DANIEL REISBERG

THE SCIENCE OF PERCEPTION AND MEMORY

A Pragmatic Guide for the Justice System

OXFORD

The Science of Perception and Memory

The Science of Perception and Memory

A Pragmatic Guide for the Justice System

Daniel Reisberg

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford. It furthers the University's objective of excellence in research, scholarship, and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press in the UK and certain other countries.

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016

Library of Congress Cataloging-in-Publication Data
Reisberg, Daniel.
The science of perception and memory : a pragmatic guide for the justice system / Daniel Reisberg.
pages cm
Includes bibliographical references and index.
ISBN 978–0–19–982696–4 (hardback)
1. Memory. 2. Perception. 3. Cognitive psychology. I. Title.
BF371.R374 2015
153.1′2024364—dc23
2014012603

9 8 7 6 5 4 3 2 1
Printed in the United States of America
on acid-free paper

For Friderike, who intelligently, insightfully, lovingly,
Improves everything that I do

CONTENTS

PREFACE

A robbery victim tries to remember who was present at the crime scene. A witness to a mugging stood 30 yards from the attack and later identifies the perpetrator from a lineup. A teenager approaches a school counselor reporting sex abuse that happened years earlier. A medical patient recalls the doctor saying that the pain in her side wasn't worrisome and, now that the tumor is much larger, she's suing. An investigation of insider trading hinges on someone's recall of exactly what was said at a business meeting.

In these and countless other examples, our ability to perceive and remember is crucial for the justice system—whether our focus is on criminal cases or civil, or, for that matter, administrative or family law. The problem, though, is that the evidence provided by perception and memory is fallible. Even under conditions we might call "optimal," our eyes can deceive us and our memories are sometimes incomplete—or (worse) mistaken. And, obviously, the justice system is often confronted with less-than-optimal conditions—people who had only a brief view of an event or a view with poor lighting; victims who are reporting precise details of conversations that took place long ago. How should we think about these points? How often do our eyes or memories deceive us? Is there some way to avoid these errors by gathering our memory-based evidence in just the right way? And, if we can't control the evidence gathering, can we at least specify the circumstances in which perceptual or memory errors are more likely and when less? Then, once the evidence is in hand, is there some means of evaluating a person's recollection to decide whether his or her report is accurate or not?

The questions in play here have been discussed within the legal community for years. Comments about memory or perceptual errors are commonplace in court rulings and law school curricula. But where can we find answers to these questions? It turns out that perception and memory are central concerns for research psychologists—scientists who investigate the functioning of the mind. I am one of these researchers, and, as a group, we're rather different from what most people think of as "psychologists." We have laboratories, not clinics; we are trained in statistics, not modes of therapy; our expertise is in research design, not clinical diagnosis. We do experiments, collect data, and draw conclusions in a fashion tightly controlled by the standards and practices of the scientific community. And, in many studies, the research has been carefully tuned to provide information directly useful for the justice system.

Results from these studies are slowly percolating into the legal world, and research psychologists are already advising the courts, consulting with attorneys, and training investigators. In many states, interview procedures with children have been designed in ways explicitly guided by what we know about children's memories and how children respond to various forms of questioning. In some jurisdictions, police use the so-called cognitive interview when questioning witnesses—an interview procedure based on what we know about adult memory. In still other cases, the courts rely on psychological research in assessing the reliability of an identification; more and more states are relying on this research in constructing their ID procedures.

We need to acknowledge, though, that the enthusiasm for this information flow hasn't been universal. Some people in the justice system argue that the scientific data aren't reliable or that the data are just not relevant to the issues at play in the courtroom. Other people are convinced there's an inherent bias in the research—and so (they claim) the science might be useful to a criminal defender but to no one else. Still others believe the science contributes little beyond common sense. Indeed, not so many years ago, I sat in a courtroom, listening as the attorneys debated the admissibility of my testimony. After hearing their arguments, the judge ruled: "I've been evaluating eyewitness evidence for my whole career without Dr. Reisberg's help, and I see no reason why I need him now." And, with those words pronounced, I was dismissed.

In this book, I'll explain why these concerns are all misguided. But I'll also do more than that: I want to lay out, in accessible form, what it is we know so that this information is readily available to judges, attorneys, investigators, and anyone else associated with the justice system. The topics I'll be discussing (and the data underlying my claims) are already covered in various publications, but often these publications are written by people without scientific training. As a result, the presentation is sometimes less authoritative, and perhaps even less accurate, than it needs to be, and, in any case, does a poor job of conveying the "scientific muscle" that sustains the authors' claims. In other cases, by contrast, the relevant information is conveyed in volumes written by the researchers themselves. These texts are excellent and certainly authoritative. But they're often dense in ways that make them difficult for nonspecialists to understand. My aim in this book, therefore, is to forge a middle path: I'll try to convey this material in a way that is sophisticated, current, and fully grounded in the best available science, but also in a way that makes the material accessible and, above all, useful.

The book has a modular structure, with each chapter tackling a well-defined topic. Each chapter is then divided into blocks, with entry tags providing information about what the block seeks to accomplish—what issue it seeks to address. My hope is that some diligent readers will read through the presentation in sequence. I have tried to set up the book, though, so that readers pressed for time, seeking help with a specific issue, can use the modular structure to find just the bits they want, treating the book as a succession of targeted handbook entries. In all cases, though, I hope the book's structure allows swift and efficient access to the information the reader needs.

In Chapter 1, Foundational Issues, I discuss the scientific roots of the work described in the book. As I've noted, claims about perception and memory come from the field of research psychology and also from neuroscience and cognitive science, and so we need to ask: What are these fields? Are their results solidly enough established and an accurate-enough reflection of real-world complexities so that they can be taken seriously by law enforcement or admitted into court?

The next two chapters provide foundation of a different sort by laying out, in general terms, what we know about perception (Chapter 2) and memory (Chapter 3), with a focus on how these crucial human capacities operate in settings of interest to the justice system. Chapters 4 and 5 then tackle a specific type of perception and memory: the perception and recognition of faces. Chapter 4 is concerned with a broad set of questions about face memory, providing the basis, in Chapter 5, for examining what we know about the identification procedures that are used (or that might be used) in criminal cases.

Chapter 6 then turns to a different type of memory and considers how well people remember voices and conversations they have heard. Memory for voices is, of course, relevant to a different ("earwitness") type of identification. Memory for conversations brings us to questions relevant to both civil litigations ("When did you hear the stock was going to be issued?") and criminal cases (for example, cases alleging conspiracy or cases trying to discern someone's intent). In this chapter, I also consider how well investigators remember their own wording when they questioned a witness—and thus whether we can count on their reports of whether they put ideas into someone's mind or led the witness.

In Chapter 7, I turn to the topic of lies. I discuss the distinction between lies and false memories and consider the factors that are relevant to deciding whether an individual is telling one or experiencing the other. The chapter also discusses the detection of lies, both by interviewers and through technical means like polygraphy or functional magnetic resonance imaging (fMRI). How accurately can we distinguish honest individuals from deceivers or true statements from lies?

Chapter 8 tackles reports of a different type: potentially false reports of one's own actions—and thus false confessions. What leads someone to make a false confession? How often do false confessions emerge? Is there a way to distinguish a false confession from an accurate one?

In Chapter 9, I change perspective. The book so far has covered perceptual and memory errors pertinent to witnesses, victims, and investigators. But there is another regard in which perception and memory matter for the legal system: jurors' memory. After all, jurors often need to recall complex arguments, evidence, or instructions, and there's typically a delay between when jurors first gain this information (perhaps an idea posed during voir dire or evidence presented early in a trial) and when they use this information in deliberation. Moreover, jurors often need to remember the source from which they gained the information: Was a specific point, raised in deliberation, actually covered in evidence? Or was the point promised during opening statements but never delivered during trial? Or perhaps was the point part of the pretrial publicity that should not be considered during deliberation at all? I'll consider the evidence pertinent to all of these issues—including the evidence that bears on whether some of the proposed aids for jurors (such as note-taking) actually have an effect.

Finally, Chapters 10 and 11 turn to children's memories. The logic here parallels that in Chapters 4 and 5, with the first chapter in each pair providing crucial background, and the latter chapter in the pair turning specifically to questions about investigative procedures. In these two final chapters, we are seeking to ask whether (or, more sensibly, when) we can count on a child's report for a crime the child has witnessed (or, worse, a crime for which the child has been the victim). My focus in this section is on how children remember and report on abuse. I consider both the memories themselves and also so-called symptom evidence (i.e., psychological patterns that might result from the abuse and which therefore might be used as evidence for the abuse). I also discuss recovered memories—memories of abuse that seem to emerge only after a delay of many years.

Before diving into all of this material, though, I want to head off a common misunderstanding: I've spoken with judges, attorneys, and police officers about these issues, and many of them are concerned that all this research tends to challenge witnesses' perception and recall. They fear, therefore, that attention to this research will encourage skepticism about witness evidence and therefore undermine law enforcement and prosecution efforts.

I argue throughout the book, though, that this concern is without basis. To be sure, there are cases in which one factor or another signals a risk of perceptual or memory error, and, in those cases, we do need to be cautious—and perhaps even skeptical—in interpreting a witness's report. In other cases, though, we might look for factors that might have pulled a witness off track and find none; in such cases, the science might be used to bolster the witness's credibility. This point rests on the fact that human perception is generally accurate (and, in fact, rather precise). Likewise, human memory is, in general, relatively complete, long-lasting, and correct. Of course, errors do arise, but the errors are the exception, not the rule, and so we usually can count on witness evidence. Moreover, the errors in perception or memory do not occur randomly; instead, known factors encourage the errors, and so, by examining a specific case, we can often determine whether errors are likely in that setting or not. Thus, part of my motivation in writing this book is to highlight the reliability of witness evidence—so that police and the courts can step away from baseless challenges to a witness's report and can, when the circumstances are right, put more weight on witness reports.

Perhaps more importantly, it's my hope that we can use the catalog of possible problems in witness evidence as a basis for improving witness evidence. Many scientists have been pursuing this effort in the collection of ID evidence—using what we know about face memory in order to obtain better identifications. A similar effort has played a central role in cases involving children: By specifying what can go wrong in child interviewing, we can assemble a list of "do's" and "don'ts" for the professionals who conduct these interviews. In this fashion, the science has already produced better procedures—codified in clear professional guidelines—that help investigators to collect evidence from children that's stronger, clearer, and less vulnerable to challenge.

On these grounds, the science I'll be describing in this book should be useful to everyone in the justice system. Sometimes the science will undermine a witness's evidence. Sometimes the science will bolster the evidence. Best of all, often the science can help us to collect better evidence. But, with all of these possibilities in view, I think it's a mistake to argue that the science favors "one side" or the other.

Of course, I understand that our courts operate on an adversarial system, and so I fear some readers will be skeptical about the optimistic views I've just expressed. Even so, it's my hope that everyone in the justice system wants to maximize the quality of the information being considered and to seek ways of distinguishing good information from bad. The science can help us move toward these crucial goals, and this book is my small contribution to this endeavor.

ACKNOWLEDGMENTS

Many people helped me with this endeavor, and I'm grateful to them all. Don Read gave me the benefit of his enormous expertise in psychology; Alafair Burke provided important insights from her perspective as an attorney and former prosecutor. Alex Houston provided fabulous help in the effort of double-checking and documenting the research data throughout the book. And then, at OUP, Catharine Carlin was wonderfully encouraging in the early steps of this project, and Joan Bossert, Louis Gulino, and Ryan Cury have provided great support in the process since then.

I'm also profoundly grateful to the many police officers, attorneys, and judges I've worked with over the years. Many of them have read sections of the manuscript; some read all of it, and all have provided comments that substantially improved the book. More importantly, all of these individuals have collectively taught me an enormous amount about the justice system, and along the way helped me to see both the strengths and the limitations of my science. And, odd though it might be, I'm also indebted to the crime victims and the defendants whose cases I've worked on; they've taught me just how important it is to get these issues right.

I'm also grateful to Jacob (a lawyer-to-be) who, with loving intent, kept pushing me to write this book, because he understood how much I needed to write it. Solomon, in turn, never lets me forget what careful scientific analysis looks like, and that helps to keep me on track. And, of course, Friderike has read every page of this volume, and her challenges and critiques were always rooted in a mix of thoughtfulness that was deeply serious and support that was wonderfully loving.

And, finally, I write in loving memory of my father. My debt to both of my parents is immeasurable. Among many other gifts, they taught me to choose my battles, but also taught me that, when the issue really matters, you have to fight for what you believe in.

CHAPTER 1

Foundational Issues

Starting in Chapter 2, we'll tackle substantive issues about how perception and memory operate and how investigations might be structured to maximize the quality of witness evidence. In this chapter, though, we'll set the broad context. We'll discuss the scientific status of psychological research—and, with that, the admissibility of this research under a *Daubert* or *Frye* standard. We'll also consider one of the common, but misguided, objections to this science—namely, that the research simply confirms common sense.

A. IS THIS SCIENCE?

The Relation Between Clinical and Research Psychology

For many people, the word "psychology" conjures up images of a one-on-one conversation between a therapist, listening attentively, and a client (they're usually not called "patients") describing life's challenges—perhaps difficulties in the client's marriage, or concerns about depression, or problems in dealing with stress. This sort of image, however, is misleading in many ways and fails utterly to convey the other face of psychology—psychology as a research discipline.

The field of psychology is, in truth, better thought of as *two* fields, one that we can broadly call "clinical psychology" and one that we can broadly call "research psychology." There are, to be sure, psychologists who live in both of these worlds (and happily so, because the two domains can help each other in important ways and, arguably, *need* each other). For the most part, though, the clinicians and researchers are distinct groups, with different types of training, different skills, and different knowledge, and these groups provide very different types of information for the justice system.

Clinical psychologists are generally trained as health care professionals, with an emphasis, of course, on mental health. These psychologists are skilled in the diagnosis, treatment, and, in some cases, prevention of mental problems. They help people who suffer from depression or anxiety; they help people to overcome compulsions or even addictions. They help people cope with stress. They can provide individualized assessment, telling you what someone's intelligence level is, or whether the person suffers from schizophrenia, or whether the person is competent to stand trial, or perhaps eligible for an insanity defense. In important ways, the work these psychologists do resembles (or overlaps with) the work done by psychiatrists, social workers, and counselors (although each of these professions is different from the others in some ways, including the educational background that can lead to each).

Research psychologists, in contrast, are trained not in diagnosis or treatment, but in the standard methods of science. They have laboratories, not clinics. They usually work in an academic setting, not in a hospital or mental health facility. Unlike clinicians (who are often licensed by a state agency), researchers have no licenses because there is simply no such thing as an officially approved "research license." Instead, researchers establish their credibility through educational and academic credentials (usually a Ph.D. in a field that involves scientific training and a post at a reputable institution). Far more important, quality control in the world of research is carried out not by some governing agency but in the fashion that's standard in the scientific world: by means of peer review. (I'll say more about peer review later in the chapter).

Many people outside of the academic world don't realize that research psychology even exists—and hence the misleading image that launched this chapter. Part of the reason is numerical: Clinicians outnumber psychology researchers by a substantial margin. And part of the reason is functional: Ordinary citizens often interact with clinicians (as therapists, counselors, and so on) but may never encounter a researcher. It's important to emphasize, therefore, that research psychology is a long-standing and well-established enterprise. Research psychology has been part of the university curriculum for a century. The *Journal of Experimental Psychology* began publication in 1916; in that year, *Psychological Review* was already publishing its 23rd volume. In more recent years, research psychologists have won a half-dozen Nobel prizes and are well represented in the U.S. National Academy of Sciences, alongside biologists, chemists, and physicists.

Psychology researchers pursue a range of detailed questions about why exactly people see what they see, remember what they remember, think what they think, and feel what they feel. The research is often multidisciplinary: To understand vision, for example, one needs to understand some fundamental points about the physics of light and the biology of the eyeball and the brain. To understand thinking, we often need to draw on insights from logicians studying deduction or economists studying choice; we sometimes test our proposals by building computer models designed to simulate a process or by comparing our claims to events in the nervous system. Indeed, this merging of methodological perspectives is the source of the term *cognitive science,* a label given to the multidisciplinary effort of asking questions about the mind. The constituents of cognitive science include psychology, anthropology, philosophy, computer science, linguistics, and *neuroscience* (which is itself an amalgam of the many disciplines aimed at an understanding of the nervous system).

Within research psychology, which specialties are likely to be useful for the justice system? *Cognitive psychologists* study perception, attention, and memory, and therefore can provide the justice system with information about these topics (topics crucial, of course, for evaluating witness evidence). *Social psychologists* study how people perceive each other, think about each other, and are influenced by each other. They can therefore provide information about witness evidence (because, after all, witnesses are usually reporting on an encounter with another human being). *Developmental psychologists* are concerned with how people change and grow across the lifespan and so can provide important information about memory in children or in the elderly, and also information about some of the important complexities of communicating with young children.

Defining Science

Should research psychology be taken seriously as a scientific discipline? The answer lies, of course, in the definition of science. Specifically, any scientific endeavor starts by formulating *empirical claims*—claims about the factual world that, ultimately, could be shown to be true or false. Science then tests those claims in a rigorous manner, using established tools and procedures. Those tests must satisfy strong requirements at every stage: in the formulation of the testable claim itself; in the design of the test procedure; and in the collection, analysis, and interpretation of data. Then, once the interpretation is in place, the entire sequence must be subjected to the critical scrutiny of other scholars, checking for flaws of any sort. It is only when the interpretation has survived this scrutiny that it can be taken seriously. And, of course, if the pattern of the data conflicts with a claim, then the claim needs to be adjusted or set aside. It is not acceptable, in the scientific world, to assert an empirical claim as true if there are no data to support the claim, and it is surely not acceptable to continue asserting a claim if there are data that contradict the claim. (For more on psychology's status as a science, see Malpass, Ross, Meissner, & Marcon, 2009.)

The Importance of Replication

It is also important that, in science, very few claims rest on just a single finding. There are several reasons for this, but one key reason is that, with all of our safeguards (checks on the reliability and validity of our measures, appropriate sampling procedures, and so on), there is always a chance that a result may emerge just as a fluke. (For more on what a "fluke" might be in this setting, see the later section on "Known or Potential Error Rate.") Science needs to guard against this possibility, and one way it does so is via *replication*—repeating an experiment (perhaps with an identical procedure, perhaps with some variation) in a new setting with a new experimenter and a new group of participants. If the result emerges again, this strengthens the claim that the earlier finding wasn't a fluke and also indicates that the result is robust enough to show up despite changes in the exact circumstances.

All of these requirements apply to any science, and they certainly apply to the study of psychology. Of course, the courts have developed their own criteria for evaluating the admissibility of scientific evidence or scientific testimony, and I will turn to those criteria in a moment. But, by any conventional definition, the inquiries conducted in psychology laboratories satisfy the requirements of "science," and, on that basis, the claims offered by psychologists, coming out of their research, deserve the same status as those offered by biology, chemistry, or any other scientific enterprise.

B. IS THE SCIENCE APPLICABLE TO "REAL-WORLD" CASES?

There is, however, an important concern attached to the courts' use of psychological research. To explain this point, let's start with a problem that scientists themselves need to address. Imagine that a psychologist wanted to learn which witnesses are more likely to have accurate memories and which are less so. As a first step toward answering this question, the psychologist might wonder, for example: Are women better at remembering events than men are? To find out, he recruits a dozen men and a dozen women, asks each to recount what he or she did the afternoon before, and then assesses these recollections for their level of detail.

This inquiry would, in fact, tell us little. As one problem, notice that this procedure provides no way of asking whether the recollections are accurate. Perhaps the women report more details, but get these details wrong. Or, as a different concern, notice that we may not have a "level playing field" here. Perhaps the women's lives actually contain more interesting, more distinctive events, and their lives are therefore easier to remember. If so, then their memories may be no better than men's, but they will nonetheless remember more because their lives are overall more memorable.

Let's emphasize, though, the role of the word "perhaps" in the previous paragraph. We said that *perhaps* the women are getting the details wrong and that *perhaps* their lives contain more interesting events. We certainly don't know that these suggestions are correct, and, indeed, we have no basis for suspecting that they're correct. But these suggestions *might be* correct, and that is enough to make the outcome of this (fictitious) experiment ambiguous. If we conducted the experiment, we would have no way to conclude anything with certainty. Or, to put the point broadly, if a result is ambiguous, it can sustain no conclusions: If there's more than one plausible way to think about the data, we don't know which way is correct, and so we can draw no claims from the data.

The logic behind these points is straightforward: Scientists would rather say nothing than say something wrong, and they would rather say nothing than say something that is unwarranted by the data. As a result, the operative rule is indeed: When there is ambiguity in the data, say nothing.

The Advantage of Controlled Settings

How do scientists—and psychologists in particular—address the concern just raised? In most cases, our path forward involves bringing the question into the laboratory: To continue with this (admittedly artificial) example, we could arrange for men and women both to witness the same event, perhaps a movie that we show them. We then let some predetermined amount of time go by before we question them about what they saw. And, because we showed them the movie, we know exactly what was in the movie, and so we can easily check the accuracy of their recall. Moreover, since our two groups saw the *same* movie, we now have a "clean comparison"—a comparison between men and women in which all other factors are held constant, thus isolating the factor (the gender difference) that we wanted to learn about.

The advantages of this approach (i.e., the benefits of scientific control) can be illustrated with numerous examples. As one case, Yuille and Cutshall (1986) interviewed witnesses to an actual crime and found that witnesses who experienced the most stress (by their own report) were also the witnesses most accurate in describing the event. This observation might imply that stress promotes memory, but there is an interpretive problem here: The witnesses who were most stressed by this crime were also the witnesses positioned closer to the crime (presumably this is why they felt the most stress), and therefore they actually had a better view than less-stressed witnesses. We need to ask, therefore, whether their better memory was a product of the stress or of their viewing opportunity. There is no way to determine this in the Yuille and Cutshall data, and that is the point here: In the laboratory, we can control circumstances and isolate the variable of interest (e.g., stress) with other factors held constant; studies outside of the laboratory, like the Yuille and Cutshall study, do not allow us this advantage. (And, by the way, data indicate that stress typically undermines memory rather than improving it; see Chapter 2.)

Addressing the Problem of Artificiality

The controls provided by the laboratory are valuable, but introduce their own problem. Returning to our example, if we know how men and women remember *in the laboratory,* does this tell us how they remember in other circumstances? (Perhaps women are more impressed by the laboratory setting and thus try harder in our tasks.) Likewise, if, in our study, we find a difference in how men and women remember the movies we show them, can we draw conclusions about how they remember other materials? (Perhaps men are less interested in cinema and so pay less attention to the film, in comparison to women.)

The point of all this should by now be obvious: Psychologists often move into controlled environments to gain scientific power, but, in doing so, they pay a price of artificiality. How can we address this concern? How can we make certain that lessons we draw from the research world are truly applicable to the "real world"?

The answer has several elements. First, we do what we can to minimize the artificiality in our studies—and so we take steps to narrow the gap between the circumstances in our experiments and the circumstances we ultimately want to understand. Moreover, we often do "hybrid" studies that are conducted in a controlled manner but carried out in a realistic (nonlaboratory) setting. Thus, studies of how stress affects memory are sometimes conducted in doctors' offices, probing the memories of patients undergoing actually stressful medical procedures. Likewise, studies of how inattention can influence perception (and therefore memory) sometimes probe how well or how poorly people remember objects in their day-to-day environment.

Second, we can use the methods of science to study the artificiality itself. As an example of this kind of testing, Haber and Haber (2001) wondered whether it matters that research participants in the laboratory know that they are in an experiment. Do they perhaps take the procedure less seriously because they know it is "only" an experiment? Or perhaps are they more attentive because they know their memories will be tested? Haber and Haber checked on these possibilities by means of a complex design that compared identification accuracy (choosing the perpetrator from a lineup) in two groups of studies: some in which people knew they were in a research experiment and some in which they did not know this. The data show no difference, suggesting that this (possible) worry is less pressing than it might initially seem.

Similarly, many studies have presented research participants with a video of a crime or a staged crime, with the participants merely bystanders to the event and not actually involved in the event. Does this contrast—bystanders versus victims—matter? If so, then this, too, would be a challenge to the realism of our studies and the value of our data. But, again, this is an issue that can be explored through research: Hosch and Cooper (1982; see also Hosch, Leippe, Marchioni, & Cooper, 1984, or, as a related study, see Ihlebaek, Løve, Eilertsen and Magnussen, 2003) compared several conditions: In one, participants were bystanders to an event; in another, the participants were fully involved in an event. As we will see in Chapter 6, this contrast does matter for some issues (thus, you are better able to remember a conversation if you were involved in the conversation rather than merely hearing it). However, the Hosch and Cooper data suggest that this concern can, in the case of eyewitness reports, be set aside.

Third, and perhaps most importantly, it is often possible to "cross-check" the data from controlled studies against data collected from real-world settings. Thus, for example, many laboratory studies show that people are less accurate when making *cross-race identifications* (a Caucasian recognizing an African American or vice versa) than when making *same-race identifications* (e.g., a Caucasian recognizing a Caucasian; see Chapter 4). Is this finding replicable in actual crime investigations? We can find out by (among other options) examining real police records and asking how often actual crime victims can identify their assailants in same-race crimes and how often in cross-race crimes. These field data are themselves open to more than one interpretation, but we can surely ask if the field data are as we'd expect them to be, based on the data from controlled studies. In fact, they generally are (e.g., Platz & Hosch, 1988), and it is this convergence between controlled studies and field studies that reassures us that our data are indeed providing a realistic portrait of how perception and memory function.

Likewise, we know that, in the laboratory, eyewitnesses often make errors, selecting from a simulated lineup someone who is not the person they saw in a simulated crime. Does this pattern reflect the reality of actual crimes? As we will see in Chapter 4, it does, and studies of actual police files suggest that real witnesses to real crimes choose someone innocent from a lineup at least 20% of the time.

This cross-checking needs to be done again and again, issue by issue, because some findings from the laboratory will turn out to be replicable in field settings whereas others will turn out not to be. Of course, for some findings, replicability is inevitable. (For example, the structure of the eyeball will be the same in the laboratory or outside of it, and so data patterns caused by this structure will inevitably hold up in the field data, just as they appear in the lab.) Far more commonly, though, we need to probe the replicability, treating any concerns about artificiality in the same way we would treat any empirical issue: as a matter to be investigated through appropriate study, using the standard methods of science.

C. IS PSYCHOLOGICAL SCIENCE ADMISSIBLE UNDER THE *DAUBERT* STANDARD?

One of the ways psychological science can be used by the justice system is via courtroom testimony delivered by a suitably chosen expert. Presumably, the expert can, in many circumstances, provide a framework that will inform the finders of fact as they seek to interpret the evidence. But this raises a question: Is this sort of testimony reliable and helpful? Or, more bluntly, is it admissible?

In U.S. federal court, the standard for admissibility is laid out in the so-called *Daubert* trilogy, a succession of three U.S. Supreme Court cases: *Daubert v. Merrell Dow Pharmaceuticals; General Electric Co. v. Joiner;* and *Kumho Tire Co. v. Carmichael.* Many states have adopted their own versions of the *Daubert* rulings; a few states have retained the older *Frye* standard (see the later section on "General Acceptance" and also Section D, targeted on the *Frye* standard).

The *Daubert* ruling has also influenced Canadian courts. The admissibility of experts in Canada is broadly governed by the standards laid out in *R. v. Mohan*, a ruling that emphasized that expert testimony must be relevant to the case and also necessary in assisting the trier of fact. Like the *Daubert* ruling in the United States, *Mohan* puts the trial judge in the position of "gatekeeper," deciding if the science is admissible or not. And, just as in the United States, Canadian courts are clear that the expert evidence should be excluded if its probative value is outweighed by its likely prejudicial impact. Some years after *Mohan*, though, the Canadian Supreme Court clarified these rules in *R. v. J. (J. L.).* Here, the Court ruled that *novel* science could be admissible, and it endorsed criteria echoing those in the U.S. *Daubert* ruling (but also noted that *Daubert* is rooted in the U.S. rules of evidence, which differ somewhat from those in place in Canada).

What is the *Daubert* standard? The court ruling emphasizes a number of criteria that define the scientific method and which make scientific knowledge "reliable." (Note, though, that the Supreme Court made it clear that these criteria are nonexclusive [and so other criteria may be considered] and nondispositive [and thus do not, on their own, settle the question].) The specific criteria listed in the *Daubert* ruling are (1) falsifiability, (2) peer review, (3) known or potential error rate, (4) existence of standards and controls, and (5) general acceptance. Subsequent rulings (and some state rulings) add a further important criterion: (6) sufficient facts or data.

Falsifiability?

The scientific community and the courts agree that scientific claims must be testable, and, for the past half-century, scientists have usually understood "testable" to mean "falsifiable." In other words, scientists are generally skeptical about the idea that a claim can be *verified*—that is, proven true. There are several reasons for this position, including the idea that the facts will always fit with a *vague* theory (because a vague theory can accommodate any fact pattern). If, therefore, we were to count a theory as "proven" when the theory fits with many facts, then we would routinely gather "proof" for vague theories—and that's an unacceptable prospect for science. (Think about how easily we could prove claims like, "Sometime soon, something interesting will happen," or "either tomorrow will be fun for you, or it won't." For theories like these, proof is cheaply acquired and of no value whatsoever.)

To avoid this prospect, scientists insist that their theories be specific enough so that it is clear which facts would fit with the theory and which facts would not. That latter clause rules out vague theories and also theories that build in too much flexibility (so that they have an escape clause no matter how the data turn out). And, of course, that clause also means that the theory is falsifiable: That is, there are facts that might emerge that would be incompatible with the theory and that would force us to abandon the theory.

To put this point slightly differently, a theory is testable only if it makes "risky predictions"—predictions that might turn out not to be true (e.g., Popper, 1963). If a theory has made these predictions and has *not* been falsified, this bolsters our confidence that the theory is correct. At that point, we would say that the theory is "confirmed." A theory that makes "risk-free predictions," in contrast, is going to fit with the facts no matter what those facts are, and hence the idea of "confirmation" loses its meaning.

These principles are certainly respected by psychology research. Our experiments do involve risky predictions. Our professional journals insist that claims be falsifiable. Indeed, the undergraduate curriculum in psychology often highlights examples of theories that are not falsifiable (Sigmund Freud's theorizing is a favorite example) and showcases for students the reasons why this status is unacceptable for a scientific claim. In these ways, falsifiability is woven into the fabric of our field—and so psychology satisfies this criterion.

Subjected to Peer Review and Publication?

Scientists use the term *research literature* to refer to scholarly articles published on a particular topic—usually articles appearing in specialized professional journals. Scientists take a result seriously only if it appears in this literature, and the reason is straightforward: Papers are published in the literature only when they have survived a rigorous evaluation process.

To understand this process, consider what happens when a researcher hopes to get a paper published. The researcher sends the manuscript to the editor of one of the field's journals. The editor, in turn, sends the manuscript to carefully selected reviewers—usually experts on the topic covered in the paper. In most cases, the manuscript is stripped of identifying information (the name of the authors, the authors' institution) before it is sent to the reviewers, to avoid bias. (And, I should mention, this latter step seems to be successful: For the vast majority of the papers I've been asked to review, I have no idea who the paper's authors are.)

The reviewers read the paper and provide a written report for the editor—evaluating the paper's strengths and weaknesses and cataloguing any concerns, whether large or small. Based on these evaluations, the editor reaches a decision about the paper's fate: accepted, rejected, accepted pending revisions, rejected with suggestions for improvements, and so on.

In this process, the reviewers are understood to be the manuscript authors' "peers" and hence the name *peer review*. The notion of "peers" reflects the idea that all scientists have equal status in this process (and so a paper will not be published merely because the author is famous or based at a prestigious institution). The reviewers are also the authors' peers in another sense: The reviewers understand the paper's methods and analysis, just as the author does; they understand the theoretical framework, just as the author does. Hence, there is no risk that the reviewers will simply defer to the author's superior knowledge. Instead, the reviewers will provide expert-level, well-informed evaluation—just what we want.

The peer-review process is arguably the only path forward for science: Who else, besides a researcher's peers, has the training and expertise with which to evaluate a perhaps technical, perhaps sophisticated paper? And, with this process, journals are quite selective; most scientific journals reject the majority of submitted manuscripts. (Rejection rates of more than 70% or even 80% are common at the field's top journals.) It seems reasonable, therefore, that the "consumers" of research (other investigators, for example) take a paper seriously when it has (and *only* when it has) survived this gauntlet.

Of course, sometimes good papers get rejected; lower quality papers sometimes sneak through. However, if a weak paper does end up in print, readers of the journal are likely to respond in some way—writing rejoinders or seeking to publish their own studies in rebuttal to the original paper. Here, the field is relying on a broader form of peer review (with the journal's readers now serving as the peers) and, for this purpose, the passage of time is important. Scientific debates can take years to unfold, and this slow pace creates an opportunity for rebuttal to an already-published paper, if appropriate, and allows other researchers a chance to publish their own work, perhaps challenging the initial finding. Indeed, we mentioned earlier that it is rare for researchers to draw a claim from a single study; it is likewise rare that researchers draw strong conclusions from a single paper. Instead, the passage of time will lead to an accumulation of papers on a topic, all of which have been through the peer-review process. If there is a consistent pattern across all of these papers, and if the claims have survived all the challenges, then, finally, the claims are accepted as firmly established.

Does the field of research psychology rely on this process? The answer is a clear "yes." Does the field, as a result, rely on a succession of specialized and sophisticated journals? The answer is again a clear "yes," and a listing of the journals can be found on the websites of the American Psychological Association, the Association for Psychological Science, and the Psychonomic Society (and others). Thus, on these criteria as well, research psychology passes the *Daubert*-defined test.

Known or Potential Error Rate?

Defining "Error Rate" in Terms of Probabilities

The *Daubert* ruling (and the various state rulings derived from it) asks whether a method has a "known or potential error rate." This notion is easily understood when evaluating a specific procedure or a specific measure: What is the error rate associated with, say, the polygraph as a lie detection device? We can, with appropriate studies, determine what percentage of the people telling lies are correctly identified as such via a polygraph examination and what percentage of truth-tellers are falsely accused. We can then use these percentages as our estimates of the *error rate* for this procedure. (We return to this example, and the assessment of the polygraphy, in Chapter 7.)

The idea of "error rate" is less transparent, though, when evaluating a broad claim resting on many different experiments. Consider, for example, a scenario in which an expert testifies that an eyewitness's *degree of certainty* provides little information about whether the eyewitness's recall is correct or not. (We'll return to this claim in Chapters 3 and 5.) In other words, the expert testifies that research has shown only a weak correspondence between someone's confidence in a memory and the accuracy of that memory. What is the error rate associated with this claim?

The answer to this question is somewhat complicated and we'll walk through the details in the next two subsections. For readers who don't want the details, however, here is a brief summary of the argument: Researchers want to avoid making claims that are not justified by evidence. (As we said earlier, scientists generally prefer to say nothing rather than say something wrong.) This sentiment leads researchers to ask, for any observed pattern in their data, "Could the pattern be a mere fluke? The product of some sort of lucky accident?" Statistical procedures allow researchers to ask these questions in a precise fashion and, in particular, allow them to calculate the probability of the results indeed being a mere fluke.

From this base, we do know the "rate of error" associated with each of our claims. More precisely, we know the probability that a claim is not warranted by the data. And, perhaps more importantly, researchers draw conclusions from the data only if this probability is very low—less than one-in-twenty when considering a single experimental result, and much, much lower when considering a fabric of results drawn from multiple experiments.

Type 1 and Type 2 Error

There are two types of errors one can make in research: concluding from the data that a hypothesis is correct when in reality it is not, and concluding from the data that the hypothesis is wrong when in truth it is correct. In most cases, researchers try to be as cautious as they can and so they generally regard the former error (technically called a *type 1 error* or, in more common terms, a false alarm) as more troubling than the latter (a *type 2 error* or false negative, or—less formally—a "miss"). Again: Scientists would rather say nothing than say something wrong. Hence, it's better to overreject claims than to overaccept; that way, you are unlikely to say something that is wrong. As a result, many statistical procedures are designed explicitly to avoid type 1 errors—the false alarms.

Therefore, when a judge in a *Daubert* hearing asks about the error rate associated with a research claim, the answer can be framed in terms of the field's statistical methods. Specifically, we can provide the judge with information about the risk of a type 1 error—the risk that the expert is offering a claim not justified by the evidence. But how is the risk calculated, and how great is the risk?

In most cases, statistical procedures begin with the *null hypothesis*—that is, the hypothesis that, in reality, there is no systematic difference between the conditions or groups being compared. The statistics are then designed to ask: If the null hypothesis were true, then how likely is it that the study would have produced the pattern of results actually observed? Or, to put this differently, the null hypothesis is essentially a claim that any differences that are observed must be the result of chance fluctuation in the data. Framed in this way, the statistics are asking: What is the probability of getting the results observed in a study purely by chance? What is the probability that the results are just a fluke?

A simple example may be useful here: Imagine that you are tossing a coin and have just observed that the last eight tosses have all been "heads." You wonder: Could the coin somehow be biased? The null hypothesis here would be that the coin is *not* biased, so that any departures from a 50–50 distribution of "heads" and "tails" is just a matter of random chance. In this case, some simple calculation would tell you that, if the null hypothesis is correct here, the probability of observing your result (eight heads in a row) is roughly 0.4%. In other words, if you did this "experiment" 1,000 times with a fair coin, you might get this result, just by chance, in 4 of those 1,000 tries. Hence, this result *could* happen by chance, but it is the sort of observation that would happen only rarely by chance. You are therefore likely to conclude, since you got this result on your very first try, that the succession of eight "heads" is probably not the result of mere chance, and thus you would conclude that your coin is indeed biased.

The logic is the same when scientists evaluate their data. Statistical procedures are used to calculate the exact probability of getting the observed results just by chance, just by a lucky accident. This probability is described as a "p-value," and a p-value must be calculated for every comparison a researcher wishes to make. (Thus, a single experiment will often require many p-values if the researcher wishes first to compare *this* aspect of the data and then *that* aspect, and so on.) If the p-value is high (i.e., a high probability), then there is a substantial likelihood that the data might be the product of mere chance—and, if so, we draw no conclusions from the data. Conversely, if the p-value is low, then it's quite unlikely that the results came about by chance, and so the risk of a false claim—a type 1 error—is correspondingly low.

How are p-values calculated? The mathematical details aren't necessary here, but, in general, three considerations are in play in these calculations. First, how large a difference is there between the conditions (or groups) being compared? (This difference is referred to as the "effect size," and, all things equal, larger effect sizes are less likely to come about by chance and so yield smaller p-values; we'll have more to say about the term "effect size" in a later section.) Second, how consistent are the data? (All things equal, data patterns that are more consistent—less variable—are less likely to come about by chance and so, again, yield smaller p-values.) Third, how many observations did the researcher make? How many points of data are there? (All things equal, data patterns based on more observations are less likely to come about by chance, and so, once again, yield smaller p-values.) Let's note, though, that these three considerations can be traded against each other. (If the effect size is large and the variability is small, you need fewer observations. If the variability is large, you either need a larger effect size or more observations. And so on.) But, in any case, it is a straightforward matter, for a given comparison, to calculate the relevant p-values and thus to determine the danger that the result might be a matter of sheer chance. We take the result seriously only when that probability—the level of danger—is very, very low.

So What is the Error Rate?

Scientists want the risk of a type 1 error to be low—but how low is "low enough" to justify drawing conclusions? By convention, when evaluating a single study, researchers employ the criterion that we alluded to earlier, a criterion defined as "$p < .05$." In other words, researchers will draw conclusions from an observed difference only if the probability of obtaining this result just by chance is less than 5%. Or, turning this around, researchers will draw conclusions only if the odds are 19-to-1 that mere chance wouldn't produce the result.

This 5% rule is, however, merely the cut-off for acceptability, and many results in the literature have much lower p-values associated with them. (Thus, the research literature often contains results that are reliable at a level of $p < .01$ or $p < .001$, and thus results for which the risk of a type 1 error is 1.0% or even 0.1%, respectively.) In addition, we need to be clear that the $p < .05$ rule describes how researchers think about single, isolated results—the product of just one comparison, in one study. This is crucial because, as we've said, scientists are rarely guided by a single result. Instead, scientific claims rest on broad patterns of interlocking results, with different studies replicating an original finding, evaluating concerns about that finding, or testing claims about the mechanism that led to the finding, and so on. As the fabric of evidence grows, the likelihood of the results being a matter of chance necessarily gets lower and lower. On this basis, the actual risk of error for a research claim is much lower than 5%. (We will return to this point when we turn to the statistical procedure known as meta-analysis.)

Where does all of this leave us with regard to the *Daubert* standard? The rate of error for a scientific claim can be construed as the likelihood that the claim might not be justified by the research evidence. In other words, the Court's request for a rate of error can be addressed with the researchers' calculation of the probability of a type 1 error. This error rate is known because it is the result of precise calculations. And the error rate is guaranteed to be low because results will be taken seriously by the scientific community only when the results "pass" this statistical test.

Existence of Standards

Training in Research Methods

Another criterion named in the *Daubert* ruling is the existence of standard research methods, and this criterion is easily met by research psychology. The methods of the field are routinely covered as part of the educational curriculum—both at the undergraduate level and in graduate training. Indeed, this training usually begins with the very first "Intro" course; most textbooks for this course include a chapter on research design and statistics. In Ph.D. programs, courses in methodology are nearly universal.

Measuring Mental Attributes

In most studies, psychologists use measurements that are straightforward and objective. For example, if we are studying the relation between identification accuracy and a witness's degree of certainty, we can assess identification accuracy by simply noting whether the witness chose the suspect from the lineup or did not; we assess degree of certainty by asking the witness to provide us with a number—perhaps a percentage, with 0% indicating a guess and 100% indicating total confidence.

In some cases, though, research demands the construction of novel measures. For example, we will see in Chapter 8 that the risk of false confession is elevated if a suspect happens to have a lower-than-average level of *intelligence* or a higher-than-average level of the personality trait of *compliance*. Of course, research on these points is possible only if we have a means of assessing someone's level of intelligence or compliance, and, in addition, we obviously need some basis for claiming that these assessments are trustworthy.

Psychologists have well-defined procedures for developing and evaluating a measurement scale (whether we are considering a scale that assesses some mental *capacity,* like intelligence, or one that assesses a *trait,* like compliance). We also have clear standards that these scales must meet before they are used in research. Among other points, a measurement scale must be *reliable*—that is, consistent in what it measures. In some cases, the focus is on consistency *within* the measuring tool (are *these* questions really measuring the same capacity as *those* questions?); in other cases, the focus is on consistency *across time* (if I test you today, will I reach the same conclusion as I did based on last week's test?). Perhaps more importantly, a measurement scale must also be shown to be *valid*—that is, measuring what it purports to measure. The test's reliability and validity are central aspects of the test's *psychometric properties,* and these properties must be assessed and documented before any measurement scale can be used.

Enforcing the Standards

We've now said that psychology—like any science—has clear standards for how research should be conducted, but how are those standards enforced? Part of the answer is already in view: via the journals' peer-review process. Reviewers are certainly alert to possible problems in a study's design or statistical analysis, and papers will not make it into print if problems are detected.

However, we should acknowledge that, in a small number of cases, unscrupulous researchers have fabricated data, and so bogus (but seemingly sensible) papers have made it into the journals. In other cases, researchers have become mercenaries, and so will (among other options) testify in court in a fashion that serves a defendant or a plaintiff but that deliberately misrepresents the scientific evidence.

When the research community detects these cases, we deal harshly with them. If a professional journal learns that a paper is based on some sort of false reporting, the paper is removed from the body of scientific evidence: The paper is officially retracted, and subsequent researchers will not rely on that now-retracted paper as a source of evidence. In addition, a variety of sanctions are likely so that the researcher is, in effect, cut off from the field. Thus, if a researcher is found to be dishonest, that researcher will no longer receive funding from granting agencies, will no longer be asked to serve as a reviewer for the journals, will no longer be invited to contribute chapters to edited volumes, will no longer be invited as a speaker, and more. In many cases, the researcher's home institution will conduct its own investigation and offer some form of censure, including the option of dismissal from the institution. Overall, the research community puts a high value on its own credibility and the credibility of its results and reacts strongly to anyone who puts that credibility in danger.

General Acceptance?

In federal court, the *Daubert* criteria replaced the older *Frye* standard, but, even so, the *Daubert* ruling incorporates the notion that was central for *Frye*—the idea that "general acceptance" of a scientific claim is an indicator of reliability. This notion is discussed in full in Section D, but, overall, there is no question that the methods and procedures of research psychology are generally accepted, and, as already discussed, are enshrined in the curriculum and peer-review process. But what about individual results or individual claims? For these, general acceptance cannot be taken for granted and has to be determined on a case-by-case basis. As it turns out, there are straightforward means through which we can assess the degree of acceptance, within the scientific community, for this or that claim, and we will consider these means in a moment.

Sufficient Evidence?

Scientific claims must rest on good methods *and* good evidence drawn from those claims. After all, no matter how careful one's procedures, a claim cannot be taken seriously if the evidence for it is thin. But this point—like the issue of general acceptance (see Section D)—must be evaluated on a case-by-case basis. Overall, though, let's note that there is an enormous amount of research relevant to the topics covered in this book. For example, Cutler and Penrod (1995, p. 68) estimated some years ago that eyewitness psychology had been the subject of more than 2,000 research publications; the number has obviously grown since then. There are similarly large quantities of research on children's memories, the detection of lies, and so on.

In addition, the studies discussed in this book, focused on questions of immediate interest to the justice system, are rooted in a far broader context, one created by many other studies that collectively provide the methodological and substantive foundation for the data described here. Some of these other studies are focused on basic operating principles of perception and memory (and research on these broad topics has been ongoing in hundreds of laboratories for well over a century). Other studies examine the functioning of the nervous system and how the nervous system supports perception and memory. Still other studies probe the statistical and methodological assumptions that underlie our research. Studies of eyewitness memory must be consistent with the evidence provided by all of this research, and if (say) a claim about eyewitness memory conflicted with claims about memory in other domains, this would be a puzzle demanding immediate attention. More strongly, the fabric of evidence in all of these arenas provides a large, mutually supportive set of scientific facts, and it is this full set of facts that makes us confident the science is working properly.

D. IS THE SCIENCE ADMISSIBLE UNDER A *FRYE* STANDARD?

General Acceptance of Method

The notion of "general acceptance" is one of several *Daubert* criteria, but was, of course, central for the older *Frye* criteria still employed by many states (including California, Florida, New York, and Washington). As discussed in Section C, the notion of general acceptance is straightforward if we are asking whether research psychology's *methods* are generally accepted. To be sure, there are occasional debates about methodology—for example, whenever a new method is introduced. And, even for established methods, there are occasional challenges; indeed, any vital science encourages these challenges as a way of always remaining open to criticism and always open to improvement in its methodological claims. But, as discussed earlier, the central methods of psychological research are widely accepted, invariably covered in the curriculum, documented in textbooks at various levels, and so on.

General Acceptance of Empirical Findings

What about psychology's *findings,* or its *empirical claims?* Are these generally accepted? Many psychologists view this as an issue to be settled through research. In other words, we can, in effect, measure general acceptance. Specifically, we can poll the investigators who have expertise on a particular topic or (more broadly) poll investigators who are familiar with the relevant data and in these ways find out what proportion of these investigators agree with a particular proposition and how many of them regard the proposition as well established.

This sort of survey data has been collected for some research topics, including research on the polygraph (Iacano & Lykken, 1997) and also research on eyewitness testimony (e.g., Benton, Ross, Bradshaw, Thomas, & Bradshaw, 2005; Kassin, Ellsworth, & Smith, 1989; Kassin, Tubb, Hosch, & Memon, 2001). These surveys tell us, not surprisingly, that investigators regard some points as well-established (and so there is general acceptance of these points) and regard other points as uncertain and in need of further study. Hence, general acceptance has to be considered on a point-by-point basis.

Let's be careful, though, not to overinterpret these surveys-of-experts. As one concern, these surveys provide a snapshot of a particular historical moment and so reflect researchers' beliefs at the time the survey was conducted. We need to be alert, therefore, to the possibility that subsequent data may strengthen a claim that the researchers regarded as not-yet-established or perhaps undermine a claim that the researchers thought was secure.

In addition, these surveys (like any research study) can be scrutinized—and sometimes criticized—by other researchers and, of course, by attorneys as well. Did the survey include the proper selection of experts? Was there perhaps some bias in how the experts were chosen? Were the questions phrased appropriately? Questions like these can be and sometimes have been raised in challenge to these "general acceptance" surveys. (For an example of these challenges, see Bailey & Mecklenburg, 2009.)

It is therefore prudent to check the surveys against other forms of evidence. As one option, we can assess general acceptance by counting up papers in the literature: How many support a proposition? How many—through argument or data—challenge the proposition? In addition, psychologists have issued a number of white papers on topics relevant to the justice system. In any field, a *white paper* is intended as an authoritative report, usually commissioned by some well-established organization, designed to summarize evidence and educate readers about a particular topic. In psychology, white papers have been commissioned by a number of our professional organizations. For example, the American Psychology-Law Society (AP-LS) assembled a blue-ribbon panel of experts in 1998 to summarize, in a wide ranging and authoritative way, the state of the art on eyewitness identifications (Wells et al., 1998); the AP-LS commissioned another white paper, in 2010, to summarize the available research on police-induced (false) confessions (Kassin et al., 2010b).

As yet another indication of "general acceptance," the Association for Psychological Science (APS) publishes a journal entitled *Psychological Science in the Public Interest*. The APS describes this journal as publishing reviews that are "written by blue ribbon teams of specialists representing a range of viewpoints, and are intended to assess the current state-of-the-science." (This quotation appears alongside the masthead in each issue of the journal.) Articles in this journal have covered the research on lie detection (Vrij, Granhag, & Porter, 2010), eyewitness evidence (Wells, Memon, & Penrod, 2006), adolescent decision making (Reyna & Farley, 2006), and the psychology of confessions (Kassin & Gudjonsson, 2004). These reviews articles are both a valuable source of information for anyone in the justice system and also provide a prima facie basis for claiming that these results are broadly enough accepted to have appeared in these authoritative reviews.

Finally, various professional organizations, including the American Psychological Association and the American Psychiatric Association, have filed amicus briefs that are directly relevant to issues described in this text. These briefs are available online, indexed by issue,[1] and thus provide a valuable resource for litigators. Of course, these briefs advocate for the organizations' own position, but, even so, they are subjected to a rigorous review process relying on the broad expertise of the organizations' membership and are carefully documented. Thus, the existence of these briefs provides yet another indication of the general acceptance of the relevant findings.

1. For example, see http://www.apa.org/about/offices/ogc/amicus/index-issues.aspx

E. DOES THE RESEARCH GO BEYOND COMMON SENSE?

The *Daubert* ruling requires more than scientific reliability; the ruling also makes clear that scientific testimony can be admitted at trial only if the testimony will "assist the trier of fact to understand or determine a fact in issue." There are several ways to unpack this requirement, but one element is straightforward: The testimony won't "assist the trier of fact" if the testimony simply reiterates what the trier of fact knows already.

Likewise, Canada's ruling in *R. v. Mohan* emphasizes that expert testimony will be admissible only if the testimony is "necessary" in assisting the trier of fact. Of course, the testimony will not be at all necessary if it simply reaffirms propositions the trier of fact already understands and accepts. So here, too, we need to ask: Does psychology research provide information that goes beyond common sense?

We all have a lifetime's experience in perceiving, and this experience makes it plain that our perception generally serves us well—and so we manage to catch the softball that's tossed to us or steer our car through traffic without suffering some vehicular catastrophe. But, in addition, we all know that our perception sometimes fools us—and so we know what it's like to see a friend down the block but then discover, when we approach, that the person is someone else altogether.

Likewise, a lifetime of experience has told each of us that our memories are usually accurate—and so we find our car just where we remember parking it, and we have no trouble remembering where we spent our last birthday or the plot of our favorite movie. But we also know that our memory sometimes lets us down: Sometimes, we try to recall a name, or a fact, and we "draw a blank." And, occasionally, our memories are just mistaken: You vividly recall mailing the letter but find the envelope still in your pocket. You recall the dinner party one way, but your spouse recalls it very differently; one of you, it would seem, must have it wrong.

On these grounds, no one needs research to establish both the broad accuracy of perception and memory and the risk of error. If these were the points offered by expert testimony, then the testimony surely would be unnecessary, covering points already within the ken of the ordinary juror. Indeed, one prosecutor in Oregon commented, "I have never called a witness as an expert on memory myself; there are 12 of them sitting in the jury box. Someone testifying that 'memory can be false' seems to be akin to saying that most people have noses on their faces" (Robben, 2012).

Let's be clear, however, that no expert would testify simply to make the point that "memory can be false." That is indeed a matter of common sense. On many other issues, though, common-sense ideas about perception and memory turn out to be *mistaken*—markedly overstating some truths and understating others and, in many cases, offering claims that have no basis in fact whatsoever. The blunt reality, therefore, is that research can offer claims that are a substantial improvement on common sense.

Examples illustrating these points are easy to find. For example, jurors surely know that memory errors occur and may also understand that these errors can sometimes be "planted" through suggestive questioning. (A question like, "You did see the gun, didn't you?" may lead someone later to recall a gun, even if there was none.) However, jurors almost certainly underestimate the risk of this sort of memory contamination. Indeed, in one study, college students were surveyed about their perceptions of various risks (Wilson & Brekke, 1994). These students were largely unconcerned about the risk of someone biasing their memory with leading questions. According to the survey results, they regarded this risk as roughly equivalent to the risk of someday being kidnapped by space aliens. As it turns out, though, the students' assessment was mistaken: Studies make it plain that just a word or two of leading can produce memory errors in roughly a third of the people questioned (see Chapter 4). Surely, the danger of extraterrestrial abductions is much lower than this.

As a different example, many people have the view that significant emotional events are somehow "burned" into the brain and hence never forgotten. Thus people announce with conviction that "I'll never forget the events of 9/11" or " . . . the day I got married" or ". . . what he looked like when he pulled the trigger." There is a kernel of truth in these claims because these significant, emotional, distinctive events do tend to be well-remembered. But, here, too, common sense is in tension with the facts. As we will discuss in Chapter 3, it is easy to document large-scale errors in these singular, significant memories. The accuracy of these memories, in other words, is far from guaranteed, and the "burned into my brain" notion is therefore just wrong.

In fact, this point can be made more broadly—leading us to another contrast between common sense and scientific evidence. People routinely offer descriptions of their own memories: "I'll never forget how she . . ." or "I remember it as though it's yesterday . . ." or, more bluntly, "I'm certain that what happened was . . ." In other words, people offer assessments of their *degree of certainty* or their *confidence* that a particular memory is correct, and they plainly regard these assessments as useful: They are more likely to voice or take action on memories recalled with certainty compared to memories that they're less sure about. Likewise, several studies indicate that jurors are more likely to believe a witness who is confident in his or her recall compared to a witness who hedges or expresses doubts (see Chapter 9). Indeed, judges themselves often highlight the witness's confidence as an indicator of the witness's memory accuracy. For example, in one prominent ruling (*Neil v. Biggers*, 1972), the U.S. Supreme Court explicitly noted that "the factors to be considered in evaluating the likelihood of misidentification include . . . the level of certainty demonstrated by the witness at the confrontation."

But here, too, common sense pulls us off track. We'll review the relevant data in Chapter 3, but for now we note simply that, in many circumstances, there's little correspondence between witnesses' degree of memory certainty and their memory accuracy. Thus, if we follow the dictates of common sense and put more faith in memories expressed with confidence, we will routinely choose to disregard accurate memories and routinely put our trust in misleading information.

Surveys Documenting Common Sense

These broad points—about the limitations of common sense—can be confirmed more formally. In many studies, researchers have surveyed individuals in the jury pool to find out what they believe about perception and memory. Other surveys have targeted other groups, and so we have information available to us about "common-sense beliefs" among police officers, attorneys, and trial judges—with a specific focus on what these individuals know about eyewitness memory, the impact of leading questions, factors making false confessions more likely, and more. In all of these cases, we can then compare these common-sense beliefs to the results of careful scientific research, and thus we can find out in a direct fashion whether the science is (or is not) within the ken of the nonscientists.

It cannot be surprising that the views of all these groups are aligned with the science on some points (e.g., Desmarais & Read, 2011). After all, ordinary experience surely teaches us *something* about how our eyes and ears work and how often our memories are correct. Moreover, the research community and the mass media do what they can to publicize the results of scientific studies, and so nonresearchers are certain to learn more as time goes by. (As one paper put it, any "evaluation of lay knowledge has a limited shelf life"; Desmarais & Read, 2011, p. 208.) But it is also true that jurors, police officers, and professionals in the justice system have numerous beliefs about memory that are patently inconsistent with what we know to be true, based on careful science. In short, then, the claim that psychological research reiterates common sense is an empirical (i.e., testable) claim, and, on a range of topics, this claim is demonstrably false.

We have already mentioned some examples—the common-sense underestimation of the risk of false memory or the common-sense overreliance on degree of certainty as an index of memory accuracy. Here is a different sort of example: Many people seem to believe that the eye can sensibly be compared to a camera and memory to a video recorder. As we will see in Chapters 2 and 3, these comparisons are misleading in crucial ways and lead to a number of errors concerning the accuracy of memory and the ways in which we might evaluate or elicit memories. Misleading or not, though, these comparisons are often endorsed. In one survey of 1,300 potential jurors in Washington, D.C., 48% thought that memories were indeed like video recordings (Schmechel, O'Toole, Easterly, & Loftus, 2006). Similarly, Simons and Chabris (2011) surveyed almost 2,000 individuals and found that 63% believed that memory works like a video recorder. Their survey also indicated that 55% of the respondents believed memory could be enhanced through hypnosis (not true); 48% believed that memory is permanent (again: not true). (Also see Wise, Safer, & Maro, 2011; for some possible complications about surveys like these, however, see Alonzo & Lane, 2009.)

What about other topics? In a 1992 survey, potential jurors were asked whether they agreed with a series of 21 claims about eyewitness reliability; the participants offered claims out of step with the available science for 15 of the 21 topics (Kassin & Barndollar, 1992). Likewise, Benton et al. (2006) surveyed 111 jurors in Tennessee and found gaps between juror knowledge and the science on many issues. Fewer than half of the jurors, for example, had views in line with the science with regard to the importance of pre-lineup instructions, the poor relationship between someone's degree of certainty and the likelihood of memory accuracy, or the often-documented difference between "same-race" and "cross-race" identifications. In fact, jurors disagreed with the experts on 87% of the issues examined in this study. (For other surveys revealing common-sense beliefs about witness memory, see Deffenbacher & Loftus, 1982; Noon & Hollin, 1987; Seltzer, Venuti, & Lopes, 1990; Schmechel et al. 2006.)

Likewise, a survey by Henkel et al. (2008) asked what ordinary citizens believe about confessions and false confessions. In the survey, many citizens (correctly) acknowledged that false confessions can and do arise, but then endorsed a number of false claims about *when* and *how* false confessions emerge. And it's not just ordinary citizens who have these beliefs: As part of the interrogation process, many police officers rely on the so-called Behavioral Analysis Interview (BAI) for detecting lies (see Chapter 8). Many of the central ideas that underlie the BAI are fully in tune with common-sense notions about how people telling lies are likely to behave (e.g., Masip, Herrero, Garrido, & Barba, 2011), but, unfortunately, these notions are out of step with the facts—a finding that reflects poorly both on common sense and on the BAI itself.

The Bias Built into Common Sense

Jurors' beliefs are not just out of step with the science; more strongly, common-sense beliefs reveal a consistent *bias*. Overall, jurors tend to overbelieve eyewitnesses and to overrely on confession evidence. In one study, for example, researchers presented a sample of registered voters with crime scenarios, each of which was based on a previously conducted empirical study; the study participants were asked to estimate the likelihood of a correct identification by the eyewitnesses in each of these scenarios (Brigham & Bothwell, 1983). Overall, 84% of the respondents overestimated the accuracy rates. Moreover, the amount of overestimation was substantial; in one of the scenarios, for example, only 12% of the eyewitnesses had made a correct identification; the voters, in contrast, given the particulars of the scenario, estimated that 71% of the IDs would be correct.

Related data come from studies in which research participants are presented with a case and asked for a "verdict." In one study, participants were provided only circumstantial evidence in a case summary; 49% of the participants voted to convict. When a single, vague eyewitness account was added to the evidence, the conviction rate jumped to 68% (Sigler & Couch, 2002).

More positively, though, jurors do shift their views once they have received new information. Thus, in one study, mock jurors who had heard expert testimony spent considerably more time discussing eyewitness identifications during jury deliberations (Hosch, Beck, & McIntyre, 1980). In a different study, exposure to expert testimony resulted in increased juror attention to identification conditions and also to better post-trial knowledge of the factors influencing identification accuracy (Cutler, Penrod, & Dexter, 1989; for more on this broad topic, see Hosch, Jolly, Schmersal, & Smith, 2009; Leippe & Eisenstadt, 2009; Leo & Liu, 2009; Read & Desmarais, 2009; Schmechel et al., 2006).

The Limited Value of Anecdotal Evidence

Many factors contribute to this contrast between common sense and scientific evidence, but part of the explanation lies in the fact that common sense is often informed by *anecdotal evidence*. This term refers to evidence that is informally collected, stored only in memory, and conveyed (as an "anecdote") only in broad form.

With just one observer and no documentation, there is obviously room to question whether an anecdotal report is true at all. If we get past that issue, we still need to ask why a particular anecdote is recalled and reported. An obvious possibility here is that the anecdote is recalled because it "stands out" in memory—presumably because the remembered episode is somehow distinctive. But, if so, then we probably should not draw general conclusions from what is apparently an unusual case.

Anecdotal evidence often takes the form of a "man who" story: "What do you mean that cigarette smoking causes cancer? I know a man who smoked 8 packs a day and lived to 103." "What do you mean that suggestive questioning can mislead a witness? I recall a witness who was asked countless leading questions, but wouldn't budge in her story." Or, as a variant: "What do you mean that pretrial publicity prejudices a jury? I remember a case in which there was voluminous inflammatory publicity, but the jury still acquitted the guy." Reports in this form are common but suffer from obvious problems (they may or may not be accurately recalled; they rest on single cases that may or may not be representative, etc.). Hence, these reports cannot be persuasive.

F. THE POWER OF META-ANALYSIS

How can we avoid the perils associated with anecdotal evidence? More broadly, how can we make sure that scientific claims are an improvement on common sense? Part of the answer, as we've already seen, lies in systematic data collection, objective recording of the data, and conclusions that reflect *all* of the data (and not just a few notable cases). But part of the answer, in addition, lies in collecting a *lot* of evidence—to make sure a pattern is reliable, to make sure the data pattern emerges in diverse settings.

Once we've accumulated lots of data, though, how exactly do we combine the results to examine the overall picture? For many years, this combination was achieved through a *literature review*—a qualitative cataloguing of the available evidence, typically published in one of the professional journals that specializes in publishing review articles. However, in recent years, psychologists have shifted to *quantitative* reviews, relying on a method known as *meta-analysis*—literally, an analysis of other analyses.

A meta-analysis begins by defining its *inclusion criteria*—clearly stated, carefully justified rules for determining which studies will be included in the analysis. Once the inclusion criteria are laid out, all studies that meet these requirements must be included. In this way, meta-analysis rules out any sort of "cherry picking" in which one might favor results consistent with one's beliefs and disregard contrary findings.