who symbolizes women in general—is described as having introduced evil (sin) into the world. Although many people might interpret that story differently today, in the Victorian Age this was the most common interpretation (Gildiner, 1977). (In fact, it is worth noting that Eve's "sin" was her desire to eat the fruit from the tree of knowledge. This has been regarded as evidence of women's inferiority or even dangerousness when white, wealthy men control not only religious institutions but also the production of "knowledge" by the institution called Science.)

Religion had for centuries been the authority on human nature. In the nineteenth century the scientific method became a popular way to find out about the world. When people wanted to know about the nature of the world and about the right ways to behave, instead of going directly to their religious leaders for directions about how God wants people to act, they turned increasingly to scientists for answers. Science came to be highly respected, and it has even been said that in some ways it *replaced* religion (Young, 1971).

For the past two centuries, most scientists in the Western hemisphere have come from cultures characterized by certain powerful beliefs, including the belief in the intellectual inferiority of women and of anyone who was not white, and this has profoundly affected the directions research has taken. In fact, most scientists have themselves been members of the privileged class, race, and sex, so it is not surprising that the majority have chosen research questions that have helped to perpetuate the view that members of their group are superior. For example, instead of setting out to investigate *whether* males have superior intellectual abilities, scientists have tended to help maintain the status quo by trying to determine *why* males are intellectually more capable.

❖ THE SEARCH FOR PROOF OF WOMEN'S INFERIORITY

The Great Brain Hunt

Let us look at how scientists in the nineteenth century set about their search for the answer to the question of why women are intellectually inferior to men. As we do this, let us keep in mind that then, as now, researchers may base their research questions on assumptions that they simply believe to be true, without questioning those assumptions—in this case, that women are intellectually inferior. When research is based on an unchallenged assumption

it is extremely unlikely that the results of that research will ever lead anyone to question that assumption; for instance, when researchers ask simply *why* it is that women are not very smart, the kinds of studies they design are likely to produce information that seems either to support their assumption or to shed no light on it, but not to disprove it.

One popular notion in the nineteenth century was that women were less intelligent than men because women's brains were smaller than men's. A scientist named George Romanes (1887) claimed to have proven this. Women have smaller heads, he said, therefore they must have smaller brains, and therefore they are less intelligent. Those who wished to believe in females' inferiority thus believed that data had been found to explain it. Rarely did anyone even raise such questions as, "If a woman had a very large head, would we say she is smarter than most men?" Romanes's claim was eventually discredited (Gildiner, 1977).

The work of Romanes illustrates yet another important point: When "data" seem to confirm what is already believed, not only is the assumption underlying the research hardly ever questioned but also people go on to construct elaborate theories based on that assumption and those data. For instance, as a result of Romanes's conclusions, other writers then went on to say that women's lesser intelligence was actually necessary for the survival of the human species, since women—having no intellectual interests—would be free to devote their energies exclusively to bearing and raising children (Mobius, 1901). Then, as with many theories, a considerable number of scientists and laypeople became so fascinated by the survival theory about sex differences in intelligence that they were even *less* likely to devote their energies to going back to Romanes's study and thinking critically about how he conducted it and the merits of his underlying assumption. This process continues today, as we describe in Chapter 4, on spatial abilities.

The long life of beliefs based on Romanes' research *after* the research had been discredited reflects another characteristic of the way people treat scientists' claims about their research: Many scientists and laypeople (and, today, the media) become intensely interested in an issue, believe a report of some bit of research about that issue, and then lose interest in it (Davidson, 1991). If, later on, the research they had read about is discredited, they may have become so accustomed to believing in that early research that they do not invest the mental and emotional energy necessary to revise their belief. This is particularly true when the earlier research seems to confirm what they, for their own personal and/or political reasons, *wanted* to believe. In this way, although the claims about brain size determining sex differences in intelligence were later discredited, many people continued to believe both in females' intellectual inferiority and in Mobius's "explanation" of the survival value of that inferiority.

Some researchers, however, rather than clinging to belief in the Romanes research, clung to their basic assumption about the intellectual inferiority of females but decided it would be important to find out where, other than in simple brain size, that inferiority was based. There then followed decades of what may seem to us today to be amusingly misguided experiments—

researchers comparing one aspect or segment after another in the brains of the two sexes, desperately seeking difference.

After most people gave up on believing that bigger brains were better, some scientists, still assuming women were the less intelligent sex, proposed that the sex difference might be *relative* brain size, that is, the size of the brain relative to the size of the body. But this effort backfired—relative to body size, it was learned, women's brains were actually *larger* than men's.

So, researchers gave up on the whole-brain theories, but the quest continued. If it wasn't due to the size of the whole brain, scientists speculated, perhaps there was a difference in the size of one *part* of the brain, whichever part might be the crucial seat of intelligence. This, too, turned out to be a fruitless direction for the researchers' purposes, however. Scientists checked one segment—or lobe—of the brain after another, expecting that each part would be larger in men than in women. In no case was the expectation borne out.

Today, we might think that such studies are no longer done, especially since it is well known that chemical and electrical changes, not sheer volume and weight, are the keys to brain functioning. Furthermore, it is now known that intelligence does not lie in only one site in the brain. Rather, each of a vast number of different parts of the brain is related to one or more of a vast number of intellectual abilities. However, in 1987, Dr. Ruth Bleier told an important story to the American Association for the Advancement of Science. She described a piece of research that had been published in the highly respected journal *Science* (deLacoste-Utamsing & Holloway, 1982), in which it was claimed that the *splenium,* the back part of the membrane which separates the two hemispheres of the human brain, was bigger in females than in males. The authors of the study suggested that this sex difference might explain females' supposedly inferior spatial abilities (see Chapter 4 for a detailed discussion of this topic). Bleier and her colleagues carefully read that study and found that it was filled with major flaws, not the least of which was that only nine males' and only five females' brains had been studied. As part of her presentation, she showed a slide on which pictures of the splenium from a number of males were displayed in one column, and those from a number of females in the other. She asked the audience to determine by looking which were the males' and which were the females', or which column had clearly larger segments. It was clearly impossible to tell from simply looking at them. Bleier explained that she and her colleagues then looked at the spleniums from a much larger number of people of both sexes, had the segments carefully and objectively measured, and found no sex difference at all. When they submitted their article about this work to *Science* magazine, it was turned down on the grounds that it was too "political."

This illustrates another important point: Historically, whether in the nineteenth century or in our own time, research that supports the beliefs of the people in power is likely to be readily and unquestioningly accepted as legitimate research; however, research that might lead people to question those beliefs is scrutinized for methodological flaws and is dismissed as "motivated by political aims." Of course, research that supports the status quo may also be motivated by political aims, but that does not tend to be recognized, because those who control the political and scientific arenas simply feel that it confirms what they already "knew" to be true.

During the nineteenth century, as one by one the parts of the brain were found not to differ in females and males, scientists reached further to try to find the physical location of what they believed to be males' superior intelligence. For instance, they measured the length of the spinal cord, putting together convoluted explanations for how a sex difference in spinal cord length might be related to intelligence. And this was done in all seriousness.

In this research enterprise, scientists not only measured parts of the body but also administered various tests to people of both sexes. When females performed better than males on such tests, it seems that the researchers had to find some way to transform that information into further "proof" of males' superiority (Caplan, 1989). Romanes (1887), for instance, conducted a study in which he found that women could read faster and more accurately than men. Instead of simply concluding that women could read faster and more accurately than men—or even that, in this respect, women might be more intelligent than men—two prominent scientists of the era accounted for the sex difference by saying that the ability to read is coupled with the ability to lie, and women are better liars than men (Lombroso & Ferrero, quoted by Ellis, 1934). In this way, a finding of a female superiority in a skill was transformed into evidence of females' moral inferiority.

Not only data-gathering research but also theories were powerful tools in the nineteenth century for justifying the privileged positions of well-to-do, white males. We shall now look at what was probably the most influential of those theories.

Social Darwinism

A major theory that was used to "explain" women's intellectual inferiority is *Social Darwinism,* using Charles Darwin's claim that, as species evolve, the individual animals and humans that survive tend to be those that are the best suited to their environments. This was his concept of *survival of the fittest.* Social Darwinists reasoned that, therefore, whatever survives, including social and political structures, individuals, and aspects of human personality, must be the fittest. Thus, they said, what exists today must be the best possible state of things.

Since women were already considered less intelligent than men, it was argued that such a sex difference was necessary for the survival of the species (for instance, so that women could put all of their energy into bearing and raising children). Similar reasoning was used to justify a myriad of factors that had actually been imposed by society, not determined by evolution. For instance, the intense social pressure on women to behave passively did tend to make many women rather passive. But Social Darwinists then claimed that this was a biologically based trait, necessary for encouraging women's sexual passivity and receptivity, so that they would become pregnant and thereby help the species to survive.

Also related to Social Darwinism was the *maternal instinct* theory, according to which women have an innate desire to take care of children, while men do not. Therefore, goes the theory, men can develop other abilities, like intelligence and perseverance, while women must concentrate on nurturing and protecting. Even if there is some maternal instinct in human females, that would hardly explain women's "lesser intelligence." In fact, certain intellectual capacities are enormously helpful in making women (or men, for that matter) better nurturers, protectors, and conveyors to the young of information that can help them survive and flourish (Ruddick, 1989).

Still another theoretical tack related to Darwinism was a notion called *morphological infantilism.* Proposed by Darwin, it is the idea that women, being smaller than men, are morphologically (physically) more like infants and children than are men. Some Victorian theorists speculated that, therefore, women are less intelligent than men but more intelligent than infants and children. This is equivalent to saying that men must be more like gorillas than are women, because men are hairier: It might even be valid, but what reason is there to believe it? The reason for using the notion of morphological infantilism to "explain" women's allegedly inferior intellectual capacity was, again, to justify depriving women of legal, economic, and political power. The Social Darwinists who promoted morphological infantilism as applied to women used the same notion for keeping Whites in power over Blacks. Black people, they argued, are physically more similar to apes than are white people and are therefore less intelligent. Clearly, Black women were the group most demeaned by this sexist and racist theory.

Morphological infantilism and other Social Darwinist theories might seem to be another of the quaint nineteenth-century theories that seem shockingly prejudiced as well as groundless to us today. But in our own era, Philippe Rushton (1989) has attracted wide media coverage with his claims that such characteristics as brain size and numbers of offspring prove that "Orientals" are more advanced on the evolutionary scale than Whites, and Blacks are less advanced than both. Two of the key components in Rushton's argument are that the number of offspring a woman produces and the duration of her pregnancies are signs of her place on the scales of intelligence and evolutionary advancement. Thus, women's bodies again become a prime focus for arguments about which humans are inferior to which others. Although his work has been criticized in great detail for its extremely poor methodology and its racist and sexist qualities, some people nevertheless want to believe he is right.

❖ SUMMARY OF SOME PROBLEMATIC PATTERNS

In this chapter, we identified some of the problematic patterns of scientists' behavior in studies of the history of research that persist even today. These include:

1. beginning with a biased assumption (e.g., that males are more intelligent than females)
2. failing to question the assumption(s) underlying the research (e.g., failing to question whether the predominance of males in high academic and political positions is proof of males' greater intelligence)
3. asking research questions based on that assumption (e.g., "Is men's greater intelligence due to their bigger brains?")
4. when results of a study do not support the assumption, continuing to avoid questioning the assumption (e.g., if men's brains turn out not to be larger than women's, relative to their body sizes, then not questioning whether men are more intelligent)
5. misinterpreting research results that seem to contradict the assumption. (Thus, what had been considered a desirable characteristic—such as reading quickly—is portrayed as an undesirable one, or one that *leads to* trouble)
6. failing to question the evidence for, the logic of, and the damaging consequences of theories

Our predecessors in the history of research on sex differences certainly made major—and often damaging and oppressive—errors in conducting, interpreting, and theorizing about their investigations. That historical perspective is helpful to keep in mind as we turn in the next chapter to some specific methodological errors often made by researchers and then move in the following chapters to an examination of the research on particular topics.

References

Bleier, R. (1987, February). *Sex-differences research in the neurosciences.* Paper presented at the symposium on Bias in Sex-differences Research, American Association for the Advancement of Science Annual Meeting, Chicago.

Caplan, P. J. (1989). *Don't blame mother: Mending the mother-daughter relationship.* New York: Harper & Row.

Davidson, K. (1991, January 20). Nature vs. nurture. *San Francisco Examiner Image.* pp. 10–17.

deLacoste-Utamsing, C., & Holloway, R. L. (1982). Sexual dimorphism in the human corpus callosum. *Science, 216,* 1431–1432.

Ellis, H. (1934). *Man and woman: A study of secondary and tertiary sexual characteristics.* Cambridge, MA: The Riverside Press.

Gildiner, C. (1977). *Science as a political weapon: A study of the nineteenth century sex differences literature.* Psychology Department, York University, Downsview, Ontario.

Mobius, P. J. (A. McCorn, trans.) (1901). The physiological mental weakness of women. *Alienist and Neurologist, 22,* 624–642.

Romanes, G. J. (1887). Mental differences between men and women. *Nineteenth Century, 21,* 654–672.

Ruddick, S. (1989). *Maternal thinking: Toward a politics of peace.* Boston: Beacon.

Rushton, J. P. (1989, January). *Evolutionary biology and heritable traits (with reference to Oriental-White-Black differences).* Paper prepared for presentation at American Association for the Advancement of Science convention, San Francisco.

Young, R. M. (1971). Evolutionary biology and ideology: Then and now. *Science Studies, 1,* 177–206.

CHAPTER THREE

USING SCIENTIFIC METHOD TO STUDY SEX AND GENDER

Science is a method of asking questions and trying to find the answers. There are many ways to ask a question. As noted in Chapter 2, many centuries ago, people looked to religion for answers to their questions. They went to religious leaders or studied legends for explanations and for the truth. Nowadays, most people believe that scientists are the truth-knowers, and as a result, many people have great respect for the scientific method. Unfortunately, it is extremely hard to do flawless research about human behavior, so it is hard for scientists to find the Truth. In this chapter, we look at some of the most common kinds of errors scientists make in conducting their research, errors which make it difficult to judge how close to real knowledge their research has brought them.

The term *scientific method* is defined in the *New Webster's Dictionary of the English Language* as a research method involving the definition of a problem and the drafting and empirical testing of the hypotheses by gathering data. In other words, the scientific method is a way to conduct research to find out about something by using a plan. It is commonly believed that the scientific method is an objective way to find the truth—that it is not affected by scientists' beliefs, feelings, or biases. The scientific method is said to produce results that are reliable. If a scientific experiment is conducted well, then the results can be replicated or reproduced in another study, and this suggests that such results are true. If the results are not reproducible, then how can we tell which set of results is closer to the truth? A key feature of the scientific method is the careful documentation of *every* step of the procedure, so that *anyone* can reproduce the experiments to test the original results.

In reality, however, very few scientific studies are ideal; many things can go wrong and thus give a distorted picture of the topic being studied. In the area of research on sex and gender, research errors leading to distortions have been extremely common; as a result, we have often seen a very inaccurate picture of the similarities and differences between the sexes. It has been like looking in a curved mirror. In each of Chapters 4 through 11, some of the most important research mistakes related to particular topics are described.

We shall introduce you to some very common methodological errors so that they will be familiar to you by the time you reach the specific topic chapters. (We will not describe some of the more technical errors or those involving statistics, however; for some excellent sources of more information on this topic, please see the Suggested Readings section at the end of this book.) We hope that you will not feel so disheartened after reading about sources of error that you will want to ignore all research forever. It is our hope that with knowledge about common sources of error, you will be able to take into account the errors you find before you draw conclusions about a study. In other words, it is not always necessary to conclude that a study is worthless because it is not perfect, but your knowledge of its limitations should help you decide how to interpret it, the extent to which you can rely on its data, how important it is, and so on. Problems arise not so much because experiments are biased but because researchers and laypeople do not take those errors and biases into account in evaluating them. The interpretation of a study should take *every* aspect of the experiment—*including its limitations*—into consideration.

For purposes of illustration, we shall here use sex differences in strength as an example. In order to understand what can go wrong in using the scientific method, let us follow the steps of a scientific study.

❖ 1. CHOOSING WHAT TO STUDY

The first step in *any* scientific study is to decide what you want to find out. Scientists don't randomly choose what they study. Scientists are human and tend to study what interests them. Often, this means that they have strong needs to prove that something is true or false, and those needs can affect the way they ask the research question: for instance, they might study the question "Why are women so weak?" rather than "Under what circumstances can people become as strong as possible?" As noted earlier, a scientist may or may not be aware of these needs, and probably most do not purposely bring their biases into their research planning. But whether these biases and motives are conscious or not, and whether they are intentionally or unintentionally brought into the activity of choosing a research question, this is *very* different from the picture of the objective scientist. Most scientists have beliefs, hypotheses or predictions about the outcomes of their studies; as we shall see later, the researcher's beliefs and predictions can heavily influence the outcome of the study.

❖ 2. DETERMINING EXACTLY WHAT YOU'RE LOOKING FOR

You must define what you are looking for. This is extremely important. For example, if a researcher set out to explore sex differences in something we'll call *flugenransk*, but didn't define it, you would have absolutely no idea whether or not the study proved that the sexes differed in *flugenransk*. Without a clear and adequate definition, you cannot be sure that the tests you've chosen actually measure what you are trying to measure. Let's say that a researcher, Dr. Wright, decides to study sex differences in strength. There is more than one meaning of the word *strength*. One is physical ability, another is mental power, another is endurance (physical or mental), and another refers to smell or taste. Maybe Dr. Wright is trying to study sex differences in physical ability, or physical endurance, or mental ability, or emotional resilience, or perhaps she is trying to test whether women or men emit a stronger odor after exercise. Most likely she intends to study one of the earlier items on the list, but it is unscientific and inefficient not to specify that precisely.

With more complicated concepts, precise definitions are even more crucial. For instance, if you are studying "intelligence," you must say *exactly* what you mean by intelligence, because it has been defined in many ways, which include an astonishing variety of abilities (such as the ability to take in information, the ability to learn concepts, the ability to memorize, the ability to re-create with a pen and paper exactly what you see, etc.). If we don't know exactly how a researcher is defining a concept, we simply don't know what is proven by the results of the research.

Furthermore, to answer a scientific question, we must have a clear and precise question so that it can be answered with clarity and precision. There are a number of different questions Dr. Wright could be asking about sex differences in strength, for example, "Are there sex differences in strength?" or "Under what conditions are sex differences in strength the greatest?" or "What conditions cause sex differences in strength to disappear?" or "Are there sex differences in *every* kind of strength?" If she doesn't state her question accurately, the same problem arises as with failing to state the precise definition: She can't tell if she is actually testing her question, and she won't know if she was successful.

❖ 3. DESIGNING THE RESEARCH

The third step is to design some kind of method for gathering relevant information. Many things can go wrong here.

First, *the method must relate back to the research questions or hypotheses and/or theories on which the study is based.* If they don't, your conclusion will be irrelevant to your hypotheses. To take an extreme example, let's say that

Dr. Wright hypothesizes that males will be physically stronger than females, and she uses the following method:

1. She has each member of a group of people eat an apple.
2. She asks each one if they enjoyed eating the apple.

No matter what her results are, they're irrelevant, because they have nothing to do with her hypothesis about strength. Even if the males *did* enjoy the apple more than the females did, it wouldn't tell us anything about their physical strength. Of course, errors in research methods are usually much more subtle, but you must be careful to ensure that your method relates to your hypotheses.

Second, *certain methodological errors can skew the results.* Some of these can appear in any methods, while others are inherent only in certain *kinds* of methods. Following are some examples:

1. Experimenter Bias. Experimenter bias involves the intrusion of the researcher's beliefs or hopes into the actual study. Experimenter bias in the plan of a study can lead to very distorted results. For instance, suppose Dr. Wright believes and hypothesizes that males are physically stronger than females. That might lead her to test her hypothesis by comparing the number of males who have cargo-loading jobs to the number of females in such jobs. She would then be ignoring the fact that most of the carrying of toddlers and groceries (which also requires great physical strength) is done by females. In any experiment, it is almost impossible to eliminate bias completely. Therefore, it is important both to acknowledge that and to keep trying to be as objective as possible.

2. Errors in Cross-Sectional Research. If you want to measure change over time, there are two practical ways to go about it, each of which is problematic. One of these is *cross-sectional* research, which involves testing people from different age groups at the same time. Suppose Dr. Wright wants to find out whether physical strength diminishes more with age in males or in females. Using a cross-sectional approach, she tests females and males from different age groups (e.g., 10- to 20-year-olds, 21- to 30-year-olds, and so on) and notes the changes for each sex across the age range. Suppose that she finds that the older males are weaker than the younger ones but for females, strength doesn't vary from one age group to another. She would not be justified in concluding that males grow weaker over time until she could prove that the males of different ages had all had the same experiences at the same times in their lives. For instance, perhaps the teenage males in her study are required by a new Board of Education rule to take more stringent physical education courses than the men now in their twenties were required to take as teenagers. If that difference in life history distinguished men in the different age groups from each other, then we would not necessarily expect that the current teenagers will be as weak as the current 20-year-olds when the teenagers reach their twenties. Therefore, it would be wrong to conclude that each man grows weaker as he grows older.

3. Errors in Longitudinal Research. The other way to measure change over time is by using the *longitudinal* method, which involves measuring the same people several times as they grow older. For instance, instead of the cross-sectional method, Dr. Wright might decide to measure the individual strength of members of a group several times over a period of 10 years. She might find that as her subjects become older, the females seem to lose more strength than the males. She might conclude that, as people get older, women tend to lose their strength faster than men. However, life history can confuse the issue here, too. What if, during the course of the study, it becomes fashionable for women to appear as slim and unmuscular as possible? Then, some women would probably exercise less and, therefore, their strength would diminish from lack of use, not from age. It is extremely difficult, with this kind of method, to determine whether a pattern changed because the people grew older or because some other factor was involved.

4. Pretest/Posttest. Often, researchers will want to explore the effect of a certain type of treatment on a group of people. One method commonly used is the pretest/posttest method, in which the subjects are given a certain test before and after the treatment. If their test scores change, it is assumed to be due to the treatment. The problem with this method is similar to that for a longitudinal study, because even if there *is* a difference in the scores for the two tests, in most circumstances it is extremely difficult to be sure that the *treatment,* rather than some other factor, led to the change. For instance, suppose Dr. Wright decided to study whether there is a sex difference in the body's ability to become stronger from increased exercise. One way to test for this might be to measure the difference between people's strength before and after two months of intensive weight training. The problem is that something that happened during the two months could influence the outcome. What if, at that time, an advertisement appeared, encouraging men to take steroids? Then, the steroids would increase the men's ability to become stronger but would not affect the women. If Dr. Wright were unaware that the advertisement had appeared, she could not take its effects into account in interpreting the results of the research and might mistakenly conclude that men become stronger from exercise than women do. It is next to impossible to make sure that between a pre- and a posttest nothing will happen except what the researcher wants to change. This is less problematic when the time between the two tests is short.

5. Maturation: Maturation of the people being studied can introduce confusion. Consider the pretest/posttest example. What if people just change over time? What if, as time goes by, men just get stronger faster than women, naturally, even *without* the extra exercise? To determine whether extra exercise or maturation led to the result, the researcher needs to use a *control* group as well as the *experimental* group. She would simply test the control group at the beginning and the end of the week during which members of the experimental group do extra exercise. Any difference between the control and experimental groups is then assumed to be due to the extra exercise.

6. Test/Retest: There is yet another problem with the pretest/posttest method, but it usually applies to a somewhat more complicated kind of test. It has been shown that people tend to do better the second time they take a test than the first time. This may be because they get used to the format of the test, they don't waste as much time figuring out how to do the test, or they simply become used to the test and are therefore less anxious. No matter what the reason actually is, it is important to know that people tend to do better the second time they take a test. So, if an experimental group scores higher on a posttest than a pretest, it may be due to the second-test effect instead of the experimental treatment (such as extra exercise or training). One way of getting around this problem is to use a control group (i.e., these people would take the test and retest *without* receiving the training). Then, any test/retest effect would be seen in the controls.

7. Order Effect. Often, researchers study how a group of factors affects a group of people. However, the order in which the people are exposed to the factors can affect the outcome of the study. For instance, say Dr. Wright designs a complex test of physical strength, involving almost every muscle in the human body, in order to be able to test for *overall* strength. If she always started each subject with pushups and then situps, the outcome could be different than if she reversed the order. What if, for example, males are better at pushups than situps, and women are better at situps? Then, the women would be more tired after having had to do the difficult pushups and would have less energy for the easier situps, while the men would have plenty of energy for the task that is harder for them. It is always necessary to vary the order of items in a study to counteract this effect.

❖ 4. CARRYING OUT THE STUDY

As you can see by now, it is virtually impossible to develop a perfect research method. This is especially true for research in psychology, because it's much harder to do totally controlled experiments on human behavior than on a chemical in a test tube, for example. So if a method *seems* perfect, beware—there may well be some hidden problematic factors. But even if it *were* possible to *design* a perfect study, a great deal can go wrong as the researcher carries out the research. Descriptions of some of these pitfalls follow. Although many of the mistakes will seem glaringly obvious, the fact is that they are frequently made and not accounted for.

1. Accuracy of the Instrument. If the instruments used are not very accurate, the results will also not be very accurate. For instance, if Dr. Wright decides to measure her research participants' strength by seeing how hard they can push down on a scale, if she uses a scale that is normally used to measure the mass of a horse, then she will find very little difference among her subjects, whereas if she uses a more sensitive scale, she will be able to measure the differences more accurately. This principle does not apply only to physical instruments; for example, if a math test does not include math problems of a wide variety of types and of degrees of difficulty, the researcher cannot make valid claims about what the results show about sex differences in math performance.

2. "Mortality." If a researcher starts out with a certain number of participants in an experiment, but some drop out (for whatever reason), this could seriously affect the results. This dropout is called participant "mortality," regardless of the reasons for the dropout. For instance, suppose Dr. Wright decides to study the effect of extra exercise on the strength of males versus that of females. Her method is to develop a weight-training program for all the participants and then to measure changes in strength. Now suppose that one-third of the females drop out of the experiment. What if those females dropped out because the weight-training program was too hard for them? These people could be the weaker ones, who might have shown more improvement in strength than the others. Dropout often occurs for some significant reason that distorts the results in a systematic way. Therefore, any mortality at all in a study should make you think twice about accepting the results.

3. Self-reported Observations. Sometimes a researcher bases a study on participants' reports or observations. Suppose Dr. Wright decides to measure sex differences in strength by asking people how much weight they can lift. Asking them might seem like a good way to determine how much weight they can lift, but the results from the experiment still won't be very accurate. Maybe some subjects don't like to admit that they're not strong. If males are more embarrassed about seeming weak than are females, then the males would appear to be stronger than they were. Depending on the purpose of the research, self-reported observations may give the most accurate or relevant information, but care must be used in interpreting them. It is too easy to make a mistake. You must consider all possible reasons that researchers might distort or fudge their results (intentionally or unintentionally).

4. The Participants Know Too Much. Sometimes the participants figure out why they are being tested, and as a result, they—purposely or unintentionally—act differently, thereby changing the outcome. Other times, they *think* they know what the researcher is looking for but they are wrong. As a result, they also act differently and skew the results. And although occasionally researchers will ask the participants, after the experiment, how much they had known, usually that does not happen.

5. Sampling Error. Often it is impossible to test *everyone* in a group of people you wish to study. Therefore, you may need to test only a *sample* of the total population and *assume* that they will probably perform roughly the same as the whole population. For instance, if Dr. Wright wanted to find the difference in strength between the average woman and the average man, she couldn't *possibly* test *all* the men and women in existence. Instead, she might choose five people of each sex at random, test them, and assume that they represent the general population. But what if they *don't* represent the total population? What if she just *happened* to pick the five strongest men in the world? Or what if she happened to pick five women, each of whom is a full-time mother or has a demanding paid job and has no time for exercise? Then, her results wouldn't accurately represent all women and men. What if she used a hundred people instead? The chances that *they* accurately represent the total population would be higher, and therefore, her results would probably be more accurate. If she used a million people, they would most likely provide an even more accurate estimate. In summary, if you don't test the whole population, you can't be absolutely certain that your results are accurate, but the larger the sample, the more accurate your results are likely to be.

6. Experimenter Bias. If Dr. Wright expects certain results and is the person who is doing the actual data gathering, then she may tend to make errors that conform to her expectations. For instance, if she is measuring how many kilograms of potatoes people can lift, and she assumes that men can lift heavier loads of potatoes than women can, consider what may happen when it comes time for her to weigh the potatoes. As she weighs a sack of potatoes lifted by a woman and she thinks the scale may read either 39 or 40 kilograms, her expectations may lead her to record the weight as 39—and to do the reverse if she were weighing the same sackful for a man. Sometimes it is possible to avoid this problem by altering the design of the experiment slightly—for instance, she might ask someone else (who doesn't know the sex of weight lifter) to read the scales.

❖ 5. INTERPRETING THE RESULTS

Consider the story about a misguided fellow trying to teach his students about interpreting scientific research. He showed his students a normal frog and said, "Jump!" to the frog. It jumped. Then, demonstrating how one records scientific observations, he wrote on the blackboard, "Frog with four legs jumps." He then cut off one of the frog's legs and said, "Jump!" It jumped—less gracefully, of course. This time he wrote on the board, "Frog with three legs jumps." He cut off a second of the frog's legs, then a third, and after each time gave the same command. Each time the frog jumped, although more awkwardly each time. Accordingly, he wrote on the board the lines "Frog with two legs jumps" and "Frog with one leg jumps." Finally, he cut off the fourth leg and said, "Jump!" When the frog failed to jump, the teacher instructed the students: "Frog with no legs becomes deaf!"

The final step in any scientific study is interpreting the results. In many ways, this is the most important step. If there were any possible sources of error in the study, here is where they must be reported. The researcher must take into account *every* detail of the study and come up with an interpretation related to what *actually* happened. There are several ways this might *not* come about:

1. Cause and Effect Problems. If one event, A, was found to happen only when another event, B, happened, you might conclude that B caused A or that A caused B. These are valid *possibilities*, but you must always remember that there is a third possibility: that a further event, C, might have caused *both* A and B. For example, suppose that Dr. Wright observed that women who ate healthier foods were stronger that women who did not. She might conclude that the type of food women eat affects their strength, or that being strong causes women to eat healthier food, but there is a third possibility—that an external event (e.g., the women's attitude about general health) affects both their strength and their diet. She must take into account all three possibilities.

2. Different Interpretations. There are usually many different ways to interpret any result. This doesn't mean necessarily that *all* of them are right or that *none* of them is right. It just means that if you find certain results and think of *one* way to interpret them, you must remember that it isn't necessarily the *only* reasonable way. For instance, suppose Dr. Wright sets up an experiment in which women and men are given a very heavy weight and a skateboard and are told that they must transport the weight a certain distance. Now suppose that most of the women use the skateboard to help them move the weight, while most of the men don't use the skateboard. Dr. Wright might conclude that the men are stronger, that the women just aren't strong enough to move the weight without the help of the skateboard. Or, she might conclude that the women are lazier than the men. Or, she might believe that the men's need to be macho would prevent them from using the skateboard. Or, she might say that the men have weaker eyes and don't notice that the skateboard is there. Or, she might decide that the men aren't smart enough to figure out how to use the skateboard to make the chore easier for them. In other words, no matter what we conclude from this experiment, all we can *really* be sure of is that more of the women used the skateboard than the men. We can make up interpretations, but we must remember that the *real* reason for the results could easily be something that has never occurred to us.

❖ 6. META-ANALYSIS: COMBINING STUDIES

An increasingly popular technique is *meta-analysis*, which is a way of using statistical methods to combine a large number of studies and analyzing their results as a group. Some people have claimed that this is useful, because, for example, you need not ignore studies that were conducted on a very small number of people. Some also say that, although Study A may suffer from Methodological Problem X, Study B from Problem Y, and so on, if you combine many studies, their good parts probably tell something important. Although there are some mathematical formulas that can help to minimize the effects of the problems, if we combine many studies, each of which is flawed in some ways we know about and probably in others we haven't yet recognized, then several problems arise:

1. It is hard to know whether the problems cancel each other out or just add up to more problems.
2. Because the statistics are fairly sophisticated, we may assume that they will yield some important truth. But both simple and sophisticated mistakes can be made using sophisticated techniques.
3. As we have seen in this chapter, even *defining* what we are studying and choosing the *ways* to study it can be difficult. If each study (or most) in a meta-analysis suffers from some lack of clarity of definition, if different ways are used for measuring the behavior, or if somewhat different forms of behavior were studied, then we are combining apples, oranges, walnuts, and maybe ironing boards! It doesn't make logical sense to analyze such a mixture as though the elements were nearly the same.

Now you are familiar with many of the research problems that plague the study of sex and gender. In the following chapters, you will see how these and other kinds of problems arise in the research and lead to mistaken impressions about people of both sexes. Remember that our aim in this book is to focus on the limitations of research, the factors that should make us slow to accept scientists' claims about their studies. But that does not mean that no research is ever helpful in moving us toward important knowledge. For instance, although longitudinal studies have the kinds of drawbacks we have noted, if they are carefully planned, if the data are appropriately analyzed, and if the results are responsibly interpreted, they can provide useful information (such as the relationship between certain food elements and disease). Sometimes, there is no substitute for research in answering important questions. But these questions must be carefully considered, and the research should be conducted with an acute awareness of its limitations.

CHAPTER FOUR

SEX DIFFERENCES IN SPATIAL ABILITIES[1]

"Women can't read road maps. They're just no good at doing spatial tasks," people often say. The claim that females have inferior spatial abilities is usually based on such tests as map-reading, maze-drawing, and picturing how block structures would look if they were turned around. The belief in this alleged inferiority has been used to justify keeping girls and women out of advanced science and mathematics classes and out of such careers as engineering, scientific research, architecture, building construction, various forms of navigation including piloting aircraft, and map development and design.

Especially when a claim about sex differences has such extensive impact on people's education and career tracks, it is important to look carefully at whether that claim has any basis in fact. To consider the topic of sex differences in spatial abilities, we shall be asking three sets of questions:

1. What are spatial abilities?
2. *Are* there sex differences in spatial abilities? If so, how significant are they?
3. Why have people believed that there are such big differences in these abilities?

❖ WHAT ARE SPATIAL ABILITIES?

Trying to define and describe spatial abilities is not like trying to define and describe a table. If five people are sitting at a table, they can all agree that

1. Portions of this chapter were excerpted from Paula J. Caplan, Gael M. MacPherson, and Patricia Tobin, "Do sex-related differences in spatial abilities exist? A multilevel critique with new data," *American Psychologist, 40* (1985), 786–799.

they are sitting at a wooden piece of furniture that has four legs and a horizontal top. *Table* is something which clearly exists and can be easily identified by anyone. The concept of *spatial abilities* is a totally different matter: If you ask five so-called "experts" how they define or identify spatial abilities, you will probably get five somewhat different definitions, and within each definition you will likely get vague or confusing terms.

Let us look closely at some examples of these definitions and then think about what we learn from them. We shall go into a great deal of detail, because in order to know what we are dealing with when we talk about spatial abilities, we must first determine whether we even know what spatial abilities are.

Harris (1978) has written:

> Spatial ability has been variously defined: "to move, turn, twist, or rotate an object or objects and to recognize a new appearance or position after the prescribed manipulation has been performed" (Guilford, 1947); "to recognize the identity of an object when it is seen from different angles" (Thurstone, 1950); "to think about those spatial relations in which the body orientation of the observer is an essential part of the problem" (Thurstone, 1950); "to perceive spatial patterns accurately and to compare them with each other" (French, 1951). Each characterization implies mental imagery, but of a distinctly kinetic rather than a static kind. (p. 405)

Lips, Myers, and Colwill (1978) wrote: "Spatial abilities are those that enable a person to locate an object in space, mentally rearrange objects, recognize shapes, and so on. This broad class of abilities is tested using tasks such as block design, jigsaw puzzles, mazes, and matching forms" (p. 156).

Now, what can we say about these definitions? Two of the three features of spatial abilities as defined by Lips et al. do *not* involve mental rearrangement or rotation, which was at the core of the definitions Harris considered. Their use of the phrase "and so on" adds nothing that helps us define or identify spatial abilities. Maccoby and Jacklin (1974) reported that "spatial ability, even more than verbal or quantitative ability, is difficult to define" (p. 91). They then, however, listed a host of factors that they said might be included in spatial ability(ies). As possible components they cited such skills as identifying which direction a sound came from, realizing that an object remains the same size even though it looks smaller when far away, recognizing an object by touch even when it has been turned around, and choosing which shapes can be turned around and fitted together to form a specified figure. They went on to consider the following skills as possible further components: the abilities to say how one part of a gear system moves when you turn another part in the system; to say how many sides of a pile of blocks someone can see from a perspective different from your own; to perform the Block Design subtest on the most frequently used intelligence test for children; and accurately to do mazes, certain kinds of puzzles, and two tests called the Embedded Figures Test (in which one tries to identify certain figures within a design made up of other figures) and the Rod-and-Frame Test (RFT; in which one is asked to sit in a chair whose position can be changed, to look at a tilted picture frame, and then to adjust a rod so that it remains straight). Various other authors (such as Cooper & Shepard, 1973; Harshman, Hampson, & Berenbaum, 1983) have attempted to analyze the term *spatial abilities* into its components and have come up with other, somewhat different lists.

In the midst of this perplexing array of approaches to analyzing the general term *spatial abilities* into its possible components, the term *spatial visualization* is often proposed as one component. It seems to mean trying to picture something, but as MacFarlane-Smith (1964) has written, no practical definition of spatial *visualization* has been generally accepted. That was true when MacFarlane-Smith's article was published in 1964, and it remains true today.

When different researchers use different definitions and different tests but all claim to be measuring something called *spatial abilities,* and when each research team then gets a different result, it looks as though there are totally conflicting findings about the same ability or set of abilities. In fact, however, each team is studying a different ability or set of abilities.

Since we know that that happens, we ought to realize that there is no way we can conclude on the basis of these kinds of research that "males are better than females on spatial tasks." However, that is exactly what has been claimed. Why is that? Historically, here is what has tended to happen. Research Team A decided to examine sex differences in spatial ability, chose Test 1 as what seemed to them to be a reasonable test of spatial ability, and found that boys did better. Research Team B decided to examine spatial ability using Test 2 and found no sex difference. Research Team C chose Test 3 of "spatial ability" and found that males did slightly better at some ages but not at all ages. Reviewers have tended to summarize results A, B, and C as "Males are better spatially than females," and experimental results that are not consistent with that summary (like those of Research Team B) have been easily overlooked.

Because of these problems with definitions, some tests have been called spatial abilities tests even though they appear to have little or nothing to do with what most people would consider to be *spatial abilities* (Caplan, MacPherson, & Tobin, 1985). For some tests of *spatial abilities* the score is determined more by nonspatial than by spatial factors. For example, the Rod and Frame Test is sometimes claimed to be a test of spatial abilities, but it has been suggested (Sherman, 1978) that scores on this test may be importantly affected by assertiveness and by fear or uneasiness (the latter because the test has often been given to females by male experimenters in a darkened room). Furthermore, although it is claimed that males perform better than females on the RFT, when a human figure is used in this test instead of an abstract rod, no sex difference in performance appears, and when the participants are told the RFT is a test of empathy or ability to understand how other people feel, females perform better (Naditch, 1976).

Keeping in mind the fact that there is so much confusion about whether there is such a thing as *spatial ability*—and if there is, whether it is one ability or many—let us now look at the magnitude of the alleged sex differences in this research.

❖ EXTENT OF SEX DIFFERENCES

To hear many educators and laypeople talk, one would think that, whatever spatial abilities might be, males are substantially superior in that arena. However, most studies actually yield no difference at all. Furthermore, when sex differences have appeared, they have had the following characteristics:

1. They are small (Kimball, 1981).
2. The overlap in males' and females' scores is great (Kimball, 1981).
3. The differences are unreliable, so that when a test is given several times, sometimes a sex difference appears and sometimes it does not (Annett, 1980; Foley & Cohen, 1984).
4. When a sex difference appears, it is almost always around or after adolescence (Caplan et al., 1985).
5. Given the huge sex differences in socialization related to "spatial abilities"—such as children learning early that girls cannot read maps, do math and science, estimate distances, and so on—it is amazing that "spatial abilities" tests do not yield enormous, reliable, lifelong differences (Caplan et al., 1985).

Several of these points deserve further discussion. For instance, in regard to #1, the small size of sex differences when differences do appear, Hyde (1981) has pointed out that these differences are so small that they account for only 1–5 percent of the variance in scores (Jacklin, 1979). Without going into detailed statistical theory, what this means is that, if you wanted to predict whether someone would score high or low on a spatial abilities test, knowing their sex would give you only between 1 and 5 percent of the information you would need in order to make a correct prediction. In other words, sex may sometimes play a role in one's score, but that role is tiny.

In regard to #3, the unreliability of sex differences, it is important to be aware that the differences appear more common than they are because data that show no-difference findings are often ignored. In fact, many no-difference studies are never accepted for publication, simply *because* they show no differences. But even when a no-difference result is published, it may be ignored. As Wittig (1979) has pointed out, it is commonly noted that Baughman and Dahlstrom (1968) found a male superiority in the spatial relations subtest of Thurstone's Primary Mental Abilities Test (Thurstone, 1963). However, that finding applied to a group of 437 White children. The finding that there were no sex differences for the 642 Black children in the same study is not commonly cited. This is a striking example of the way that sexism and racism can combine to result in the suppression—and eventual invisibility—of important information. A further example of the invisibility of crucial information is Parlee and Rajagopal's (1974) observation that one study often cited as evidence of a male spatial superiority (Dawson, 1967a, 1967b) in fact employed only male participants!

In regard to point #4, since those sex differences that do appear are not usually found until adolescence, it is extremely doubtful that innate factors are involved.

❖ BUILDING THEORIES ON SHIFTING SANDS

Based on the kind of flimsy evidence just described, theories have been created to "explain" the "sex difference" in "spatial abilities." It is irresponsible and damaging to create a theory that has little or no foundation in data, because once a theory exists, both scholars and laypeople tend to assume that the theory is based on solid evidence. Then, rarely does anyone check to see whether there are data that justify the theory. Interestingly, the theorists have tended to claim that these so-called sex differences are innate, inevitable, unchangeable.

Let us look at two theories and their major flaws, which show that, even if they had been based on solid evidence, the theories themselves don't hold water.

1. Genetic Theory. According to the major genetic theory about sex differences in spatial abilities, these abilities are sex-linked through the X chromosome. The theory is somewhat complicated, but if it were true, then mothers' and sons' scores on a spatial abilities test (if a legitimate spatial test could be found) should be more similar to each other than mothers' and daughters' scores, and there should be no correlation at all between fathers' and sons' scores.

In fact, however, the data do *not* fit this pattern, but an important reviewer of the literature (Harris, 1978) first presents the data and then claims that this genetic theory *is* true. He says, "the model of spatial ability as a recessive sex-linked trait can stand" (p. 449). If one didn't look at the actual numbers in the data but only at theorists' and reviewers' claims, one would assume that the genetic theory was well supported.

2. Brain Lateralization Theories. Many different theories about alleged sex differences in spatial abilities have been proposed. These have been based on the assumption that sex differences in the brain give rise to differences in performance on "spatial" tests. A revealing fact is that these theories conflict with each other, so they cannot all be true. Some theorists (e.g., Levy, 1970) claim that spatial abilities are superior when based in one hemisphere of the brain, whereas others (e.g., Buffery & Gray, 1972) claim that they are better when based in both hemispheres. To complicate matters even more, McGuinness (1980) has suggested that "*both* hemispheres appear to operate in all tasks" (p. 244).

Finally, as a number of scholars have pointed out (see Caplan et al., 1985, for a review), in the area of research on brain functioning, many studies are poorly done and/or are carried out on atypical populations (e.g., people whose brains are diseased or damaged), and a few studies can be found to support just about any theory.

The great variety of activities known as spatial abilities have in common the fact that they are activities at which males are considered to excel. If tests of spatial ability included the ability to judge how much flour is in a cup, or how to use a dress pattern in sewing, the results might look quite different. Spatial abilities may very well be based on stereotypically male abilities. The wonder is that in spite of this, the few differences found between females' and males' spatial abilities are small and unreliable. If scientists were to stop using the concept of spatial abilities, it would create an environment in which they could take a fresh look at individual people's varied abilities to learn and at blocks to their learning—and that would be helpful to both sexes.

References

Annett, M. (1980). Sex differences in laterality—meaningfulness versus reliability. *The Behavioral and Brain Sciences, 3,* 227–228.

Baughman, E. E., & Dahlstrom, W. G. (1968). *Negro and white children: A psychological study in the rural south.* New York: Academic Press.

Buffery, A., & Gray, J. (1972). Sex differences in the development of spatial and linguistic skills. In C. Ounsted & D. Taylor (Eds.), *Gender differences, their ontogeny and significance* (pp. 123–158). Edinburgh: Churchill Livingstone.

Caplan, P. J., MacPherson, G. M., & Tobin, P. G. (1985). Do sex-related differences in spatial abilities exist? A multilevel critique with new data. *American Psychologist, 40*(7), 786–799.

Cooper, L. A., & Shepard, R. N. (1973). Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual information processing* (pp. 75–76). New York: Academic Press.

Dawson, J. L. M. (1967a). Cultural and physiological influences upon spatial-perceptual processes in West Africa (Part I). *International Journal of Psychology, 2,* 115–128.

Dawson, J. L. M. (1967b). Cultural and physiological influences upon spatial-perceptual processes in West Africa (Part II). *International Journal of Psychology, 2,* 171–185.

Foley, J. E., & Cohen, A. J. (1984). *Gender differences in cognitive mapping.* Paper presented at the Toronto Area Women's Research Colloquium, Toronto, Canada.

French, J. W. (1951). The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monographs* (No. 5). Chicago: University of Chicago Press.

Guilford, J. P. (1947). *Printed classification tests* (Army Air Forces Aviation Psychology Research Program Report No. 5). Washington, DC: U.S. Government Printing Office.

Harris, L. J. (1978). Sex differences in spatial ability: Possible environmental, genetic, and neurological factors. In M. Kinsbourne (Ed.), *Asymmetrical function of the brain* (pp. 405–522). New York: Cambridge University Press.

Harshman, R. A., Hampson, E., & Berenbaum, S. A. (1983). Individual differences in cognitive abilities and brain organization, Part I: Sex and handedness differences in ability. *Canadian Journal of Psychology, 37,* 144–192.

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using Ω^2 and *d. American Psychologist, 36,* 892–901.

Jacklin, C. N. (1979). Epilogue. In M. A. Wittig & A. C. Petersen (Eds.), *Sex-related differences in cognitive functioning* (pp. 357–371). New York: Academic Press.

Kimball, M. M. (1981). Women and science: A critique of biological theories. *International Journal of Women's Studies, 4,* 318–338.

Levy, J. (1970). Information processing and higher psychological functions in the disconnected hemispheres of human commissurotomy patients (Doctoral dissertation, Pasadena California Institute of Technology, 1970). *Dissertation Abstracts International, 31,* 1542B.

Lips, H., Myers, A., & Colwill, N. (1978). Sex differences in ability: Do men and women have different strengths and weaknesses? In H. Lips & N. Colwill (Eds.), *Psychology of sex differences* (pp. 145–173). Englewood Cliffs, NJ: Prentice-Hall.

Maccoby, E. E., & Jacklin, C. N. (1974). *Psychology of sex differences.* Stanford: Stanford University Press.

MacFarlane-Smith, I. (1964). *Spatial ability: Its educational and social significance.* San Diego: Knapp.

McGuinness, D. (1980). Strategies, demands, and lateralized sex differences. *Behavioral and Brain Sciences, 3,* 244.

Naditch, S. F. (1976). *Sex differences in field dependence: The role of social influence.* Paper presented at the meeting of the American Psychological Association, Washington, DC.

Parlee, M. B., & Rajagopal, J. (1974). Sex differences on the embedded-figures test: A cross-cultural comparison of college students in India and in the United States. *Perceptual Motor Skills, 39,* 1311–1314.

Sherman, J. A. (1978). *Sex-related cognitive differences: An essay on theory and evidence.* Springfield, IL: Charles C Thomas.

Thurstone, L. L. (1950). *Some primary abilities in visual thinking* (Report No. 59). Chicago: University of Chicago Psychometric Laboratory.

Thurstone, L. L. (1963). *Primary mental abilities test.* Chicago: Science Research Associates.

Wittig, M. A. (1979). Genetic influences on sex-related differences in intellectual performance: Theoretical and methodological issues. In M. A. Wittig & A. C. Petersen (Eds.), *Sex-related differences in cognitive functioning* (pp. 21–65). New York: Academic Press.

CHAPTER FIVE

ARE BOYS BETTER THAN GIRLS AT MATH?

Mathematics is a science, not an art. In math, you are either *right* or *wrong.* In subjects such as English and even in chemistry or physics, an answer can be partly right, but in beginning math and much of high school math, an answer is either right or wrong; you can check your answer in math to see whether it is correct. This makes math a unique arena in which to study sex differences. For instance, girls are more likely than boys to be taught to seek adults' approval (see Caplan, 1973, for a review), and one way to win approval is to give a teacher the right answer to a question. Girls may be more upset than boys by being asked questions that have single, correct answers, rather than, say, being asked to describe a character from a novel, since for the latter you can be partly right, and it is hard to be totally wrong. If a group of students take a math test, and the boys score higher than the girls, you might conclude that the boys were simply better at math. However, it is possible that the girls were worried about giving the wrong answer, and so they were less likely to try some items. This would mean that the boys were not necessarily better at that particular math skill even if their scores were higher.

Much of what we found in Chapter 4 with regard to spatial abilities also applies to the study of mathematical abilities: Most of the research yields either small sex differences or no sex differences, and the differences that do appear don't tend to emerge until around adolescence, after girls and boys have been exposed to many years of socialization about which sex is supposed to be good or bad at which school subjects. Furthermore, the results of some studies (e.g., Decore, 1984) show that females' grades in mathematics courses are actually *higher* than those of males. Decore, for instance, found that between 1970 and 1982 at the University of Alberta, females' grades in both elementary and intermediate calculus were nearly always higher than those of males. And Hanna (1988) reports, based on her study of math ability in 18 different countries, that "gender-related differences in achievement vary considerably both within and among countries" (p. 14).

Since it is so generally believed that not only are boys and men superior at math but also that this alleged difference is *innate*, it is important to look at a number of socialization factors—at least within North America—that would tend to enhance male students' math performance and interfere with that of female students. Eccles and Jacobs (1987) found that their research indicated that junior and senior high school students' grades and the likelihood that they would even enroll in math courses are more influenced by social and attitudinal factors than by their actual ability to do mathematics. One of the best-known of these factors is math anxiety, which has been shown to be higher in girls than in boys (Eccles & Jacobs, 1987). It is interesting that students' math anxiety does not seem to be based very much on how well they have done in math in the past. In other words, girls have greater anxieties about their math ability, but this is not because their ability is inferior. Math anxiety *is* related to the grades students get in math courses and their plans to take more math courses in the future. Other social and attitudinal factors include parents' belief that math is harder for girls than for boys (Eccles & Jacobs, 1987); the tendency for fathers to help their children with math homework more than mothers do (Meece, Parsons, Kaczala, Goff, & Futterman, 1982); the greater preponderance of men than women as teachers of advanced math courses (Meece et al., 1982); the stereotyping of math textbook materials and math games as more appropriate for boys than for girls; teachers' higher expectations for boys than for girls in terms of math performance (Meece et al., 1982); and teachers' tendency to spend more time instructing and interacting with boys than with girls in math courses (Meece et al., 1982). The production in 1992 of talking Barbie dolls that complained about math being difficult was a recent, glaring example of the persistence of such stereotyping.

Probably the most influential work on sex differences in math has been done by Benbow and Stanley (1980, 1983), and it illustrates some of the most common kinds of methodological problems in research on sex differences in math, so we shall examine one of their most important studies in some detail. One major reason for its importance is that it was widely publicized in the media. Some of the typical headlines (cited by Eccles & Jacobs, 1987) were:

> Are Boys Better at Math?
> (*New York Times*, December 7, 1980)
>
> Do Males Have a Math Gene?
> (*Newsweek*, December 15, 1980)
>
> The Gender Factor in Math. A New Study Says Males May Be Naturally Abler Than Females
> (*Time*, December 15, 1980)

Clearly, the media took the Benbow and Stanley research very seriously. Those headlines strongly suggest that boys are actually better at math than girls. Furthermore, people have believed for a long time that males are superior to females in mathematical ability, so the interpretation of Benbow and Stanley's results agreed with the accepted outlook. If we look deeper, however, we find that the reality of their work doesn't match the headlines. Moreover, the flaws in Benbow and Stanley's research are typical of the majority of sex-difference studies of math.

Benbow and Stanley (1980, 1983) studied the scores that Grade 7 and Grade 8 gifted students achieved on the mathematics portion of the Scholastic Aptitude Test (SAT-M), a test widely used to help determine who is admitted to college. Students scoring in at least the top 2–5 percent of any standardized math achievement test were invited to take the SAT. They came mainly from the Middle Atlantic area, although later on, some students from elsewhere in the United States were included. Nearly 50,000 students accepted the invitation. Benbow and Stanley found that, overall, the boys achieved higher scores than the girls. They therefore concluded that boys have greater "math reasoning ability."

There are several major errors in that research. Some of these are embedded in the design of the study, while others are just wrong interpretations of the results. Each will be discussed in detail, but briefly, they are:

1. **Measuring Math Reasoning Ability.** The researchers used the SAT-M as an indicator of "math reasoning ability," even though this test is not an accurate indicator of math aptitude.

2. **Obtaining a Uniform Sample.** We can only reasonably conclude that a difference exists between the sexes if the groups are identical in all other ways. The researchers stated that the boys and girls in the study had equal amounts of formal education. This may be true, but even when the subjects spend the same number of hours in the classroom, many factors are involved in learning other than simply the quantity of time spent in the classroom. For instance, having heard that "girls aren't very good at math" or "girls who are good at math aren't very feminine" could have important effects on students of both sexes. Futhermore, in keeping with traditional female socialization, more intelligent girls than boys may have had too little self-confidence to accept the invitation to take the SAT and participate in the research.

3. **The Power of Suggestion.** The researchers did not consider the fact that the students' expectations about their own performance, as well as other people's expectations of them, might have affected their performance on the SAT-M.

4. **From Specific to General.** The researchers wrote as though the results from their study would apply to all females and males, everywhere. This is not a valid assumption.

5. **The Unjustified Claim That Males' Superiority Is Innate.** Nowhere do they cite conclusive evidence of this.

❖ MEASURING "MATH REASONING ABILITY"

Benbow and Stanley wanted to compare the "mathematical reasoning ability" of boys and girls. They neglected, however, to define the phrase, so we don't know what they intended to study. How can we know whether or not the SAT-M accurately measures what they call "mathematical reasoning ability"? Since the acronym, SAT, stands for Scholastic Aptitude Test, perhaps we were meant to assume that Benbow and Stanley felt they were measuring aptitude. However, scores on the SAT-M are influenced by many factors other than pure aptitude. If you ask a person to solve a problem, but the problem requires the person to know the quadratic formula, it is impossible to solve the question without that knowledge. Then, does the question measure aptitude or achievement? It means nothing about the person's ability to solve the problem if they don't know the formula.

One possible use of a test of mathematical ability might be to predict how well a student would do in college math courses; however, the test Benbow and Stanley studied is not very useful in making such predictions. Slack and Porter (1980) found in their research that high school math grades, and even math achievement test scores, were more reliable than SAT-M scores for predicting a student's math achievement in college. Furthermore, Fox and Cohn studied students in junior high school and found that the girls' SAT-M scores were unreliable predictors of their achievement later in school.

Fox, Tobin, and Brody (see Kolata, 1980) interviewed many of the girls in the Benbow and Stanley study and found that a great many of them did not want to participate in accelerated math classes. They were afraid that their peers would think of them as "different," and they thought that the accelerated classes were dull and that the boys in the classes were "little creeps." Although the researchers did not interview the boys in the same way, their results suggest that girls believe it is not socially acceptable or desirable for them to do well in math; this belief could certainly hinder their math performance, especially since girls are more likely than boys to seek social acceptance (Caplan, 1973).

❖ THE PROBLEM OF OBTAINING A UNIFORM SAMPLE

Benbow and Stanley called what they found a "sex difference" in math. If the girls and boys they tested were identical in every way except for their sex, it would be fairly safe to assume that something about maleness and femaleness led to the difference in math scores. But it is *not* legitimate to make that assumption if the girls and boys differ in some way besides their sex. One of Benbow and Stanley's major assumptions was that all of their students had the same amount of formal education. Their reason for believing this was that every student was in Grade 7 in a U.S. school. However, the issue of the quantity of educational experience is much more complex. Grade 7 girls and boys

don't necessarily receive the same amount of formal education, even when they are in the same classes (Eccles & Jacobs, 1987). Leinhardt, Seewald, and Engel (1979) found that by Grade 7, math teachers have spent up to 36 more hours instructing boys than girls. With less exposure to math, it is easy to see how girls might have less desire to study math.

Aside from simply the number of hours spent with each student, there is also the factor of how the teacher treats the child. Stanley himself (reported by Holden, 1987) noticed that females are more oriented toward social interaction and aesthetics, while males tend to be more oriented toward the quantitative, abstract, "power and control" (p. 661). If he is right about this, then maybe math teachers tend to teach in a style that appeals more to males than to females. This could easily explain the discrepancy in females' and males' scores. In fact, Patricia Casserly (reported by Kolata, 1980) studied 20 schools in which the members of both sexes scored equally on the math achievement tests and found they had several common features; for instance, the math teachers of these students communicated a love of and enthusiasm for math. This may have enhanced their interpersonal connections with students—a factor that the girls might have found particularly encouraging (Gilligan, 1982).

People don't learn from formal education alone; therefore girls' and boys' SAT-M scores might have been affected differentially by experiences outside the classroom. Someone who plays math-related games will be expected to learn a lot more about math than someone who doesn't. It has been shown that boys are more likely to be involved in mathematical games and math-related activities, and to read more math-related books than are girls (Astin, 1974; Fox & Cohn, 1980; Leinhardt, Seewald, & Engel, 1979).

In several ways other than biological sex, then, the girls and boys in the Benbow and Stanley study may well have differed from each other, and those other ways could certainly have led to a sex difference in the sexes' average math scores. This raises the possibility that boys may not be innately better than girls at math—as the headlines have implied—but simply have more experience with math.

❖ THE POWER OF SUGGESTION

In our society, boys are expected to be better at math than girls. This expectation could heavily influence the results of the SAT-M. If you lead a person to expect something, they tend to interpret whatever happens as bearing out what they were led to expect. For instance, teachers who are told that a child is not very bright tend to notice things that confirm that expectation (Rosenthal & Jacobson, 1968).

In the same way, children who are told "you cannot do math" tend to come to believe that. (Of course, usually these messages are more subtle, but just as powerful.) Then, whenever they are confronted with a mathematical problem, they are likely to conclude automatically that they cannot solve the problem and, therefore, they are less motivated to try or persist. After studying the various influences on students' math grades, Eccles and Jacobs (1987) concluded that the strongest influence on a student's math ability was how their mother thought they would do.

Miele (1958) found that as children get older, the difference in scores for boys and girls on the Wechsler Intelligence Scale for Children (WISC) becomes greater and greater (with boys doing better). Also, in her review of the literature on sex-difference research using the WISC, Attard (1986) concluded, "It appears that, on the whole, no gender differences are evident on the arithmetic sub-test [of the WISC and WAIS] up to approximately age 16" (p. 14). These results seem to reinforce the hypothesis that sex differences in math result at least partly from other people's influence; as children approach age 16, they accumulate more and more years of exposure to the idea that boys are better than girls at math. Since people in our society tend to *believe* that boys are better than girls at math, it ends up appearing to be true on test results.

❖ FROM SPECIFIC TO GENERAL

Finally, Benbow and Stanley did not take into account what are called *sampling errors.* Since they studied a sample of about 50,000 Grade 7 students in the United States, even if their results were valid, they should only be assumed to apply to Grade 7 students in the United States. It is quite possible that even the results that Benbow and Stanley produced would be different for Grade 1 students or college students or 40-year-olds or 80-year-olds. It is also quite possible that the results would be different in another country, since Hanna (1988) and Schildkamp-Küngider (1982) tested tens of thousands of students from all over the world and found that in some areas, girls got the higher scores, while in other places, boys did.

Even if Benbow and Stanley's results had been otherwise accurate, they would still only apply to the specific people they tested. And it must be remembered that the people they studied were a *highly* selected group: They represent not students in general but only students who had scored in the top 2–5 percent on one of several math tests *and* who accepted the invitation to participate in the study. If, for instance, there is no sex difference in math ability for 95–98 percent of students, then Benbow and Stanley's claim to have found sex differences in math is a serious distortion.

If there *were* compelling evidence that most or all boys have better math abilities than most or all girls, then it might have been reasonable to consider adjusting our education system and our way of thinking accordingly (for example, having teachers spend more time teaching girls). However, if there *is* no sex difference, or only a small, unreliable, and late-developing one or a difference for only a small fraction of people, then it is dangerous to talk about "a sex difference"; to talk in that way leads people to believe that girls just

can't do math. Indeed, that claim *has* been made, and many females who might have done quite well in a math career—in teaching of math; in accounting; in statistics, surveys, and poll-taking; in computer-related fields have therefore not pursued one.

As a result, for sex-difference studies about math, as for all sex-difference research, it is essential to be aware of possible sources of error, since these distort our view of the truth about females and males.

❖ THE UNJUSTIFIED CLAIM THAT MALES' SUPERIORITY IS INNATE

When Benbow and Stanley's work was reported at the 1986 American Association for the Advancement of Science meeting, Benbow claimed that hormonal differences lead to males' greater proficiency in math. Naturally, the media eagerly reported this story. What they did *not* mention was that hormonal levels of the students in their study were never measured, thus making Benbow's claim entirely unjustified (Caplan, 1987). This is a particularly important issue, since when there is, or seems to be, a biologically based and innate difference such as a hormonal one, people are likely to assume that little or nothing can be done to reduce the supposed inferiority of one sex.

A careful exploration of the nature of Benbow's claim about a hormonal basis for males' alleged superiority in math is useful because it reflects so many of the errors that can occur when theory and research are not thought about carefully. Although the following discussion is very detailed and complicated, it is worth going through, because it illustrates how an unfounded theory can be used as a basis for assumptions, predictions, and hypotheses. Then, data are gathered on the basis of those assumptions and hypotheses, and authors tend to try to interpret them in a way that supports the shaky theory. It is very important to remember that, *once data have been gathered to test a theory, the theory often comes to seem to be true, even if the data do not support the theory particularly well.*

In a paper titled "Extreme mathematical talent: A hormonally induced ability?" (1987), Camilla Benbow and Robert Benbow presented their "yes" answer to the question in the title of their paper. As you will see, their argument is very roundabout and complicated, and there are problems every step of the way. It is based on the unsupported theory of two other researchers.

Benbow and Benbow (1987) noted that Geschwind and Behan (1982) had reported that left-handed people are more likely than right-handers to suffer from immune disorders, learning disabilities, and migraines, and that Geschwind and Behan "hypothesized" that this was due to high levels of the "male" hormone, testosterone. Benbow and Benbow (1987) failed to mention that this claim by Geschwind and Behan has been vigorously criticized and has not actually been proven true by solid research. The Benbows suggested that testosterone slows down the development of the left hemisphere of the brain, so that the right hemisphere compensates by growing stronger, and that this improves mathematical abilities. Therefore, they concluded, excellent math students should have more immune problems and be more likely to be left-handed than would the general population. They decided to test those speculations on a group of students, but for no apparent reason they left out migraines. So, they were investigating the *implications* of only *part* of a theory, and the theory itself was not well supported in the first place. Furthermore, as we shall see, the students they studied were a highly unusual group.

The Geschwind and Behan theory was based partly on the idea that the immune disorders result from the effects of testosterone on the immune system's thymus gland. However, Benbow and Benbow (1987) cite no evidence for this idea. The theory was also based partly on the idea that testosterone slows down the development of the brain's left hemisphere, so that the right hemisphere compensates by growing stronger. However, Benbow and Benbow (1987) cite no evidence for this idea either.

What about the Benbows' speculation that mathematical tasks are better carried out by the right than the left hemisphere of the brain? They cite no evidence for this claim but simply assert that this is "considered to be" the case. In fact, however, many aspects of math involve the ability to think analytically, which in most people is located in the left hemisphere, and other aspects of math involve the ability to deal with spatial relationships of the kind that in most people are housed in the right hemisphere. Thus, it is just too simplistic to say that math tasks should be better performed when the right hemisphere is doing them.

Even *if* all of the claims and speculations by both pairs of authors had been proven to be true, then one would expect *most* males to be far better at math and much more likely to develop immune disorders and migraines and to be left-handed than most females. But that is certainly not the case.

Based on all of these unproven propositions, the Benbows *speculated* that, since males tend to have more testosterone than females, *some* males would be both left-handed and skilled at math, due to the hormone's effect on the right hemisphere, and that same testosterone would affect their thymus gland, so that they would have immune disorders. Next, they predicted that, in their very special, unusual group—the most extremely skilled math students (they had scored 700 or more on the college entrance SAT-M *before age 13* and were 1 in 10,000 students!)—left-handedness and high mathematical *reasoning* ability would be correlated with each other. They did not explain why they chose to look at math reasoning rather than at any other math abilities, and nowhere in their paper did they present any evidence that math reasoning is more likely to be affected by testosterone or by hand preference or by the brain hemispheres than any other math ability.

The Benbows claimed that they had supported their hypothesis when they found that, in their highly unusual group, there were about twice as many left-handers and twice the frequency of allergies (an immune disorder) as in the general population. However, as every introductory psychology student learns, left-handers are more common in a wide range of unusual populations, including prisoners and students at Harvard University. Therefore, it is extremely difficult to know how to interpret yet another example of a high incidence of left-handers in an extreme population. And as for the unusually high frequency of allergies in the top math students, so little is understood about allergies themselves and about possible effects of hormones on allergies that it is premature to make too much of that finding. Furthermore, a carefully done study would include investigation of the whole spectrum of immune disorders, when we have been given no reason to believe that only one specific type would be affected by testosterone levels.

Then we might wonder how the Benbows might explain why only *some* members of their highly selected group fit the pattern that their questionable theory predicted. In fact, they again plunge into speculation, suggesting that *those* students *might have been* exposed before birth to higher than normal testosterone levels. Do they present any data to support this claim? Their argument here becomes quite strange and again convoluted. They had no proof that these students had had such prenatal exposure, but they hauled out the finding that they were more likely than most students to have been born during months that have more than 12 hours of daylight per day. Then, they stated, "Daylight affects pineal gland secretion, altering the level of melatonin, which in turn has an inhibitory effect on reproductive hormones" (pp. 150–151). In other words, daylight affects Factor A, which affects Factor B, and *that* can reduce the hormone level. Aside from the sheer length of this unproven explanation about how top math students *might* have been exposed to high levels of testosterone, their reasoning is simply wrong. If more daylight is supposed to *reduce* the reproductive hormones, then these students should have had *less* testosterone, not more, than most students. And according to the Benbows' own (unsupported) chain of reasoning, lower testosterone levels should lead to *poorer* mathematical abilities.

If you look back at the headlines cited earlier in this chapter, it may seem surprising that the public could be presented such claims when they are based on highly speculative theories, research on extreme groups of people, and just plain poor reasoning. However, such presentations are not uncommon. When some journalists hear what seems to be a "hot" story, they do not stop to learn whether or not there is any scientific basis for it.

We hope that, through the scrutiny of the range and variety of problems in the Benbow and Stanley study, you have some sense of the complexity and difficulty of the field of sex differences in mathematics. This sense should be helpful to you as you read other research or plan your own.

References

Astin, H. (1974). Sex differences in mathematical and scientific precocity. In J. Stanley, D. Keating, & L. Fox (Eds.), *Mathematical talent: Discovery, descriptions, and development* (pp. 70–86). Baltimore, MD: Johns Hopkins University Press.

Attard, M. (1986). *Gender differences on the Arithmetic and Coding subtests of the Wechsler Intelligence Scale for Children.* Unpublished master's thesis, University of Toronto.

Benbow, C. P., & Benbow, R. M. (1987). Extreme mathematical talent: A hormonally induced ability? In David Ottoson (Ed.), *Duality and unity of the brain* (pp. 147–157). London: Macmillan.

Benbow, C., & Stanley, J. (1980, December 12). Sex differences in mathematical ability: Fact or artifact? *Science, 210,* 1262–1264.

Benbow, C., & Stanley, J. (1983, December 2). Sex differences in mathematical reasoning: More facts. *Science, 222,* 1029–1031.

Caplan, P. J. (1987, February). *Do sex differences in spatial abilities exist?* Presented at the Symposium on Bias in Sex-Differences Research, American Association for the Advancement of Science, Annual Meeting, Chicago.

Caplan, P. J. (1973). *Sex differences in determinants of antisocial behavior.* Unpublished doctoral dissertation, Duke University.

Decore, A. M. (1984). Vive la différence: A comparison of male-female academic performance. *Canadian Journal of Higher Education, 14*(3), 34–58.

Eccles, J. S., & Jacobs, J. E. (1987). Social forces shape math attitudes and performance. In M. R. Walsh (Ed.), *The psychology of women: Ongoing debates* (pp. 341–354). New Haven: Yale University Press.

Fox, L., & Cohn, S. (1980). Sex differences in the development of precocious mathematical talent. In L. Fox, L. A. Brody, & D. Tobin (Eds.), *Women and the mathematical mystique* (pp. 94–112). Baltimore: Johns Hopkins University Press.

Geschwind, N., & Behan, P. (1982). Left-handedness: Association with immune disease, migraine, and developmental learning disorder. *Proceedings of the National Academy of Sciences, 79,* 5097–5100.

Gilligan, C. (1982). *In a different voice: Psychological theory and women's development.* Cambridge, MA: Harvard University Press.

Hanna, G. (1988). Mathematics achievement of boys and girls: An international perspective. In D. Ellis (Ed.), *Math 4 girls* (pp. 14–21). Ontario Educational Research Council.

Holden, C. (1987, May 8). Female math anxiety on the wane. *Science, 236,* 660–661.

Kolata, G. B. (1980, December 12). Math and sex: Are girls born with less ability? *Science, 210,* 1234–1235.

Leinhardt, G.; Seewald, A.; & Engel, M. (1979). Learning what's taught: Sex differences in instruction. *Journal of Educational Psychology, 71*(3), 432–439.

Meece, J. L.; Parsons, J. E.; Kaczala, C. M.; Goff, S. B.; & Futterman, R. (1982). Sex differences in math achievement: Toward a model of academic choice. *Psychological Bulletin, 91,* 324–48.

Miele, J. A. (1958). Sex differences in intelligence: The relationship of sex to intelligence as measured by the Wechsler Adult Intelligence Scale and the Wechsler Intelligence Scale for Children. *Dissertation Abstracts International, 18,* 2213.

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom: Teacher expectation and pupils' intellectual development.* New York: Holt, Rinehart & Winston.

Schildkamp-Küngider E. (Ed.) (1982). *International review on gender and mathematics.* ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, OH.

Slack, W., & Porter, D. (1980). Training, validity, and the issue of aptitude: A reply to Jackson. *Harvard Educational Review, 50*(3), 392–401.