Statistics with Confidence

An Introduction for Psychologists

Michael J Smithson

Statistics with Confidence

Sage Foundations of Psychology Series

Statistics with Confidence

Michael Smithson

SAGE Publications London • Thousand Oaks • New Delhi

First published 2000

SAGE Publications Ltd
6 Bonhill Street
London EC2A 4PU

SAGE Publications Inc
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
32, M-Block Market
Greater Kailash - I
New Delhi 110 048

British Library Cataloguing in Publication data

A catalogue record for this book is available from
the British Library

ISBN 0 7619 6030 9
ISBN 0 7619 6031 7 (pbk)

Library of Congress catalog card number 97-061783

Typeset by Keyword Publishing Services Ltd
Printed in Great Britain by The Alden Press, Oxford

To my daughter, Kia

'The touchstone of knowledge is the ability to teach' (Auctoritates Aristotelis)

Contents

	Preface	xi
1	**Uncertainty and Psychological Research**	**1**
	What are we doing here and why?	1
	The approach and structure of this book	2
	Some advice and support materials	4
	Paths to knowledge or belief	5
	Uncertainties in research	11
	Types of ignorance and uncertainty	13
	Quantifying and counting	20
	Questions and exercises	25
2	**Variables and Measurement**	**26**
	Observational and measurement strategies	26
	Basic concepts of measurement	29
	Measurement validity and error	32
	Types of measurement and varieties of data	37
	Discrete and continuous variables	38
	Levels and types of measurement	40
	Missing data, absent people, and the fallacy of false precision	48
	Questions and exercises	50
3	**Exploring, Describing, Displaying, and Summarizing**	**52**
	Exploration, description, display, and integrity	52
	Describing data from a nominal variable	54
	Bar charts versus alternatives	58
	How to lie with graphs: Lesson 1	61
	Describing an ordinal variable	63
	Describing a quantifiable variable	64
	Histograms versus alternatives	66
	How to lie with graphs: Lesson 2	70

Clear vision 71
General strategy 72
Summarizing distributions: Central tendency 72
Summation notation 76
Variability and dispersion 77
Choosing summary statistics 81
Questions and exercises 83

4 Research Design and Probability **87**
Validity in research design 87
Statistical inference 90
Sampling methods 94
Experimental design, randomized assignment, and statistical inference 103
An introduction to probability 106
Compound events and conditional probabilities 111
Probability distributions 116
Where have we arrived and where are we going? 126
Questions and exercises 127

5 Sampling Distributions and Confidence Intervals **131**
Estimating a population parameter from a sample 131
Sampling distributions for proportions 135
Confidence intervals for proportions 140
Confidence intervals for means 149
Confidence intervals for means and proportions with larger samples 157
Confidence intervals and summary statistics 162
How much confidence? 169
Human intuition about confidence levels 171
Answers to the confidence exercise 172
Questions and exercises 173

6 Statistical Models and Significance Tests **175**
Statistical models of a population statistic 175
Making decisions with confidence intervals: Significance tests 179
Making the wrong decision: Type I and Type II errors 185
Confidence intervals and the *t*-test 192
Differences between models and effect sizes 198
Confidence level, precision, and effect size tradeoffs 202
Questions and exercises 209

7 Predicting a Quantitative Variable from a Categorical Variable: The *t*-Test and Analysis of Variance **211**
Comparing means from two independent samples 211
Comparing proportions from two independent samples 220
Multiple confidence intervals 223
Comparing variances from two independent samples 230
The analysis of variance 234
The size of effect in ANOVA 241
Questions and exercises 249

8 Quantitative Predictors: Regression and Correlation **252**
Predicting a quantitative variable with a quantitative predictor 252
Evaluating linear regression models 262
Assessing the accuracy of prediction 269
More about the correlation coefficient 273
Evaluating the magnitude of a correlation 276
Correlation, prediction, determinism, and causation: A mildly philosophical interlude 283
Regression, ANOVA, and the general linear model 286
Questions and exercises 289

9 Predicting Categorical Variables: Contingency Tables and Chi-Square **293**
Tabulations and contingency tables 293
Working with one categorical variable 299
Working with two categorical variables 308
Chi-square, effect size, and association 313
Using percentages and proportions to interpret associations 317
The odds ratio as a measure of association 325
Questions and exercises 331

10 More than Two Variables: A Peek at Multivariate Analysis **334**
Concepts for multivariate analysis 334
Statistical and experimental control 343
Introduction to two-way ANOVA 347
Model comparison and *F*-tests for two-way ANOVA 353
Questions and exercises 362

11 Putting Statistics into Perspective **365**
Measurement, data, and statistics 365
Review of statistical inference 368
Statistical inference without random sampling or randomized assignment? 370
Statistical confidence and models 372
Effect sizes and effects models 376
Hypothesis testing and accumulating knowledge 379

References **382**

Answers to Selected Questions and Exercises **386**

Appendix. Statistical Tables **409**

Index **433**

Preface

This book is intended for students in the behavioral and social sciences who are undertaking their first course in statistics. It covers descriptive and inferential statistics, for one and two variables (with some material on multivariate techniques), and treats all statistical concepts at an introductory level. I have oriented the book to students with no more background than high-school algebra. This is not a mathematical book. The emphasis here is on understanding fundamental concepts and being able to use data-analytic techniques to enhance and extend our mental powers. There are no proofs or derivations, and the formulas or computations that have been included are intended to enhance conceptual understanding.

Additionally, unlike many statistics textbooks for psychology students, this book uses psychology and related disciplines extensively for examples and to inform the discussions of statistical concepts and techniques. In my view, psychology and the social sciences have a great deal to contribute to our understanding of how we make sense of data, construct scientific accounts, and debate those accounts throughout the course of our research and teaching. The perspectives offered by these fields also keep instructors and students alike mindful of the fallibility and tentativeness of any research method (statistical or otherwise), or controversies and changes in methods, and the realization that no method is set in stone.

Chapter 1 introduces statistics as a way of managing statistical uncertainties, which is only one kind of uncertainty in research. This chapter also poses the idea that gaining or imposing certainty in one area may require various kinds of tradeoffs. One of the most common tradeoffs is a repeated motif throughout this book – whereas greater precision makes it more likely that we will be wrong, more vagueness renders our statements less informative, diagnostic, or predictive.

Chapters 2 and 3 emphasize exploratory data analysis (EDA), in the sense of graphical and tabular techniques for displaying data, and 'packages' of summary statistics such as the five-number summary. Students are advised that EDA comes in two flavors: data reduction (summarizing, truncating, or combining data) and data enhancement (elaborating, explicating, or dissecting data). Chapter 2 addresses the basic concepts involved in measurement and the construction of variables, providing students with a framework in which to understand various kinds of data and measurement issues. Chapter 3 then presents techniques for EDA, including descriptive statistics and data displays.

Chapter 4 introduces statistical inference, staying with the concepts of random sampling and randomized assignment, the fundamentals of probability, and building up probability distributions from there. This chapter also informs students about different sampling designs, and the fundamentals of designing experiments.

Chapters 5 and 6 present confidence intervals as a way of displaying the range of plausible values for a population parameter, given the data at hand. These two chapters set this book apart from most other introductory statistics textbooks, because they present a confidence interval framework in lieu of the traditional significance-testing approach. They also set the standard for the treatment of most of the inferential techniques dealt with in subsequent chapters.

The main goal of this book is to enable students to understand and make clear, sensible statements about data and to know what degree of confidence can be ascribed to those statements.

In addition to the confidence interval framework, I emphasize two related themes through this book:

- Measuring the size of an effect and establishing benchmarks for effect sizes, rather than focusing on Type I error rates. I find that concepts such as statistical power can be conveyed clearly and simply by using confidence intervals and effect sizes.
- Model-comparison as a way of assessing the relative merits of competing models or hypotheses. Hypothesis-testing (as in significance tests) is presented as a special case, starting with Chapter 6.

Chapters 7, 8, and 9 introduce bivariate statistical techniques, starting with the between-subjects *t*-test and ANOVA (Chapter 7), then proceeding to correlation and regression (Chapter 8), and thence to chi-square and odds-ratio techniques for categorical data (Chapter 9). Chapter 10 provides a glimpse into multivariate analysis, with an emphasis on two-way analysis of variance. Chapter 11 concludes the book with an overview and guides for choosing appropriate statistical techniques.

At this point, I feel compelled to offer a few remarks concerning my reasons for adopting a confidence interval framework. Quite simply, on all counts it seems clearly superior to the traditional significance-testing approach. Persuasive arguments to this effect have been offered for many years in psychology. Early examples include Rozeboom (1960) and Meehl (1967). The groundswell of authoritative opinion against significance testing and in favor of confidence intervals mounted to a tidal wave by the 1980s and early 1990s (e.g., Oakes, 1986; Hunter & Schmidt, 1990).

And still, instructional and editorial policies largely clung to significance testing. Frustrated commentators such as Oakes (1986: 68) asked why it hadn't been abandoned long ago, and yet a large-scale survey of American graduate psychology programs around that time (Aiken *et al.*, 1990) found little evidence of change. Recently, calls for banning significance testing altogether have appeared in high-profile journals (e.g., Hunter, 1997; Schmidt, 1996).

A task-force on this topic in the American Psychological Association (APA) published its report on the APA website in 1996, recommending substantial reforms in statistical analysis, but stopping short of stipulating a ban on significance tests. Their chief recommendations have been taken up in this book:

- More extensive descriptions of data (i.e., means, standard deviations, sample sizes, five-point summaries, box-and-whisker plots, other graphics, and descriptions related to missing data as appropriate);
- Routine reporting of both direction and size of effects as well as their confidence intervals.

Rather than banning significance tests, I think a healthier approach involves teaching confidence intervals and model comparisons, while presenting significance testing as a special case so that students can read the older literature. Even from a hypothesis-testing viewpoint, there are advantages to a confidence-interval-based approach. Confidence intervals alert us to all the null hypotheses we can and cannot reject. Power is much more easily taught with confidence intervals. That said, there is also clearly a need for statistical methods instruction to move away from significance testing so that the next generation will not repeat and perpetuate our errors. This book is my response to that need.

Finally this book was written to be used with computers, but it is free of computer-specific material. There are three reasons for this. First, even the most user-friendly software still requires getting used to, and examples based on computer statistics packages tend to be opaque to students who haven't become acquainted with the packages concerned. Second, software and operating systems are changing much more rapidly than the 'half-life' of a book, and it is far easier (and less expensive) to update computer-based material on electronic media than in a new edition of a book.

Third, one of the real advantages of computer-based educational materials is their interactivity, which lends itself to both exploratory intuition-enhancing materials as well as an endless supply of practice problems. Rather than having the book bear the entire burden of instruction, I have elected to 'export' many exercises and instructional enhancements to electronic media.

For this reason, the book contains only a small set of exercises at the end of each chapter, with more problems and exploratory modules available on the electronic media accompanying it.

As far as possible I have endeavored to make the computer-based extras platform-independent. For the most part, that entails Windows and Macintosh versions of files and executables. These materials include the following:

- Data-files referred to in the book, in ASCII, Excel, and SPSS formats;
- Answers to exercises and problems in HTML so they may be viewed by most web browsers;
- StatPatch, a suite of tutorial and exploratory modules;
- Demos, a suite of Excel wordbooks also for tutorial and/or exploratory purposes.

I also maintain a website associated with this book. It may be found at http://psy.anu.edu.au.staff/mike/Statbook/TOC.html. It contains more supplementary material, corrections to any errors found in the book, and links to other helpful sites and resources.

Acknowledgements

This book benefited enormously from input and assistance generously provided by many colleagues, and I am indebted to them. Valuable suggestions and advice for the book and statistics teaching have come my way from Natalie Albon, Neil Black, Michael Cook, Geoff Cumming, Suzy Dormer, Margaret Foddy, Wolfgang Grichting, JoAnn Lukins, Elinor McKone, James Neill, Peter Raggatt, Paul Reser, Judy Slee, Susan Smithson, Neil Thomason, Ross Wilkinson, and several cohorts of second-year Psychology students at The Australian National University and James Cook University. I would also like to thank Cobie Brinkman, Michael Cook, Bernd Heubeck, Jeff Ward, and the Australian Consortium for Social and Political Research for their generosity in providing me with data-sets, examples, and key illustrative materials.

My greatest thanks are owed to those who read drafts of the entire manuscript: John Beale, Alex Haslam, Craig McGarty, and Michelle Ryan. Michelle read two or three drafts of every chapter, and John shouldered the burden of indexing. The impact from these four colleagues went far beyond mere proofreading and error-spotting. Any parts of this book that read clearly and communicate effectively probably do so because I took their advice. On the other hand, any errors, lapses, or omissions remain my responsibility.

Michael Smithson, Canberra, December 1998

Uncertainty and Psychological Research

1

CONTENTS

What are we doing here and why?	**1**
The approach and structure of this book	**2**
Some advice and support materials	**4**
Paths to knowledge or belief	**5**
Uncertainties in research	**11**
Types of ignorance and uncertainty	**13**
Quantifying and counting	**20**
Questions and exercises	**25**

What are we doing here and why?

Not many students enter into psychology with a burning desire to study statistics. Most of us are really interested in the subject matter of psychology: people. We want to find out about the secrets of human behavior and mental life. So, it is quite reasonable for us to wonder why we have to study statistics. After all, it has a world-wide reputation for being both boring and difficult.

A traditional short-term answer is that we need some understanding of statistics in order to read psychological research of the sorts that we encounter in other courses. Those of us who end up doing any psychological research need to understand statistics fairly well. This answer seldom satisfies everyone, and some of us go on to ask why psychology, of all disciplines, should seem so preoccupied with statistics. This, too, is a fair question, whose answer I hope will become clear long before you have reached the end of this book. For now, it will have to suffice to say that statistical methods are ways of coping with some of the uncertainties in psychological research and theory.

Another traditional answer is that learning about statistics and data analysis will provide you with very marketable, portable skills that not only equip you well to go on in psychology but in many other areas as well. I regularly receive letters, emails, or phone-calls from students who are kind enough to let me know that they got their present job or promotion partly because they were the only applicant who could handle data. Even those who claim they had forgotten everything say that having done statistics before gave them the confidence that they could do what was required of them. It is certainly the case that in recent times many types of employment have come to require data-analytic skills, familiarity with information technology, and a capacity to conduct research on people – and a course in statistics contributes directly to all of these.

The best, most long-term reason for studying statistics, however, is the same as for studying psychology or anything else at a university. It hands us one of the keys to improving how we live. We all must live amidst risks, for instance, and these days we are compelled to take account of more risks than at any time in the past. Assessing risks without any understanding of statistics is akin to flying blind. Probability and statistics comprise the best frameworks invented for dealing with uncertainty and assessing risks. They enable us to think about uncertainty and risk in ways that are totally unavailable to anyone who knows nothing of probability or statistics. Statistical knowledge also equips us with a special critical acumen that can help us to see through scams and falsehoods generated by those who juggle numbers and graphs. In fact, one of the first things you will learn in this book is some techniques for lying with statistics, so that you may know a statistical falsehood when you see one.

The approach and structure of this book

The main goal of this book is to enable you to make clear, sensible statements about data and to know what degree of confidence can be ascribed to those statements. Accordingly, in the first few chapters we will be focusing on ways of thinking about, describing, and presenting data; and from Chapter 4 onwards we will deal with techniques for making inferences and estimates from data and assessing how much confidence we can have in them.

This is not a mathematical book, although it does contain some formulas. There are no proofs or derivations, and what computations there are have been included to enhance your conceptual understanding. The emphasis throughout is on concepts, along with practical knowledge of how and when to use statistical techniques and what they can and cannot tell us. New terms and concepts are set out in **bold**. They are recapitulated at the end of each section. Likewise, questions and problems are included at the end of each major section, with answers to questions with **bold** numbers provided in the Answers section at the end of this book.

As much as possible, the examples in each chapter have been based on real research from a variety of psychological topics. Although that sometimes entails digressions by way of explaining what the examples are about, it should more than make up for that in the extra understanding of how statistics is used in genuine psychological studies. Likewise, the exercises and problems have been made as realistic as possible.

This first chapter begins with an overview of scientific research and how it is related to other ways of knowing, learning and discovering. We then move on to sections dealing with uncertainties in research and issues that arise when we count or quantify things. The themes that run throughout this material are the place statistics has in psychological research, what it is designed to do, and some important considerations about when and how to use statistics.

Chapter 2 concerns measurement and variables. It starts with strategies for observing and measuring aspects of human behavior and mental life. Concepts of validity, reliability, and error in measurement are introduced, with the understanding that people are reactive and intelligent and therefore cannot be studied in quite the same ways as we would study inanimate things. The 'technical' portions of this chapter introduce you to the different kinds of data and ways of working with them, as well as issues raised by missing or incomplete data.

Chapter 3 builds on the material in Chapter 2 by elucidating basic principles of data description, exploration, and summarization. It also leads you through alternative ways of displaying data, including ones that can mislead or misrepresent. This is the chapter that deals with descriptive statistics, which are quantities for summarizing and characterizing larger collections of data.

Chapter 4 is where we begin to work with inferences, or plausible statements about numbers. It is not enough to just describe a particular research finding; we wish to be able to generalize from it and also infer possible explanations of the phenomena being studied. This chapter thereby introduces the idea of inferring a characteristic of a population from a random sample of data, and the related concept of randomized assignment in experimentation as a basis for inferring an experimental effect. The basis for these inferences is probability, so the second half of Chapter 4 explains the basic aspects of probability theory and how it can be applied to statistical inference.

From here on, it becomes difficult to describe chapter contents in much detail because they involve increasingly many terms that have not been introduced yet. Chapters 5 and 6 are in many ways the backbone of this book. Chapter 5 introduces the idea of a *confidence interval*, which is an interval around an estimate that is associated with a certain level of confidence. You have probably encountered intervals around estimates before (e.g., 'The latest prediction is that the unemployment rate will be somewhere between 8% and 10% next year'). The same general idea applies to estimating a numerical characteristic of a population when we have only a representative sample from that population. For instance, if we take a random sample of a thousand people from the city of London and measure their heart-rates, the average heart-rate of the thousand will not necessarily be exactly the true average of the entire London population. How far away might it be, and with what probability? Chapter 5 enables us to answer that kind of question.

Chapter 6 applies these concepts to *hypothesis testing* and *comparisons between alternative models* of underlying reality. These extensions provide the keys to understanding a great deal of experimental psychological research. Chapters 7, 8, and 9 bring those concepts into the realm of predicting one variable from another, which is one of the Holy Grails of scientific research. Chapter 10 briefly introduces some additional concepts needed for psychological research that involves more than one predictor at a time. Finally, Chapter 11 provides a review and retrospection on the material covered in this book.

Some advice and support materials

Students encountering statistics for the first time often worry about whether they will be able to understand the material or keep up with the pace of instruction. They fear that their lack of training in mathematics will jeopardize their chances of doing well. If you feel you don't have a strong mathematics background, do not worry: that will not stand between you and understanding the material in this book. For one thing, most of your fellow students and your instructors are not mathematically inclined either. For another, this is most certainly not a course in *mathematical* statistics, although it introduces and uses some concepts from that field. Instead, it is aimed at a practical and conceptual understanding of what statistics is for, how to use it wisely, and how to interpret what others have done with it. There are *some* arithmetical concepts and techniques that you need to be reasonably familiar with, but they involve nothing more than basic arithmetic and secondary school algebra. With a bit of patience and effort, you will find that things will come more easily as time goes on.

Perhaps the most important thing is if you don't understand something in the book, *please ask someone*. Don't worry about sounding stupid; there is no such thing as a 'stupid question' in statistics. In picking up a foreign language you have to ask people what this means or how to say that, and the same is true of statistics and research. Study the material with a friend as much as possible. Often two heads really are better than one, and even when you are the person doing the explaining you'll find that you learn more from having to articulate your knowledge.

The second thing you need to bear in mind is that studying a research methods book is a lot like reading in a foreign language for the first time. Take things slowly and make sure you understand the meaning of each symbol, new term, or formula before going on to the next bit. Try to do small amounts of study regularly rather than cramming just before an exam or assignment is due. If you find yourself getting lost, make a note of where you got lost and why before going to someone for help.

Something else that you can do to aid your learning is to make use of available support and resources. The electronic media accompanying this book provide helpful materials in addition to the questions and problems at the end of major sections and chapters. First, there are data-files corresponding to the appropriate problems and most of the main examples. These are in SPSS, Excel, and ASCII formats, so they should be readable by most statistics packages.

Second, there is a suite of tutorial modules collectively called *StatPatch* and *Demos*, and you will see references to them throughout this book. StatPatch is a mix of exploratory and problem-generating modules designed to build understanding and intuition in statistics. Demos is a collection of Excel workbooks, serving much the same purposes. One advantage that they have over textbook problems is that these modules generate infinitely many problems as well as providing immediate feedback, so you can learn at your own pace and have as much practice or exploration as you wish.

Finally, I maintain a website associated with this book. It may be found at http://psy.anu.edu.au/staff/mike/Statbook/TOC.html. It has links to other sites and helpful resources for psychology students studying statistics and research methods. While I would have liked to include web addresses in the book, I decided not to, mainly because many addresses change fairly often and new resources appear all the time. You are more likely to obtain the latest sites and correct addresses if I maintain them on the web.

Paths to knowledge or belief

Researchers are sometimes called 'knowledge workers.' This phrase suggests that their main contributive goals are adding to what we know or believe about the world, and correcting erroneous beliefs. Since many psychologists and other kinds of researchers claim to be using 'scientific methods' in their pursuit of knowledge, it is worthwhile briefly considering what they mean by this claim and how scientific methods are related to and distinguished from other ways of acquiring knowledge.

The following list is not exhaustive, nor is it the only possible list of different methods of knowledge acquisition. It is partly based on debates in the philosophy and sociology of science over whether science is an institution that is truly distinguishable from other institutions, and whether it has a defensible claim to superiority over those other institutions in getting us closer to truth or even reducing untruth. Fascinating as those debates are, we will not go into them here. Instead, the intention is to provide a rough guide to various paths to knowledge and belief, including scientific research:

- Personal (first-hand) experience
- Authority and/or consensus
- Intuition
- Common sense and tradition
- Rationalism and reasoning
- Scientific methods

DEFINITION

Personal experience encompasses events that we describe with phrases such as 'I saw it with my own eyes' or 'hands-on.' For many people, first-hand personal experience is synonymous with reality-testing. It is virtually impossible for us to see ourselves holding false beliefs here and now; the best we can do is to realize retrospectively that we once held a belief that we now consider false. This 'blind-spot' points towards one of the main drawbacks to reliance on personal experience, namely that without adequate precautions and comparisons with others' experiences, we may easily be led astray.

For one thing, personal experiences are necessarily very circumscribed and may not even comprise a representative sampling of the totality of experiences. Our experiences of blind people, avalanches, dugongs, and snowflakes encompass only a tiny and unrepresentative fraction of the blind people, avalanches, dugongs, and snowflakes to be found anywhere and for all time. Nevertheless, we sometimes overgeneralize on the basis of our experiences, as in making inferences to the entire population of dugongs from the only one we ever saw.

Worse still, we may be deluded or fall prey to illusions in our own experiences. You undoubtedly already know that your senses (sight, for instance) can be fooled by a magician or an optical illusion. We suffer from cognitive illusions as well, some of which we will become acquainted with in this book.

DEFINITION

Personal experience is nevertheless a crucial component of any scientific method, because scientific methods are grounded in empiricism. **Empirical methods** are those based on first-hand experiences of the world, so personal experience is a necessary component of those methods. **Empiricism** is a doctrine that ascribes superior truth-status to things that have been directly observed or manipulated over things that cannot be observed or manipulated. Most scientists are empiricists of one kind or another.

There are prescriptions in the scientific versions of personal experience that distinguish them from the usual versions. Most importantly, a scientist is supposed to adopt a stance of **impartiality** (or **disinterestedness**) towards all competing opinions or theories, including their own. That does not mean they cannot have values or pet ideas, although many writers confuse impartiality with the notion of being value-free (whereupon they rightly contend that no one is value-free and then wrongly conclude that scientists cannot adopt an impartial stance). It does mean that a scientist should take precautions in their research so that someone with different values and opinions could repeat their investigations and arrive at the same conclusions. An experiment set up so that the experimenter is 'blinded' with regard to which subjects have been assigned to which treatment condition is an example of such a precaution. Another example is designing a study expressly for investigating conditions under which the scientist's theory should fail if it is incorrect.

In direct contrast with personal experience, using **authority or consensus** as a path to knowledge entails relying on second- or third-hand accounts of others' experiences. Authorities are sources with high status in our eyes. Parents, teachers, scientific experts, and religious leaders are examples of authorities. So are encyclopedias, scientific journals, television news programs, and computer programs. When every relevant authority agrees on a proposition (e.g., 'the world is round, not flat'), we have a consensus that makes that proposition appear indubitable. It is not difficult to see that the vast majority of what we think we know is based on appeals to authority and/or consensus. DEFINITION

Authorities can, of course, be wrong. A Dean of the Harvard Medical School was renowned for declaring to incoming first-year students that before they graduated they would have to commit some 40,000 'facts' to memory. Within 10 years of their graduation, about half of those 'facts' would be shown to be wrong. Unfortunately, he was fond of concluding, we never know which half. We have no way of knowing which of today's authoritatively established truths will become tomorrow's laughingstock. Moreover, the greater the authoritative consensus behind a belief, the less likely anyone will buck the tide to find out whether it is wrong after all.

Scientific methods rely on appeals to authority, and agreement among relevant authorities is a legitimate goal in scientific work. All scientists are members of one or more scientific communities and none of them remain uninfluenced by those communities. Scientific communities have norms and institutions that many have argued make them less likely than other communities to fall prey to a misleading authoritative consensus. A **norm** means a usual or expected practice, rather like a custom. One of the most popular and also widely criticized lists of scientific norms is Robert K. Merton's (1973). I have added one more to his original four (Honesty). DEFINITION

- **Universalism**: Research and theory are to be judged on their own merits, regardless of the scientist's gender, ethnicity, creed, political affiliation, or any other characteristic. Blind peer review of research papers is an example of this norm in action, since the authors of the paper and the reviewers are unknown to each other.
- **Organized skepticism**: All ideas and evidence should be carefully scrutinized and subjected to skeptical inquiry. No results or conclusions should

be accepted other than provisionally, and even then subject to replication by other independent researchers.
- **Communalism**: Scientific knowledge should be shared freely with everyone. Proprietary secrecy is contrary to this norm. Where ethically possible, research practices, processes, data, and other 'raw' instruments or products should be publicly available for scrutiny.
- **Disinterestedness**: Alternative ideas are to be considered and tested on an equal footing with one's own, in such a way that someone with other views could repeat the tests or investigations and arrive at the same conclusion.
- **Honesty**: Cheating or dissembling is an especially strong taboo in scientific communities, so much so that an instance of it may result in banishment or ostracism.

This list of norms has provoked heated debate, both about whether scientists really adhere to them and whether they should. While there are plenty of counterexamples against each of these norms (e.g., instances of prejudice, discrimination, credulity, secrecy, or fraud among scientists), defenders of the scientific community point to the institutional practices that embody them and observe that to the extent that anyone adheres to those norms they are adopting a scientific outlook and attitude (cf. Grinnell, 1987).

Now let us turn to **intuition**. In one sense, having an intuitive understanding of or belief about something entails not being able to ascribe that understanding or belief to a legitimate basis. Another sense of this term refers to the sudden, blinding insight that seems to arrive from nowhere. Both of these meanings amount to tacit knowledge, knowing something without knowing how we know it. It is here, perhaps, that scientists most closely resemble everyone else. While some famous scientists have written popular accounts of having flashes of intuition and while many scientists prize good intuition as highly as the rest of us, they also happily confess that they don't know how it happens either! There is a widely held view among scientists that intuition alone is not sufficient to justify an idea, but that is not news to most people.

There is one respect in which scientists may diverge somewhat from popular views about intuition and common sense. They tend to be fascinated with research outcomes that fly in the face of intuition or common sense. It is possible that the fascination with counter-intuitive findings simply reflects a shrewd judgment that such findings are unlikely to have been discovered before and quite likely to advance one's scientific career, but there seems to be more to it than that.

Common sense and traditional truths are repositories of second- and third-hand knowledge loosely organized into theories and explanations of how the world works. Common sense is certainly a good place to begin but may be a bad place to end for scientific research. Moreover, psychology probably has one of the most difficult relationships of any discipline with common sense. The main reason for this is that most of us are pretty good common-sense psychologists, at least within our own cultures. Otherwise, we could not make our way through everyday life. In contrast, most of us are rather poor common-sense chemists and very poor common-sense subatomic particle physicists. Fortunately, we have little need to depend on our common sense in those areas. We can leave them to experts.

Psychological research often is accused of not going any farther than common sense while taking much longer to get there. There are two lines of defense against such accusations. One is that common sense contains mutually contradictory propositions that are not recognized as contradictory because people use them at different times. It is not difficult to think of opposing proverbs that demonstrate this, for instance:

- Look before you leap, vs. He who hesitates is lost.
- Opposites attract, vs. Birds of a feather flock together.
- Absence makes the heart grow fonder, vs. Out of sight, out of mind.
- Many hands make light the work, vs. Too many cooks spoil the broth.
- It's never too late to learn, vs. You can't teach an old dog new tricks.
- No one is an island, vs. We die alone.

Haslam & McGarty (1998) make amusing and instructive use of the third pair of proverbs to demonstrate how one might build up a research program to investigate which one is correct under various conditions. The other line of defense refers back to scientific norms of organized skepticism and disinterestedness. No matter how many people have endorsed a common-sensical assertion and no matter how long it has been believed, if it has not been properly tested then it is not scientific knowledge.

Finally, we turn to rationality and reasoning. **Rationality** involves adherence to a system of reasoning (usually standard logic). A popular view of science and, to a greater extent, mathematics, is that it relies heavily on logical reasoning and thereby rationality. While scientific research does make use of logic, logic is by no means sufficient on its own. Traditionally, rationality (along with rationalism) has been linked with knowledge and certainty. While ancient canons of rationality comprised substantive contents and told people what to believe, those versions were gradually supplanted by procedural and algorithmic prescriptions. Instead of directing people to specific conclusions, modern versions of rationality tell them how to reach conclusions. That is why most widely accepted versions of rationality boil down to some kind of logical consistency and coherency. DEFINITION

What is **rationalism**? It amounts to faith that rationality is the 'best' guide to decision making. Anything else (i.e., the nonrational, irrational, or anti- DEFINITION rational) is considered to be worse. Rationalists are anti-Heraclitans, which means they think there is sufficient regularity and stability in the universe for us to learn generalizable lawlike properties of it. They share this view with many empirical scientists (and much common-sense reasoning as well!). If we do not have a learnable world in some minimal sense, then rationality has no use. The usefulness of logical consistency assumes predictable, stationary relations among things in the real world. So does much scientific research. Some of the debates about whether psychology can or should be a science hinge on just this issue.

Where does statistics fit into all of this? Statistics is the offspring of a liaison between empiricism and rationalism. Statistical techniques are derived from general frameworks for understanding empirical data, so statistics and empirical research go hand-in-hand. Statistical models are based on theories of probability that, in turn, have some rationalistic and mathematical foundations. The marriage of empiricism and rationalism has not always been a peaceful one, and there are competing theories of statistics and probability. The versions we will use in this book are the most popular in psychology and work quite well under a wide range of conditions, but it is always wise to bear in mind that they are not the only approaches that could be used.

SUMMARY

The alternative **paths to knowledge** and belief reviewed in this section include:

- Personal (first-hand) experience
- Authority and/or consensus
- Intuition
- Common sense and tradition
- Rationalism and reasoning

Science makes use of all of these, albeit in ways that differ from their uses in everyday life.

Scientific communities have **norms** that many have argued make them less likely than other communities to fall prey to a misleading authoritative consensus:

- Universalism
- Organized skepticism
- Communalism
- Disinterestedness (or impartiality)
- Honesty

SUMMARY

Scientific methods are grounded in **empirical methods**, based on first-hand experiences of the world.

Empiricism is a doctrine that ascribes superior truth-status to things that have been directly observed or manipulated over things that cannot be observed or manipulated.
Rationality involves adherence to a system of reasoning (usually standard logic).
Rationalism is a faith that rationality is the best guide to decision making.

Statistics and probability are a combination of empiricist and rationalist ideas.

Uncertainties in research

The phrase 'psychological research' claims a large and diverse terrain, perhaps larger and more diverse than at any time in the history of the discipline. Pick 500 psychologists at random, ask them to describe how they do their research, and the answers will probably include experiments, surveys, case studies, in-depth interviews, ethnographies, test construction, discourse analysis, and computer simulations. You might encounter some who are doing historical or archival investigations, or even archeological studies.

It may seem as if there are no concepts or methods shared by all of these approaches to studying human beings. There are certainly practitioners of specific approaches who say that their approach has absolutely nothing in common with others. Nevertheless, there are good reasons to be suspicious of this sort of territorial statement and to think that researchers may share a few common bonds after all.

First, all researchers engage with the unknown in one sense or another. They begin by claiming that there really is something new under the sun and they are going to return from their voyaging to tell us something about it. Accordingly, they grapple with uncertainties, trade in novelties, map uncharted seas, and make discoveries. For all researchers, ignorance and uncertainty are both friend and foe, sometimes simultaneously. Without ignorance or uncertainty, there is nothing new to discover and the research game is over. In the grip of ignorance or uncertainty, however, the researcher is seldom in a position to demonstrate or prove anything conclusively. The physicist Richard Feynman captured this essential characteristic of scientific work in his 1955 address to the American National Academy of Science:

> The scientist has a lot of experience with ignorance and doubt and uncertainty, and this experience is of very great importance, I think. When a scientist doesn't know the answer to a problem, he is ignorant. When he has a hunch as to what the result is, he is uncertain. And when he is pretty

> darned sure of what the result is going to be, he is in some doubt. We have found it of paramount importance that in order to progress we must recognize the ignorance and leave room for doubt. Scientific knowledge is a body of statements of varying degrees of certainty – some most unsure, some nearly sure, none absolutely certain. (Feynman, 1988: 245)

Second, all researchers are members of one or more research communities. These are collections of people who agree sufficiently with one another to be able to share a conceptual framework, but whose discourse within that framework is characterized by vigorous argument, disputation, and conflict. Like ignorance, disagreement is both friend and foe to the researcher. Researchers crave consensus, but only up to a point. Complete agreement is a disaster for research, because no one is able to move outside the accepted way of thinking and there is nothing genuinely creative going on. Disagreement, while essential for motivating research, is often also agonizing for the researcher. Given the stylistic conventions of the time, a medical researcher turned philosopher of science, Ludwik Fleck, expressed this very well in 1935: 'At the moment of scientific genesis [discovery], the research worker personifies the totality of his physical and intellectual ancestors and of all his friends and enemies. They both promote and inhibit his search.' (Fleck, 1935/1979: 95.)

Third, all researchers make mistakes, both in their own eyes and the eyes of others. Here again, error is friend and foe. Anyone who gets it right the very first time really has learned nothing new. Making an error and realizing that it is an error are necessary components of any learning process, and therefore any process of discovery or creation as well. Again, Fleck is right on the mark: 'Discovery is thus inextricably interwoven with what is known as error. To recognize a certain relation, many another relation must be misunderstood, denied, or overlooked.' (Fleck, 1935/1979: 30.)

This does not mean, of course, that making any kind of mistake leads to discovery or learning. It *does* mean that reluctance to take a step for fear of making a misstep will surely impede discovery and learning. All researchers strive against indoctrinated fears of failure, error, and ridicule, much of it traceable to years of formal education that has rewarded them only for correct answers to problems whose solutions already are known. We can always dream of a system of education that does not penalize students for making mistakes! On a slightly more sober note, we can reward ourselves and others for risking productive and interesting mistakes, along with careful descriptions of them and our current states of ignorance. The University of Arizona's Medical Curriculum on Ignorance (Kerwin, 1993) is a salutary (and, alas, nearly solitary) example of a curriculum that invites students to describe and study not only knowledge but also what they don't know, and sustains ignorance as an object of study throughout their entire degree program.

This book is about uncertainty in psychological research and some widely shared methods for understanding and coping with it. It is also about how to make productive mistakes by taking strategic risks in designing and conducting research. That said, this book does not cover anything like the full gamut of research styles and techniques, the varieties of uncertainty, or their sources. That would require many books. None the less, there are some reasons behind the choices of research styles and uncertainties that inform this book's core.

To start with, we can place this book's focus in the context of various kinds of ignorance and uncertainty. The following list is adapted from Smithson (1989) and divides ignorance into two major chunks:

Types of ignorance and uncertainty

1. Distortion
   - Qualitative: Confusing one thing for another.
   - Quantitative: Systematic inaccuracy.
2. Incompleteness
   - Absence: Missing information.
   - Uncertainty: Indeterminate information.
     - Probability and statistical uncertainty: Likelihood of an event.
     - Ambiguity or vagueness: Multiple possible meanings or a range of values.

The first chunk, **distortion**, is usually taken to be some kind of systematic descriptive error. Its qualitative version consists of **confusion**, mistaking one thing for something else (as in a misdiagnosis), and its quantitative version consists of **inaccuracy** (as in a miscalibrated weighing scale). DEFINITIONS

The second chunk, **incompleteness**, refers to information that is **absent** (missing) or **uncertain** (indeterminate). Indeterminacy of information is then subdivided into two categories: probabilistic and ambiguous or vague. **Ambiguity** and **vagueness** refer to ways in which information can be blurry, have multiple interpretations, or have shades and degrees of meaning. **Probability**, on the other hand, refers to the likelihood that something will happen.

Psychological research (indeed, perhaps all research) necessarily traffics in all of these kinds of ignorance and uncertainty. In psychology, problems of distortion are usually the province of measurement and ascribing meaning to our observations. We will introduce some of the basic concepts of measurement in Chapter 2, and discuss some issues concerning distortion in measurement there. Entire textbooks are devoted to measurement, however, and that is not the primary focus of this book (see Kaplan & Saccuzzo, 1989, on psychological testing and measurement, for example, or Foddy, 1993, on designing questions for surveys and experiments).

Incompleteness, on the other hand, is mainly the province of data-analytic techniques, particularly statistical techniques. Most of this book focuses on incompleteness, especially probabilistic or statistical uncertainty. Uncertainty is generally held to be more difficult to deal with than distortion, and less likely to be eliminated even from the best research. However, uncertainty can be described, sometimes quantified and estimated, and even manipulated in the service of the researcher. In this book we will encounter these three strategies for managing uncertainty many times.

Another viewpoint on uncertainty in research emerges once we distinguish among the sources of uncertainty that become salient during the research process. One way to understand this is to begin with a schematic guide to that process. Like almost any schematic, the one in Figure 1.1 is oversimplified. It begins with the researcher defining a topic and research goals, developing questions and/or hypotheses, and then going on to design the study, collect and analyze data, and finally interpreting the findings and revising what is known about the topic on the basis of those findings.

The feedback loops indicated in this figure are not the only possible feedback effects, and to some extent the revisability of the earlier stages depends on the kind of research being conducted and the norms of the research community involved. A rigorous experiment, for example, designed to test very specific hypotheses, leaves little room for the researcher to revamp those hypotheses in midstream. None the less, the feedback loops represent the fact that the research process may not be a one-way sequence but can involve successive iterations, whereby the researcher oscillates between stages until satisfied enough to move on.

DEFINITION

The six stages in this schematic also provide convenient labels for sources of uncertainty. **Topical uncertainty**, to begin with, concerns how the researcher is to describe the object of their investigations. Consider psychological research on affect or emotions. Are we studying emotional traits such as temperament, or more temporary states like moods, or even briefer episodes?

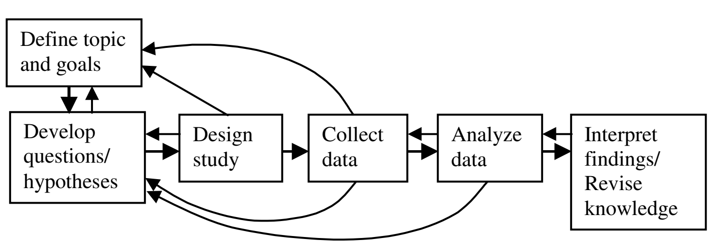

FIGURE 1.1 Research process

Are emotions best thought of as physiological, somatic, or socially based? What does the topic of emotions include and what does it exclude? Until we have at least tentative answers to questions such as these we cannot be sure about our topic.

Investigative uncertainty concerns research questions and hypotheses. The kinds of questions that are sensible to ask about a topic are influenced to some extent by prior views and assumptions about the topic. For instance, if we assume that emotions are primarily products of physiological processes, then the most obvious starting-point would include aiming at an accurate description of the physiological concomitants of each distinct emotion. If, on the other hand, we assume that emotions are mainly interpretive constructs that arise from social interactions, then inquiring about physiological states and arousal would seem less relevant. DEFINITION

Methodological uncertainty refers to the design of the study and whether it will suit our purposes, answer our questions, or test our hypotheses. This term is taken from Haslam & McGarty (1998, Ch. 11) but used somewhat more broadly here. They distinguish between two kinds of methodological uncertainty, whereas I will use three: DEFINITION

- Design uncertainty
- Internal uncertainty
- External uncertainty

Design uncertainty concerns the overall method to be used in the study. Should we set up an experiment? A survey? What about a qualitative field study? **Internal uncertainty** refers to whether the researcher can interpret the outcomes of the study correctly, and **external uncertainty** refers to whether the study's results can be generalized to other populations or settings. These concepts may seem quite abstract now, but they will become clearer in the chapters to follow. For the present time, let's consider a few examples of research that has involved methodological uncertainty. DEFINITIONS

One of the chief concerns for anyone using more than one method to study something is that the methods might produce conflicting findings. Researchers must then sift through the studies and findings for possible explanations of the differences among them. For example, self-report or 'subjective' measures may differ from 'objective' indicators, as is often the case on topics such as risk or quality of life. A researcher faced with this situation will attempt to weave an explanation for it into a general account of risk perception, or health consequences of perceived versus objective quality of life. This explanation might refer to how people's perceptions and attitudes differ from so-called 'objective' measures, or to differences between the phenomena being studied, or even effects of studying the same phenomenon in different ways.

As an example of the latter case, in a study in which I was involved (Smithson *et al.*, 1991) we found that survivors of suicide attempts were much more likely to admit to consuming alcohol at the time of their attempt if we asked them directly about it than if we left them to tell about the event in their own words. Any reasonable attempt to explain this contrast would require a theory of how people reconstruct such memories and then edit them into narratives about themselves.

A much earlier example is Hovland's (1959) review of the experimental versus survey research evidence on attitude change subsequent to exposure to a change-inducing message. He found that surveys were less likely to find evidence of attitude change than experiments. He ascribed this divergence to a tendency of the surveys to attract people who already favored the view advocated by the message, the shorter time-intervals used in experimental studies, and differences in the kinds of attitudinal issues used.

Sometimes the issues involved in comparing different methods may simply be too complex or researchers may be too biased in their views for a resolution to be achieved. Consider, for instance, the contrast between one group of primarily quantitative studies of mental patients and another group of mainly qualitative studies reviewed by Weinstein (1979) whose review is discussed at length in Bryman's (1988) textbook. The first group used structured survey questionnaires with rating scales, multiple-choice questions, sentence-completion tests, and the like. For the most part, these studies reported that mental patients had favorable attitudes towards their institutions, benefited from and even enjoyed hospitalization. The qualitative studies, on the other hand, used unstructured interviews with patients who in turn answered questions in their own words, observation by researchers masquerading as patients, and observations on hospital wards. These studies consistently found evidence of debasement and oppression by hospital authorities; and patients' feelings of anxiety, boredom, powerlessness, and betrayal.

Weinstein's review provoked controversy over how best to account for these different findings. He found fault with the methodologies of the qualitative studies. Critics of his review responded that Weinstein had tried to force a comparison between two relatively noncomparable sets of studies. They pointed to differences between the two groups on the admission status of the patients, and whether patient experiences or patient attitudes and outcomes were the object of investigation.

As Cook & Campbell (1979: 66) observe, relying on just one method to study a phenomenon when we know little about it lays the researcher open to accusations of 'mono-method bias.' The challenge in using more than one method is to do so in a strategic and even-handed fashion so that contrasting findings suggest new avenues for research and ways of integrating the findings. That way lies progress in any field.

The major portion of this book is devoted to **statistical uncertainty**, which is uncertainty related to the analysis and, to some extent, interpretation of data. **Descriptive statistics** characterize the data themselves, and so **descriptive uncertainty** concerns those characterizations. Suppose you have taken an exam in a cognitive psychology class, and the instructor is about to distribute the exam results. Before doing so, she mentions that the class average score was 64%. How well would this describe each student's score, including your own? The less variability in the scores, the closer the average would be to describing individual students' scores. DEFINITIONS

Inferential statistics, on the other hand, are used for drawing conclusions about populations or underlying processes from the sample of data at hand. **Inferential uncertainty**, therefore, arises when we are not sure what kinds of statistical inferences we can make from our data. For example, the instructor might wonder whether your class has scored higher than last year's class, whose average was only 59%. She would realize that the difference between class averages of 59% and 64% might occur simply by chance, and she would use inferential statistics to address that possibility. DEFINITION

Finally, **interpretive uncertainty** arises when, despite having good data and sound statistics, we are still unable to decide between competing interpretations for what we have found. Suppose the instructor finds out that your class probably did perform better on the exam than last year's class. Is your class more intelligent? Did they work harder? Did she do a better job of teaching? Were the exams equivalent, or was this year's exam easier? These are plausible alternative explanations for her findings, and she would want to eliminate all but one of them if possible. DEFINITION

Given all of these different sources of uncertainty, research might appear to be a very daunting enterprise. Most research, however, does not deal with all of them simultaneously. In fact, researchers routinely distinguish one kind of research undertaking from another in terms of which of these uncertainties are being dealt with. The labels we will work with here are 'exploratory,' 'descriptive,' and 'explanatory' research (Neuman, 1997, also uses these).

Exploratory studies deal primarily with topical and investigative uncertainties. If we do not know anything about a topic, if little or nothing has been written about it, then we need to refine our comprehension of it and develop questions that may be used to guide future research. Until topical uncertainty has been reduced to some extent, little progress on the other sources of uncertainty can be made. This is not to say that topical uncertainty must or can be completely eliminated. Some of the most interesting topics in psychology, creativity and consciousness being two examples, still are quite vague and sharply disputed even though they are the objects of long-running mature research programs. DEFINITION

DEFINITION **Descriptive** studies focus mainly on investigative uncertainty, although the research may also end up dealing with methodological and descriptive statistical uncertainties. The goal of description is to provide an accurate portrayal of the phenomena that leads to further questions, hypotheses, and eventually explanations and theories. Developing a better way of measuring anxiety would be a good example of a descriptive study. Descriptions may be in either qualitative or quantitative form, and quite often the researcher will use this research to organize understanding of the phenomena.

As you might already have imagined, some studies can be both exploratory and descriptive. In the mid-1980s I worked with a former stomatherapist nurse, Therese Turner, who wanted to do her Honours research project on how colostomy patients managed their stigmatized condition in everyday life after their operations. Since a colostomy entails rerouting the colon so that it empties involuntarily into an external plastic bag instead of via the rectum, people who have had a colostomy are often at risk of public embarrassment. A search of the literature at the time revealed almost no relevant studies, so she elected to conduct a descriptive study based on in-depth interviews of former colostomy patients. She began by asking them what they thought were the major problems they faced and how they dealt with them. The matters raised by these people in the interviews generated further questions, and she returned to her informants for additional information. In the end, their accounts provided many fruitful suggestions for future research as well as advice that could be provided to such patients before and after the operation.

Given a topic about which something is known and some descriptions of it, we tend to wonder why it is so. In **explanatory** research, the principal objects are reasons, causes, and interpretations. Explanatory studies therefore concentrate on statistical and interpretive uncertainties. They often test hypotheses or theories. We conduct such studies when we already have a good idea of the nature of our topic and what methods to use in studying it. Experiments are perhaps the best examples of explanatory research, because they require enough prior knowledge about a phenomenon to be able to manipulate some aspects of it in order to observe the effects that follow.

DEFINITION

A number of concepts and terms have been introduced in this section, some of which may seem abstract and unfamiliar. If you can bear with it, these ideas will become clearer and form the basis for a genuine overview of psychological research that will stand you in good stead, not just for learning the material in this book but for understanding the diverse kinds of research throughout psychology.

SUMMARY

Research of any kind has three things in common that are both friend and foe:

1. Dealing with ignorance and uncertainty,
2. Disputation and conflict within a framework shared by other researchers, and
3. Learning and discovery through errors.

The kinds of uncertainty dealt with in research include the following:

1. **Distortion**: systematic error.
   - **Confusion**: Mistaking one thing for another.
   - **Inaccuracy**: Systematic miscalibration.
2. **Incompleteness**: Missing or indeterminate information.
   - **Absence**: Missing information.
   - **Uncertainty**: Indeterminate information.
   - **Probability** and statistical uncertainty: Likelihood of an event.
   - **Ambiguity or vagueness**: Multiple possible meanings or a range of values.

Sources of uncertainty in research arise at each of its six stages:

Topical uncertainty concerns how the researcher is to describe the object under investigation.

Investigative uncertainty concerns the nature of research questions and hypotheses.

Methodological uncertainty refers to the design of the study and whether it will answer our questions or test our hypotheses.

- **Design uncertainty** concerns the overall method to be used in the study.
- **Internal uncertainty** refers to whether the researcher can interpret the outcomes of the study correctly.
- **External uncertainty** refers to whether the study's results can be generalized to other populations or settings.

Statistical uncertainty concerns the analysis and, to some extent, interpretation of data.

- **Descriptive statistics** characterize the data themselves, and so **descriptive uncertainty** concerns those characterizations.
- **Inferential statistics** are used for drawing conclusions about populations or underlying processes from the sample of data. **Inferential uncertainty** concerns what kinds of statistical inferences we can make.

Interpretive uncertainty arises when the researcher is unable to decide between competing interpretations of the research findings.

Quantifying and counting

Since statistics are closely allied to quantification and counting, we should examine both of those practices before sailing off into areas where we take them for granted. We will start with counting, since that is the more venerable of the two and easier to conceptualize.

Counting assumes that the things being counted all belong to the same category. Its main advantage over using words is obvious once we grant that assumption. Saying that 'many' people in the class are right-handed is a nearly useless description compared to saying that 72 out of 83 are. Moreover, we can perform arithmetic operations with counts that are impossible with linguistic terms. Numbers and mathematics are not arbitrary social conventions. They have been successful because they help us think more clearly about certain things. To get a quick appreciation of this assertion, try multiplying 'three hundred and twenty-five by one hundred and twelve' versus 325 by 112, or try doing division with Roman numerals (if you are curious about other counting systems, take a look at Barrow, 1992).

Before counting behaviors, manifestations of cognitive processes, or the like, we need to be sure that they really do belong to the same category. For instance, consider the act of choosing the correct alternative on a true–false question in an exam. If we count the number of people who chose that alternative, we are lumping together those who knew the answer and those who happened to guess it. There may be no harm in counting how many got the question right, but we would be mistaken if we went on to say that was the number of people who knew the answer.

Quantification involves assigning numbers to distinguishable observations. While some concepts such as speed, duration, or length seem 'naturally' quantifiable, many psychological concepts provoke debates over whether they are quantifiable and if so, how best to quantify them. We will explore concepts in this book that inform those debates. When it is successful, quantification has much the same advantages as counting. It enables us to say not just that change or differentiation has taken place, but *how much* of it has occurred.

There are some popular arguments against quantification and counting, and we should examine them before moving on. One of the most pervasive is that 'reducing people to numbers' is anti-humanistic. It degrades people by ignoring their uniqueness as individuals. It is true that quantifying and counting require that we lump people together in some respects, thereby ignoring unique features. All general descriptions and theories do this. However, careful description and measurement never degrades anyone or anything. Also, words can just as easily and far more tellingly debase people by distorting or glossing over important characteristics.

Another related argument is that many important, observable things cannot be quantified. Characteristics that are not quantifiable tend to be ignored or discounted in favor of those that are quantified. There is some truth to this argument too, but the fault does not lie with quantification itself. After all, even qualitative characteristics ultimately must be codified, summarized, and counted once sufficiently many instances of them have been collected. Problems about quantification arise mainly when numbers become separated from their contexts. Good researchers know that every numerical datum has a context that needs to be considered before combining it with other numerical data, such as realizing when an EEG is showing 'artifact', that a rat went to sleep in the middle of a maze-running trial, or when a child is not paying attention in a reaction-time task.

A more general overview of the tradeoffs involved here might refer to 'data reduction' versus 'data enhancement.' Ragin (1994: 92) provides slightly different terms, and aptly observes that data reduction enables a researcher to see the big picture at the expense of attending to details, while data enhancement provides surrounding contextual information about the data that enables the researcher to better understand a particular case. **Data reduction** entails *combining* or *truncating* data. In order to perform either operation, we must treat the data as if they are combinable or comparable, or as if their qualitative differences are irrelevant. **Data enhancement** entails elaborating a set of data by *dissecting* it into components, or *supplementing* it with related data. Data enhancement involves an assumption that each datum is unique or distinctive in some relevant way that needs further explication. DEFINITIONS

For instance, suppose we take weekly measurements of 100 people's levels of self-esteem using a well-established self-esteem index that consists of 15 questions, for a period of 11 weeks. An example of data reduction via combination would be averaging each person's self-esteem scores over the 11-week period. An example of reduction by truncation would be to rank their scores from lowest to highest, and use the middle (6th-ranked) score as their 'typical' score, ignoring all the other scores above and below it.

Likewise, an example of data enhancement via dissection would be breaking each score into the responses people gave on every one of the 15 questions in the self-esteem index. Data enhancement via supplementation, on the other hand, might consist of having people keep a diary of self-esteem-influencing events that would then be listed along with their score for the week (e.g., being reprimanded at work, or winning a ribbon in a local fun-run). Another kind of supplementation would be asking people to record their thoughts, feelings, and reasons for responding to the self-esteem questions, so that we have an elaboration of their accounts of the meanings behind their responses.

Many of the statistical techniques covered in this book have been designed to effectively condense or reduce data in various ways. In Chapter 3, for example, we will explore various kinds of summary statistics (such as the average, or arithmetic mean) that reduce a collection of scores to a few pieces of information about the properties of those scores. Other techniques, mainly those concerned with various ways of displaying data in graphs or tables, involve data enhancement as well as reduction. Although Ragin (1994) is oversimplifying somewhat, there is some truth in his claim that quantitative research techniques are mostly data condensers.

The tradeoff between these two ways of treating data is fairly obvious. Reducing large volumes of data to a few pieces of information is grist for any scientist's mill because it is compatible with the scientific goal of generalizable explanations and theories. A general theory effectively tells us that we may treat numerous specific cases as if they are essentially identical. When appropriately and intelligently applied, data reduction can reveal hidden order, regularity, or relationships among data in a powerful and even elegant fashion.

On the other hand, reducing means combining or truncating information and therefore ignoring it. The researcher who condenses data thereby risks ignoring important details or distinctions among particular cases. If taken to extremes, data reduction techniques can 'reduce people to numbers' by omitting crucial information about where the numbers come from or what they mean. In a somewhat different sense of the word, Dennett (1994) coined the term 'greedy reductionism' to refer to excessive reductionistic ambitions. Researchers can sometimes get carried away by the power of their data-reducing techniques, especially since the advent of computers. So can consumers of research. A friend who is an analyst in a large government department concerned with health and safety has repeatedly told me that she is always under pressure from her superiors to 'boil it down to *one* number.'

Data enhancement provides ways of grounding information in context. Even quantitative data may require data enhancement in order to be properly understood. For example, consider the effect of an income increase of \$100 per week on someone with a \$150 per week income versus someone whose income is \$10,000 per week. Or compare the student who has scored 70 on an exam with 70 on both the 'technical' and 'conceptual' components, with another student whose score of 70 is the average of 95 on the technical and 55 on the conceptual components.

The disadvantages and pitfalls of data enhancement are twofold. First, the researcher may become overwhelmed by elaboration and thereby unable to see the forest for the trees. This is simply the opposite side of the reductionism coin as outlined earlier.

Second, inappropriate or irrelevant contextual distinctions can mislead us into separating data that should be combined. This is a somewhat subtler point that the first one, but a simple example can illustrate it. Shoe-sizes are numbered according to different conventions in the U.S. than they are in Australia. I wear a size 11 shoe if I purchase it in the U.S. but only a size 9 if I buy it in Australia. A survey of shoe-sizes with samples from both Australia and the U.S. would therefore require that we record where the respondent's shoes were purchased. However, if the survey were restricted to Australia, then separating shoe-sizes by the state in which they were purchased would be irrelevant.

When should we choose data reduction or data enhancement? There is no simple answer. It depends on the researcher's goals and what is already known about the area. One strategy that is frequently used is to begin at one extreme (either reduction or enhancement) and then work back towards the other as far as is needed. To conclude this section, here is an example of a debate that is frequently found in psychology, namely whether a psychological concept should have more than one dimension or not. The concept in question is risk.

The editorial in a recent issue of the *Royal Statistical Society News* (October 1998) bemoaned the fact that people choose their risks in an 'irrational' way. According to the editor, people 'refuse to engage in activities which have known, but quite negligible, risks yet fearlessly participate in those whose dangers are orders of magnitude greater' (p. 1). He recounted the solution put forward by the past president of the RSS, which was that a 'Richter-type' scale of risk be constructed whereby people could compare known risks in a systematic way. A very similar proposal was made by the mathematician John Allen Paulos in his book, *Innumeracy* (1988). Such a scale would be an example of data reduction, since it would collapse all risk evaluation down to one dimension.

Eating a peanut-butter sandwich every day for one month, working in a coalmine for a few hours, and living next to a nuclear power plant for five years all involve an increase in risk of death of about one in a million, so if we were weighing risks on just that basis, we should equally value these three. But most of us do not. A large empirical research literature demonstrates that people perceive and evaluate risk along several dimensions. The list below contains the influences on risk preference identified in studies of risk perception (Otway & von Winterfeldt, 1982):

1. Involuntariness of exposure to the risk.
2. Lack of personal control over outcomes.
3. Uncertainty about probabilities or consequences of exposure.
4. Lack of experience or familiarity with the risk.

5. Difficulty in imagining consequences.
6. Delayed effects.
7. Genetic effects.
8. Catastrophic size of consequences (either geographically or numbers of people affected).
9. Benefits are not visible.
10. Benefits go to others but not oneself.
11. Human-caused rather than naturally caused.

The Richter-type risk scale does not provide a valid way of characterizing people's valuations of risk. It does, however, provide a worthwhile benchmark against which to compare how people do evaluate risks, because it orders risks along the continuum that we would use if probability of injury or death were our only concern in risk assessment. Given a person who has accurate information about such probabilities, we may use the Richter-type scale as a way of determining whether they are evaluating risks solely on the basis of probabilities. If we find that their preferences for risks disagree with the rank-ordering of those risks on the scale, then we know that we need to take more than just probabilities into account when attempting to describe people's risk preferences.

The moral to this example is that data reduction and enhancement can work together in getting a start in an unknown area. The field of risk assessment began with attempts to 'reduce' risk perception to a one-dimensional scale and the failure of that simple model stimulated the search for additional dimensions. This is an example of a pattern commonly encountered in scientific research, namely beginning with a simple, reduced model and then complicating it as necessary to fit the phenomena. The reverse process also can be found in some areas, whereby researchers begin with elaborate data enhancement and then systematically eliminate unnecessary features until they arrive at a more parsimonious model.

SUMMARY

Quantifying or even counting should not be undertaken without first considering whether these are sensible given possible arguments to the contrary.

Counting assumes that the things being counted all belong to the same category. *Quantification* involves assigning numbers to distinguishable states.

- **Data reduction** entails *combining* or *truncating* data.
- **Data enhancement** entails elaborating a set of data by *dissecting* it, or *supplementing* it with related data.

Two common strategies in research are to begin with data reduction and then enhance as much as necessary; or to start with data enhancement and then reduce as much as possible.

Questions and exercises

Q.1.1. Which are examples of scientific, rationalistic, intuitive, and authoritative methods of gaining or creating knowledge?

(a) Using a voltmeter to see whether your torch battery is flat.
(b) Your doctor says you have dermatitis, and you decide that you have dermatitis.
(c) Figuring that because only dogs make barking sounds in your neighborhood, the source of the barking sound outside your front door is a dog.
(d) Even though you haven't got a formal definition of creativity, you know it when you see it.
(e) Looking up the meaning of an English word in the *Oxford English Dictionary*.
(f) The curried chicken was too hot last week, so you try using one half-teaspoon less Madras powder this time.

Q.1.2. Give two examples of statistical descriptions.

Q.1.3. Give two examples of statistical inferences.

Q.1.4. Suppose a psychological researcher points out that everyone sees the same distinct bands in the rainbow regardless of how they classify or name colors, and uses this observation as an example of categorization that is independent of culture. Another psychologist argues that this phenomenon is not categorization at all. What kind of uncertainty is involved here?

Q.1.5. Give an argument for why happiness should be measured on a single scale, and an argument for why it should be measured on two separate scales (one 'positive' and the other 'negative').

2 Variables and Measurement

CONTENTS

Observational and measurement strategies 26

Basic concepts of measurement 29

Measurement validity and error 32

Types of measurement and varieties of data 37

Discrete and continuous variables 38

Levels and types of measurement 40

Missing data, absent people, and the fallacy of false precision 48

Questions and exercises 50

Observational and measurement strategies

Why should we concoct systematic strategies for observing or measuring anything? There are at least four compelling reasons. First, all of us have only very limited first-hand knowledge of anything about the world, including human existence. The vast majority of what we think we know or believe is based solely on second- and third-hand accounts by authorities such as parents, teachers, and the media. Often, that is the best we can do. Nevertheless, without first-hand experience, second- and third-hand accounts require us to make assumptions about their truth-status. Even our own experiences are sometimes of doubtful pedigree – all of us are potentially fallible observers and recorders, to say nothing of memorizers. Moreover, our first-hand experiences are not just haphazardly constrained, but systematically truncated by social conventions, matters of interpersonal attraction, and political and other instrumental agendas. To gain a wider experience of what a representative sample of people thinks, feels, or does is no small undertaking and requires far more time, strategic work, and resources than most of us can bring to bear.

A second reason is that systematic measurement strategies and instruments are ways of extending our senses. We cannot directly see electrical activity in the brain when someone is listening to music, but we can measure that activity at least indirectly by using sensing and imaging techniques beyond our own capacities. Likewise, we cannot hang around a temple for 500 years and count the number of worshippers, but we can arrive at the scene 500 years after the temple was built and measure the wear on the steps at its entrance. Finally, while we do not have immediate access to any animal's preferences for one food over another, we may infer those preferences by watching which foods the animal selects when it is given a choice.

A third reason for strategic measurement refers to the demands on our time and attention. We cannot pay attention to everything, and we do not have much time. Nevertheless, we may make some careful choices about what to pay attention to, how to set about it, and how much time to devote to it. Depending on the researcher's goals and theoretical orientations, some kinds of observation and measurement are more relevant, important, or probative than others.

Suppose you are investigating the mental processes involved in reading, and you are debating with your colleagues about whether the capacity to recognize the meaning of a word operates independently of the capacity to recognize how it should sound. Then you should be very interested in finding people who have suffered head traumas that have left one capability intact but not the other, since that would unambiguously support the separate capacity theory. Unfortunately for opponents of that theory, head-trauma victims who have lost both capabilities do not provide unarguable support for their position. Likewise, you should be more interested in native readers of languages such as Chinese or Japanese, where the symbols representing meaning are sometimes separate from those representing sound, than readers of languages such as English, where the same symbols do both kinds of work.

A fourth reason for strategic observation and measurement is overcoming or guarding against biases and hidden assumptions. At one time some people claimed that scientific measurement could be freed of bias and uninfluenced by values. These days, some people have gyrated to the opposite extreme of claiming that all measurement is inherently biased and driven by researchers' values, ideological orientations, and preconceptions. Neither of these positions is true, and both invite intellectual laziness born of complacency in the first case and nihilistic relativism in the second. A more viable position that also gives us something to work with is that at least some biases can be identified and many of those can be overcome, even though we must bear in mind that there probably is no infallible method for doing so. Two kinds of perennial biases are those that direct our attention toward certain phenomena and away from others, and those that compel us to explain certain events but not others. One thing that makes these biases important is that they can entrap nearly everyone regardless of their ideological orientation or even their motives for doing their research.

We shall turn first to attentional biases. In 1986, the American space-shuttle 'Challenger' exploded shortly after take-off, killing all on board. A key cause of this tragedy was the failure of 'O-rings' that held the booster's fuel tank to the rest of the rocket. A *post-hoc* investigation revealed that these O-rings tended to fail when the temperature outside the rocket fell below a certain level. Why didn't the highly trained engineers who designed and tested the rocket figure this out beforehand? It turned out that they had considered the possibility that the O-rings might be sensitive to temperature. They had even examined the relationship between the number of O-ring failures and temperature for all previous space-shuttle flights when one or more O-rings had failed. What they had omitted to do was check the temperatures involved with flights where *no* O-rings failed. Had they done so, they would have seen that all those flights had temperatures above a critical level.

This is an example of our predilection for being much better at detecting the presence of something than its absence. We are greatly inclined to see objects and not the spaces around them – indeed, a famous basic exercise for beginning artists is to learn to see and draw those spaces – and we, along with other animals, learn much better from a cue linked to the presence of something than a cue linked to an absence. Thus, a common mistake in early medical and clinical psychological research was to focus exclusively on the clinical cases instead of also studying people who did not have the clinical condition.

Let us also consider a related bias that befalls even trained researchers. Suppose each of the cards in Figure 2.1 has a number on one side and a letter on the other, and someone tells you 'If a card has a vowel on one side then it has an even number on the other side.' Which of the cards *must* you turn over in order to decide whether the person is right? Try deciding which cards these would be before reading on.

Now imagine that you are a forensic criminal psychologist specializing in serial murderers, and your experiences in the field have inspired a hypothesis: 'All serial murderers kill domestic pets before turning to killing people.' If you were going to investigate whether this hypothesis is true, which of the following kinds of people would be most important to incorporate in your study? Try rank-ordering them from first to fourth most important, before reading on.

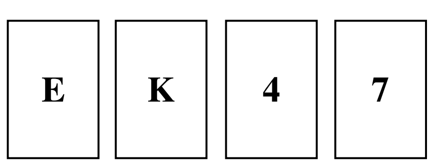

FIGURE 2.1 Card Task

- Serial murderers
- People who are not serial murderers
- People who have killed pets
- People who have not killed pets

If you are like most people, your response in the card task was to choose to turn over E and 4. Relatively few people give the logically correct answer, which is E and 7. The reason this answer is correct is that turning over the E card will disconfirm the rule if you find an odd number on its other side, and turning over the 7 will disconfirm the rule if you find a vowel on its other side. The other two cards are not relevant, since the rule states 'if vowel then even number' (and not, for instance, 'if even number then vowel'). The card task was first studied by Wason & Johnson-Laird in 1972 and is a classic example of the **confirmation bias**, namely the tendency for people to attend to instances that confirm their hypotheses or beliefs and ignore instances that might disconfirm them. Likewise, to rank pet-killers as more important to study than people who have not killed pets would be to fall prey to this bias. DEFINITION

Now let's briefly consider explanatory biases. Perhaps the most widespread is what might be called the **expectancy bias**, whereby we feel compelled to explain the unexpected but not the expected. Events that follow what we believe is the 'natural' or 'typical' order do not seem to require explanation, but 'unnatural' or 'atypical' ones do. Thus, we have seen many more studies of the causes of homosexuality than of heterosexuality, and one of the reasons is that a large and influential sector of Western societies perceives homosexuality to be 'unnatural'. Furthermore, we tend to think that atypical phenomena require special explanations over and above those for typical phenomena. Thus, from time to time people have proposed special theories of criminal behavior or of psychological disorders that are largely separate from theories of ordinary behavior or nonpathological psychology. The seductive nature of explanatory biases owes a great deal to the fact that if any phenomenon is merely what we expected it to be or what seems natural to us, then that is because we already have a theory about that phenomenon. However comfortable with our theory we might be, we should be prepared to subject it to as much rigorous testing as any other, and so we should be willing to investigate the ordinary, mundane, or expected along with their extraordinary counterparts. DEFINITION

Basic concepts of measurement

Whenever we measure anything, it is reasonable to distinguish between the record we make of this activity and the thing we are measuring. The record incorporates our measurements and any other relevant information; the thing

DEFINITION we are measuring often is called a 'construct.' A **construct** may be an abstract concept such as verbal latency, information search strategy, anxiety, or appetite; but it may also be something that is quite tangible, such as reaction time, the trajectory of eye movement, galvanic skin response, or the amount of food consumed in a day. Some constructs, moreover, may serve as **indicators** of other constructs so that by knowing about one construct we obtain information about another. Thus, we might use reaction time as an indicator of latency, eye movement to indicate something about information search strategy, galvanic skin response to indicate nervousness, or daily food consumption to indicate appetite. Indicators that involve recording something about the real world are often called **operationalizations** of their respective constructs, and if we are convinced that one or more indicators truly represent everything DEFINITION about a construct, then we claim that they **measure** that construct. These two terms are important because they involve strong claims on the part of a researcher. Anyone who says they have *operationalized* a construct is claiming that they have a way of observing that construct's real-world manifestations, and if they further say they can *measure* the construct then they are claiming access to the totality of its real-world manifestations.

DEFINITION A **variable** is an operationalization of a construct that can take on different values or states for different people (or even for the same person on different occasions). What is or is not a variable depends on the population being studied. If we are studying a group of 10-year-olds' reading ability, age is not a variable in that study. But if we are studying reading ability in children from an entire primary school, then age is a variable. Moreover, we must specify the conditions under which the construct can vary. Some constructs may change over time, or differ across people, or across situations for the same person.

As with most other aspects of research, deciding what will be treated as a variable is a matter of judgment and may be controversial. Gender, for example, is not a variable for most people during their lives, but it can be (transsexuals are those whose gender has varied at least once). In most, but not all, research on people it is sensible to treat gender as varying across people but not over time for the same person. However, a current issue in the psychology of the self is the extent to which an individual's personality or self-concept can vary throughout one's lifetime or even fleetingly from one situation to another.

All measurement strategies require the researcher to take at least some theoretical stances and risks, because in the absence of any theory the researcher has no idea what can or should be measured, let alone how to set about it. If we think that personality is fully formed and constant throughout adulthood, then the idea of measuring personality traits makes good sense. On the other hand, if we think that personality changes from one situation to another, the most we can hope for is to measure personality states at various points in time.

Even constructs themselves may be altered or redefined as theoretical frameworks develop. For instance, definitions of 'mental disorder' changed sufficiently throughout the 1970s that homosexuality was eliminated from the list of disorders in the more recent editions of the *Diagnostic and Statistical Manual* of the American Psychiatric Association.

It should not be surprising, then, that psychologists can differ dramatically on what is possible or appropriate to measure, and what the same measurements signify. Radical behaviorists, for instance, claim that we have no access to anyone's internal mental states or processes, so it is impossible to measure any such thing as a belief, attitude, or emotion. Their focus is exclusively on behavior and its causes, and unlike cognitive psychologists they do not infer cognitive states from behaviors. Likewise, where cognitively oriented researchers argue over whether various paper-and-pencil tests reveal people's underlying beliefs or attitudes.

The main point here is not simply that there is no such thing as theory-free measurement, although that is a popular view among psychologists and social scientists. Almost any measurement can be interpreted via many theories regardless of its origins, so that the same measurement may be given different meanings or evidential status by researchers using different theoretical perspectives. In fact, many controversies in psychology really are propelled by disputes over what certain data or measurements mean and imply. When conducting research or reading others', it is important not only to understand the researcher's theoretical standpoint on what is being measured, how, and why; but also how other theoretical perspectives would answer those questions and what uses they would make of the same measurements.

SUMMARY

The **confirmation bias** is a tendency to attend to instances that confirm prior hypotheses or beliefs and ignore instances that might disconfirm them.
The **expectancy bias** is the inclination to explain the unexpected but not the expected, or the unnatural but not the natural.

A **construct** is a concept, usually a characteristic or property, that underlies measurement.
A construct is an **indicator** of another construct if knowing about one provides information about the other.
Indicators are **operationalizations** of their respective constructs if they give us access to some aspect of their real-world manifestation.
If an operationalization captures the totality of a construct's real-world manifestation, then we claim that it **measures** that construct.
A **variable** is an operationalization of a construct that can change.

Measurement validity and error

Nearly every psychological perspective or theory incorporates stipulations of what are valid ways of measuring psychological phenomena and what constitutes measurement error. **Measurement error** occurs whenever measurements are influenced by something other than what the researcher intends to measure. **Validity**, on the other hand, is a general term denoting the extent to which measurement is not contaminated by error. Researchers who use randomized assignment in experimental studies or random sampling from populations distinguish between two kinds of measurement error. The first kind, **systematic error**, refers to influences on measurement that contain regularities and therefore bias the measurement outcomes. The second kind, **random error**, is not regular and therefore does not bias measurement outcomes, but nevertheless renders them less precise.

DEFINITIONS

DEFINITIONS

As an example of systematic error, suppose we have a test of verbal intelligence that is written in Russian. For those of us who do not read and write Russian fluently, this test will systematically underestimate our verbal intelligence. An example of random error would be a test of verbal intelligence in the test-taker's native language, comprising 25 questions randomly chosen from a large bank of such questions. By luck of the draw, some of the 25 questions will be easier than average and some will be more difficult, but the error in this case is distributed randomly throughout the test.

DEFINITION

A related term that is often used in psychological research is **reliability**, the extent to which measurement is free of random error. That is, a measure is reliable if it produces the same result every time under identical conditions. Although reliability and validity might seem similar at first, they are not synonymous at all. Our verbal intelligence test consisting of randomly chosen questions may be valid but it will not be perfectly reliable because of random error. On the other hand, we can have perfectly reliable measures that are nevertheless invalid. A verbal intelligence test in Russian administered to people who know no Russian at all will be very reliable, since each person will get a low score no matter how many times they take the test without learning more Russian in the meantime. No reasonable assessor would claim such a test is a valid measure of verbal intelligence for those people.

There is a large variety of systematic errors that psychologists have to contend with, and because validity is such a huge topic, they often subdivide it into several specific kinds of validity to be dealt with more or less separately. We will discuss errors first, and then return to the different kinds of validity. In their influential textbook, Rosenthal & Rosnow (1991) divide systematic errors into those arising in such a way as to not affect the respondent's responses (which they call 'noninteractional') and those that do ('interactional'). **Noninteractional errors** usually are caused by the researcher or

DEFINITIONS