

EXPERIMENTAL
DESIGN
FOR BIOLOGISTS

ALSO FROM COLD SPRING HARBOR LABORATORY PRESS

An Illustrated Chinese-English Guide for Biomedical Scientists

At the Bench: A Laboratory Navigator

At the Helm: A Laboratory Navigator

Lab Dynamics: Management Skills for Scientists

Lab Math: A Handbook of Measurements, Calculations, and Other Quantitative Skills for Use at the Bench

Lab Ref: A Handbook of Recipes, Reagents, and Other Reference Tools for Use at the Bench

Lab Ref Volume 2: A Handbook of Recipes, Reagents, and Other Reference Tools for Use at the Bench

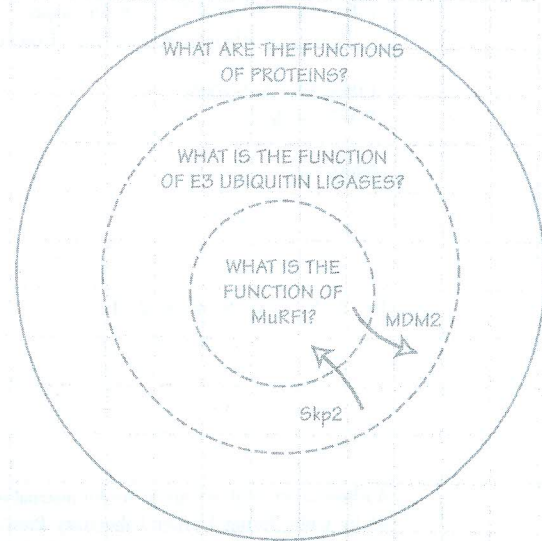
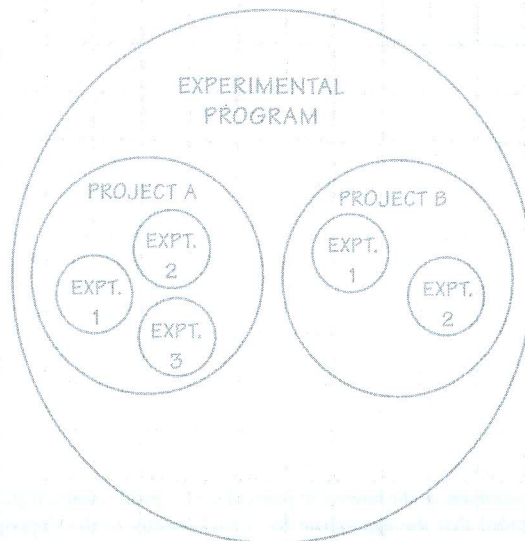
Laboratory Research Notebook

Safety Sense: A Laboratory Guide

EXPERIMENTAL DESIGN FOR BIOLOGISTS

David J. Glass

Novartis Institutes for Biomedical Research



COLD SPRING HARBOR LABORATORY PRESS
Cold Spring Harbor, New York

Experimental Design for Biologists

All rights reserved
© 2007 by Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York
Printed in the United States of America

Publisher	John Inglis
Acquisition Editor	David Crotty
Development Director	Jan Argentine
Developmental Editors	Siân Curtis, Maria Smit, and Virginia Peschke
Project Coordinator	Inez Sialiano
Permissions Coordinator	Maria Fairchild
Production Editor	Rena Steuer
Desktop Editor	Compset Inc. and Susan Schaefer
Production Manager	Denise Weiss
Cover Designer	Lewis Agrell

Photograph of Karl Popper ©Lucinda Douglas-Menzies/National Portrait Gallery, London.
Photographs of Francis Bacon and Rene Descartes are reproduced, with permission, from the Collections of the Library of Congress.

Library of Congress Cataloging-in-Publication Data

Glass, David J.
Experimental design for biologists / David J. Glass.
p. cm.
Includes bibliographical references and index.
ISBN-13: 978-0-87969-735-8 (alk. paper)
1. Biology--Mathematical models. 2. Experimental design. I. Title.

QH323.5.G565 2006
570.72'4--dc22

2006023665

10 9 8 7 6 5 4 3 2 1

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Cold Spring Harbor Laboratory Press, provided that the appropriate fee is paid directly to the Copyright Clearance Center (CCC). Write or call CCC at 222 Rosewood Drive, Danvers, MA 01923 (978-750-8400) for information about fees and regulations. Prior to photocopying items for educational classroom use, contact CCC at the above address. Additional information on CCC can be obtained at CCC Online at <http://www.copyright.com/>.

All Cold Spring Harbor Laboratory Press publications may be ordered directly from Cold Spring Harbor Laboratory Press, 500 Sunnyside Blvd., Woodbury, New York 11797-2924. Phone: 1-800-843-4388 in Continental U.S. and Canada. All other locations: (516) 422-4100. FAX: (516) 422-4097. E-mail: cshpress@cshl.edu. For a complete catalog of all Cold Spring Harbor Laboratory Press publications, visit our World Wide Web site <http://www.cshlpress.com/>.

*For my nieces
Molly Jane and Madeline Rose
who provide a reason to ensure that the
future will yield great discoveries
and to my teachers
Charles Cantor, Lex van der Ploeg, Arg Efstratiadis,
Steve Goff, and George Yancopoulos
with gratitude and the hope that a little has rubbed off*

ant-
pyright
r infor-
the

oratory
and
cata-
s.com/.



Contents

Preface	ix
1 Defining the Experimental Program	1
2 The Hypothesis as a Framework for Scientific Projects: Is Critical Rationalism Critical?	7
3 Scientific Settings in Which a Hypothesis Is Not Practical	17
4 The Problem/Question as a Framework for Scientific Projects: An Invitation for Inductive Reasoning	21
5 What Constitutes an Acceptable Answer to an Experimental Question? ...	33
6 How Experimental Conclusions Are Used to Represent Reality: Model Building	35
7 Establishing a System for Experimentation	51
8 Designing the Experiment: Definitions, Time Courses, and Experimental Repeats	57
9 Validating a Model: The Ability to Predict the Future	67
10 Designing the Experimental Project: A Biological Example	75
11 Experimental Repetition: The Process of Acquiring Data to Model Future Outcomes	103
12 The Requirement for the Negative Control	117
13 The Requirement for the Positive Control	133
14 Method and Reagent Controls	149
15 Subject Controls	157
16 Assumption Controls	171

17	Experimentalist Controls: Establishing a Claim to an Objective Perspective	183
18	A Description of Biological Empiricism	191
19	A Short Synopsis	199
	Index	201

183
91
199
201



Preface

THIS BOOK IS A PRODUCT OF A SHORT COURSE on Experimental Design that I developed while still working at Regeneron Pharmaceuticals and I would like to express my thanks to all who gave feedback on that course. Like many ideas, the decision to turn the course into a book came about after a little too much wine with dinner—a dinner with fellow scientists Dan Finley, Fred Goldberg, and Allan Weissman—at which we were discussing both the odd fact that experimental design was not commonly taught to prospective biologists in graduate school, and the obvious discontinuities between the demands of Critical Rationalism as written and the way that science was actually practiced.

Of course, this book still would not have been produced were it not accepted by a publisher. Therefore, I am extremely grateful to David Crotty at Cold Spring Harbor Laboratory Press for agreeing to take the project on behalf of the Press. Siân Curtis then edited the manuscript in an extremely able fashion, with the help of Ginger Peschke and Maria Smit. The project was overseen by Jan Argentine and her colleagues at the Press. Thanks very much to Siân and Jan for their steady feedback and enthusiasm. Thanks so much also to Rena Steuer for expert production guidance, to Susan Schaefer for typesetting, and to Denise Weiss for her design expertise.

A great deal of help in writing this was lent by Kumar Dharmarajan. Kumar is a former student and intern in my laboratory, who was a medical student at Columbia when much of the book was written. He was thus able to take on the role of the “prospective audience” and gave invaluable feedback on each chapter, identifying passages that were unclear and asking questions that helped in the rewriting. Thanks very much to him for spending so much time on this project. Brian Clarke in my lab also read a large chunk of the manuscript and highlighted some sections that needed clarification.

Woody Fu is a former student who now has many jobs, including artist and cartoonist. He provided three cartoons for this book, and they came out so well that I am sorry we did not make greater use of his talents.

Thank you very much to the folks at Novartis, where I now work, for supporting this project. Novartis places a great emphasis on continuing education, and it was a

comfort to find that this project invoked enthusiasm in the company. Special thanks to statisticians Leah Martell and Mathis Thoma for taking the time to read a large portion of the text for potential statistical misstatements. However, having said that, if there are any remaining problems, they are my responsibility alone. It would be great to hear from readers, including feedback regarding problems with the text so that these could be corrected in future editions. My e-mail address is expdesignbiology@gmail.com/. Feel free to drop me a line.

DAVID J. GLASS



Defining the Experimental Program

AT MEDICAL SCHOOL, STUDENTS ARE TRAINED in the basics of the medical craft: how to perform a physical examination, interpret symptoms, approach a patient, and perform a differential diagnosis. At law school, the initial course load is designed to steep the novice in the forensic approach: how to write a brief, follow legal procedures and prepare contracts, how lawyers are expected to conduct themselves in various situations, and how to think like a lawyer.

Now here's an amazing fact: At this writing, in graduate schools throughout the United States, budding Ph.D. candidates receive almost no formal instruction in experimental design, despite the fact that the effective design, conduct, and interpretation of experiments is critical to their success as scientists. A quick survey of the curricula for Ph.D. students in biology, biochemistry, and related disciplines, in even the most established universities, reveals courses devoted almost entirely to factual information. There is no emphasis on the formal experimental processes required to unmask the previously unknown, and no discussion of how these processes derive from various—and sometimes dissenting—theories of epistemology.¹ Currently, the one notable exception to this lack of procedural education is training in statistics, although such training is sporadic and varies greatly depending on the scientific discipline. However, although an understanding of statistics can be tremendously helpful in the development of effective experimental design, both apprentices and more experienced scientists would benefit from a broader understanding of different experimental frameworks and their implications.

But isn't a formal procedural education dispensable for a practicing scientist (who is, after all, more concerned with empirical data than with philosophical theory)? Is it not obvious what is required to perform an experiment? Isn't the scientific method quite well established, developed over hundreds of years of philosophical reasoning, and distilled into consistent scientific practice? The answer to all of these questions is "no."

¹Epistemology is the study of the grounds of knowledge, with particular reference to its limits and validity; scientific epistemology is the subject of this book, addressed on a practical level so as to be useful to the working scientist.

The purpose of this book is to alert the scientist that the way in which a question is framed can affect the choice of experiment and the understanding of its outcome; therefore, a methodology of experimentation is presented to guide the scientist toward informative results.

The idea that philosophical frameworks can make the difference between the success and failure of an experimental approach to discover anything useful may be met with some resistance. In their formal education, many scientists purposefully shy away from “goosey” philosophy in favor of hard science, thinking that, because they are gaining empirical results that are anchored in physical reality, and that “the data are the data,” it is impossible to alter a result by changing a mind-set or by choosing a different experimental approach. The devil, though, is not in the experimental results themselves, but rather in the choice of which results the scientist deems relevant and the subsequent interpretation of those results. That is why the experimentalist must establish a firm framework for the experiment and have a clear understanding of where and when unwarranted assumptions may bias data and cause misinterpretation of results.

Let's return to the question of the “well-established” scientific method. Currently, the dominant scientific approach, in which experiments begin with a written statement of fact or “hypothesis,” is actually less than 100 years old and can be traced directly to the theories of the philosopher Karl Popper (Fig. 1). Popper's major idea (called “Critical Rationalism”) was that, although it is very hard to prove a notion true, it is quite simple to prove a statement false. He thus advocated framing experiments with a falsifiable hypothesis. Furthermore, he believed that the use of the hypothesis would be helpful in steering away from a perceived problem of inductive reasoning, or reasoning from experience. The problem with experience, according to Popper (and philosophers such as Immanuel Kant, David Hume, and Bertrand Russell), was that personal experience was limited. Therefore, reasoning from a limited data set of

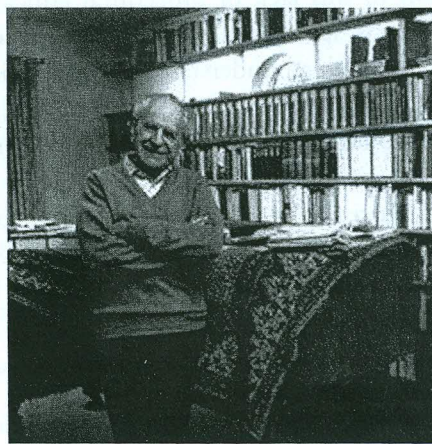


Figure 1. Karl Popper.

personal experience, and making laws or conclusions from that data, would lead to errors. In addition, according to Hume, inductive reasoning, as when one applies past experience to the future, is flawed. So, instead of trying to prove a notion true, which might take many repetitions and multiple modes of analysis, Popper thought it better to try to prove the notion false. In this approach, a single demonstration that a statement was false would be sufficient to disprove the idea. (These few sentences summarize several volumes and hundreds of pages of Popper's writings. So, yes, they are a simplification. As will be repeated several times, in various ways throughout the chapters, this book is geared to getting the scientist to the bench as soon as possible, armed with the necessary critical framework to do useful experiments. The relevant philosophy will be touched on only to the degree necessary to help the scientist to the bench and to provide the interested individual with the hints necessary to research the relevant references.)²

Popper's ideas have become so ingrained that the National Institutes of Health (NIH) currently recommend that all grant requests begin with a testable hypothesis, although, interestingly, the NIH seems to believe it all right to try to prove a hypothesis true, in direct contradiction to Popper's theory.³ That is quite an astonishing measure of success for a single substrain of philosophical thought, and it seems to imply that the hypothesis is the only acceptable way in which to conduct scientific inquiry. However, that is not the case. The practical effects of Popper's falsifiable hypothesis idea are discussed further in this volume in great detail, along with potential alternatives.

Karl Popper was not the first philosopher of science; alternate approaches to scientific experimentation have been proposed. Most historians seem to agree that the modern scientific method is founded on two different approaches, one proposed by Francis Bacon (Fig. 2) and the other by René Descartes (Fig. 3). Although it is beyond the scope of this book to review these writings extensively, the reader should be aware that it was Bacon and Descartes who argued that useful scientific knowledge can be gained from a series of inductive (Bacon) and deductive (Descartes) approaches. (This is, of course, another gross simplification, but be assured that both Bacon's and Descartes' arguments filter through most of this book—indeed, in many ways, the arguments made here are a return to the middle ground between Bacon and Descartes, in part as a corrective step away from Popper.)

Thus, it seems that current approaches to the scientific method are not necessarily based on firmly fixed philosophical ground and that therefore, some reassessment may be useful. We must take a step away from the laboratory to regard the world of science from the perspective of an observer of the process of experimentation. What do we

²But, just in case you need a single reference, try *The logic of scientific discovery*, by Karl Popper, 2002. Routledge, New York.

³"Many top-notch NIH grant applications are driven by strong hypotheses rather than advances in technology. Think of your hypothesis as the foundation of your application—the conceptual underpinning on which the entire structure rests. Generally applications should ask questions that *prove* or *disprove* a hypothesis rather than use a method to search for a problem or simply collect information." (Taken from the NIH Web site on grant applications [http://www.niaid.nih.gov/ncn/grants/plan/plan_c1.htm]).



Sir Francis Bacon

Figure 2. Francis Bacon.



Figure 3. René Descartes.

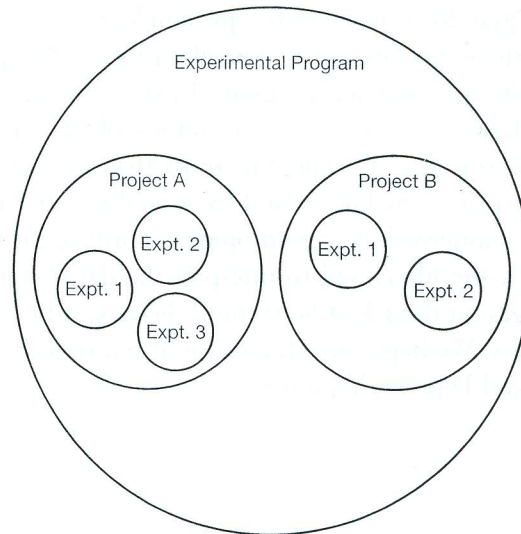


Figure 4. Venn diagram illustrating the “experimental program,” which encompasses all that the scientist does to understand a previous unknown. It may involve learning about science, training, reading about the subject in general, and performing actual experiments. The scientist’s “project” is the experimental effort to answer a particular series of questions aimed at a particular unknown. Experiments (in the example, project A includes expt. 1, expt. 2, and expt. 3; project B includes expt. 1 and expt. 2.) are then performed in an attempt to either answer the question or falsify the hypothesis.

see? What is the first step toward discovering something that is currently unknown? Let’s call this effort the “experimental program,” as illustrated by the Venn diagram in Figure 4. The experimental program encompasses all that the scientist does to establish as fact a previous unknown. It is the universe in which the scientist operates,

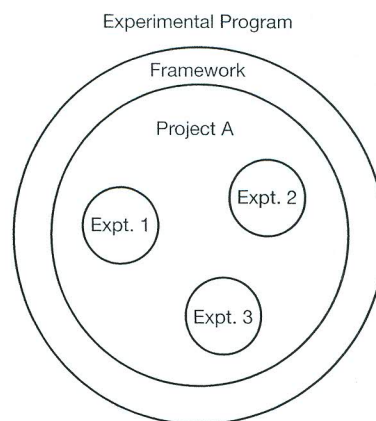


Figure 5. Venn diagram illustrating the experimental framework within the experimental program. The scientist’s “project” may be framed by a hypothesis to be falsified or a question to be answered. The framework is critical in determining how the scientist approaches the project and how the data resulting from individual experiments are interpreted.

or, to put it more grandly, it is scientific epistemology.

When a scientist is working on a particular project, that project may be thought of as a subset of the experimental program. In this context, a project is a particular scientific inquiry designed to understand a circumscribed issue. Many scientists might be unaware that their project is limited by more than the field of inquiry (the apparent border of each project in Fig. 5) and by something that may be thought of as a "framework." The framework is a philosophical construct; it is the approach used to address the project, the filter used to interpret the data from each experiment. The framework impinges on the scientific method, because different frameworks demand different approaches. We begin our discussion of frameworks in Chapter 2, with the current favorite, Karl Popper's hypothesis.



B
for
h
bod
hy
o
that
with

Con
es
m
exam
w
ob
ally a
so e
pr
usual
N
m
unpr
a v



The Hypothesis as a Framework for Scientific Projects

Is Critical Rationalism Critical?

BEFORE DISCUSSING THE HYPOTHESIS (and thus Critical Rationalism) as a framework for experimentation in detail, it must be recognized that the practicing scientist may have no choice but to offer a hypothesis. As discussed in Chapter 1, certain funding bodies (and reviewers, and even some scientific journals) require the articulation of a hypothesis as a justification for experimentation. However, by recognizing the pitfalls of the Critical Rationalism framework and understanding the alternatives, it is hoped that the scientist will be able to apply a properly critical approach, even if working within this paradigm.

THE HYPOTHESIS HAS THE SAME GRAMMATICAL STRUCTURE AS THE CONCLUSION

Consider the following hypothesis: The sky is red. First, note the structure of the hypothesis. It is a declarative statement. Because the hypothesis is required to be falsifiable, it must take the form of a declarative statement, which can be falsified. A question, for example, would not be falsifiable, and therefore could not function as a framework within Critical Rationalism. If the hypothesis is not falsified upon experimentation, the obvious conclusion might be that the hypothesis has been proved correct. This is actually an error against which Karl Popper warned. He argued that it is impossible to prove something definitively true. Yet the simple fact is that to proceed with an experimental program, conclusions must be made and therefore the failure to falsify an experiment usually results in a conclusion. In this case, the conclusion is "The sky is red."

Note that the conclusion is identical in structure to the hypothesis. This makes it more difficult for the scientist to distinguish between the statement operating as an unproven hypothesis (to be tested for falsification) and the statement functioning as a verified conclusion.

**EXPERIMENTS ARE AIMED AT CONFIRMATION, NOT FALSIFICATION,
BECAUSE THE HYPOTHESIS INSTILLS A REQUIREMENT TO
MEASURE THE "POSITIVE" RESULT**

Certain experimental norms have crept into the standard method. At first glance, these appear to be helpful. One example is the use of scientific controls to provide a built-in mechanism by which to deny a hypothesis.

If we return to our hypothesis that the sky is red, it might appear that this statement could be confirmed simply by measuring "red." Yet even the most inexperienced of investigators knows that a negative control is required that will include a means of measuring the absence of red or "not red." Therefore, we can measure red and not red, but to test our hypothesis, should we measure anything else? There is no obvious need for this; certainly nothing in the hypothesis demands a measurement of anything besides red and not red.

Here is an experimental design to test the hypothesis that "the sky is red": First, the scientist sets up the system and proves that red can be measured accurately. An object known to be red, such as a rose, is used as a positive control. A second object that is known to be not red, such as a green emerald, is used as a negative control. A device that we'll call the "red-ometer" measures radiant light, is thus calibrated, and is now capable of two readouts: red and not red. If red is detected, a large "confirmation" flag appears.

Having prepared the experimental system, the scientist then points the redometer to the sky and leaves it running for 1 hour. After 1 hour, nothing has happened: The flag has not been raised and therefore a negative result is recorded. To ensure that "red" was given a chance, the time frame for the experiment is increased to 4 hours. But, because the experiment was started at noon, sunset does not occur within the time frame of the experiment and another negative result is recorded. The scientist then reasons that the machine may be insufficiently sensitive and also notes that a day is 24 hours long. So, the experiment is repeated over a 24-hour period. After a full day of measurements, the scientist returns to the redometer, and sees that the "confirmation" flag has been raised (because the red light of dawn and sunset occurred during the experimental time frame). The experiment is declared a success because a positive result has been achieved.

The scientist repeats the experiment every day for a month and notes that the color red has been measured on every single day except for one (a particularly cloudy day). The scientist calculates statistics¹ on the number of times red was detected (29 out of 30) versus the number of times not red was detected (0 out of 30, because the sky is never emerald green) and finds that red detection reached statistical significance. It is therefore concluded that the sky is red.

¹A detailed discussion of statistics is beyond the scope of this book. Although the argument is that even the best statistics cannot rescue a poor framework, it should be noted that, when correctly applied, statistics can be your friend and can stop you from making inappropriate assumptions or changing your parameters in midstream.

primer → positive experiments
do positive results

You may object to the above narrative as an elaborate straw man. But is that (at least in part) because you know that the sky is usually blue (during the day) or black (at night, if you don't count the stars, the moon, or radiation from the planets), and you therefore realize that a plethora of additional controls would have helped to establish appropriate caveats to the sky is red conclusion? This is exactly the purpose of the scenario—to set up a discussion of those additional controls. The scenario also demonstrates that the hypothesis framework may obscure the need for those controls, because of the circumscribed nature of the operative declarative statement. Remember that in most circumstances, the scientist does not know the “real” answer and therefore the only actual guide for what needs to be tested is the hypothesis. So although it may be obvious that, in the sky is red case, the scientist should really measure the full spectrum of light to see the bigger picture, we might be less sure of what to test if the hypothesis was “activation of the NF- κ B² pathway results in inflammation.” Without prior knowledge of the NF- κ B pathway, we have no indication of whether the hypothesis is correct, even though equivalent issues exist, as in the sky is red scenario. For example, the data that might be obtained using this second hypothesis will vary dramatically, depending on the cell type used (activation of NF- κ B in the liver leads to inflammation, whereas activation in skeletal muscle does not). The results will also differ depending on the criteria used to define inflammation. The hypothesis as written, however, does not demand experiments that will delineate the tissue-specific responses or even a firm definition of inflammation. Therefore, by analogy with the scientist who extended the time frame of the experiment until a positive result was obtained, the scientist testing “activation of the NF- κ B pathway results in inflammation” may feel it appropriate to keep looking for a “sensitive” system, that is, a system that produces “positive” data. Also, just as in the sky is red scenario, where the simple fact of the matter is that the sky is sometimes red, in the second case, the simple fact is that NF- κ B seems to induce inflammation in multiple cell types. The point here is to highlight the different emphasis placed on data obtained from positive instances and on experimental settings that yield negative results. In the case of the second hypothesis, because NF- κ B responses are cell and situation specific, a filter that scores only positive (inflammation) results will provide a picture as incomplete as the experiment that looks at the sky and records only red results.

A second point should be made about the NF- κ B hypothesis. The criteria for a positive readout, that is, what constitutes inflammation, might become looser if negative data are achieved. For example, if there are five markers³ of inflammation, yet

²NF- κ B is the name of a protein. The fact that this protein is probably unknown to most readers of this book is precisely the point: The hypothesis is relied upon much more strongly when dealing with an unknown, and thus assumes greater stature. Here, the scientist lacks the outside knowledge that helped to expose the problems with the sky is red hypothesis. For those who are curious, the NF- κ B protein functions as a “transcription factor,” meaning that it induces a particular set of genes to be transcribed into messenger RNA.

³In biological terms, a marker is a diagnostic criterion. For example, pain in the right lower abdominal quadrant is a marker of appendicitis. A high serum glucose level is a marker of diabetes. Elevated levels of the protein IGF1 are a marker of increases in growth hormone.

only two are activated by NF- κ B, positive data exist only for the two markers that are activated. For the other markers, the question of whether the test was correctly designed remains, and the scientist may feel justified in reporting only the positive data. In this case, statistics can come to the rescue. Almost every proper statistical analysis forces the experimentalist to define, in advance, the criteria that indicate a positive effect. However, if the hypothesis really is influencing the interpretation of what are and what are not positive data, the reliability of the three negative markers as indicators of the effect under investigation may now be questioned.

We must stop here to highlight this last point. If an experiment generates some positive and some negative data, the process of weighing these readouts differently may actually cause the scientist to redefine the issue being studied. That is a dangerous pitfall, but it can be avoided simply by establishing success criteria before experimentation (see Chapter 12). We now discuss why this filtering activity may be more likely to occur within the Critical Rationalist framework.

The Hypothesis Establishes a Binary Distinction between Data Types, Positive and Negative, that May Be Used as a Filter

Returning to “the sky is red” hypothesis, it is clear that in designing the redometer, the scientist established a data filter. The measurement system is capable of recording only red and one other color, whose only relevant characteristic is that it is not red. Lest you think that the deck has been stacked in the example by choosing green as the negative control instead of black, realize that even if the experiment had been performed using black as the negative control, the observation that the sky is black may still not have been interpreted as a positive result. The critical point here is that the scientist must set up a special positive data flag for a red readout. A filter set in place for the result to be perceived as positive is a very common feature of experiments. Even though black may be measured for the entire evening, the machine uses black only as a control against red. Therefore, the majority case of a negative finding (the sky is black) is inadvertently discarded in favor of the minority case of a positive finding.

A real-world example of a positive data filter can be illustrated by considering the hypothesis that caffeine causes cancer. First, an epidemiological study may be performed to determine whether individuals who drink coffee, tea, or cola drinks (all of which are caffeinated) have a greater incidence of a particular type of cancer than individuals who did not ingest caffeine. Let’s say that 100 different types of cancer are surveyed. Even if a statistical difference is found in only one instance, for example, pancreatic cancer, that one cancer type will be read as positive, because that was the case in which a statistically significant difference was determined. Data from the other 99 types of cancer will be labeled as negative. Thus, the scientist might conclude that the hypothesis that caffeine causes cancer was proven, at least in the case of pancreatic cancer. Or, more carefully, the conclusion might be that drinking caffeine increases the chance of an individual developing pancreatic cancer, but not other types of can-

cer. This, by the way, is another instance in which a thorough knowledge of statistics must be brought to bear. A good statistician would know that any conclusion at all from these data is inappropriate, because the study performed is really a survey, and no firm hypothesis is in place for each cancer type. Unfortunately, inappropriate conclusions are routinely made from such surveys because, for example, the structure of the hypothesis “caffeine causes cancer” established the binary positive/negative distinction between achieving a statically significant correlation and failing to achieve such a correlation. Because a positive result was achieved in only one out of the 100 cancers tested, that one cancer is highlighted, and the headline reads that caffeine may increase the risk of pancreatic cancer, not that caffeine fails to cause 99 types of cancer.

But what if caffeine does, in fact, increase the risk of pancreatic cancer, but does not increase the risk of the other types of cancer? That is certainly possible, and that is what one might conclude from the above study. However, the scientist must avoid making such a conclusion until a second, follow-up study has been performed. If the second study is also done in the Critical Rationalist framework, the hypothesis would be that caffeine causes pancreatic cancer. In the first setting, pancreatic cancer was found to be statistically significant in a broad survey of 100 cancer types. Although I will avoid a detailed discussion, a tiny flavor of statistical analysis may be helpful here. If the appropriate statistical test of data from the first study⁴ resulted in a probability or “*p*” value of 0.05 (the usual statistical threshold for significance), there is a 5% chance of a random finding of significance. Thus, in a survey of 100 tumor types, it would not be surprising if one individual cancer scored as a random positive; indeed, in the 100 cancers surveyed, one might expect to find five random positives. However, if the second study is focused on a single type of cancer, the chance of that one type of cancer scoring positive at random is again only 5%. The chance of the same cancer type randomly scoring positive in two distinct studies is 0.05×0.05 , or 0.25%, meaning that a single repetition of the experiment can dramatically increase confidence in the finding. (Replication of experiments is discussed in detail in Chapter 8.) The point to highlight here is that the positive filter from the broad hypothesis resulted in a focus on one cancer type out of 100, rather than the more likely conclusion that caffeine does not increase the risk of cancer. To believe the conclusion that caffeine increases the risk of pancreatic cancer, one must also conclude that something in the pathways that leads to pancreatic cancer is uniquely sensitive to caffeine. But, stranger things have happened than the finding that a particular carcinogen results in a distinct type of cancer (e.g., asbestos leads to mesothelioma). Thus, a rejection of the positive finding is also inappropriate. It just needs to be treated appropriately, and any conclusion based on the positive filter should be deferred until a confirmatory study that is based on a much more narrow and explicit hypothesis is performed. In the confirmatory study, the hypothesis must seek to falsify the

⁴Probably, an ANOVA analysis.

particular case that initially yielded a positive result. Critical Rationalism can be successful if it is followed rigorously.

There are, of course, different approaches to the "caffeine causes cancer" hypothesis. In a second experimental design, instead of an epidemiologic study, the scientist may decide to perform a pharmacologic experiment and establish a paradigm in which rats are fed increasing amounts of caffeine and studied for 1 year. At concentrations that are equivalent to those ingested by even the most dedicated coffee-drinker, the rats fail to develop any more tumors than their caffeine-free littermates.

Once again, the experimenter is faced with a quandary: Should the negative data be accepted or should the experiment be taken further? The scientist may decide to proceed, reasoning that it usually takes years of exposure to environmental insults to cause cancer and therefore, to achieve the desired effect,⁵ the dose must be increased. So the rats are given 100 times the normal daily intake of caffeine, every day for 1 year. In addition to becoming extremely edgy, these caffeine-treated animals do indeed develop tumors at a higher rate than the untreated controls, and this higher incidence turns out to be statistically significant. The scientist may conclude that caffeine can induce cancer and argues that the higher doses used were necessary to model the long-term exposure caused by decades of caffeine consumption in humans.

Again, it is certainly possible that both of these conclusions (that caffeine causes cancer and that the 100-fold increase in dose accurately models a longer-term dosage at normal levels) may be proven correct. But are these valid conclusions from the data obtained here? And did the hypothesis framework of the experiment push the scientist to increase the dose until a positive result was achieved?

It is also possible that caffeine at normal doses does not cause cancer, whereas caffeine at much higher doses, e.g., 100 times the normal dose, does trigger tumorigenic processes. This scenario introduces Chapter 16, which discusses designing accurate models. Sometimes it is impossible to have an experimental setting that mimics the behavior of a gene product at physiologic levels, and sometimes, in contrast, it is indeed possible to produce extremely useful data by dramatically increasing a dosage to see what the agent being tested can do. But these steps should not be performed because of a predetermined "need" to prove an unproven hypothesis. The experiment must accommodate those processes that cause cancer at normal environmental concentrations, and it may therefore be unhelpful to have filters in place that seek to establish causality based on a predetermined declarative statement, for example, that a particular mechanism is causative. In other words, if we conclude that caffeine causes pancreatic cancer, the hypothesis framework does not seem to require a further search for more likely causes of pancreatic cancer. On the contrary, it seems to remove such a search from the discussion (why look for other causes when the hypothesis is already proven?).

The following is a second real-world example of how a hypothesis can establish an inappropriate filtering system. Let's use the hypothesis that activation of enzyme X

⁵Note the phraseology "desired effect." Does a scientist have any business desiring one effect over another?

caffeine (normal dose) has no effect
→ increase tumor
caffeine (100x use) 1 goden u
→ human
over exposure
200

normal - pharmacologic dose
100x use

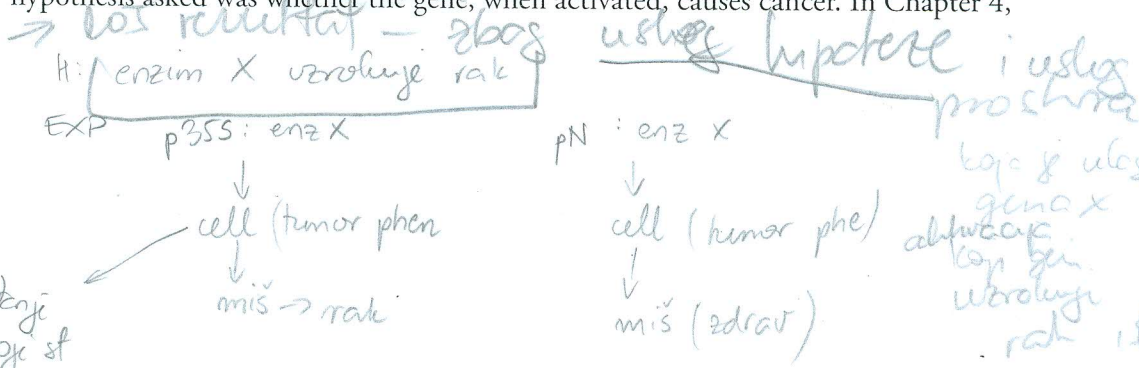
an be
 esis.
 t may
 h rats
 that
 ts fail
 data
 ide to
 lts to
 sed.
 for I
 indeed
 ence
 te can
 long-
 causes
 losage
 data
 entist
 reas
 igenic
 curate
 the
 indeed
 what
 of a
 com-
 s, and
 ased
 sm is
 r, the
 uses
 n the
 lish
 me X
 ther?

causes cancer, which has the virtue of being based on a previous finding. Suppose that a scientist cloned an activated form of enzyme X from a colon tumor. In testing the hypothesis, the scientist made a genetic construct, encoding a mutant form of enzyme X that is always "on" (a constitutively active mutant). As a negative control, the scientist made a second genetic construct, encoding the normal form of enzyme X. The vectors encoding both the mutant enzyme X and the normal enzyme X were inserted into cells. The cells that received the constitutively active mutant did indeed produce a cancerous phenotype and, when injected into mice, caused tumors. The cells that received the normal enzyme X did not become cancerous, nor did they cause tumors in mice. The experiment was repeated ten times, always with the same result. This all sounds good, and I must admit that if these data were produced in my lab, I would probably be pretty excited, at least at first. But let's be a little more critical.

konsh
 kontrol
 super

What if I told you that the cells used for the experiment don't normally express enzyme X? What if I went further and told you that the cells that normally express enzyme X react differently when they receive the constitutively active mutant? This second cell type reacts by differentiating into neurons and does not develop tumors. Now where is our hypothesis that enzyme X causes cancer? Perhaps now we don't believe that the hypothesis is correct. But, wait again. What if we performed even more experiments and discovered that colon cells, which don't normally express the gene encoding enzyme X, start to express the enzyme inappropriately under certain conditions, and under such conditions, enzyme X does indeed induce cancer in these colon cells? We might now propose a new hypothesis that enzyme X causes colon cancer. Can we be comfortable with this more circumspect hypothesis? After all, the gene was originally cloned from a colon tumor. But, let's say that 100 different colon tumors are surveyed and enzyme X is not found to be activated in any of them. Therefore, even though enzyme X, when mutated, is capable of causing colon cancer, it does not seem to be a very common mediator of the disease. Finally, a new discovery is made: In brain tumors, when cells express enzyme X, it is more likely that the tumors will be cured by differentiating them into normal neurons, a transformation that is achieved by treating the tumors with a second protein that activates enzyme X. Thus, we've gone from a hypothesis that focused the scientist on the idea that enzyme X induces cancer, to findings that showed that stimulation of enzyme X may actually block cancer. Was the hypothesis-formulating process a help or a hindrance in this experimental program? The hypothesis did force experiments that were geared to test whether the enzyme could induce tumors. However, nothing in the hypothesis demands an analysis of whether the tumorigenic cells would normally express the gene product. Nor does the hypothesis require the scientist to determine in which cells the gene is normally active, nor how its activation affects those cells. Finally, the hypothesis does not require a survey of naturally occurring tumors, to determine whether the enzyme is activated in those cells. The only question that the hypothesis asked was whether the gene, when activated, causes cancer. In Chapter 4,

redovstava
 eksp
 alobi rene
 proshkanya



- 1) Što ako normalis uspore ne ekspm raji X
- 2) Što ako znamo da u stanicama koje ekspm raji X nando nastoji neuronj alobi magj p355

it is discussed whether alternate frameworks are more helpful in demanding effective experimental designs.⁶

The above scenarios show that the hypothesis, by its nature, may establish a positive/negative filtering system, through which relevant data may inadvertently be discarded as “not positive” and through which positive data may be inappropriately weighted.

THE SCIENTIST IS NOT REWARDED BY DISPROVING A HYPOTHESIS, BUT IS REWARDED BY “CONFIRMING” A HYPOTHESIS

The above scenarios identified instances in which an experimental setting might be tweaked until a positive result is achieved. In imagining such a scenario, one need not impugn an individual’s integrity or accuse them of ill intent; the change in standards for the readout might occur because a negative result leaves the scientist with the nagging issue of whether the experiment was correctly designed. In other words, the scientist may have a built-in incentive to keep changing experimental components until a positive result is achieved. If this incentive is perceived, it must be guarded against, and raising this possibility will perhaps alert the scientist to the issue. Remember that a good knowledge of statistics can be helpful, because most statistical frameworks require the scientist to determine the criteria for a positive result in advance of such studies.

Another set of factors may encourage an individual to chase “positive” data. Because we are dealing with science, and therefore empiricism, one might like to imagine that human motivations and psychological influences are easily dissectible from the experimental program. However, to control against this “human element” in science, several obvious motivating factors need to be placed on the table. Here is a fact: Scientists are rewarded far more for confirming a hypothesis than for disproving it (even though a hypothesis is not supposed to be confirmed). High-profile (and even low-profile) journals generally do not publish negative data, and indeed, why should they? Who wants to read about an idea that turned out to be incorrect? Furthermore, why should funding bodies reward someone for putting forth a hypothesis—even a very clever one—that turned out to be wrong? And why should a university’s department chair grant tenure to an investigator who has not been “successful” in finding evidence to support a hypothesis?

It may be surprising to the novice scientist to realize how far-reaching the hypothesis framework is. Once set up as the norm by which an experiment is judged,

⁶By the way, enzyme X actually exists. The receptor tyrosine kinase TrkA was cloned out of a colon tumor, where it was found to be activated via a mutation. It is not obvious, however, that TrkA is a common mediator of colon cancer. When overexpressed in fibroblasts, TrkA can transform them into a cancer-like phenotype. However, when activated in a neurogenic PC12 cell, TrkA induces differentiation into neurons. Thus, although this story is complicated, it illustrates the point that a simple hypothesis may hinder rather than help uncovering such a picture, because it establishes an unhelpful positive/negative framework.

it has become the norm by which the scientist is judged, and the positive/negative binary may be extended to the evaluation of the scientist. Alternate frameworks may be less subject to this problem, as we shall see.

THE SCIENTIST MAY INVEST EMOTIONALLY IN PROVING A HYPOTHESIS CORRECT, WHEREAS ALTERNATE FRAMEWORKS MAY BE MORE RESISTANT TO THE SCIENTIST'S MINDSET

Much has been written about the influence of psychology on the experimentalist, but little may have filtered down to the bench worker and the individuals designing experiments and interpreting data.⁷ However, if potential emotional influences are recognized, they may be more easily removed.

If, by its nature, a hypothesis is a construct of the scientist, the scientist may be proud of that construct. And because, by its nature, the hypothesis has the same grammatical structure as the conclusion, the scientist's emotional investment in the idea may further blur the tipping point between the proven and the unproven, or encourage the scientist to confirm the hypothesis.

Also, if the principal investigator in a laboratory establishes a hypothesis, and instructs a young graduate student to study that hypothesis, there is at least some concern that the student may be biased in favor of the hypothesis as "truth." This may not be a problem if the experimental readout is objective, that is, not subject to interpretation by the data gatherer, but if there is a component of subjectivity to the readout, special care must be taken to insulate the observer from the influence of the hypothesis. The power of "blind" observations is discussed in Chapter 12. For now, be aware that although it is possible to control for investigator bias and for psychological influences of the framework, unless the issues are recognized, the appropriate controls are rarely included in the experimental design.

OTHER FRAMEWORKS MAY BE RELATIVELY INSULATED FROM THE PROBLEMS OF THE HYPOTHESIS

As an experimentalist, you may disagree completely and aggressively with the last few points that argue that the hypothesis can inappropriately create the incentive to amass data in its support. But what if an alternate framework existed in which any experimental result achieved would be evenly weighted? What if there was a way to frame experiments in which all readouts were deemed "positive," and in which all reproducible readouts actually yielded useful data? Would that not be preferable to the Critical Rationalist approach, even if you believe that Critical Rationalism works well?

⁷However, Popper (*The logic of scientific discovery*, 2002, Routledge, New York) actually dismissed psychology as being irrelevant to the logical testing of an idea. He focused instead on the psychology underlying how an idea springs to mind as the place at which psychology might come into play, not the psychology that may be operative and induces the defense of an idea. One might note the irony in this omission.

The following chapters suggest an alternative framework and even point out settings in which hypotheses are actually unworkable, thereby establishing a need for an alternative.

SUMMARY AND CAVEATS

Criticisms have been leveled against Critical Rationalism in this chapter, and yet this philosophical approach has been dominant during perhaps the most fruitful period of scientific exploration ever known. Therefore, it is obvious that the hypothesis does allow useful experimentation. However, experienced scientists will recognize instances in which unhelpful interpretations can result from inappropriate use of the hypothesis, and may therefore be forewarned of the potential problems instilled by the Critical Rationalist framework. The next chapters offer an alternative to Critical Rationalism, in which hypotheses can still be made, but in which either the problem or the question govern the experimental framework. We then approach the very large issue of experimental controls.

Before moving forward, the reader should be aware that there are many philosophical critiques of Karl Popper and of Critical Rationalism. Much of the criticism against Popper is made on issues peculiar to philosophy, such as whether he is consistent in his criticism of inductive reasoning and whether he in fact allows inductive reasoning to pollute the purity of his hypothesis. In contrast to those critiques, I will argue that inductive reasoning is what happens in the most productive experimental systems, and the fact that it sneaks into Critical Rationalism is probably this approach's saving grace. A philosopher reading this book may find such comments excessively rash, but they must consider the vast differences between present-day science and science conducted in the days when Hume, Kant, and Russell rejected induction as based on scanty experience (i.e., noting that empiricism was not particularly empirical). In modern science, with the tools that are available, empiricism can be based on sampling huge amounts of data. For example, experiments that were once dismissed as random fishing expeditions may now constitute comprehensive analyses. In such a setting, empiricism may actually be truly comprehensive and, therefore, the criticisms applied to inductive reasoning in the past may be less applicable today.⁸

Further objections to inductive reasoning are discussed in Chapters 9 and 18.

⁸However, those criticisms certainly need to be recognized, because they are helpful in limiting the scope of conclusions drawn from inference. We take note of past objections throughout the book.



Scientific Settings in Which a Hypothesis Is Not Practical

IN THE 1990S, VERY LARGE BIOLOGICAL RESEARCH PROJECTS were initiated that seemed to defy the requirement for a hypothesis. For example, the sequencing of various genomes—the process by which scientists determined the entire DNA sequence for a particular animal, plant, or species of bacterium—occurred without obvious hypotheses in place.

Could there have been a proper, falsifiable hypothesis framing the sequencing of the human genome? We now know that this genome consists of approximately 3 billion base pairs,¹ with each base comprising one of the four genetic letters, or nucleotides: G (guanine), A (adenine), T (thymine), and C (cytosine). Once the sequence was known, particular regions could be compared to the sequences of known genes, and, using established rules of gene expression, the set of potential gene products could be predicted with increasingly reliable accuracy.

Suppose a hypothesis framing the original sequencing project had the following structure: “Having the complete sequence of the genome will reveal genes that are involved in human processes.” That statement is simply impossible to falsify, because the genes in the human genome are clearly all involved in human processes. The statement, therefore, is not a hypothesis—it is a rationale, a wish.

Would it have been helpful to hypothesize that a particular number of genes in the human genome were related to a particular known gene? Such a hypothesis would indeed be falsifiable. Thus the hypothesis that “there are ten genes in the genome related to insulin” is valid in a Critical Rationalist framework because it can be subjected to falsification. But, although such a hypothesis could be used as an excuse to sequence the genome, would anyone seriously pursue such a project on this basis? (Imagine undertaking a project costing billions of dollars and countless people-hours to test the hypothesis that there are other genes related to insulin in the genome, or even to test much broader hypotheses, such as “genes exist that are related to

¹Human Genome Project Information, Oak Ridge National Laboratory at http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer/prim1.html/.

(a particular family of) receptors in the genome.” Such hypotheses do not pass the straight-face test.)

It may be helpful to recall what actually happened in the Human Genome Project. When one of the major players in the project, Dr. J. Craig Venter, testified before the United States Congress on his goal to sequence the genome, he did indeed feel the need to offer a hypothesis. He stated, “It is our hypothesis that this approach will be successful.”²

Thus, in the case of the Human Genome Project, it seems especially clear that a hypothesis was neither necessary nor would it have been especially useful in conducting the largest and most labor-intensive biological project in human history, and yet extremely valuable information was gained by sequencing the genome.

Was any framework necessary, or actually in place, for sequencing the genome? Certainly something was needed to motivate the work required to get all of that sequencing done; humans do not randomly gather data to perform a project—the data are gathered for a reason. But, as we will now see, motivations are not the same as frameworks.

In the case of the human genome, it was recognized that the need to clone genes often delayed projects for a long period of time, and resulted in ad hoc and limited sequencing efforts. Dr. Venter had personal experience with such a limited sequencing project and wanted to get the job done faster. Perhaps he became engrossed in the technology for its own sake, viewing the project as a great problem to solve, possibly in the same way that a climber views a large mountain. However, at the time, given the large amount of money required for the project, no one could say that they were sequencing the genome “because it was there.” They needed a rationale, and one rationale used to raise funds was that understanding the genome would speed the search for cures to human disease.³ This might have actually been the sincere rationale for many (or all) of the individuals working on the Genome Project. Certainly the investors who put money into genomics companies in the late 1990s probably did so not because they thought sequencing the genome was cool (although it was), but rather because they expected the results to be helpful (which they were).

Could the rationale that understanding the genome would speed the search for cures to human disease be reframed as a hypothesis? Actually, no, because it cannot be subjected to falsification. We cannot go back in time to see whether cures would have been found just as rapidly without the genome sequence. The rationale is just that—a goal.

Here we are colliding with the psychology of motivation, which is distinct from any of the frameworks that are useful for scientific exploration. Goals such as curing cancer, helping humankind, getting tenure, making money, or becoming famous

²Prepared statement of J. Craig Venter, Ph.D., before the Subcommittee on Energy and Environment, U.S. House of Representatives, Committee on Science, June 17, 1998, at http://www.house.gov/science/venter_06-17.htm/.

³*The genome war: How Craig Venter tried to capture the code of life and save the world*, by James Shreeve, 2004. Knopf, New York.

might help to motivate a scientist, but they are not useful frames for a project, because they do not guide experimental design or data interpretation. This is not to say that goals should be dismissed or ignored, but they do not function as frameworks.

Was there a recognizable framework for the Human Genome Project? Simply put, it was the genome itself, the great unknown, that was the problem for those interested in obtaining the sequence, and this problem required a solution.⁴ This problem/solution binary may function more obviously as a frame for an experimental project if it is restated as a question. For example,

What is the sequence of the human genome?

This question, and the whole issue of questions as frameworks, is discussed in Chapter 4. For now, it may be useful to present some of what happened after the human genome was characterized, because the history helps to illustrate where hypotheses can be proposed and, therefore, what is required in advance of a hypothesis.

USEFUL FORMATION OF A HYPOTHESIS REQUIRES PRIOR KNOWLEDGE OF THE SYSTEM

On obtaining the human genome sequence, and comparing the new DNA data to previously obtained sequences of well-characterized genes, investigators immediately realized that new genes could be identified by their sequence similarity to genes with known functions, and it was thus possible to guess the probable function of many of the newly identified genes. For example, if four genes were known to encode a particular family of proteins called receptor tyrosine kinases, and the genome contained 20 more sequences that were similar to these known genes, it could be predicted that the newly found sequences might encode novel receptor tyrosine kinases.

Those predictions could then be phrased as hypotheses, for example, "gene 278 is a receptor tyrosine kinase." This hypothesis is perfectly testable, and it is quite a reasonable example of the Critical Rationalist framework. One could express the protein encoded by gene 278 and determine whether it functions as a tyrosine kinase.⁵ So now we're back in the realm of useful hypotheses. It was the enormous data-gathering event (the genome sequencing) that occurred between the point where a hypothesis was not practical and the point at which a useful hypothesis could be formulated. Therefore, something about the data gathering that occurred by sequencing the genome was actually necessary to formulate hypotheses on the basis of the human genome.

Although this appears to be perfectly logical, it is actually a problem for the Critical

⁴The problem/solution dichotomy has itself been subjected to criticism, especially by the philosopher Martin Heidegger, but his ideas are not addressed in this book.

⁵Tyrosine kinases place phosphate groups on tyrosine. A kinase is an enzyme that adds a phosphate group and a tyrosine kinase adds the phosphate to the amino acid tyrosine. Receptor tyrosine kinases are tyrosine kinases that reside at the surface of cells and bind secreted growth factors. These molecules have a very well-defined sequence structure, with each family member containing conserved amino acid motifs.

Rationalist, because a hypothesis, strictly speaking, should not be based on induction.⁶ Indeed, a major motivation in the formulation of the hypothesis framework was to steer away from the problem of inductive reasoning (reasoning based on experience).

But the Human Genome example provides a powerful illustration that a hypothesis cannot be formulated without a preexisting data set that is first understood on some level. Similarly, the statement “The sky is red” cannot be formulated without first at least understanding that the sky is colored and that the colors exist in a known spectrum of light. The available colors from which to choose potential shades for the sky is exactly analogous to the available sequence of the human genome.

Something else needs to be recognized about the Genome Project. Despite the fact that it was done without a hypothesis, it was in fact accomplished and provided useful information that accurately⁷ reported the sequence of a human genome. This example of a huge project that was not at first amenable to a clear-cut hypothesis may prove the point that hypotheses are not required at all (if a hypothesis was unnecessary when the project was very complicated and resistant to guesswork, why would it be required to solve a simpler issue?).

SUMMARY

Although in Chapter 2 it was argued that the hypothesis framework contained significant pitfalls, these arguments did not show that useful experimentation could be performed in the complete absence of a hypothesis. The Human Genome Project provides at least one example of a type of experimental project that is actually resistant to a hypothesis, and helps to show that it is possible to characterize the unknown without first guessing the possible outcome and that this can be done on a very large scale. If that is possible, one must wonder why smaller-scale projects, for which guesses may be made, need be framed with a hypothesis.

⁶This is a rather difficult area. Popper wrote positively about contrasting “better” theories, but at the same time he seemed to reject empiricism as a basis for those theories. See, for example, Popper’s *Conjecture and Refutations: The Growth of Scientific Knowledge*, 2002 (Routledge Classics, London).

⁷Not perfectly; sequencing errors continue to be corrected.

4



The Problem/Question as a Framework for Scientific Projects

An Invitation for Inductive Reasoning

WHAT SORT OF TOOL MIGHT RESCUE our first hypothesis that “the sky is red,” and force us to perform the various experiments needed to yield the relevant data? Consider for a moment how information is gathered in everyday life. Most people, who do not consider themselves “scientists,” gather useful information by asking questions.

A MODEL OF THE INDUCTIVE FRAMEWORK: HOW A SYSTEM OF SEQUENTIAL QUERIES CAN BE VALIDATED

Imagine that you are a tourist, visiting New York City, and you arrive at Columbia University, on 116th Street and Broadway. You want to get to Carnegie Hall, but you do not know how to get there. What do you do? You probably would not choose to take would be to construct a hypothesis, stating that “taking Broadway south and making a right on 57th Street will get me to Carnegie Hall”¹ and then seek to falsify that hypothesis, gradually discarding routes until you hit upon the correct answer. We will call this the “hypothesis falsification” approach. An alternate approach would be to ask the question “How do I get to Carnegie Hall?” (Fig. 1). If you can then find a New Yorker willing to answer your question,² you might be given three different answers. You may then ask which is the best route, but you will have to provide your parameters for “best.” For example, would you like to pass by Lincoln Center or do you simply want to get there as quickly as possible? The New Yorker may then be able

¹Actually, you need to make a left on 57th Street.

²Without first saying, “practice, practice.”

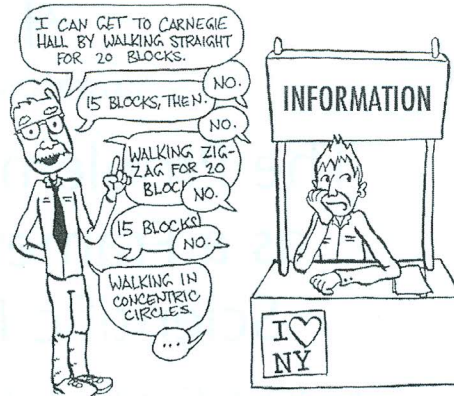
Hypothesis
TestingQuestion/
Answer

Figure 1. Cartoon illustrating that an effort to get from point A to point B (in this case, to arrive at Carnegie Hall [point B] from elsewhere in New York City [point A]) is very difficult if hypothesis testing is used as intended—to falsify particular hypotheses. One can get a series of “no” answers, without moving forward. In contrast, if the same project is framed with a question, much more rapid progress may be made.

to give you a reasonable route, taking into account the time of day, but (and this is important) the advice that you get will be limited by that New Yorker’s knowledge of the possible alternatives. It was this limitation of individual experience that was responsible for some of the early criticisms of inductive reasoning.³ Indeed, a small number of prior observations can be labeled as an inappropriate filter; you can’t accurately advise a tourist on the fastest possible route to Carnegie Hall until you have tried all of the routes available, timed them, and done so repeatedly.

However, perhaps some of those early objections to inductive reasoning are less valid today, given the power of the available technology to include large data sets. For example, rather than walking up to a New Yorker to ask for directions, you might now be able to use a computer program that can access data for every street in New

³We are now bumping up against some of the great philosophers of Western civilization, including Immanuel Kant (writing in *Kant’s Theory of Knowledge*), David Hume, and Bertrand Russell. Hume and Kant are discussed more directly in Chapter 9.

⁴This
e: a
⁵T,
for th
when
pr e
ca s
tion t

York City; if you did that, the frame of reference used to answer your question may include an extensive set of available routes between Columbia University and Carnegie Hall, and every potential route may be factored into the analysis. The ability to perform a study on such a representative data set is actually an important development, to which we will return when dealing with more biology-based issues. The ability to access very large samples significantly reduces the problem of limited experience, minimizing the chance that the data set is unrepresentative.

Continuing with the mapping scenario, say that you have decided to use a computer program to figure out how to get from Columbia University to Carnegie Hall in the fastest way possible. Like the approach of asking the *New Yorker* for directions, your first step with the computer program would be to ask a question. Let's say that the computer program is written to first input the query and then call up prior measurements of the time it took each of a thousand cars to drive on every street between Columbia University and Carnegie Hall (this information is part of a large database at the disposal of the computer program).⁴ In addition, the program monitors traffic alerts and eliminates those streets that are currently impassable. A calculation is then made and an output produced. The output can be thought of as a prediction based on prior experience—the fastest route traveled by those cars included in the data set. But more accurately, the output is simply a conclusion, and it is an open question as to whether it is correct to make a prediction based on that conclusion. Whether the program accurately predicts the fastest route at any particular time would need to be subjected to experiment; you would then validate the system by asking if the program was accurate. You would test the accuracy by having identical agents travel the suggested route versus alternative routes and time the trips. If the prediction of the program was consistently accurate (in a statistically significant manner), its methodology would be validated. If the program were inaccurate (i.e., if some other routes were consistently faster than the one given by the program), you would need to go back to the drawing board and perhaps add even more data, repeating the validation experiment until you had an accurate program.

The metaphor of the computer-mapping program serves to illustrate how inductive science can be done. An experimental program is designed to answer a question, such as "What is the fastest route from point A to point B?" A data set is gathered and potential answers are made, using this data set. The answers are validated according to their ability to predict the future in a statistically significant manner. The success at predicting the future is the criterion for determining whether the answer to the question is "accurate."⁵

⁴This is not the way in which established mapping programs are written. It is merely a metaphor for how experiments may be done via induction.

⁵This ability to accurately predict the future will be seen as the "demarcation" that made Karl Popper argue for the use of a hypothesis. Popper said that a hypothesis is necessary because it can be easily disproven, whereas it is much more difficult to prove that something is correct. He also argued that the ability to disprove the hypothesis provides a clear demarcation. However, I will argue that an answer to a question that can successfully predict the future, in a statistically significant manner, provides just the type of demarcation to which Popper referred.

t Carnegie
ig is used as
ing forward.
de.

nd this is
ledge of
that was
d a small
it accu-
you have

are less
sets. For
ou might
in New

including
and Kant

PREDICTING THE FUTURE VALIDATES BOTH INDUCTION AND HYPOTHESIS TESTING

Let's say that you want to know whether you have accurately answered the question "What is the fastest way to get Carnegie Hall?" If you think that the criticisms leveled against Critical Rationalism were nonsense, you can still invoke a hypothesis, stating that "The route that my computer program suggests is the fastest route to Carnegie Hall." You must then try to disprove that hypothesis by sending agents out along the chosen route who will compare it with alternative, control routes, and then time the routes. Such an approach could lead you to "the truth," if you do not fall prey to any of the pitfalls of hypothesis testing. Alternatively, the accuracy of the program can be tested simply by asking the question "Is the suggested route the fastest way to get to Carnegie Hall?"

Because a question was asked, an answer would then need to be derived by using experimental tests to determine whether the suggested route is indeed the fastest. In both cases, there would be a result that could be subjected to statistical analysis.

COMPARING THE DIFFERENT APPROACHES

To formalize what we have said so far, let's take the two paradigms one at a time.

In the Critical Rationalism mode, the approach is to

1. Decide on an experimental project.
2. Make a hypothesis.
3. Subject the hypothesis to falsification.
4. Get a result.
5. Determine whether the result is accurate by repeating the hypothesis test.

In the question/answer mode, the approach is to

1. Decide on an experimental project.
2. Ask a question.
3. Get an answer.
4. Determine if the answer is accurate by asking whether the derived answer is the same when the question is asked again.

At first glance, the two methods might seem quite similar because they both require a predictive experiment to be performed. Therefore, the issue is whether the question/answer method proposes any advantages to the scientist over the hypothesis/falsification method.

THE QUESTION AS A FRAMEWORK FOR SCIENTIFIC PROJECTS

To contrast Critical Rationalism to the proposed question/answer method, it may be useful to revisit some of the criticisms of the hypothesis and determine whether the question/answer method escapes those criticisms.

Let us return to our original hypothesis.

The sky is red.

How do we recast this hypothesis into a question to be used as a framework for experimentation? The first impulse may be to simply transpose the statement into a simple binary query:

Is the sky red?

This question can be labeled “close-ended” or “binary” because it seems to demand a “yes” or “no” answer, although the answer can be qualified.

In the mapping program example used earlier, the broad, open-ended question “What is the fastest way to get to Carnegie Hall?” was answered by breaking the problem down into thousands of close-ended binary questions, following the general structure “Is route X the fastest way to get to Carnegie Hall?”, “Is route Y the fastest way to get to Carnegie Hall?”, “Is route Z the fastest way to get to Carnegie Hall?”, etc. The binary queries serve as subsets within a systematic method to answer the broad, open-ended question “What is the fastest way to get to Carnegie Hall?”

Without the broad, open-ended question, there is still no proper framework for the experimental project. Asking any one of the binary questions does not force any of the other questions. For example, without the frame “What is the fastest way to get to Carnegie Hall?” one is not necessarily required to compare the answer to the question “Is X the fastest way to get to Carnegie Hall?” with that of “Is Y the fastest way to get to Carnegie Hall?” In other words, there is a greater chance that the experimentalist will fall into the trap of limited inductive experience by comparing X to an improperly restricted data set without the frame of the broad, open-ended question.

In the case of the experimental project to determine the sky color, the open-ended question format, which may be used to frame our experimental project, is

What color is the sky?

DERIVING THE OPEN-ENDED QUESTION

As we noted, if one starts with the hypothesis that the sky is red, the most obvious question to be derived from this statement would be “Is the sky red?” This close-ended question fails on its own to demand iterative queries covering the other shades

of the rainbow; for example, you do not need to ask “Is the sky blue?” if you start with the question “Is the sky red?”

However, if you begin with the experimental project to determine the colors of the sky and you are charged with framing that project with a question, then the experimentalist would begin with the open-ended question “What color is the sky?” rather than ask about a specific color. Indeed, if we take a step back, forget that hypotheses ever existed, and instead just begin with the initial project (to determine the color of the sky), one would object to the binary question “Is the sky red?” as a bizarre leap to a particular conclusion—why would you ask the single question about a particular color if you were charged with establishing a framework for the project for the accurate determination of sky color? You would ask the question “Is the sky red?” only within a series of questions. That series would need to be framed, and the frame would be the open-ended question.⁶

Understanding that the open-ended question does not derive easily from the hypothesis may be seen as yet another criticism of the hypothesis. In other words, the hypothesis by its nature is a leap to a particular, and this leap is necessary to establish a framework that can be subjected to falsification. In contrast, the sequential-question method does not have such a requirement. Here, the scientist is rewarded for asking a broad question that can be subdivided into answerable yes/no (binary) queries. The broader the net of the open-ended question, the more likely a subset question will yield a “yes” answer.

THE BINARY YES/NO AND OPEN-ENDED QUESTION HAVE DIFFERENT GRAMMATICAL STRUCTURES IN COMPARISON TO THE CONCLUSION

The question “Is the sky red?” has a different grammatical structure than the potential answers of “Yes, the sky is red”; “No, the sky is not red”; or “The sky is red at dawn and at dusk.” Remember that in the case of the hypothesis “The sky is red,” the conclusion had the identical structure:

Hypothesis (unproven): *The sky is red.*

Conclusion (proven): *The sky is red.*

Compare that to the question/answer framework:

⁶You may wonder why that is the frame. Remember that the definition of a framework is a philosophical construct for the experimental project that forces the scientist to perform the experiments necessary to determine “the truth,” which we define here as the experiments necessary to predict the future with statistical accuracy. As will be shown, the open-ended question gets you to where you want to go. If it is still not obvious that you would end up with “What color is the sky?” perhaps certain rules are in order, such as, for example, you must choose as your frame an open-ended, rather than a binary, question.

Question (unproven): *What color is the sky?*

Answer (proven): *The sky is blue.*⁷

It is quite helpful that the format for the rationale that motivates the experiment differs from the analytical output derived from the experimental data; the question framework forces the scientist to confront the fact that, before experimentation, the answer to the question is unknown. Put another way, the question emphasizes that the scientist is operating in the absence of a proven answer, in question mode. The question is being asked to illuminate a state of ignorance. This emphasis is in firm opposition to the hypothesis. The hypothesis suggests that the scientist has at least enough knowledge of the forthcoming results to produce a statement that will be (the scientist hopes) unfalsifiable. Stated more honestly, the hypothesis often suggests that the scientist has brilliantly derived the correct statement of reality in the absence of experimentation, which will later manifest the scientist's brilliance.

THE QUESTION DOES NOT INSTILL A REQUIREMENT FOR A PARTICULAR RESULT

In contrast to the Critical Rationalist framework that requires the hypothesis be held up for falsification, therefore placing the scientist in a situation of tension between that requirement and his desire to have a "proven" hypothesis, the open-ended question does not create any tension between the experimentalist and the result. Remember the question "What color is the sky?" In the framework of this question, any answer obtained by measuring the wavelength of light reflected by the sky will be "positive data." There is no need to "filter" the answer through a preestablished, set point for success (such as proving that the sky is red). Any data could be seen as "successful" (see Fig. 2).

The realization that one can formulate an experiment such that any result can be deemed "positive" is actually of great importance to the practicing scientist. The biggest fear for the experimentalist is that time will be wasted on chasing a hypothesis that turns out to be false. There is no shortage of young scientists who leave the profession because they failed to obtain "positive data" in frameworks that are biased toward failure. Pity the poor graduate student charged with proving the hypothesis

⁷Obviously that is not a complete answer, but we'll see shortly how one gets to the rest of the answer. For now, the example simply serves to illustrate the difference in structure between the question and the answer, and why that is useful to the scientist. If the reader thinks that I am "gaming" the discussion by having the answer to the question be more specific and accurate than the conclusion to the hypothesis, recall how these components were derived—the conclusion to the hypothesis is less accurate because of the biasing effect of the predisposing hypothesis. If you simply ask the question as stated, it would be more difficult to end up with an answer such as "The sky is red." Arguments supporting this contention are advanced further in the text.

↓
Showway



Figure 2. Cartoon illustrating that the hypothesis may induce a scientist to filter for “the positive response.” In the desire to actually prove a hypothesis correct, as opposed to falsifying it, the scientist will wait until the hypothesis “It’s dark outside” is verified, even if it did not describe the situation for which it was originally constructed.

“The sky is brown.”⁸ How different the student’s future might be if the question were “What color is the sky?”

DOES THE QUESTION/ANSWER FORMAT GUARANTEE THAT THE EXPERIMENTALIST WILL ARRIVE AT THE “CORRECT” ANSWER?

Not necessarily. If this book had been written in the 1970s, the initial prescription being offered here, to frame experiments with broad open-ended questions, may have been dismissed as unworkable; at that time, it was much harder to amass comprehensive data sets relating to biological questions. A big attraction of the hypothesis is that it allows the scientist to focus on the reagents at hand. For example, what if the scientist had the ability to measure only “red” or “not red”? In that situation, the open-ended question would demand the purchase of additional equipment. In a bio-

⁸Except, perhaps, the graduate student who happens to work in Los Angeles, California, where the smog currently renders the sky a particularly unpleasant shade of brown.

ne biološki tako široko
plastično levo

FOKUSIRANJE
U SKLOPU
GLAVNOG
PITANJA
UNUTAR GLAVNOG - HAPA POD PITANJA

logical context, broad, open-ended questions such as "What gene products, when inhibited, can cure cancer?" might be regarded as unhelpful. Such a framework demands a huge series of binary questions, and few labs, if any, are equipped to answer them. However, as will be shown later in this chapter, broad questions can be subdivided into more tractable, smaller-scale, open-ended queries that are approachable in the most modest of laboratories. Furthermore, even if the scientist is not capable of responding completely to an open-ended question, and even if this inability may constrain the project to a less-than-complete answer, the broad, open-ended question still has several benefits as a framework in this context.

!

NA KAVU BINARNH
podrazumeva

The Broad, Open-ended Question Defines the Scope of the Problem

Even if a scientist cannot test every gene to establish whether, when overexpressed, it can cure cancer, it is still helpful for the scientist to realize that testing whether a single gene has such properties involves grappling with only a very small percentage of the total possibilities. In mathematical terms, say that only a single gene out of 35,000 total genes has the ability to cure cancer when overexpressed. In that setting, the odds are that any particular gene that a scientist tests has only a 1/35,000 chance of giving a positive readout. Therefore, a hypothesis such as "Gene 99 will cure cancer" is doomed to failure, whereas the question "Which gene will cure cancer?" at least allows the scientist to wrestle with the scope of the problem and to continue until a usable methodology for surveying all human genes is developed.

The Open-ended Question Does Not Restrict the Scientist to Studying a Particular Subset

- ne ogranicuje

Even when a particular open-ended framework is less than helpful in guiding the scientist to the "right" answer (because it is too large), at least it does not get in the scientist's way. For example, sometimes a hypothesis may actually be deemed restrictive. In "The sky is red" framework, one might view measuring blue as irrelevant. In contrast, given the question "What color is the sky?" one would be forced beyond the binary query of "Is the sky red?"

The Open-ended Question Informs the Scientist that Experimentation Is Being Performed in a Position of Ignorance

- biti HIPOTEZA
daje daje da
da
odgovor
na
pa umjesto
direktnog
& potvrdi

This is the format issue discussed previously. The grammatical structure of the question, and the requirement to formulate the framework using this structure, shields the scientist from the idea that the answer is already known. (If the answer is known, why is the question being asked? If the question is being asked, the answer must be unknown, even if the scientist can make a pretty good guess as to what the answer will be.)

- POZICIJA IGNORANTNJA
u exp.

ponse."
wait until
was origi-

were

option
have
mpre-
sis is
f the
n, the
bio-

he smog

The Open-ended Question Forces the Formulation of Close-ended Questions

The mass of close-ended questions that help to define the scope of the problem allows the development of better methodologies. To answer the close-ended question of whether the sky is red, the equipment must be capable of those two analyses and no more. However, to answer the open-ended question, "What color is the sky?" the equipment must be upgraded to test for all colors in the visible spectrum.

The Open-ended Question Does Not Give Differential Weight to Any Particular Answer

In a question/answer mode, the scientist is not rewarded by discovering that the sky is yellow, as opposed to blue. This is in contrast to the situation in which the advance standard is established that an idea is only proven correct if the experiment yields a particular result. If I ask you what flavor ice cream you like, I probably will not care whether you like chocolate or cherry-berry. However, if I opine in advance that you like vanilla, then I have a stake in your answer, and I may want to argue with you once you've proclaimed your love for chocolate.

The Question Framework Guarantees a "Successful Experiment" as Long as the Technology is Available to Produce Answers

If a scientist cannot answer the question "What color is the sky?" after performing an experiment, that scientist should not adopt the question as his/her experimental project. If only a single particular color can be measured, that produces a low likelihood for success, given the existence of the broad question. Thus, the scientist is shielded from plunging into a low-percentage project and is forced by the breadth of the framework to develop the tools and methodologies needed to tackle a particular problem. An example of this is the question "What is the sequence of the human genome?" This question demanded dramatic new developments in sequencing methods. Had the ability to sequence whole chromosomes not been developed, the question could not have been answered. Of course, when the question was first asked, these methods were not available, and this highlights the next benefit of the Problem/Question approach.

Particular Questions May Actually Stimulate the Development of New and Useful Technology

This may also be true of particular hypotheses, but open-ended questions seem, by their nature, to be more likely to stimulate innovation to deal with the resultant broad framework.

the type
rewards

prediction
adjective

→ guaranteed
w/ specific
exp

hpi
What is the
seq of the
human
genome
- raising
NEW
TECHNOLOGY

primer - o podeli prostora ali i suzavajanje pitanja

to keep open mind of researchers a ne podbeci

ASKING MORE DISCRETE QUESTIONS

simplifikacija
zbog exp
organizacije

Open-ended questions need not be broad or all encompassing. The reason the broad question is suggested as the framework is that recognizing the breadth of the issue being studied helps to inform the researcher of the need to keep an open mind and avoid assuming that the world is composed of nails simply because the experimental system is limited to a hammer. Having made that point, it should also be noted that one need not jump from "What genes when mutated cause cancer?" to the binary question of "Does gene X, when mutated, cause cancer?" Given that there is often a technological barrier to being truly comprehensive, it is often (or almost always) necessary to carve out a more restricted landscape within the issue. For example, one may limit "What genes, when mutated, cause cancer?" to "What genes, when mutated, cause pancreatic cancer?" Or, if one is particularly interested in gene X, one can first ask the open-ended question "What is the function of gene X?" and then intersect this issue with the cancer issue (Fig. 3). This will eventually result in the open-ended question "What is the role of gene X in cancer?" Note that the answer "There is no role for gene X in cancer" is still valid, even though this particular question formulation does seem to suggest that gene X does have a role in cancer. Assumptions can slip into questions in addition to hypotheses, but we hope that Figure 3 helps to illustrate that the issue is much broader than the circumscribed question "What is the role of gene X in cancer?" Also, one can create a procedure in which the two sets are not made to intersect until some data demonstrate that they do intersect (Fig. 4).

Put another way, why would one ask the question "What is the role of gene X in cancer?" unless there was some prior knowledge gained, perhaps from previous exper-

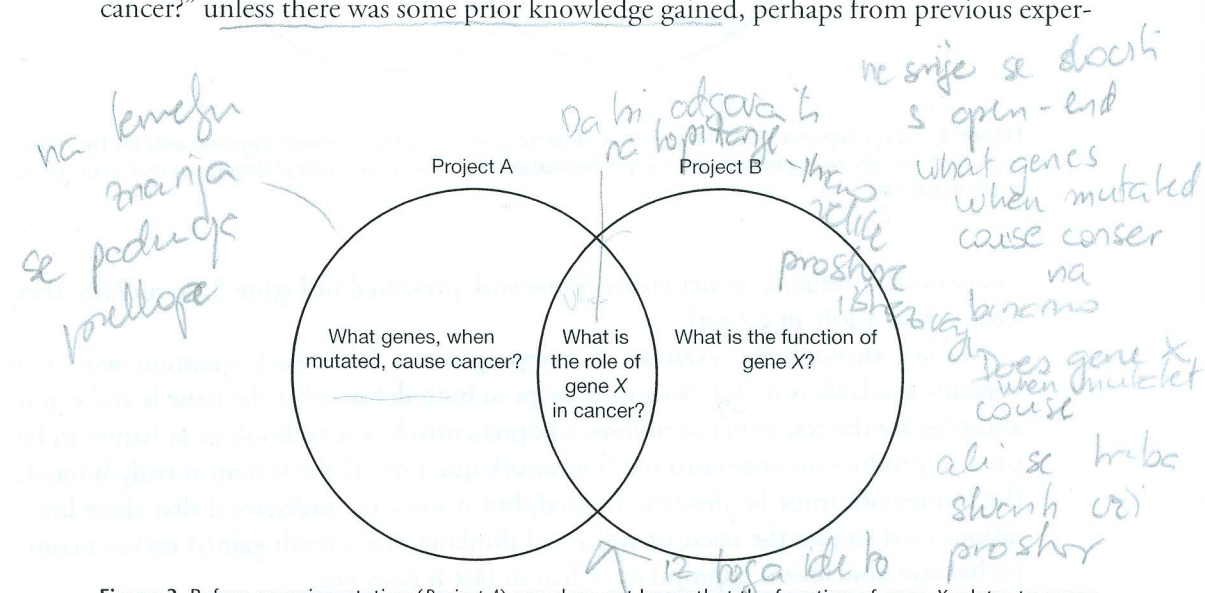
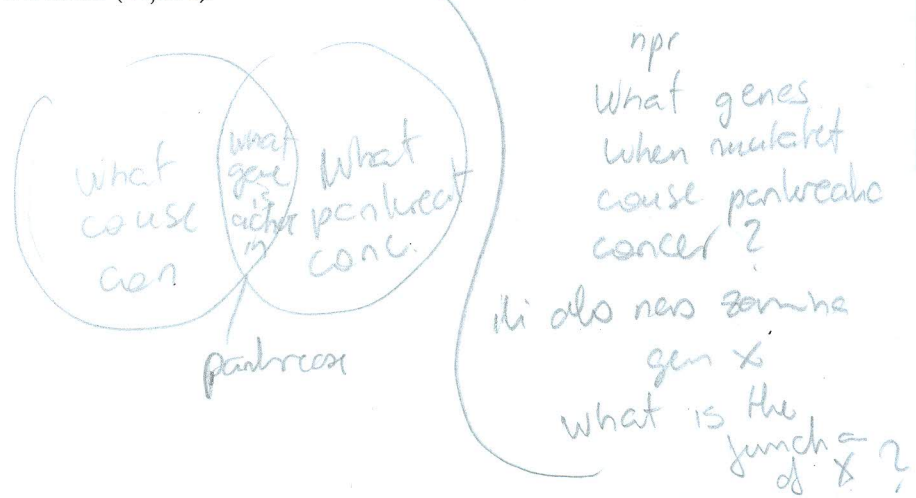


Figure 3. Before experimentation (Project A) one does not know that the function of gene X relates to genes that, when mutated, cause cancer. But after experimentation showing that gene X has a role in cancer, one can link the two spheres of knowledge and access information concerning genes that cause cancer to assess the function of gene X in cancer (Project B).



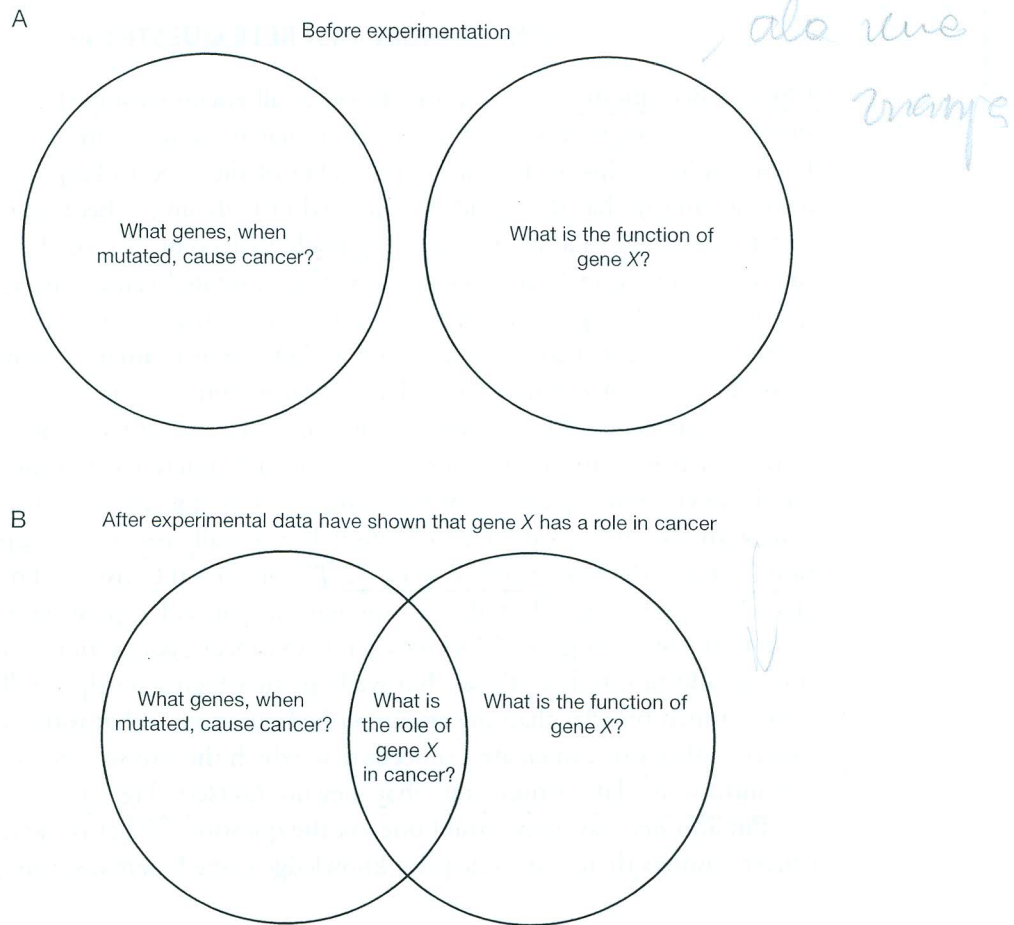


Figure 4. Venn diagrams illustrating (A) two discrete questions before experimentation and (B) the intersection of two discrete questions after experimentation (i.e., after experimental data show that gene X has a role in cancer).

imentation conducted under either framework presented in Figure 3, to indicate that gene X had a role in cancer?

Going through the exercise of mapping out a framework question with the reagents available is useful, because it helps to both define what the issue is and is not and educate the researcher as to how, comprehensively, the technology at hand can be used to produce an answer to the framework question. If the system is truly limited, the framework must be similarly limited, but it must be understood that these limitations exist so that the scientist will avoid thinking that a result gained makes a comprehensive conclusion appropriate, when in fact it does not.

5



What Constitutes an Acceptable Answer to an Experimental Question?

WHAT ANSWER WOULD BE SCIENTIFICALLY ACCEPTABLE in response to the query “What color is the sky?” Does this question immediately suggest a result structure that conforms to reality?

Suppose that to answer this question, a scientist conducts an experiment at noon and does everything as previously outlined in Chapter 4, including asking a series of close-ended questions (Is the sky blue?, Is the sky green?, Is the sky yellow?, Is the sky red?) and measuring all of the wavelengths of visible light, and then gets the following answer:

The sky is blue.

The scientist then goes further and asks a follow-up question:

Is the sky blue?

The investigator performs the full-spectrum measurements ten more times, and, to limit experimental variation, takes each set of measurements at noon on successive days. The answer to the follow-up question turns out to be “Yes, the sky is blue” 27 times out of 30 (it was cloudy on days 6, 7, and 14, and the sky was therefore gray). The difference between the number of days on which the sky was blue and the number on which it was gray, or “not blue,” is statistically significant, and the blue readout yields the same answer as the answer to the initial question.

But did the scientist successfully answer the question? According to the criteria for determining a validated answer, the question may well have been answered. The answer to the initial question allowed the scientist to ask a prospective question, and that question produced the same answer as the initial question in a statistically significant manner.

The scientist may thus believe that a “validated answer” was achieved. However, the reader may object that because the scientist failed to take any nighttime measurements, or measurements at any time other than noon, we cannot really conclude that the experiments produced a complete answer to the question. But the experiment did yield an answer that can predict the future. If the criterion for a successful model is its ability to predict the future in a statistically significant manner, the model that “the sky is blue” will be successful as long as the measurement is taken every day at noon. The model will only be called into question on the day that the scientist perhaps has insomnia and thinks to take a measurement at midnight—that experience will cause a revision of the model. This is the normal process of scientific exploration, wherein an answer that was thought to be definitive is called into question by a new experience. This does not mean it is wrong to say that the sky is blue; however, the midnight measurement educates the scientist to see that the scope of the question forces a caveated answer after a new discovery. Once the scientist thinks to measure the sky’s color at midnight, the model shifts to “the sky is blue at noon, but not at midnight.”

This process of amending an answer introduces the concept of “model building,” which is explored more fully in Chapter 6. For now, the point should be made that an answer to an experimental question is deemed to be “correct” if it provides a “model” that represents reality. A model can be said to represent reality if it accurately predicts the future. For example, if the “model” of the sky shows the sun shining brightly and illuminating a blue sky with occasional white clouds, and if we wake up the next day and see a picture that coincides with our model, we will believe the model to be accurate. Similarly, if a model of gravity shows an object when it is released falling to the ground, and every time we drop an object it indeed falls to the ground, then we will trust our model of gravity.

A model is thus a representation of what the scientist believes to be reality and its accuracy rests entirely on its ability to demonstrate what will happen in the future. If we drop a leaf to the ground and the leaf is blown by the wind up into the air, we will have learned about a force that may oppose gravity. This may cause the model of gravity to take such opposing forces into account.

It may be unsatisfying that a model may be accepted as “accurate” even if it is incomplete, but as we will see later, the demand for universality is not easily applied to biological settings. Therefore, one may be satisfied at accurately characterizing a part of the problem, as long as the limitation of the answer is recognized.



How Experimental Conclusions Are Used to Represent Reality

Model Building

FOR A SCIENTIST TO ACCUMULATE KNOWLEDGE in a useful way, that is, to answer the question that functions as the framework for the experimental project, two things must happen upon the completion of the experiment:

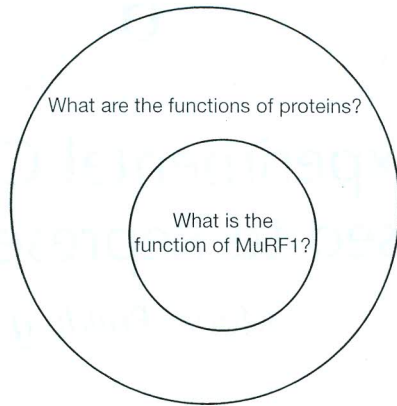
1. The experiment must be interpreted and a conclusion derived.
2. The conclusion must be placed into the context of previous experiments to either confirm or perturb the understanding of the subject. This contextualizing process, which can be called “model building,” is the process of formulating the answer to the framework question.

Because a hypothesis is formulated in advance of experimentation, it is a guess or a speculation. In contrast, a model is built after the experiment has been done and is therefore based on accumulated data. This is the critical distinction between a model and a hypothesis, which might otherwise be thought to accomplish the same task. A model is so named because it is designed to represent the pertinent features of the subject being analyzed, thus providing a representation of the answer to the framework question. Model building is a process of induction, an empiricism-based, bottom-up exercise in developing an understanding of accumulated facts.

AN EXAMPLE OF MODEL BUILDING

Start with an Open-ended Framework Question

To give an example for how a model might be built, we will analyze an experimental project that was actually conducted. Lest the reader get lost in the details or wonder why a more symbolic representation is not being offered, be assured that even though this is a “real-world” example, it functions as a metaphor for how models in general can be built.



Dadele o prot MuRF1 u znanu
 bas nista vege od da je
 prot i zato ga stavljaju
 u knjig u koga su i sve
 ostale grupe proteina ali
 kao zaseben podskup.
 Zbog toga ni ne moze postaviti
 close-ended question
 kao Da li MuRF1 radi X?
 Takvi ras pitanje ionako
 odgovaralo samo na X pa
 drugje moze biti prihvatljive
 ne bi
 in zapanje

Figure 1. Before experimentation, MuRF1 is known to be a protein, but, if no data exist to tell what type of protein MuRF1 might be, the entire set of proteins is equally available as context for MuRF1 function. This is a lot of information to go through and thus of limited use.

Also, as we shall see, overgeneralization makes it difficult to grasp some of the systemic issues that perturb the experimental process. It is only when grappling with actual experimental details that one discovers a general process that can be used to build a useful model, i.e., one that functions as the answer to the question framing the experimental project.

Let's say that a scientist wants to understand the function of a protein called MuRF1.¹ This scientist frames the experimental project with the following question:

What is the function of MuRF1?

Because this is the framework question and no experimentation has been done on the protein, the scientist has no model for the function of MuRF1. In other words, no sufficient experiential data exist for the initiation of an inductive process concerning MuRF1. The key word in that last sentence is "sufficient." MuRF1 is a protein and, because proteins as a class have been subjected to experimentation, there does exist a pool of knowledge that will be tapped for induction. However, before the first conclusion, that pool is inaccessible because it is not yet clear how MuRF1 intersects with the set of known information concerning all other proteins, except to note that MuRF1 is a protein and, therefore, any information concerning MuRF1 will fit within the subset of data relating to the function of all proteins (Fig. 1).

So, we begin as a *tabula rasa*, a blank slate with regard to MuRF1 in particular, but we understand that MuRF1 belongs to a certain class: proteins. However, because we have no relevant data on MuRF1 or its functions, we do not yet know how MuRF1 relates to other proteins. For example, with regard to Figure 1, it would be impossible to know if proteins outside of the MuRF1 circle have relevance to MuRF1's function; it could be that MuRF1 is unique and therefore other proteins are not relevant to how MuRF1 works. Although this state of ignorance may appear to

¹As mentioned, this is an actual protein. But MuRF1 and the descriptions of the experiments done to examine it also function as a metaphor for any experimental project.

be a disadvantage, it can also be a position of privilege because, before the first question is asked, there is no opportunity for bias. In this setting, the scientist can be open to any possibility when answering the framework question.

Consider this “privileged state” further and recall the discussion in previous chapters. First, contrast the position of the scientist confronted with the open-ended question “What is the function of MuRF1?” to the scientist who frames the experimental project with a close-ended question, such as “Does MuRF1 do *X*?” It should be clear that the close-ended question does not allow for the unbiased “position of privilege,” because the jump to *X* without any background data will bias the scientist to (1) start the project by examining *X* and also (2) possibly finish the project with this examination, because the only requirement is to study *X*. We won’t belabor the point by discussing the potential hypothesis, that MuRF1 does *X*; the issues surrounding such a construct have already been outlined.

Let’s return to the “privileged state,” before the first experiment is done and after the open-ended question “What is the function of MuRF1?” has been asked. As noted, at this point the possibility of building a model to represent a potential answer does not yet exist because the scientist lacks the necessary experiential information to begin even the first bit of induction concerning MuRF1 in particular—no experiment has been done, no conclusions have been drawn, and no data have been accessed concerning proteins in general. Because conclusions are the bricks with which the model is erected, it is impossible at this time to start building. So what does the scientist do?

FORMULATING THE FIRST EXPERIMENTAL QUESTION BY ACCESSING THE INDUCTIVE SPACE

How does one progress from the framework question, which establishes the boundaries for the experimental project, to an experimental question, which delineates the requirements for a particular experiment? What might be a useful first question that will focus the scientist’s experiments, so that useful data can be derived that will help to build an answer to the framework question?

As was noted above, MuRF1 is not the first protein ever to be studied; thousands of proteins’ functions are already known. These proteins have particular structures and contain certain subsections called domains that have been shown in previous cases to presage particular functions. Thus, perhaps the scientist’s first question should be

Does MuRF1 resemble any proteins of known function?

This question invites a particular process of comparing the structure of MuRF1 to the structure of other proteins, resulting in a set of proteins that resemble MuRF1 in different ways. On a more general level, what the scientist is doing (by comparing MuRF1 to other proteins whose function is known) is attempting to access relevant data on related subjects to gain an understanding of the current subject (Fig. 2). Contextualizing the specific protein MuRF1 in the broad group of all proteins allows

experimenta ho
Da li to
strukturn
slicucom
i domenc
na neki
poznati protein?

it type
unction.

stemic
experi-
odel,
ject.
called
ion:

ie on
rds, no
arning
and,
exist a
onclu-
h the
MuRF1
in the

icular,
because
how
uld be
nce to
s are
dear to

amine

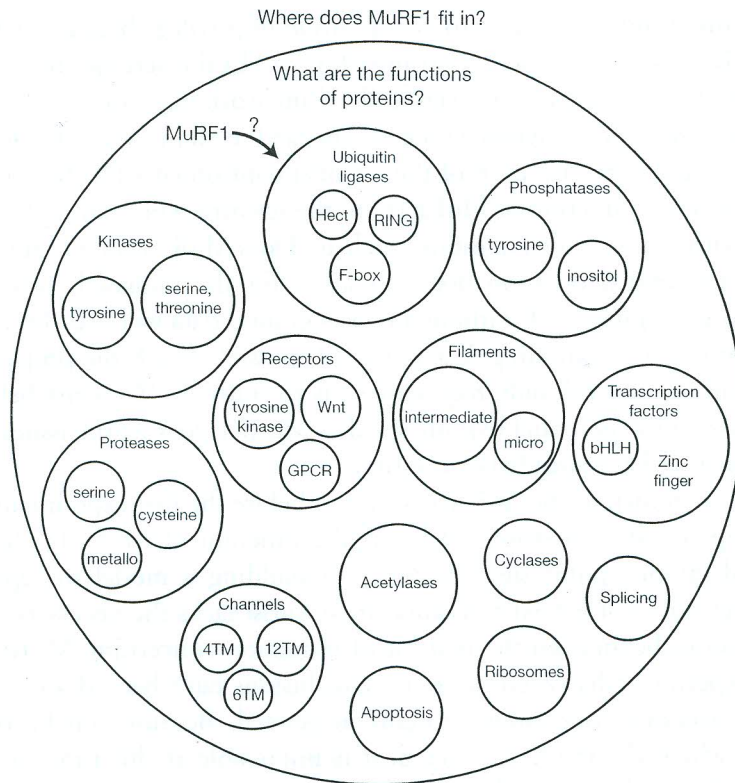


Figure 2. Once the broad set of proteins has been subdivided into different categories of proteins, by understanding structure/function relationships, a more focused question can be asked such as “What type of protein is MuRF1?”

MuRF1 to function as a metaphor for general study; the same process occurs regardless of the study subject. Contextualizing the specific subject in the set of data that is relevant for that subject is labeled here as accessing the “inductive space.”

INDUKCIJA
INDUKTIVNI PROSTOR

THE NECESSITY OF BACKGROUND INFORMATION

Establishing the Broader Inductive Context that Shapes the First Experimental Question

Is it always true that potentially relevant questions about a particular experimental subject can be suggested by referring to a related, but previously studied, subject? If one consults *Philosophical Explanations*, by Robert Nozick,² the answer to this question is yes, because even if the inferred information from the known subject turns out to be wrong in its application to the new subject, discovering that will still be helpful;

²“I believe increased understanding can be produced even by an explanation known to be false; seeing what in principle could give rise to a phenomenon illuminates some of its aspects by the way it latches onto these.” R. Nozick. 1981. *Philosophical Explanations*. Harvard University Press, Cambridge, Massachusetts.

INDU

understanding that $A \neq B$, even though A and B share some characteristics, will aid in the isolation of the relevant differences between the known subject and the unknown subject.³ Furthermore, the question/answer process uses the inductive space in a particular way, as a basis upon which to ask a question, not to "deduce" an answer.

As noted earlier, one of the reasons earlier philosophers shunned inductive reasoning was the concern that selective experience would be biasing and therefore produce flawed conclusions. But the process suggested by the question/answer methodology is that the scientist use prior knowledge, not to formulate a conclusion about a particular unknown, but as a basis upon which to ask a question about the unknown. This is the relevant difference between the question/answer methodology and Critical Rationalism: The critical rationalist uses prior knowledge to frame a hypothesis, whereas in the question/answer framework the scientist uses prior knowledge to ask a question.

In the case of the specific example of MuRF1, the scientist does not use the established knowledge concerning proteins in general to deduce the function of MuRF1. Rather, he/she accesses the known data to ask a question about MuRF1. From the answer to this question, the scientist will then ask the question again. If the repeated question resulted in the same answer as that for the initial question, a model will be constructed that functions as the conclusion—the answer to the experimental question. This type of reasoning that uses specific findings to build a model, which can then be used to generate larger principles, is what is meant by induction.

WHAT IF THE SCIENTIST IS CONFRONTED WITH A SUBJECT THAT IS DISSIMILAR TO WHAT HAS BEEN PREVIOUSLY STUDIED?

It is increasingly rare, especially in biology, for a scientist to encounter an experimental project for which no relevant data exist to initiate the process of asking questions, but it is important to consider this "state of nature" situation to ensure that the model-building prescription derived may be generalized. Analyzing the "state of nature" helps in understanding how one establishes the initial set of inductive space that is required to formulate questions within the experimental framework.

When confronted with a subject that has no relevant antecedent models that could be used to frame experiments, a set of questions that can be thought of as "descriptive" need to be asked. The first time a protein was isolated, the issue was simply to determine what it was. Later, it could be discovered that proteins served particular functions. Next, experiments were conducted to relate individual protein structures to those discrete functions. It took several decades of descriptive and then correlative analyses to build the context necessary to ask a question such as "What is the function of MuRF1?"

The set of information gathered from answering broad, descriptive questions helps to define the initial inductive space. For example, the following list of potential

³For example, just because two proteins start and finish with identical amino acids doesn't mean that they have the same function. It is thus useful to learn when similarities are relevant and when they are not.

exp pitanji
ne donosi
zabljutek

MURF1 na
sedeu boi
na U-ligane

a sta
da ne vidim?

sto als
radimo
na nistem
sto ne maemo
poverati mi s
ome
parahm

teins, by
"at type

gard-
hat is

imental
If one
ion is
it to be
lpful;

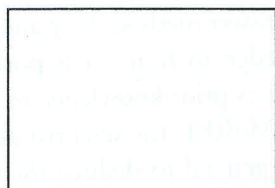
ing what
these."

questions could have been asked about proteins when no prior knowledge concerning them existed (the questions are restated in a more general way in parentheses, to show that they may be applied to other subjects besides proteins) (Fig. 3):

1. Of what is the protein comprised? (What is the composition of the unknown, X ?)
2. What form(s) does the protein take? (What is the structure of the unknown, X ?)

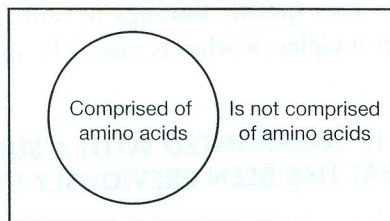
A State of nature

What is this thing?



B Discovery: The "thing" is comprised of amino acids

What is this "thing" that contains amino acids?



C Defining peptide structure and naming the "thing" that is comprised of amino acids: "Proteins"

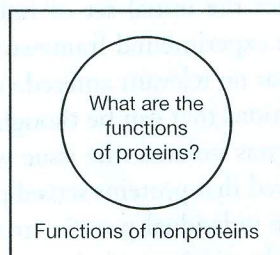


Figure 3. If no information is available about an object of study, there is no inductive space and thus no way to categorize the object. Therefore, one must start with basic experiments to categorize the "thing" being studied (A). Once some information becomes available, such as the fact that the "thing" has amino acids (B), one can then establish a category for it: proteins (C). Once the "protein" category has been established, that category can be used to study the broad class of "things" that fall into the set of "proteins," and information about one protein may be found to be relevant for another. This categorization also allows the scientist to exclude the class of "things" outside of proteins (nonproteins; C).

3. What is the function of the protein? (What does X do?)
4. How does the protein structure imbue protein function? (Does the function relate to the structure of X ? If so, how?)

These questions might have been used to interrogate disparate unknowns such as gravity, light, matter, the Earth, space, the atmosphere, the human body, etc. When analyzing the questions, a gradual accumulation of data becomes clear, starting with structural information and proceeding to a mechanistic understanding of how the subject functions. From this information, the scientist can progress to more specific experimental projects, for example, from a general scheme for how proteins behave, to particular examinations of individual proteins and their functions in the context of their environment. This class of investigations can categorize the experimental project “What is the function of MuRF1?”

ESTABLISHING THE INDUCTIVE SPACE, WHICH INFORMS AN EXPERIMENTAL PROJECT

At this point, it should be understood that surrounding the framework question “What is the function of MuRF1?” exist background data concerning proteins in general. The question “What is the function of MuRF1?” cannot be asked before the scientist has asked a question such as “What is the structure and function of proteins in general?” and has built a sufficient model to address the broader question. In fact, it may be realized that the framework question “What is the function of MuRF1?” could be viewed as a subset of the broader question “What is the function of proteins?” (Fig. 1). Although this last question is not a useful experimental project for any one scientist, it may be helpful to realize that one question is the subset for the other, because the broader question helps to define the inductive space, which is relevant to the experimental project.

The Unknown Subject Versus the Alien Subject

The question concerning MuRF1 should now be appreciated as functioning in the more common setting at which the scientist does not start from the “true” beginning, but operates in a background of accumulated knowledge. Perhaps a quick additional illustration will amplify this point. Compare the following two framework questions:

1. What is the function of an elephant’s trunk?
2. What is the function of a trunk-like structure found on an alien from outer space?

In framing experiments to answer question 1, the scientist does not need to go into an extensive study on the very nature of the elephant, because that is already

cern-
ses, to

m, X ?)

, X ?)

I thus no
“thing”
amino
s been
of “pro-
tion also

known. One can infer that the elephant has a heart, blood, a liver, a brain, a mouth, and nasal passages, and that these structures work as previously determined. Also, the trunk is physically located near known structures, the mouth and the nasal passages. Thus, inductive space is available from which one can proceed with questions about the elephant's trunk (see Fig. 4).

However, in framing experiments to answer question 2, one needs to do all of the work that was already done on the elephant and other animals, during previous centuries, to get to a place where question 2 can be answered. Thus, even though a model cannot be built to represent the function of the elephant's trunk until the first experiment is done, relevant experience still needs to be appreciated; such experience allows the scientist to ask questions.

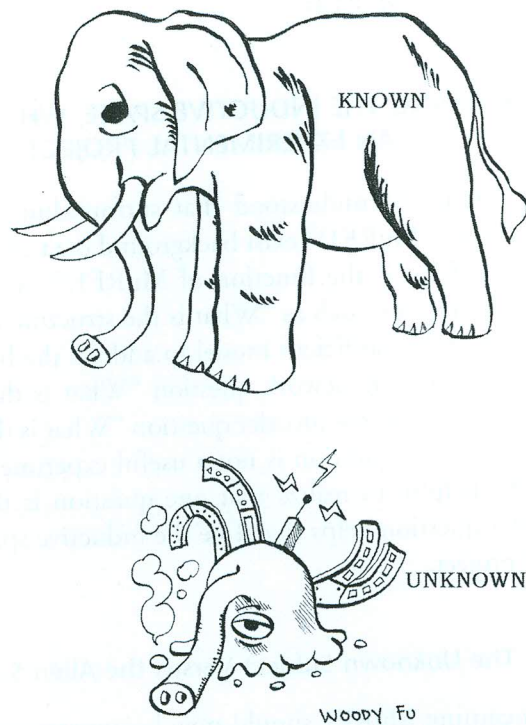


Figure 4. This cartoon illustrates the usefulness of accessing the inductive space. Assume that one does not know the function of an elephant's trunk; it is therefore unknown. However, if one knows the functions of the other parts of the elephant, enough context is established to start asking relevant questions about the trunk; for example, because it is positioned at the same place as a nose, one might think to ask whether it performs the same function as a nose. This process of resorting to prior knowledge and using it to posit broader knowledge is called inductive reasoning. In the method advocated in this book, one does not use induction to establish a theory, but rather to ask a question. It is only by answering the question and determining the answer through experimentation that one can build a model as to the function of the trunk. In contrast to the case of the elephant, if a structure that resembles a trunk is attached to an alien, for whom no information is available, many more basic questions need to be asked before one can develop enough of a knowledge base to determine the function of the trunk-like structure.

Returning to the Experimental Project and the First Question

Now that the concept of inductive space has been further established, the scientist can return to the first question concerning the function of MuRF1. The specifics of MuRF1 are explained to illustrate how the scientist may proceed from initial categorizing questions to particular experimental questions, conclusions, and thus a model. The first question seeks to categorize MuRF1 in the context of previously studied proteins:

Does MuRF1 resemble proteins of known function?

The question can be formulated as a bioinformatics computer query, in which a program such as BLAST⁴ is used to compare the MuRF1 protein sequence to sequences of other proteins. The scientist performs the BLAST search and obtains a list of several hundred proteins that share a particular domain with MuRF1.

The scientist investigates further, searching the published literature for proteins that possess this shared domain, and learns that many of these proteins belong to a class of proteins called “E3 ubiquitin ligases.” These proteins can instigate the cellular breakdown of distinct protein substrates by causing the addition of several copies of a small peptide, called ubiquitin, to the substrate. Once at least four ubiquitin molecules are ligated to a protein, it can then be targeted for proteolysis by a large protein machine called the proteasome.

Thus far, by virtue of a primary sequence comparison, the answer to the question “Does MuRF1 resemble other proteins of known function?” is “Yes, MuRF1 resembles certain E3 ubiquitin ligases.” Operating in either a question/answer framework or Critical Rationalism, the scientist has not shown that MuRF1 actually functions as an E3 ubiquitin ligase. Learning that MuRF1 and the E3 ubiquitin ligases have this particular domain in common is not sufficient to establish the function of MuRF1.⁵ To do that, one must perform an experiment.

Many other more detailed, structural questions could be asked about MuRF1; entire fields of study are devoted to protein structure. For example, MuRF1 could be crystallized and understood on a molecular level using X-ray diffraction. MuRF1 could also be analyzed structurally in distinct biochemical contexts, such as by determining the proteins that bind to MuRF1. All such investigations help the scientist to understand MuRF1’s appearance at an increasingly precise level; the purpose of these studies is to allow the scientist to be increasingly specific in

⁴BLAST is an acronym for Basic Local Alignment Search Tool. See, for example, <http://www.ncbi.nlm.nih.gov/BLAST/>.

⁵This might be a disputable point, and it illustrates the difference between deduction and induction. For example, if 1 million proteins that possessed the protein domain in question were analyzed and each time this domain occurred the protein turned out to function as an E3 ubiquitin ligase, one might infer that possession of the structure was sufficient to assign function. However, even in that case, you would be left with the question concerning MuRF1, which would then require an experiment to produce final proof.

MuRF1
ima zafedichu
domen
E3 / 150000

ali
na ne
gaa
nito o fut

outh,
so, the
ages.
about
of the
vious
ough a
first
ience
e does
ows the
relevant
might
nowl-
vocate
it is only
ould a
esem-
uestions
rtion of

defining which proteins are “truly related” to MuRF1, that is, those that continue to share relevant features even in this more highly determined context.

Proceeding from Initial Structural Questions to the Second Set of Questions that Interrogate Function

After obtaining the structural information that MuRF1 appears similar to E3 ubiquitin ligases, the next step is probably obvious. If the scientist were operating in the Critical Rationalist framework, the following hypothesis would be proposed: MuRF1 is an E3 ubiquitin ligase. Even in this example, we see again how induction sneaks into Critical Rationalism; the hypothesis could not have been made until at least some background data were available, and this experiential data seems to be necessary to guide the formation of the first hypothesis.⁶ In the question/answer framework, a new question can be asked:

Does MuRF1 function as an E3 ubiquitin ligase?

Notice that this is a binary question. The answer is either “yes” or “no.” But wouldn’t this type of question work the same way as the question “Is the sky red?” Remember that in the current instance an overarching framework question is in place (What is the function of MuRF1?) that operates in a manner similar to the overarching question “What is the color of the sky?” in the previous example. Therefore, determining whether MuRF1 functions as an E3 ubiquitin ligase should not impede further explorations in other areas.

Performing the Functional Experiment and Using the Derived Data to Build the Model

The new subset question “Does MuRF1 function as an E3 ubiquitin ligase?” now forces the scientist to perform an experiment to determine whether MuRF1 has E3 ubiquitin ligase activity. One detail the reader needs to know is that many E3 ubiquitin ligases, in addition to ubiquitinating⁷ other molecules, can also ubiquitinate themselves. Because the scientist does not yet know the actual substrates for MuRF1, this first functional experiment simply determines whether MuRF1 can self-ubiquitinate when placed in an appropriate experimental context or assay.

Suppose that such an experiment is conducted and MuRF1 is found to be ubiquitinated when placed in an auto-ubiquitination assay.⁸ The scientist now has the first conclusion based on a functional experiment within the experimental program “What is the function of MuRF1?” At this point it may be useful to recall the four-step

⁶If you doubt this, ask yourself this question: What hypothesis would have been possible in proposing a function for MuRF1 before structural information was available?

⁷Ubiquitinating refers to the process of ligating the peptide ubiquitin to another protein.

⁸The relevant design and controls for experiments are discussed in following chapters.

approach for proceeding in the question/answer mode (see Chapter 4). The previously suggested scheme was to

1. Decide on an experimental project.
2. Ask a question.
3. Get an answer.
4. Determine whether the answer is accurate by asking whether the derived answer represents the answer to the question when it is asked again.

Perhaps at this point, with an actual example in hand, this outline can be refined to

1. Decide on an experimental project.
2. Ask a broad question to frame the experimental project.
3. Ask a subset question to frame a specific experiment to start deriving data that will be used to answer the broad question above.
4. Get an answer to the subset question.
5. Determine whether the answer is accurate by asking whether the derived answer represents the answer to the same question when it is asked again.
6. Use the answer to build the model.
7. Ask a new subset question.

With this scheme, perhaps it is now clearer as to how the scientist may proceed. Because only one functional experiment was done (answering the question “Does MuRF1 function as an E3 ubiquitin ligase?”), a repeat form of that question can now be asked:

Does MuRF1 function as an E3 ubiquitin ligase?

The format of the second question is identical to the previous question, but now the scientist has a “stake” in the answer, because previously the answer was “yes.” In this second experiment, an expectation is involved, and this expectation should be a legitimate concern to the scientist. The main virtue of this process, however, is that the expectation of a particular answer occurs only after data have already been gathered. Also, the experiment was designed before the first datum was obtained. Therefore, it should be emphasized, and this is quite an important point, that in the repeat phase of the experiment, the exact same methodology as was used the first time should be used again. If this is done, the scientist will be relatively insulated from making any alterations to “better guarantee” the same result as was achieved in the first experiment. Having noted this issue, we can modify our process even further:

1. Decide on an experimental project.
2. Ask a broad question to frame the experimental project.

3. Ask a subset question to frame a specific experiment to start deriving data with which to answer the broad question above.
4. Obtain an answer to the subset question.
5. Determine whether the answer is accurate by asking if the answer derived represents the answer to *the same* question, when it is asked again, *using the same methodology as was used the first time*.
6. Use the answer to build the model.
7. Ask a new subset question.

The scientist proceeds to ask the repeat question “Does MuRF1 function as an E3 ubiquitin ligase?” and does the experiment again, in exactly the same way as before. Once again, the answer is “Yes, MuRF1 functions as an E3 ubiquitin ligase.”

BUILDING THE MODEL

Only now, after the repeat functional experiment has been done, can the first step be taken to build a model that is designed to answer the question “What is the function of MuRF1?” A validated, data-founded conclusion has been drawn, which can therefore be used to construct the first draft of the model.

The fact that the first conclusion drawn was that “MuRF1 functions as an E3 ubiquitin ligase” should not imply that this conclusion is immutable, merely that it is deemed accurate, based on the available data, and exists within an experimental approach which, as we shall see, can allow for modifications of the derived model if new conclusions are made that cannot coexist with the previous understanding.

Transforming Conclusions into a Unified Model

The first experimental conclusion, that “MuRF1 functions as an E3 ubiquitin ligase,” is next applied to the framework question “What is the function of MuRF1?” The scientist now has a relevant conclusion that responds to this framework question. The model, as it exists at this point, can be stated in the following way:

MuRF1 functions as an E3 ubiquitin ligase.

Because the scientist has only one conclusion after the first experiment, the structure of the model is identical to the structure of the single conclusion.

Each New Conclusion Allows the Scientist to Focus on Particular Inductive Space, which Allows New Questions to Be Asked

The scientist is now ready to ask a second question to further address the framework question “What is the function of MuRF1?” It should be recalled that the first question

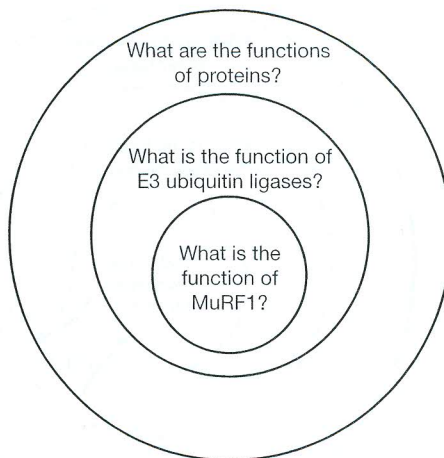


Figure 5. Once it is determined that MuRF1 belongs to the subcategory of proteins called E3 ubiquitin ligases, a much smaller set of proteins are deemed to be relevant inductive space: the E3 ubiquitin ligases. This realization allows the scientist to make faster progress, because he/she can now focus on the relatively smaller set of E3 ligases for hints on how one might study MuRF1. Notice, however, that one is still operating in the broad set of proteins. Therefore, if it is found that MuRF1 has an additional function outside of its E3 ubiquitin ligase function, relevant information is still available to access.

“Does MuRF1 function as an E3 ubiquitin ligase?” was made possible by drawing on available data for other proteins and then comparing MuRF1 to those known proteins. In this way, the basis for induction⁹ could be established.

Now, with the model in place that MuRF1 functions as an E3 ubiquitin ligase, it should be noticed that the inductive space surrounding MuRF1 has been perturbed. Before, there was a vast amount of general knowledge concerning protein structure and function, a great resource—but also a vast forest of data—and one from which it is difficult to derive pertinent questions concerning any one protein. However, with the first discovery about MuRF1 function, the scientist can ask further questions by drawing on all of the specific data concerning other known E3 ubiquitin ligases.

The known data concerning other E3s constitute a subset of the greater inductive space that is composed of all data on all proteins (Fig. 5). The new, restricted inductive space brings the data concerning E3s into much sharper focus; it is like looking at New York City from 1000 feet versus from the Moon—it is much more obvious how different parts of New York City relate when the observer can focus better. Similarly, when the inductive space is focused on a particular class of proteins, such as E3s, as opposed to including all proteins, other questions that could be asked about MuRF1 become much clearer.

⁹What is advocated in the question/answer methodology is that MuRF1’s resemblance to other E3 ubiquitin ligases should be used to set up a question, which then demands an experimental proof. Then, a general conclusion is gradually constructed (the model) after accumulating specific data.

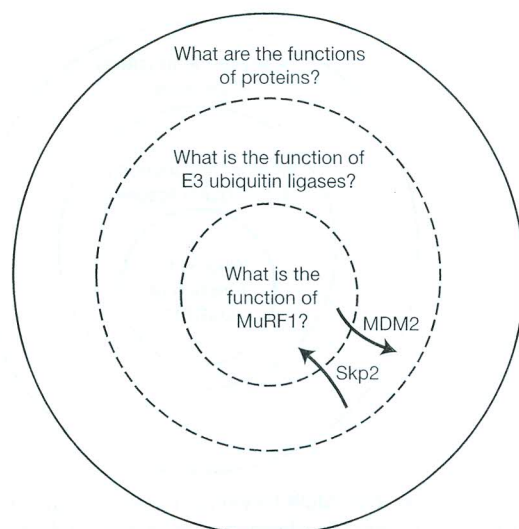


Figure 6. As one refines the inductive space, the amount of information that needs to be immediately accessed becomes smaller and is thus more accessible. For example, once it is clear that MuRF1 belongs in the set of E3 ligases, data on other E3 ligases immediately become available to aid in the work on MuRF1. This allows for much faster progress than before it was realized that MuRF1 was an E3 ligase; in the former instance, all the knowledge on proteins in general may have been viewed as equally important to MuRF1.

SMALLER INDUCTIVE SPACES ARE OPPORTUNITIES FOR RAPID EXPERIMENTAL PROGRESS AND POTENTIAL FILTERS THAT CAN CAUSE THE SCIENTIST TO MISS IMPORTANT FINDINGS

This new, more focused inductive space, the set of information concerning known E3 ubiquitin ligases, constitutes a subset of information about proteins in general and offers the scientist a more easily accessible set of specific and pointed questions concerning MuRF1 function (Fig. 6). For example, all known E3 ubiquitin ligases cause the ubiquitination of discrete and specific protein substrates. Two such E3 ubiquitin ligases are the proteins Skp2 and MDM2. Both Skp2 and MDM2 physically interact with their substrates and simultaneously bind to other protein machinery necessary to trigger substrate ubiquitination (Fig. 6). Given this model for how the known E3s, Skp2, and MDM2 function, the next obvious question about MuRF1 function would be the following:

What are the protein substrates for MuRF1?¹⁰

This new question is one of many now obvious questions that can be asked about MuRF1, given the discovery that (1) MuRF1 is an E3 ubiquitin ligase and (2) E3

¹⁰The critical reader might point out that this new question has an unproven premise, that MuRF1 does in fact have protein substrates. It would take analyzing all extant proteins in the genome to demonstrate that MuRF1 has no substrates at all; however, it should be realized it is possible to reach the conclusion that MuRF1 has no protein substrates from the question as posed.

ubiquitin ligases have particular substrates (Fig. 6). The importance of the model and the inductive space that it establishes should therefore be appreciated. They allow the scientist to make rapid progress in understanding something that is unknown, but that falls within recognized inductive space.

Although the focusing of the model or increasing the resolution of the model is, on a practical level, critical to the scientist, it is also a potential problem for the question/answer method. Remember that one of the major reasons for the proposal of Critical Rationalism was to grapple with the filtering effects of limited experience. Just as the scientist might reject directions from Columbia University to Carnegie Hall from any one New Yorker, so too the scientist may be concerned that the model of MuRF1 function might itself become a filter for further experimentation. In other words, once it is established that MuRF1 functions as an E3 ubiquitin ligase, would the scientist stop trying to determine functions for MuRF1 that are distinct from its E3 activities?

Examine the structure of the model that

MuRF1 functions as an E3 ubiquitin ligase.

This is a declarative sentence and can be seen to be limiting if it is thought to exclude other functions for MuRF1. However, notice an additional feature of the model: It is not exclusive. Now, the importance of the framework should therefore be appreciated to a greater degree. The framework question

What is the function of MuRF1?

may prevent the scientist from allowing the model to act as a filter. More specifically, it is the inductive space made available by the framework question that prevents inappropriate filtering (Fig. 6). In this case, that means the entire set of data available on known proteins or the model of how proteins in general function.

For now, it should be understood that the scientist has built a first model and succeeded in going from a broad question to a series of specific questions, allowing efforts to be focused on a particular subset of inductive space that, in turn, allows for the rapid production of new experimental data.

...diately
...longs in
...MuRF1.
...former
...MuRF1.

...n E3
...al and
...con-
...cause
...iquitin
...eract
...ssary
...n E3s,
...ction.

...bout
(2) E3

...does
...onstrate
...ion that



Establishing a System for Experimentation

THE SCIENTIST IS NOW READY TO BEGIN to think about experimental design. The issue of experimental design is the subject for much of the remainder of this book, although the philosophical points outlined in the first chapters are referenced to explain the need for certain experiments.

Let us return to the framework question "What color is the sky?" To answer this question, the scientist must develop an experimental system that will allow the relevant experiments to be performed. For example, the scientist may wish to approach the framework question by asking a series of binary questions, for example, interrogating each color in the visible light spectrum.¹ To accomplish this task, the scientist must validate that the relevant instrumentation to be used is capable of measuring these colors. Furthermore, there must be a way of knowing if the relevant colors are not measured.² Therefore, before the scientist ever points an instrument of color measurement toward the sky, he/she must first prove that the answers that the instrument yields are validated. This process, called "establishing the system," is a step that unfortunately is often ignored in many laboratories. However, establishing the system is absolutely critical to assuring the success of the experimental program. How can the scientist be sure of data achieved with a particular reagent unless the reagent is first validated as producing results "as advertised"?³

¹The scientist is thus accessing the "inductive space," obtained from other situations in which color is measured, as well as a knowledge of what color is, what the visible spectrum is, etc. These all function as "background" for the approach taken, but the existence of prior knowledge should be recognized, because it is the basis from which the experimental approach is derived.

²This need is explained later in the chapter.

³The issue of belief versus skepticism, and what is required for belief, is a "big" topic in its own right. Scientific epistemology is based on a process of empiricism, which is itself founded on a trust in what one sees. If one is interested in philosophical questions of belief versus skepticism, a good place to start is Robert Nozick's *Philosophical Explanations*, 1981. Harvard University Press, Cambridge, Massachusetts. (See Chapter 3.)

DESTABLISHING THE
SYSTEM

THE NEED TO VALIDATE THE SYSTEM USED IN AN EXPERIMENT

Imagine a motorist driving along a highway at the posted speed limit of 100 kilometers/hour. However, despite the driver's care, she is pulled over by a policeman who monitored her velocity with a radar gun. The radar gun established that the car was traveling at 115 kilometers/hour. When the policeman tells the driver she was driving above the speed limit, she angrily denies this, because her instrumentation gave a different readout.

Each individual in this scenario can be thought of as an observer making a particular measurement, using a distinct piece of instrumentation. Both people are asking the question "How fast is the car going?"

The following issues arise from this example:

1. Different methodologies used to measure an event can produce disparate results.
2. The only way to know which methodology is accurate is to validate the methodology with an independently verified reference.
3. The need to conduct such a verification project defines the need for "controls."

Most scientists think of controls as being embedded within each individual experiment, as a way of providing necessary comparisons to understand the data produced. But before the first experiment is conducted, a series of controls must be in place to trust that the data produced are accurate. To understand this, let's return to the driving example above and imagine the scenario if no policeman was present. If the motorist was simply driving in the absence of an external monitor (the policeman), she would probably believe that her vehicle was traveling at 100 kilometers/hour, because that was what her speedometer was reporting. Few motorists stop to question whether their cars are actually going at the speeds determined by their instrumentation (after all, they are motorists and not scientists). However, for a scientist/driver, it would not be enough that the speedometer is reading 100 kilometers/hour; there must be an external way of determining that the readout is accurate.

Does the policeman function as an external check of the car's velocity? Certainly, that is his purpose. However, in the case of the radar gun, its readout is just as suspect as that of the car's speedometer, unless it too has undergone a calibration exercise.

In the driving scenario, the policeman is a literal "control" on the driver; his function is to ensure that no cars exceed the speed limit. However, even though he is labeled as a control, he will fail in his function unless his instrumentation is validated. A control provides data that serve as a reference point—an object of comparison. But to function in this way, the control itself needs to be trustworthy (and the system thus needs to be validated).

The paradox, where even the control may be suspect, illustrates the problem of operating within a system without first making sure that the reagents to be used function as they should. If a policeman is using a miscalibrated radar gun, many motorists

will be falsely accused of speeding; the more inaccurate the radar gun, the more innocent people will be victimized. Alternatively, if the driver's speedometer is incorrect, she is at risk of driving faster than intended, and possibly having a life-threatening accident due to her misplaced faith in her instruments. An outside observer of this unvalidated scenario will have no way of drawing any conclusions as to who was at fault, or why an accident or a traffic-stop occurred, because they too cannot believe any of the data produced by the system.

How does the speedometer get validated? At a race track, for example, a measured kilometer may be available that is spaced out meter by meter and perhaps remeasured as appropriate, establishing a reasonable assurance of accuracy. Using this "validated" kilometer, cars can travel this distance, have their time of travel determined by a timer, and this timer can then be used as a cross-check for their speedometers. One might ask whether the timer is validated, and one can thus appreciate that there is some point between reasonable skepticism and paralyzing distrust; however, this is not an easy point to isolate. As a practical matter, in the laboratory, there is a need to validate systems that are not well established (such as newly made reagents, e.g., antibodies or DNA probes, etc.), as opposed to instruments such as balances or timers. But there are examples where "uncontrolled equipment" has been shown to result in experimental alterations (e.g., different plastics in various brands of test tubes can alter DNA transfection efficiencies). A careful scientist is always aware of this issue.

Once the speedometer is validated in a particular "test car," the radar gun can be validated; the radar gun could be used to measure the rate of speed for the car whose speedometer is now "scientifically accurate," and could thus be trusted to give an accurate readout. Once validated, the radar gun can be used to measure the speed of cars whose speedometers have not been independently checked—the validated radar gun functions as the "control."

What the reader should appreciate after critically going through this scenario is the great difficulty and care that must be used in establishing an experimental system to validate data.

EXPERIMENTAL CONTROLS

As introduced above, controls are the tools that the scientist uses to validate an experimental outcome. By unfortunate convention, controls are usually labeled as either "negative" or "positive."

The Negative Control

A "negative" control establishes a "not-X" frame of reference for the question being asked. For example, given the question "Is the sky red?" the negative control will be used to establish an instrumentation readout for the time when "not red" is measured.

In this example, the scientist will want to know that the system is capable of measuring the absence of red, otherwise no frame of reference is available to understand the “red” result. Put another way, the scientist wants to know that the instrument is operational despite it failing to measure “red,” so a measurement needs to be produced in the absence of red. What is more difficult to appreciate is that the absence of red needs to be quantitated so that one can appreciate the number of times that the positive readout was achieved in comparison to the negative.

To understand this further, recall the original scenario, in Chapter 2, in which a readout to the instrumentation was only produced if “red” was measured. In such a system, the scientist only appreciates the “positive” finding of “red,” without realizing that “not red” comprises the more common condition.

This need to measure “not red” and appreciate recognize this reading as data illustrates a problem with “negative” versus “positive” terminology, but it is beyond the scope of this book to fight that particular battle. For now, simply be aware that it is important to not only have a frame of reference that measures the “ X ” set, where X is the subject of the experimental query, but to also appreciate the “not- X ” result as relevant.

Negative controls are discussed in more detail in Chapter 12. The concept is introduced here in the context of system validation. In this instance, one must have a negative control to ensure that the instrumentation or reagents are accurate.

Consider the case of an antibody that is raised against a particular protein called “M-cadherin.” The antibody was generated to answer the question “In which tissue is M-cadherin localized?” Many scientists may go to a catalog, purchase an antibody that is advertised as recognizing M-cadherin, and use that antibody in an experiment, applying it to tissue sections to determine which tissue is positive for M-cadherin. These scientists are taking for granted that their M-cadherin antibody (1) binds M-cadherin and (2) fails to bind other proteins.

Let’s examine both assumptions and ask what would happen if either assumption is wrong. If the first assumption, that the antibody binds M-cadherin, is wrong, a “negative” result would be inaccurate. The scientist using the antibody would fail to determine which cells or tissue types make M-cadherin, because the tool they were using to measure the protein did not work.

The second assumption is that the antibody fails to bind other proteins, or, put another way, is specific for M-cadherin and only M-cadherin. If this assumption is wrong, a positive readout could be misleading, simply because the scientist cannot be sure if the positive readout is a true indicator for the protein of interest as opposed to some other, irrelevant protein.

Therefore, before the scientist ever uses his or her newly purchased antibody to answer the experimental question, he/she must ask a new question:

Is the system validated?

In the case of the M-cadherin antibody, that question becomes

Does the M-cadherin antibody work?

And by this question, the scientist is asking two other questions:

1. Does the M-cadherin antibody recognize M-cadherin?
2. Is the M-cadherin antibody specific for M-cadherin and only M-cadherin?

The M-cadherin antibody can be thought of as “the system” in place to measure M-cadherin. To validate the antibody, both questions above need to be answered. To address whether the M-cadherin antibody recognizes M-cadherin, several different techniques could be used. The most straightforward method would be to produce recombinant, purified M-cadherin protein and to determine if the antibody recognizes the recombinant material. This would be an experiment in itself, with its own negative control. In this case, the negative control would simply be the absence of M-cadherin. The scientist could thus do an experiment, with nothing in one sample and with M-cadherin in another sample, and determine whether the antibody recognizes the protein.

To discover whether the M-cadherin antibody is specific for M-cadherin and only M-cadherin, more sophisticated experiments are needed. The “gold standard” method for determining if this is valid would be to test the antibody versus all proteins other than M-cadherin. This can actually be done, now that animals can be produced that are deficient for individual genes. For example, tissue prepared from an animal that did not express M-cadherin would be an ideal “negative control” for the M-cadherin antibody, because any cross-reactivity of the antibody to tissue from an M-cadherin-deficient animal would prove that the antibody is not completely specific.

Of course, one does not always have access to ideal systems such as knockout animals for the purpose of validating reagents. One could simply survey tissues through other means, such as mRNA analysis, for example, to find a cell that was M-cadherin null (i.e., negative for the protein) and use this cell to determine if the antibody is specific. The null cell would be the negative control, and the gene encoding M-cadherin could be transfected into the null cell to test the antibody for specificity.

The Positive Control for System Validation

In the above example, in which the issue is validating the M-cadherin antibody, the positive control is the recombinant M-cadherin protein. The scientist knows that this protein exists because it is produced and purified, and can be sequenced and verified to be M-cadherin. This then becomes a “validated control,” as in the case of the speeding example, used to ensure that the antibody being tested is capable of recognizing M-cadherin.

Imagine an experiment without a positive control. Let’s say that the scientist validated a null cell to be used as a negative control and proved that the M-cadherin antibody did not recognize anything in that cell. This cell was then used as a reference

point versus samples of proteins from all other organs. The scientist uses the antibody to survey all of the samples and produces uniformly “negative” data because he/she did not detect M-cadherin in any sample.

The scientist may be tempted to conclude that the M-cadherin protein was missing from all tissues. However, let’s now add the positive control. In addition to the negative sample and all of the test samples, now the scientist also tests a sample that has had 10 μg of M-cadherin spiked into it. Once again, the scientist does the experiment and again fails to detect M-cadherin in all samples, including the positive control. Now the interpretation of the experiment changes. Instead of concluding that no tissue contains M-cadherin, the scientist must conclude that there was a problem in detection, because he/she knows that one sample had M-cadherin. This is the value of the positive control, to ensure that the subject being surveyed can be detected.

Sensitivity Controls

Between the negative control and positive control exists shades of gray. Returning to the metaphor of the car, it is obvious that an automobile can travel at many speeds. So if the question is “How fast is the car going?” one needs to know that the radar gun can measure speeds within a certain range. Also, if the policeman would write a ticket if the driver is traveling 115 kilometers/hour, but would not ticket a driver traveling 110 kilometers/hour, then it becomes important to know whether the radar gun can distinguish between speeds that differ by only 5 kilometers/hour.

This introduces a new type of control that can be thought of as a “sensitivity control.” In the case of M-cadherin, the scientist may want to know more than where M-cadherin can be found; he/she may wish to know what levels of M-cadherin can be detected by the antibody. For example, if the antibody being used can only detect 10 μg of M-cadherin protein per gram of tissue, the scientist would have no way of determining if smaller amounts of M-cadherin are expressed. Thus, even if 8 μg of M-cadherin are made in certain cells, an experiment may yield a negative readout with the insensitive antibody being used. However, if the scientist is aware of the issue, other types of experiments (such as an immune-precipitation experiment) could be done to try to concentrate the protein.

The combination of negative, positive, and sensitivity controls all help the scientist to validate the system, a necessary step in performing a successful experiment.



Designing the Experiment

Definitions, Time Courses, and Experimental Repeats

NOW THAT A FRAMEWORK FOR THINKING about experimental design and a system for validating the resulting data have been established, the scientist can open a laboratory notebook and begin to map out the first experiment. As indicated in previous chapters, the experiment may be undertaken in response to either a question or a hypothesis, and arguments were made as to why the scientist might consider the question/answer framework. The experiment needs to be designed so that its resulting data will be useful in building a model that serves to answer the question. Several examples of experiments are given in this volume. For now, an example of an experiment may be helpful in pulling together the philosophical arguments made in previous chapters and in illustrating how they can result in a practicable experiment that will generate useful data. We will use for an example of designing an experiment the question "What color is the sky?"

DEFINING TERMS

Consider again the following question, which frames an experimental project:

What color is the sky?

To answer this open-ended question, the scientist defines color as "visible color." He/she maps out the range of light wavelengths that constitute visible colors and decides to measure each wavelength in this range. By deciding to measure only visible spectra (or on a more general level, by defining a particular term), a filter is imposed—a certain limitation on experimental outcomes. This filter may or may not be appropriate, but recognizing its impact will at least allow the experiment's interpreter to understand the context in which results have been derived. In other words,

when an answer is given to the question "What color is the sky?" the informed listener will understand that the scientist means "What visible color is the sky?"¹

The scientist also needs to define the term "sky." For example, should the instrument be pointed straight up or toward the horizon? Perhaps it should span a range for positions from one horizon to the other. These decisions perturb the experimental outcome. During a series of experiments, the scientist decides that the measuring device will be systematically rotated over a 180° axis, spanning the sky from one horizon to the other, but that only a single point will be determined in any individual experiment.

With the definition of these terms, the process of experimental design has begun.

DURING THE EXPERIMENT, THE OBJECT MUST BE STUDIED UNDER REPRESENTATIVE CONDITIONS

After the scientist has translated the terminology of the experimental question, or hypothesis, into certain decisions regarding how the experiment is to be conducted, the next set of choices are made based on an understanding of what constitutes the "representative case." If an experiment is performed on a subject that operates in an atypical or nonphysiologic way, it is highly questionable as to whether the resulting data will translate to the more usual setting.

For example, in the case of the sky, as has been discussed, inductive information helps to determine the representative conditions. We know that the sky changes color during the course of a 24-hour day. We also know that it changes color under different environmental conditions, such as when it rains or when there is sunshine (or shortly after a rainstorm, as in the case of conditions favoring a rainbow). Furthermore, we are aware that different colors are visible in different parts of the sky. This inductive space (or prior knowledge) can, and indeed should, be used to design the experiment.

After accessing the inductive space (which may or may not be used to generalize a broader conclusion, based on the results of the experimental question), the scientist now decides to determine the color of the sky over 24 hours. Note that if the inductive space was not accessed, the scientist may have decided to simply point the instrument skyward and take measurements for only 5 minutes. Indeed, in the absence of any background information, there is no obvious reason why such a choice would not have been made. Prior knowledge and its influence on experimental design must be acknowledged; if the inductive space were to change, this could necessitate changes in the experimental design and require revisiting previously asked questions. For exam-

¹Although this may seem obvious in the particular example given, the reader should be aware that the "What color is the sky?" example is designed to serve as a metaphor for other experiments in which the issues are not as obvious. Many scientists may be unaware that their understanding of terms significantly perturbs the outcomes they deem relevant, and just as in the case of the scientist in the present example who chooses visible colors, scientists in other instances inevitably choose what they can "see," even though they may learn later that additional relevant factors were ignored. These later discoveries set the groundwork for reinterrogating the scientific issue.

REPRESENTATIVE
OBJET
ECS.

ple, it may be that certain planets are visible in the night sky during very short time periods during the year. Depending on how broadly the scientist interprets the question "What color is the sky?" the appearance of a red planet such as Mars may or may not be deemed relevant to the question.

Thus, by stating that an experiment should be performed under conditions representative of the subject as it exists in reality, we reiterate the importance of defining terms. We also recognize the need to "capture" the subject at multiple points to gather an understanding of which results should be seen as "usual," what the variability from that "norm" might be, and how that variability might be interpreted. In this case, the definition of representative conditions may be subject to reinterpretation and depend on the scientist's inductive space and how it evolves. This is a fact of life and the reason that scientists can design "better" experiments as they become more familiar with their systems; it is also the reason that answers and models tend to become more caveated as a system is better understood. Thus, at the first instance of asking "What color is the sky?" the instinctive answer might have been blue. But, over time, and via proper experimentation, all of the caveats and fine points of the question will come to be understood. These caveats will result in refinements in the way the question is asked, the manner in which it is understood, and the answers that are deemed relevant. For example, over time, the question "What color is the sky?" may include the subquestion "What visible color is the sky, at 10:00 a.m., while standing at a particular point, and directing one's instrument of measurement 45° upward from the horizon toward Mars on December 15, 2007?" This degree of specificity may disappoint the scientist who wants to derive a grand, all-encompassing conclusion, but this is exactly the point of this book: to disappoint the experimentalist who attempts to create a grand, all-encompassing conclusion that nevertheless does not conform to reality.

In the current example, let us stipulate that the scientist is new to the system and thus one of the only pieces of relevant prior knowledge available is that the sky color changes during the course of the day. This leads to the next decision in the scientist's experimental design: The initial experiment will be conducted over 24 hours to capture a representative sampling.

TIME COURSES VERSUS EXPERIMENTAL REPEATS

The scientist has decided to measure the color of the sky, pointing the instrument first in a particular direction for a period of 24 hours. He/she then decides to perform measurements every 5 minutes. This decision is made by again resorting to inductive knowledge. The scientist recognized that dawn or sunset could easily be missed if color were measured every hour, but he/she deems it unnecessary to perform continuous measurements (near-infinite time points make life very difficult if results are to be analyzed statistically). This type of iterative measurement, called a time course, can be recognized as an attempt to understand whether any particular point measured is representative of a general effect or whether the system fundamentally changes under

What color

Time courses



different conditions. If it is indeed found that the system is not “homogeneous,” for example, if the sky is red at one point and blue at another, the scientist must decide how to interpret the complexity. Under some definitions of relevance, a rare case such as finding the sky red at sunset might be discarded. In other paradigms, all data may be interpreted as equally relevant and a discovery of an exception to the general case may be celebrated, such as when one takes the time to notice the beauty and transient nature of a sunset.

We should pause here to note the continuous reference to prior knowledge and the fact that we usually do not know the answer when we ask a question. Therefore, it's important to consider how easy it would be to miss the dramatic change in sky color that occurs during short periods of the day in particular regions of the sky, given the choice of when, where, and how measurements are made. This should again serve as a metaphor, in this case for the care that needs to be taken in experimental design and why scientists often learn more when they go back to ask the same question again as they gather additional data about their systems.

Although statistics are not the focus of this book, their existence must not be ignored. As noted earlier, when used correctly, statistics can be extremely helpful, forcing experiments to be performed in such a way that the data generated can be proved to represent reality (by generating models that can give the scientist predictive power as to how the system will behave in the future).

One way in which statistics are extremely helpful is in forcing the scientist to make more than one determination of a particular outcome. However, this type of “repeat” is not the same as a time course. A repeat in the statistical sense occurs when instances of a particular perturbation, treatment, or event are measured more than once. The following is an example of doing repeats: Let's say that a scientist wanted to determine whether a drug caused weight loss and he/she performed a study in which many different individuals were tested. A group of 1000 might receive the potential weight-loss drug and another group of 1000 might be given a placebo. Each group has 1000 “repeats” that allow the statistician to determine the size of the effect, the standard deviation, etc.

When trying to determine the color of the sky, multiple measurements would have to be made of each 5-minute time point to obtain a statistically significant determination of sky color at that particular time. It should be reemphasized, however, that measuring sky color at 6:15 a.m. is not a repeat of the measurement at 6:00 a.m. For example, if the sky is black at 6:00 a.m., and red at 6:15 a.m. (because the sun has started to rise), the difference in color at 6:15 a.m. does not invalidate the measurement taken at 6:00 a.m. However, the difference in color between these two points complicates the manner in which the scientist may answer the experimental question. Rather, a repeat in the “What color is the sky?” question is the determination of the result at 6:15 a.m., for example, on each of several days. If the measurements are taken over the course of a week, seven repetitive measurements of that particular data point will be obtained.

HOW REPEATS AND TIME COURSES PERTURB THE EXPERIMENTAL "ANSWER"

Initially, before the time course was carried out, the scientist might have predicted a simple, one-word answer to the question "What color is the sky?" Now, with the experience of the time course, the scientist understands that the sky is not just one color; rather, it changes over time. Therefore, the scientist cannot give a one-word answer to the question. A different type of model needs to be constructed to accommodate the data, and that model will show that the sky's color changes. Returning to the question, as was already noted, if the sky color at 6:00 a.m. is different from the color at 6:15 a.m., that does not "disqualify" the finding at 6:00 a.m. However, if the color is measured at 6:00 a.m. a second time, say, the next day, and one does not get the same answer, this will establish that a firm statement of color at 6:00 a.m. cannot be given. The scientist may eventually learn that the true repeat of the 6:00 a.m. measurement of February 23, 2005 is not the measurement that was made at 6:00 a.m., February 24, 2005, but rather the measurement that will be made on February 23, 2006. We therefore learn that it is sometimes difficult to know whether an experiment is being repeated under identical conditions. But this example illustrates the need to approximate identical conditions, insofar as possible, for a "true repeat" measurement to be made. This is illustrated again later.

DEFINING THE SCOPE OF A SINGLE EXPERIMENT TO FURTHER DEFINE WHAT CONSTITUTES A REPEAT

Now that the experiment has been defined, it is clear that a single experiment encompasses multiple time points taken every 5 minutes during 24 hours. To repeat the experiment, therefore, a distinct 24-hour period would have to be used to perform the entire experiment again.

Why is such repetition required? If only a single day is measured, the scientist has only one data point for each 5-minute period analyzed. This time course can still help to get a picture of how the sky's color changes during the course of a single day. But there is no guarantee, or even reasonable likelihood, that the particular day chosen for the experiment will be representative of a "typical day," or that any one of the 5-minute time points will be representative of what usually happens at that time. Indeed, the terms typical or representative reference statistics; one does not know whether a day or a time point is typical or representative until sufficient data have been gathered to determine standard deviations and standard errors of the means, and calculate the statistical significance for particular outcomes.

Knowing that multiple measurements are required to do the appropriate statistics, the scientist decides to make measurements every day for 1 week. As we have discussed, this may not be the "best" decision, given the complexities of the question, but in this first experiment, the scientist decides that only a single week will be measured.

So far, in the course of designing the experiment, multiple decisions have been made:

1. To make measurements over a particular range of light wavelengths.
2. To point the instrument in the same direction for each measurement.
3. To measure the sky during the course of a particular time period (24 hours).
4. To make measurements at regular and particular intervals (every 5 minutes) during that time period.
5. To repeat the time course a sufficient number of times (every day for 7 days) to obtain statistical significance.

At the end of the week, the scientist will be able to apply statistics to determine whether the color of the sky at each time point is consistent (i.e., the standard deviation and/or the standard error from the mean is small or large) and whether statistically significant changes can be determined among the various time points. Now that these initial decisions on how to answer the question have been made, the scientist must put in place the relevant controls to ensure that the measurements produce an “accurate” answer.

SYSTEM VALIDATION AND THE ACCUMULATION OF CONTROLS

The types of controls listed for this type of experiment are formalized in later chapters, but for now, different controls are introduced based on the needs of the experiment so that the reader can see how experimental requirements create the demand for distinct types of controls. First, the scientist needs to have “instrumentation controls” in place to ensure that relevant light wavelengths can be measured (see Chapter 7). Briefly, positive controls for each color are gathered and the instrument is calibrated using these controls. Second, the scientist must ensure that the sky is in fact being measured; this probably requires simple observation and a positive control that the instrument is receiving data from the sky, even if it is not reading out as the particular color being interrogated. For example, if the instrument is set on “brown” and does not receive any indication that “brown” is being measured, it is important to know that the measuring system is not broken. Thus, positive controls for the measuring system might be the simultaneous measurement of blue and black (because the sky is almost always either blue or black) and such controls would prove that the system was operational. As a second positive control, to prove that the system was capable of measuring brown, an object known to be brown would be used, and the system would be tested on that object.

Notice that the selection of these particular controls is possible only after gaining some understanding of the system. This again illustrates how experiments are naturally improved as one learns more about the question. The more information that can be

accessed, the more one can put in place controls to ensure that new data can be gathered. Note, however, that this point may be deemed contradictory to the theory of Critical Rationalism, because the philosophers who belonged to this school rejected the idea of past experience as being predictive of the future. They insisted, instead, that “things could change,” and even though gravity, for example, had been extant for as long as anyone could remember, that was no guarantee of its existence in the future. This argument was a primary motivator for Karl Popper to advocate the hypothesis as a vehicle to falsify an idea: If it is impossible to use past experience to predict the future, it is impossible to say something is true as a way of saying it will happen again. However, it is the explicit project of science to deduce “laws of nature”—an understanding of how things operate—as a way of determining how they will operate in the future. This philosophical issue is discussed in more detail in Chapter 9. For now, it is important to understand that positive controls make use of prior validating experiments on a particular reagent, so that this reagent can be used to validate the current experiment.

To continue with the choice of experimental controls, it may be helpful to return to the list of decisions made for designing the first experiment, so that controls can be inserted for each point. Here is a list of decisions that the scientist made in experiment 1, with a summary of the methodology and controls to be used.

1. *To make measurements over a particular range of light wavelengths.* Each measurement of a particular wavelength will be validated with a positive control, using an object known to be a particular color, to ensure that the instrument can report the given wavelength. Each measurement of a particular wavelength will be validated with a negative control, using an object known to be devoid of that color, to ensure that the instrument does not give a false report. The integrity of the instrument will be validated by having it measure blue, black, and red at all times. If the instrument fails to measure one of these colors, the time point will be flagged so that the instrument may be inspected.
2. *To point the instrument in a particular direction.* The angle from the horizon will be measured and recorded; for example, the instrument may be set to measure at 45° from the horizon. The instrument will be kept in the same position for the full week. As a control, the instrument’s position might be triangulated using surrounding landmarks.
3. *To measure the sky during the course of a particular time period (24 hours).* A timer will be used to measure 24 hours. As a control, the scientist will regularly confirm the time using the Internet.
4. *To make measurements at regular and particular intervals (every 5 minutes) during that time period.* A timer will be used that can alarm every 5 minutes. As a control, the scientist will again verify the time using the Internet.

5. To repeat the time course a sufficient number of times (every day for 7 days) to obtain statistical significance. At the end of each day, the scientist will record the data and check off the day on the calendar. The experiment will be done the same way for each of 7 days.

GATHERING AND ANALYZING THE DATA AND INTERPRETING THE FIRST EXPERIMENT

The scientist records the color at a particular point in the sky every 5 minutes over 24 hours and repeats this process every day for 1 week. At the end of the week, the scientist maps out the sky color for each of the time points and obtains data that may look like this:

	1/1/07	1/2/07	1/3/07	1/4/07	1/5/07	1/6/07	1/7/07
6:05 a.m.	black	black	black	black	black	black	black
6:10 a.m.	black	black	black	black	black	black	black
6:15 a.m.	black	black	black	black	black	black	gray
6:20 a.m.	black	black	black	black	black	black	gray
6:25 a.m.	black	black	black	black	black	gray	gray
6:30 a.m.	black	black	black	black	gray	gray	red
6:35 a.m.	black	black	black	gray	gray	red	blue

Notice that the data are invariant for 6:05 a.m. and 6:10 a.m., but inconsistent at later time points (when sunrise begins). Experimentalists may interpret these data in several ways:

1. The data establish the different colors that are observable in the sky at different time points.
2. The data establish that the color of the sky changes over time, from black to red to blue.
3. The data establish that the sky is black at 6:05 a.m. and 6:10 a.m., and black in a statistically significant manner until 6:25 a.m. After 6:25 a.m., no firm conclusion can be made about the color of the sky.

From this particular experiment, only the third interpretation can be said to be “valid” because it is the only one that is statistically defensible.² Let’s look at the first two conclusions more closely.

²This is difficult to place in the correct context because explaining the actual statistics is beyond the scope of this volume. However, for the current example, assume that the data show that the sky is black to a statistically significant degree from 6:05 a.m. through 6:25 a.m., and that no other color reaches statistical significance.

ays) to
ord the
same

over
ek, the
: may

sistent
ita in

fferent

to red

ck in a
-lusion

l to be
: first

scope
ack to a
+tistical

1. *The data establish the different colors that are observable in the sky at different time points.* Actually, from a statistical framework, the data do not establish this. In this data set, blue is measured only once and red only twice. How can the scientist say that the blue measurement is correct for 6:35 a.m. on 1/7/07, given that this measurement is inconsistent with the other 6:35 a.m. measurements and no other corroborating data exist for the single time point? The only way the scientist can be sure that the sky was blue at 6:35 a.m. on 1/7/07 is if it was measured several times at 6:35 a.m. on 1/7/07 and if all those measurements were consistent. However, that was not the design of this particular experiment. The same is true of the red measurements; they too would fail to reach statistical significance, given this experimental design. Thus, even though the question is “What color is the sky?” and even though the scientist measured blue at one time point, the answer cannot be reported to be blue because blue was not measured at the same time point on different days.

However, the blue measurement need not be entirely discarded. The scientist may notice that the sky color seems to change in a recognizable trend, and this may form the basis of a new question.

2. *The data establish that the color of the sky changes over time, from black to red to blue.* A scientist might be tempted to draw this conclusion; however, it is not appropriate for several reasons. First, the scientist cannot yet be sure of the blue and red measurements, as mentioned above. Second, this conclusion is not responsive to the question “What color is the sky?” This is an important point and illustrates how science works. Suppose that from the observation of the change in sky color, which occurred while addressing the question “What color is the sky?” the scientist obtains data that indicate the possibility that the sky color is not fixed. An orderly transition occurs over time during the day and shifts during the week (and later, the scientist has an inkling that the pattern might follow a yearly periodicity). These first findings set up a rather exciting potential model for what may be happening in terms of the Earth’s position relative to the sun. However, it is not appropriate to draw a conclusion along these lines or to incorporate these ideas into a model until a question is asked and statistically significant, representative, and repeated data are gathered that allow such a conclusion to be drawn.

In a hypothesis-testing framework, one may use the above data to formulate a hypothesis that the sky color changes from black to red to blue, and the exact timing of these changes varies from one day to the next. This new hypothesis would then have to be tested in a new experiment designed for the purpose.

We see, therefore, how statistics can help to create defensible conclusions, how the question (or hypothesis) functions to limit the scientist to drawing conclusions that are responsive to the framework, and how to defer other conclusions for distinct experimental settings designed to address other experimental projects that may arise, given the data. Even if the scientist is brilliant and infers from the first experiment the

whole picture of how the sky changes over time, it will not be possible to incorporate that inference into the model until the data are validated through repetition and within an appropriately designed experimental project. Thus, a new project must be established, framed by questions such as "Does the sky color vary?" and the sky color measured again over a 1-week period to see whether the same results occur again. If they do, the scientist would have to perform the same experiment several more times, until the transition from black to blue and the variation of the timing of sunrise could be validated by statistics. Only then would the scientist be able to incorporate conclusions into the model that are responsive to this particular experimental question.

Without belaboring the "What color is the sky?" example, a simple perusal of a small sample of data helps to demonstrate the many ways in which the question can be answered:

1. At a particular time point, giving the range for that time point.
2. As a "norm" for a particular day, giving the most common color that was observed and validated over time through repetition (if the scientist was unfortunate enough to use this approach, a conclusion would never be reached!).
3. As a complete data report, giving the total data set as the answer.

Each of these methods of processing data may be viewed as valid, but the reader must understand that the answer as presented using method 1 may be "blue" or "black," and caveated as being valid for only a particular time, whereas the answer using method 3 will be more complex. Although it is difficult to prescribe a "best" way to present the data, it may be said that the preferred way is to give an answer that conforms to the understanding of the question and that most accurately represents the data set.

9



Validating a Model

The Ability to Predict the Future

THROUGHOUT EARLY HISTORY, PEOPLE WHO COULD predict the future were referred to as “wizards” or “magicians.” However, once empirical science developed, people realized that there were certain “laws of nature,” and that, when one understood these laws, one could do more than simply describe how reality had behaved in the past; one could also predict how reality might behave in the future. For example, nobody is called a magician or a wizard today for predicting that when an individual tosses a ball into the air, it will fall to the ground. Despite the fact that we cannot see the cosmic strings¹ that comprise gravity, we have sufficient experience and “confidence” in gravitational force—it is a “statistical truth” that when someone drops a rock, it will proceed in a predictable direction. One can say that the ability to predict what will happen to the rock is based on the knowledge gained from a “model” as to how gravity functions: On Earth, when an object containing mass is released, it falls toward the ground unless it is opposed by some other force (such as wind blowing a leaf upward, in opposition to gravity).

The ability to predict the future is the actual goal of all experimental science, although this goal is probably recognized more in some disciplines than in others. In the case of meteorology, it is the daily purpose of weather forecasters to tell us whether we will need to carry an umbrella when we leave the house the following morning. We can judge the accuracy of a weather model by determining on a daily basis how well the meteorologist foretold the day’s rain or hail on the evening news the night before. Thus, we do not trust the meteorologist as a matter of faith; rather, we act as empiricists and behave the way that we do because our experience has shown that these people can actually predict the weather. In a sense, members of the population act as scientific reviewers of an experimental model, and they indicate with their actions that they now accept the model as generally accurate because it has proven to be so.

Every day, people engage in probably unrecognized, ready examples of predicting the future, acting in accord with an acceptance of a scientific model. If a person living

¹That is a metaphor, not a literal invocation of string theory (M theory), which is a theory not easily subject to experimentation, and therefore quite problematic for the empiricist.

during the 1800s had been presented with an airplane weighing 41,000 kg and invited to trust his life on its ability to fly, his answer would probably have been “no.” Yet today, thousands of people board such vessels every day, with full confidence that they will be lifted into the air in a steady, controlled fashion. The scientific model that predicts that the plane will fly is based on experiments in aeronautics; the amassed data allow engineers to calculate the amount of engine thrust and “lift” required to take a particular object and send it skyward. The data gained from past air travel are obviously sufficient to predict the future; otherwise, we would have airplanes falling from the sky on a regular basis.

Although the goal of predicting the future is understood as a “given” in many settings, this goal is perhaps not appreciated as the method for validating all scientific models. Indeed, a built-in requirement for statistical significance is that a certain outcome be repeated a number of times, so that distinct events can be compared, in order to determine whether a statistically significant separation exists between the groups participating in a particular study. Another way of phrasing this requirement is that there is a statistical need for events in one group to behave in a similar way, and that this behavior will be sufficiently distinct from the way a second, control group behaves. We might think of this requirement for multiple confirmatory events as just another way of saying that we must be able to demonstrate that the model represents how reality will behave.

IN THE QUESTION/ANSWER FRAMEWORK, WHEN DOES ONE DETERMINE WHETHER THE FUTURE VALIDATES THE MODEL?

After the experiments that address a scientific question have been completed, a model is built and questioned for its accuracy. If the question is “What color is the sky?” and the model is a detailed timetable for sky color and how it varies minute by minute, we can test this model by asking whether the model is accurate. We might then compare the sky color the next day at various times, check our observations against the model, and gain a statistical answer to the question by noting how many times the model was correct versus how many times it was incorrect.

The ultimate test of any model is to ask whether it predicts reality. If the question is “What color is the sky?” and the model answers that the sky is blue, then, if we ask whether the model predicts reality a sufficient number of times in different experimentally relevant settings (such as during night-time hours), we will see that for more than half the time, this model is not accurate. We will thus need to caveat the model as to the times that the sky is blue. However, if we only ask the question when the sky is in fact blue, we will fail to correct the model.

The reader might object that establishing this requirement for the scientist (i.e., he/she needs to be able to answer the question “Does the model predict reality?”) is just as biasing as the need to test a hypothesis. In each case, the scientist has a stake in the outcome and may therefore be motivated to regard the data in a particular way, searching for confirmation while filtering out conflicting data. However, there are sev-

eral distinctions between the requirement to falsify a hypothesis and the need to confirm a model. The most telling distinction is that a model is built on experimental evidence. The model that the sky will be blue at noon tomorrow must be based on sufficient experimentation to justify that statement as statistically accurate. In contrast, the hypothesis must be proposed in advance of experimentation, in the absence of induction. Therefore, on a mathematical basis, there is obviously a much greater chance of falsifying the hypothesis—at least in some part—than falsifying the model, which has already been grounded on statistical confirmation. Thus, given an equally biased scientist, the odds that he/she will have a “need” to be biased are significantly reduced in the question/answer/model framework when compared to Critical Rationalism.

The second important distinction concerns the mandates of each framework. The formal requirement of Critical Rationalism is that the hypothesis be held up for falsification. As discussed previously, this creates tension between the scientist’s need to show something to be “true” and the framework’s rejection of that possibility. In contrast, the model can be accepted as “accurate.” If a model is demonstrated to be inaccurate in a particular instance, a modification or caveat of the model is generated, and this becomes the setting for new research. Thus, the model structure is more maleable than the hypothesis.

THE USE OF THE MODEL AS A BASIS FOR PREDICTING THE FUTURE

David Hume famously declared that just because the sun rises every day, there is no logical reason to believe that it will rise again tomorrow. It was out of a substantial embrace of Hume that Popper responded by offering a construct—the hypothesis—that could be held up for falsification; he virtually abandoned the idea that past descriptions of how reality operates might be representative of reality in the future. However, we argue here that it is precisely because the sun has risen every measured day² that one can construct a model that the sun rises every day. And if, when questioned for its validity, that model is proven to describe the future accurately (when the sun does in fact rise the next day), one can be rationally confident that the sun will rise again the following day.

This operational step of translating the experience of the past to a description of reality that can be verified as accurately describing the future is not the same as the failure to falsify a hypothesis. The verification of the model is an act of inductive reasoning, and that act can be validated, again, by asking whether the model is an accurate description of what will happen. The Critical Rationalist may insist that this reliance on the past is a false predictor of what might occur because nature may change at any moment. Yet the inductive scientist may be comforted by the empirical

²Or on enough days to reach statistical significance.

fact that the laws governing reality have not changed demonstrably over measured time and that models based on such experience have been shown to accurately represent the future.

For every failure to predict the future, the model can be refined until it represents what will occur with increasing accuracy. As long as a model functions to verify how the subject will behave in a demonstrably consistent manner,³ it is verified. Lest that sound circular, we must point out that statistical significance is gained only if the model is shown to be verifiable many more times than it is not, and the instances for which the model is not verified will help to define the limits of the model's predictive power.

Another example in which past experience inexorably predicts the future allows us to construct a relevant model. If you look at a large beach ball, observe that the ball is red, and then close your eyes, you might ask whether the ball will still be red when you open your eyes again. You may predict that the ball will be green, but when you open your eyes, the ball will still be red. You may repeat this activity over and over again. At some point you will settle on a model that the ball is red. Each time you question whether the model predicts the future, you will find that it does. Furthermore, if at some point in the future, you ask whether your model that the ball is red is still correct, and you open your eyes and you find the ball is green, you may wonder at this result. But then if you learn that while your eyes were closed someone doused the ball with green paint, you will not wonder for very long, and you will not consider your previous model to be somehow counter to reality. Rather, you will adapt your model, and now say that the ball is green, where it was once red. Again, this model will be verifiable for as long as the ball is, in fact, green.

Our inability to predict some events, such as the exact instance at which cancer may occur, to the same degree of accuracy as we may predict the sun rising or the color of a ball that we have observed, is probably a function of the number of variables that affect the particular subject of study and of our current understanding of the relative importance of those variables. This, again, is demonstrable by the gradual improvement in our ability to predict those individuals more likely to get cancer. Thus, each time the Critical Rationalist objects to the inductive scientist's claim that measurement X implies that Y will happen, one need only ask whether, in fact, Y happens.

The nexus at which the philosophy of science and the practice of science must meet is in understanding why science need be practiced in a certain way. The fundamental need of the scientist is to study how nature operates so that he/she is able to learn how nature will operate. Without the requirement of being able to accurately predict the future, there is no difference between the practitioner of science and the practitioner of science fiction. For example, one might say that starting 1 second from now, gravity will cease to exist. This model can be shown to be in accord with everything

³This issue, assigning a definition for what constitutes a consistent or verifiable result, is the basis for statistical determinations.

that has happened in the past, because it contains nothing that contradicts past events, and yet it can also be shown that this model does not predict the future. One might argue that predicting such a radical shift as the end of gravity is indeed a contradiction of past experience, but radical empiricists would agree that such experience should not be used inductively, that is, an empirical event should not be used to construct a general rule. However, in this volume, it is argued that the ability to show successfully that a model represents how reality will behave in the future is a critical feature of validation and provides the means to distinguish between science and science fiction. The fact that gravity does not cease to exist is the means of invalidating the model—the fact that it does not accurately represent what will happen.

If we ask whether gravity will cease 1 second from a defined point, we can show that it does not, and we can do so over and over again. The model that gravity will end in the next second therefore does not help us to predict what will happen in the future. On the contrary, it gives us information that is demonstrably false and therefore actually decreases our ability to function effectively in the future. If we had accepted the prediction that gravity would cease to exist in the following second, we might have taken that opportunity to jump off a bridge, inappropriately confident that we would suffer no ill effect. It is therefore necessary to distinguish between science and science fiction because individuals need to know the consequences of actions—both their own actions and those in their environment—to function. All of the predictions an individual may make, based on their experience of reality, are examples of inductive thinking founded on a model of how reality operates.

One additional philosophical issue should probably be discussed before reducing the philosophy to scientific practice, and that is the issue of perception. Distinct from the philosophical strain traceable from Aristotle through Descartes, which accepts the existence of external objects and our description of their innate qualities, there is a philosophy of mind that is traceable from Kant. This notes that our experience of reality is governed by the limitations of our perceptions and by the workings of our individual minds. Because perception is demonstrably flawed, as are ideas of space and time,⁴ so too is our surety of reality, and whether it may be said to have “universal laws” that exist irrespective of whether they are accepted in the same way by different individuals. Thus, how can a scientist say something will be, when one cannot describe its essence, but only our perception of it? Obviously, dealing with this level of question to philosophical precision is beyond the scope of the volume, but it must be answered sufficiently to provide traction for the current project and an assurance that the project is grounded in reason.

One can do a few quick thought experiments to demonstrate that a scientific prediction of reality is relevant beyond the predictor. For example, it would be meaningless to others if a scientist said that smoking cigarettes causes cancer, or that

⁴Although the understanding of space and time has obviously changed dramatically since Kant wrote *Critique of Pure Reason*.

HIV causes AIDS, or that load-bearing exercise increases muscle mass if all of the findings were relevant only to the scientist, as filtered by his or her perception of things, and if there was nothing innate about the causal reality that the scientist has studied. Some transitive property of the inductive act must be demonstrated. In other words, it must be shown that the experimentalist's measurements and prediction of reality have meaning in themselves, based on the essence of the subject being studied, and are not dependent on something particular to the scientist or on the way he/she perceives reality. By "transitive property of the inductive act" we also mean that the scientist's prediction can be shown to be relevant to some other individual, whether or not the individual knows of the prediction envisioned by the scientist or whether the scientist's perception of the subject diverges from that of the individual.

As for the thought experiments, first imagine that you are a sharpshooter and that you are carrying a rifle loaded with tranquilizer darts into the deepest jungles of a tropical rainforest. Eventually, you spy a group of individuals and hide yourself from view. You point your tranquilizer gun at one of the individuals who has never seen a gun before and therefore has no frame of reference to understand how such a device might operate, much less what a tranquilizer might be. Standing next to the individual is an observer; this observer is an "objective observer" and capable of describing whether an individual is awake. You take aim at the individual's leg. Before firing your gun, you ask whether your model, that the tranquilizer dart will cause the individual to lose consciousness, will be verified. You then fire your gun. Despite the gun, the dart, and the tranquilizer being outside the prior range of the individual's reality, the individual's being is still perturbed, in a manner that is consistent with the model: The individual is reduced to an unconscious state for several minutes, as confirmed by the independent observer, you, and later the victim. Thus, the dart carried its reality to the individual, despite the lack of any demonstrable common frame of reference shared between the shooter and the individual.

In the next thought experiment, let us say that you are a virologist confronting a skeptic who denies that the poliovirus causes polio. You describe all of the evidence demonstrating that poliovirus does, in fact, cause polio, but the skeptic is still unconvinced. In this case, you are not dealing with someone who merely lacks experience with your frame of reference; you are dealing with someone who actively denies that your claims are based in reality. Yet it can be shown that this active disagreement is not sufficient to decrease that person's likelihood of contracting the disease. For example, if one studied the incidence of polio in people who accepted the evidence that the poliovirus caused polio and compared this with the incidence of polio in those who denied the causality, one may not be surprised to learn that the skeptics end up with a greater incidence of disease because they have no incentive to become vaccinated against the virus. This demonstrates that an active denial of a common frame of perception is not sufficient to alter the essence of the subject being studied—in this case, poliovirus.

In the final thought experiment, imagine yourself standing next to a stone wall. You are confronted by a philosopher who denies both empiricism and that physical

of the
ion of
t has
other
of real-
, and
rceives
entist's
: not
er the

that
s of a
f from
een a
evice
vidual
ether
n, you
o lose
, and
ivid-
e indi-
7 the
nty to
erence

ring a
idence
con-
e with
t your
suffi-
if one
iovirus
l the
r inci-
e virus.
not
is.
e wall.
sical

objects or laws have an innate reality. The philosopher compares the world to the movie *The Matrix* and opines that nature's objects are merely creations of the individual mind.⁵ You respond that if this is the case, the philosopher will not mind if you push his head into the wall, because the wall is not an object unto itself, but merely a construct that may not be shared between scientist and philosopher. Either the philosopher will be silenced as to the existence of the wall or the wall will silence the philosopher, thereby establishing its existence. In this final case, it is shown that the scientist's question of the predictive nature of his/her model can be demonstrated to be relevant to an individual, even if the individual rejects the scientist's entire metaphysical framework. The most intractable skeptic still cannot escape the essence of the stone wall or the force of the scientific model verified by questioning its ability to predict the future.⁶

The inductive scientist might thus be assured that it is possible to confirm a model as being an actual representation (and a project as being relevant to other individuals) even to those who might reject the ontologic basis of such frameworks and the premises underlying the chosen experimental method.⁷

In the example of an experimental project in Chapter 10, we determine the point at which the model is required to predict the future and how this helps to validate the model.

⁵The movie-goer may recall that, in the film *The Matrix*, a spoon could be bent using the power of the mind only when the observer was in the computer-generated world of the Matrix. In the "real world," the spoon could not be bent by the mind. In fact, the observer is told that the spoon is not bending while in the "Matrix," because the spoon does not actually exist. Therefore, the observation that physical laws could be violated was proof that the observer was not observing reality; conversely, the transfer to a setting in which physical laws were now obeyed was a sign that one was in "the real world."

⁶These examples are chosen because they must demonstrate that the essence of an object can impose itself even in opposition to one's denial of that essence. This seems best illustrated when the two claims are placed in opposition.

⁷Although this posture might sound aggressive, the proper comportment of the scientist can be thought of as being most successful in constructing an accurate model when it is characterized by an acceptance of the data as they present themselves, and in the construction of a model based on the data, without imposition.

10



Designing the Experimental Project

A Biological Example

IN BIOLOGY, THE MOST FREQUENT CONDITION under which experimentation occurs is the situation where the scientist asks a question on a subject for which relevant data already exist. An earlier chapter addressed the “state of nature” condition, where a subject is tackled for which no prior information is available, and described how information is gradually accumulated, until finally, enough is known about the subject to start asking higher-level questions. A more typical setting is that in which the scientist is armed with substantive prior knowledge about the subject. From a philosophical perspective, this experimental setting is the most problematic—it is only when prior knowledge is available that one can argue whether or not (or under what circumstances) it is appropriate to accept that knowledge, and whether by doing so the results will become biased or less likely to produce a model that conforms with reality. Let’s consider the following example that will serve as a metaphor for this experimental setting and follow it through until we can derive a model that builds on (or alters) the information as it was previously understood. The ability to produce a model that allows the scientist to know what will happen in the future, when the model is tested for its accuracy, will be the criterion for determining the success of the model.

EXAMPLE OF EXPERIMENTAL DESIGN: DETERMINING EcoRI’S RESTRICTION SITE

There is a class of proteins called “restriction enzymes.” These proteins have a discrete property: They digest double-stranded DNA containing particular sequences. For example, a restriction enzyme might digest DNA that contains the sequence GCTAGC. This restriction enzyme would not, however, digest DNA that is devoid of this sequence.¹

¹Restriction enzymes are produced by bacteria and are a defense mechanism against foreign DNA (e.g., from invading bacteriophage). The enzymes recognize and cleave the foreign DNA at sequences that do not occur in the host DNA, or that are somehow modified in the host DNA, for example, by methylation.

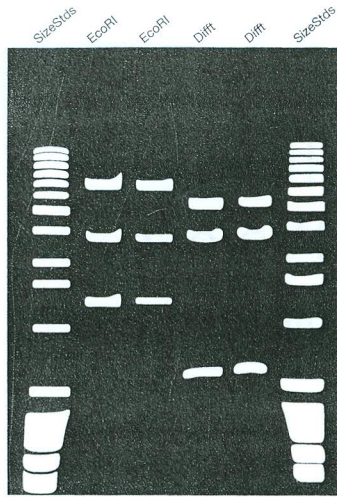


Figure 1. Electrophoresis of DNA. This process allows DNA fragments to be separated by size.

In this example of an experimental project, we follow the studies of a young graduate student named Benny, who works for a Principal Investigator named Marshall. Marshall had previously demonstrated that a protein called EcoRI² was probably a restriction enzyme by conducting experiments in which EcoRI was incubated with DNA. He discovered that the DNA was reduced from a high-molecular-weight mass of viscous goo to much smaller pieces. When Marshall separated the DNA fragments by size, using a process called electrophoresis, he could see that the DNA had been digested into many distinct pieces of discrete and reproducible size (Fig. 1). He also found that when he incubated the digested DNA pieces with a second enzyme called a “ligase,” the pieces were reassembled into high-molecular-weight DNA. This was sufficient evidence for him to construct a model for EcoRI, in which he noted that the EcoRI protein has properties consistent with restriction enzymes, because all previously characterized restriction enzymes behaved in a manner that Marshall observed for EcoRI.

These experiments were performed before Benny the graduate student entered Marshall’s laboratory. Now that Benny has arrived, Marshall has assigned him a particular project concerning EcoRI; Benny must figure out the precise DNA sequence that EcoRI recognizes and digests.

Benny is a smart and admirable individual and has therefore read the previous chapters in this book. He thus attempts to map out his Experimental Project as suggested in these earlier chapters. He sits down and tries to construct a question that will frame his Experimental Project. He writes the following question on the top of a piece of paper:

What DNA sequence does EcoRI recognize and digest?

He is not sure whether this question is correctly phrased, but he decides to map out his project by drawing a Venn Diagram. The first circle he draws confines the ques-

²This is an actual restriction enzyme.

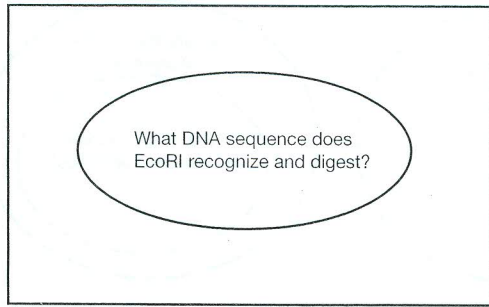


Figure 2. Benny's first question to frame his Experimental Project, using a Venn Diagram.

tion as he has written it (Fig. 2). But then, when he actually attempts to ask the subsidiary questions that will frame his first series of experiments, he gets stuck. He does not know what to do.

Benny therefore takes out a new piece of paper. Because EcoRI is a protein,³ Benny draws a large circle, and in that circle he writes "What are the functions of proteins?" (Fig. 3A). Within the first circle, he draws a second, smaller circle, and in it he writes "What is the function of restriction enzymes?" (Fig. 3B). Benny hesitates before placing EcoRI in this more restricted circle, because he does not know whether he has met sufficient criteria to call EcoRI a "restriction enzyme." He realizes he must read more about them, which he will do shortly.⁴

Let's now say that at the time Benny is doing experiments, 50 distinct restriction enzymes have been discovered and the DNA recognition sequence at which each enzyme cleaves has already been characterized. Benny adds four of these restriction enzymes to the diagram and notes that there is information on 46 more (Fig. 3C). Next to the Restriction Enzyme Circle, Benny adds in a circle for EcoRI; here, he uses a dotted line, to illustrate his desire to use information known about the other known proteins, to help figure out the answer to his question on EcoRI. This dotted-line circle contains the broader question, "What is the function of EcoRI?" (Fig. 3D). Here, Benny accesses the data that Marshall has already produced, which demonstrated that EcoRI could digest DNA into discrete smaller pieces and that this DNA could then be stitched back together by a ligase. These data allowed Marshall to construct a model for

³The skeptical reader will object that this is a premise that Benny has not proven. That skepticism, and how Benny handles it, is discussed shortly.

⁴The skeptical reader will object that Benny has accepted that EcoRI is a protein, but has deferred accepting that EcoRI is a restriction enzyme. As will be shown in this chapter, Benny has not rejected any prior knowledge; rather, he is adopting a structure that will allow him to accept any new data as "positive" and will also allow him to perturb the model, even if the new data are contradictory to the previous understanding.

If Benny wanted to make the skeptic at least a little happier, before placing EcoRI in the circle proscribed by the word "Protein," he could prove that EcoRI was a protein by digesting it with a protease and by showing that it was composed of amino acids, the building blocks of proteins. However, as was just argued, it is useless to proceed in this way, because the skeptic will never be satisfied. Benny's only obligation is to perturb the model as necessary, given the data he accumulated in response to his questions.

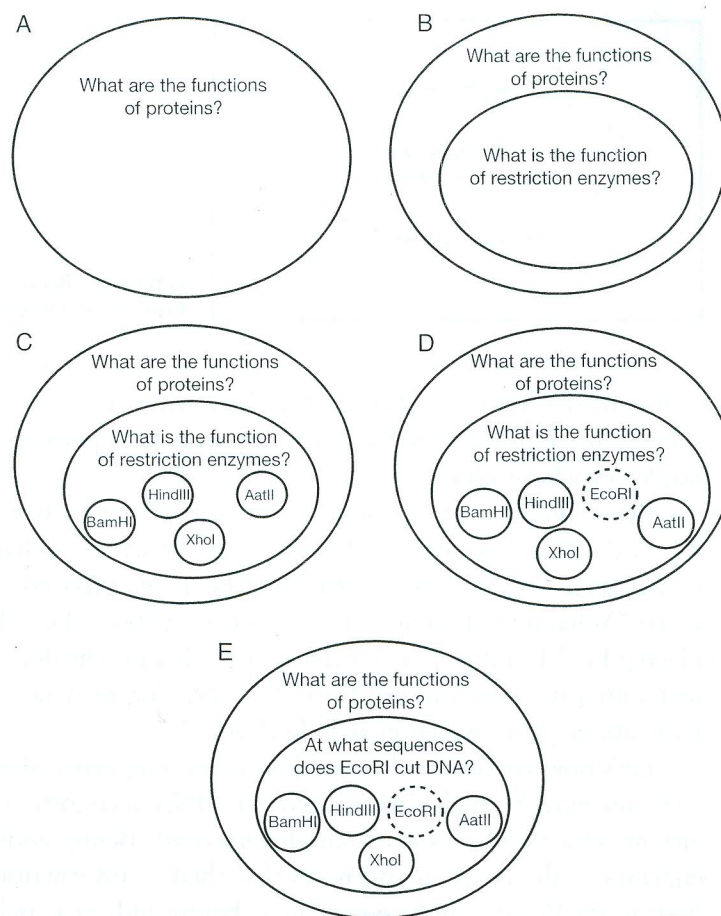


Figure 3. Process of framing the experimental project, starting with the broadest open-ended question (A), to define the inductive space, and proceeding to a more constrained area (B), to focus the issue, taking advantage of relevant data (C), and positioning the unknown appropriately (D). The question can be refined as necessary (E).

EcoRI—that it cuts DNA in such a way that a ligase can put it back together—and prompted him to ask Benny to determine where in the DNA EcoRI was cutting. Having gone through this logic, Benny now changes his question, to the following open-ended format:

At what sequence(s) does EcoRI cut DNA? (Fig. 3E)

Compare that question to the previously phrased question:

What DNA sequence does EcoRI recognize and digest?

The previously phrased question contained premises that were unproven. For example, in the previously phrased question, it is accepted that EcoRI recognizes (or binds) DNA and digests it at a particular sequence (since “sequence” is in the singular). In the new question, “At what sequence(s) does EcoRI cut DNA?” the premise of the

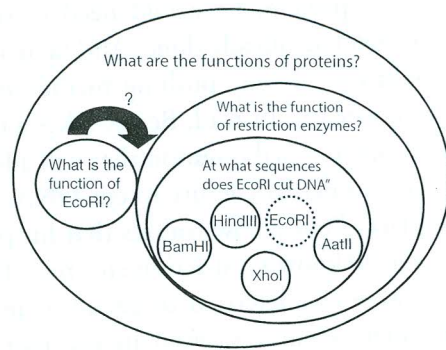


Figure 4. It is useful to remember to allow for the unexpected, such as the possibility that EcoRI might have functions besides (or other than) being a restriction enzyme.

question—that EcoRI cuts DNA—has been proven, and the scientist is now asking the follow-up question, to determine where the DNA is being cut.

The reader may think the distinction to be rather small, but it should be emphasized that care must be taken in insulating the scientist from unproven premises. These premises will cause the scientist to filter data in the same way as the unproven hypothesis.

Let's now take a second to look at the structure that Benny has put in place (Fig. 4). He has not immediately jumped to the question "At what sequence(s) does EcoRI cut DNA?" even though this is the question that Marshall assigned him to answer. Instead, Benny has written a broader, first question to frame his project "What is the function of EcoRI?" and he has placed that next to the class of restriction enzymes, because there are data suggesting that EcoRI does in fact belong to this class of proteins. His first question (What is the function of EcoRI?) will allow him to accept any evidence of EcoRI function as "positive"; he already knows that EcoRI has some function, because Marshall has at least shown that it reduces large-molecular-weight DNA into smaller pieces. His second question accepts that EcoRI cuts DNA, but he has attempted to write this question in such a way as to accept any sequence in which EcoRI cuts as "positive." He has not confined his second question by the properties of known restriction enzymes, even though he is aware of them and even though he will use methodologies that were used to study known restriction enzymes in analyzing EcoRI's sequence.

A skeptical reader might ask why Benny seems to have accepted some prior knowledge (that EcoRI is a protein that decreases the size of DNA) but has rejected other prior knowledge (that EcoRI definitely belongs in the class of proteins called Restriction Enzymes, which is what Marshall believes). Why doesn't Benny begin his project by rejecting all prior knowledge? This is the crux of the matter—Benny is not rejecting *any* prior knowledge with the structure as written; he is placing his project within a context that will allow him to move forward quickly, accepting the results that Marshall obtained. However, he is framing his project with an open-ended question that will be answerable even if it does not conform to the received model.

Simply put, scientists never reject prior knowledge when they start their projects. If they did, there would be no progress—each scientist would have to re-derive all previously gathered data before they could move forward. Therefore, if Benny were

question (A), to
king advantage
as necessary (E).

together—and
was cutting.
the following

. For example,
s (or binds)
singular). In
premise of the

forced to reject the idea that EcoRI was a protein, he would need to spend time repeating all of the experiments that Marshall had already done. And he would not be able to stop there; the true skeptic would have Benny pushing further and further backward until he has rejected all prior knowledge. Instead, Benny takes a much simpler precautionary measure; he adopts an open-ended question that will allow him to accept any answer he receives as it pertains to EcoRI's ability to cut DNA.

With this Experimental Project in place, Benny now realizes that his project has already been successfully completed for the 50 known restriction enzymes. He decides to check out how these other enzymes were characterized to get ideas on his own experimental approach. If a particular methodology worked in the past, and has resulted in an answer that has been well validated, then Benny feels he would be well served to know that in advance.

Now may be a good place to pause, to continue discussing the objections to Benny's approach. As noted, what Benny has done in constructing Figure 3, and by deciding to read about other restriction enzymes, is the sort of thing that Critical Rationalism was trying to avoid; the reason to avoid the process as outlined is that the instances of prior experimentation may have led to a picture that is either wrong or incomplete. Therefore, using prior knowledge as a guide for designing new experiments might be thought to guarantee repetition of any mistakes that were previously made.

Without repeating all of the arguments made in Chapters 1–7, we can make some quick points here regarding Benny's particular Experimental Project. First of all, it should simply be pointed out that a hypothesis that predicts that EcoRI will be found to cut DNA sequence is not practical, given that Benny does not even know the size of sequence he is looking for. There are literally millions of possibilities and, even if Benny knew in advance that EcoRI recognized a sequence composed of six DNA base pairs (bp), he could still be left with 4096 possibilities⁵ and therefore picking any of those to hypothesize as the correct cut site for EcoRI is almost guaranteed to be falsified, leaving Benny with very little progress.⁶ This is a very important point for the reader to understand: In this example, given the point where Benny steps into the project, a hypothesis simply is not practical. In this instance, the hypothesis would need to take on the following structure:

EcoRI cuts a DNA sequence.

We know that this hypothesis is actually correct, because Marshall has already demonstrated that EcoRI cuts DNA, and this action results in DNA pieces of

⁵There are four deoxyribonucleotides that comprise DNA: A, C, G, and T. Given that any of the four possible base pairs can be chosen at any of the six sites in the potential DNA recognition sequence, there would still be 4⁶ possible cut sites for EcoRI, or 4096 possibilities.

⁶One could sequentially test each of the 4096 possibilities, if one knew that the cut site was 6 bp long. But in the absence of that knowledge, when there are millions of possibilities, this sort of methodology is not practical. Even if the possible answer set was limited to 4096, why go through the iterative process that forces one to falsify several thousand possibilities, when the answer can be obtained more directly, by asking which sequence is in fact EcoRI's cut site? It will be shown that either framework, if used correctly, can arrive at the "correct" answer.

specific sizes. Therefore, this is a case that is quite often seen in the literature, where a hypothesis, for which the result has already been determined, is chosen out of necessity. Therefore, one might safely say that this hypothesis is not helpful in pushing experimentation in a certain direction; rather, it simply fulfills a bureaucratic obligation. Put another way, the failure to falsify the hypothesis that “EcoRI cuts a DNA sequence” does not, on its own, help the scientist to determine the answer to the question “What DNA sequence does EcoRI cut?” It is not until a lot more information is gathered and Benny has actually determined the answer, using his Question/Answer approach, that in the final, binary, Question/Answer stage, a hypothesis can be comfortably substituted for the “real” question. We will highlight the stage at which this could happen and point out again that this has demonstrated that inductive reasoning always seeps into the Critical Rationalist framework. Furthermore, even if a hypothesis were chosen in the present case, the methodology used would probably still be the same as that used for the known restriction enzymes. This is also how prior knowledge, and inductive reasoning, enters into Critical Rationalism; the hypothesis does not prevent the scientist from using established methodologies—in many cases, such a requirement would preclude experimentation. Finally, in the present case, Benny is not making any assumptions about the restriction site(s) of EcoRI—he is resorting to prior experience to reference successful methodologies, as opposed to possible conclusions.

We can see the prior knowledge that Benny is relying on in the circles he drew in Figure 3. To speed progress, he must know more about restriction enzymes; that is the “inductive space” he will use to ask questions about EcoRI. The only way those questions would fail to result in answers would be if Marshall had made a mistake in his conclusion that EcoRI restricts DNA. If that had happened, Benny would find it out soon enough if he used a properly validated system to do his experiments.

ACCESSING THE INDUCTIVE SPACE: LEARNING WHAT IS ALREADY KNOWN ABOUT THE EXPERIMENTAL SUBJECT

Because it was not practical to access all prior knowledge about every protein, the Venn Diagram that Benny has drawn (Fig. 3E) now focuses his attention on the particular subfield of restriction enzymes. So that the reader can follow Benny’s progress, the particular, relevant characteristics of restriction enzymes will be explained, and we will then see how this prior knowledge helps Benny to shape his experimental approach. We later ask whether the methodology used here (i.e., accessing knowledge about known proteins to set up a particular experimental method) caused Benny to miss something, by introducing the “necessity versus sufficiency” test. Here is what Benny learns about restriction enzymes:

1. Restriction enzymes cut double-stranded DNA at distinct sites. They usually, but not always, recognize DNA that has a palindromic sequence, i.e., a sequence that

reads identically (from 5' to 3')⁷ on each strand. For example,



Although not all restriction enzymes cut at particular palindromic sequences, a small minority of restriction enzymes have sites that are not palindromic.⁸

2. The number of base pairs that comprise a restriction enzyme's "cut site" varies. Some restriction enzymes recognize only 4 bp of DNA, whereas others require 6 bp or more.
3. Most restriction enzymes have only one cut site, but there are exceptions. For example, *AccB7I* cuts at the sequence *CCAXXTGG*, where *XX* represents a variable sequence. This is not very common, and even in this example, the enzyme recognizes the invariant flanking bases, and cleavage always occurs at the same place. Other restriction enzymes have a second, "less preferred" site that is recognized only under certain conditions. This is rare, but it does occur.⁹
4. Some restriction enzymes recognize a specific sequence, but cleave the DNA elsewhere. For example, a restriction enzyme may bind at *CCCGGG*, but the enzyme then reaches out and cuts the DNA 4 bp downstream from the last G, irrespective of the nucleotide it finds there. If *EcoRI* is like this case, finding a sequence necessary for *EcoRI* to function may not represent the actual cut site. It is only by obtaining this prior knowledge that Benny is motivated to rethink the structure of his question once again. His question as originally phrased was

What DNA sequence does EcoRI recognize and digest?

But now Benny has learned that the digestion site for *EcoRI* may be distinct from the "recognition" or binding site. He realizes that the original format of his question contained an unrecognized assumption inconsistent with reality in several demonstrated cases. His second attempt at formulating his Experimental Project resulted in this question:

At what sequence(s) does EcoRI cut DNA?

However, if *EcoRI* is like the restriction enzymes that cut at a site distinct from the binding site, the answer to this question will not be meaningful (because in this

⁷The notations 5' and 3' refer to the structure of DNA; some prior knowledge about DNA and basic biology is assumed in this text. Readers who lack that knowledge can refer to a chapter on DNA in a basic biology text.

⁸... which illustrates a reality: Absolutes are rare, which is another reason why models work best when they are constructed to allow for the caveats and exceptions that almost inevitably crop up.

⁹This information was not available when *EcoRI* was discovered. The example as outlined here illustrates how inductive reasoning may be accessed; this is not a historical example demonstrating how *EcoRI*'s specificity was solved, although it could serve as a metaphor for how a new restriction enzyme discovered today could be studied.

case, EcoRI will actually cut at any sequence, as long as it is 4 bp downstream from the binding site; the format of the question does not really capture this possibility). Formally put, Benny needs to define his terms and figure out what he actually wants to discover. He consults his Principal Investigator, Marshall, who confirms that the goal is to figure out *both* what DNA sequence EcoRI recognizes and where it cuts. Benny reexamines his question and decides to rewrite it in the following way (Fig. 4):

What DNA sequences are sufficient for EcoRI to cut DNA?

The concept of something being “sufficient” to result in a particular event, and why that is important, is discussed later on. In the present case, it should be clear that finding the sequence required for EcoRI to digest DNA encompasses both the binding site and the cut site for EcoRI. Note that Benny was able to grasp this level of complexity only because he had read about other restriction enzymes. As will be shown, even if Benny did not know about this particularly troublesome class of restriction enzymes, a validated system would still have allowed him to get to the “correct answer,” eventually. There is no question that he will be able to get there faster by having learned to take into account other, previously demonstrated, possibilities that he would not have thought of on his own.

5. Different restriction enzymes vary as to the type of cut they make in the DNA. For example, some cut straight through the DNA, leaving two “blunt” ends:



Other restrictions enzymes zigzag through the DNA, leaving “sticky ends,” so named because they are easy to piece back together (ligate):



The two single-stranded 5' CATG 3' sequences that are hanging off both ends can easily be stitched back to each other again, with the aid of hydrogen bonding and the previously mentioned enzyme called a “DNA ligase.” Blunt ends are somewhat harder to religate because there is no opportunity for hydrogen bonding between the ends.

Benny realizes from learning this that his question does not force him to determine the end structure of fragments digested by EcoRI. His question is not

Does EcoRI digestion leave blunt or sticky ends?

Benny asks Marshall about this. Marshall is impressed that Benny has done some reading and admits that he did not think about the end structure when giving

Benny his assignment. He tells Benny that when he is done with his first project, he can then follow up by tackling this issue.

We learn from this example that the scientist must make a choice as to which information is relevant in answering the question. This brings us back to the issue of “defining terms.” Benny could have defined the subsidiary question “What is the DNA cut site for EcoRI?” as including the question “What is the DNA structure left upon EcoRI digestion?” Instead, he decides that his project is simply to determine where EcoRI cuts, not the structure of the DNA left by the enzyme. These may be issues taken up once he answers the questions as currently stated and understood.

6. Different restriction enzymes require distinct experimental conditions in order to function. Some enzymes require 100 mM sodium chloride to cut DNA, whereas others cannot function under these conditions. Almost all restriction enzymes function better in 10 mM dithiothreitol (DTT). Most restriction enzymes function best at 37°C, but they are denatured at 65°C.

By learning these details, Benny realizes even more clearly that he needs to “establish his system.” He must make sure that he is testing EcoRI under experimental conditions that will allow it to function. Fortunately, Marshall has already figured these out for EcoRI.

The skeptic might stop again here and ask whether the finding of a particular experimental condition, under which EcoRI is capable of cutting DNA, is the same as answering the question “What DNA sequences are sufficient for EcoRI to cut DNA?” Just because an “artificial” system can be constructed that allows EcoRI to function on DNA of a particular sequence does not mean that this is the way EcoRI functions in bacteria. This is an important point, to which we return later in this chapter. Note, however, that the initial question, as stated, does not force Benny to this level of rigor. That comes later.

DEFINING TERMS

Having read about restriction enzymes, and achieved a new level of understanding about potential issues that might be considered, Benny reexamines his question to define his terms. He must do this to ensure that his experiments will produce data that will contribute to a model that is responsive to his question. Here again is his question:

What DNA sequences are sufficient for EcoRI to cut DNA?

He then defines his terms:

1. DNA is double-stranded deoxyribonucleic acid.
2. A “sequence” refers to a particular order of DNA base pairs. “Sequences” is plural in structure. However, most restriction enzymes only cut at a single DNA sequence.

project,

to which

the issue

“What is

[A struc-

simply to

e enzyme.

ated and

order to

whereas

enzymes

es func-

needs to

r experi-

s already

articular

the same

RI to cut

EcoRI to

y EcoRI

ter in this

benny to

erstanding

stion to

data that

question:

is plural

sequence.

This does not trouble Benny, because the question, as written, will allow him to find a single sequence and also force him to ask whether that single sequence is sufficient to answer the question or whether there may be more out there to discover.

3. Sufficient means “enough.” Therefore, he will need to find a sequence that comprises the DNA that allows EcoRI to operate. After Benny determines this, he can then ask the follow-up questions that discriminate between binding and cutting.
4. EcoRI is a particular restriction enzyme. The abbreviation refers to the fact that it was the first (I) restriction enzyme to be isolated from the bacterium *Escherichia coli*.
5. “Cut” means to sever DNA into smaller pieces, by breaking the phosphate ester bond that joins nucleotides in a linear strand.

USING KNOWLEDGE ABOUT THE SYSTEM TO SET UP A SERIES OF QUESTIONS AND A METHODOLOGY THAT CAN ANSWER THOSE QUESTIONS

After reading the background literature concerning other restriction enzymes, and after having defined his terms, Benny decides on a methodology based on the needs of his project. This methodology will begin with validating a system. However, to know which system he needs to validate, he must sketch out his methodology—the steps he will go through using his system—to answer the question. The methodology can be thought of as the “experimental design.” It encompasses the use of a validated system and the conduct of an experiment. To ensure that the system will be valid for the Experimental Project, it is helpful to sketch out a series of experiments. It is not necessary to design in detail all of the experiments in advance; sometimes, a particular experiment becomes necessary for which the system was not validated. In that case, the system must be validated for the new experiment. Benny decides he will need to do several things to establish his system and answer his question “What DNA sequences are sufficient for EcoRI to cut DNA?”

1. *Obtain a piece of DNA that can be digested by EcoRI* (we will call this piece of DNA “Lambda”). This will be used as the test material to discover the cut site for EcoRI. We can think of this piece of DNA as being part of the “system” that must be validated to answer the experimental question; in other words, Benny will need to prove that he can cut this Lambda DNA with EcoRI to do his experiments.
2. *Obtain a piece of DNA that cannot be digested by EcoRI* (we will call this piece of DNA “Theta”). Benny will use this DNA as part of his system as a “negative control.” He will insert potential EcoRI cut sites into Theta to determine whether he can then digest the DNA, once the test sequence has been added into it.
3. *Obtain one or two of the known restriction enzymes.* Marshall has already shown that two other enzymes work under the same experimental conditions as EcoRI,

so Benny will want to use these enzymes as “positive controls” for the buffers, water, and other materials to test EcoRI. These other enzymes will thus help Marshall to validate Benny’s system. This is probably the “core” validation step; he must prove that he can actually use EcoRI to cut the Lambda DNA, so he will need to have controls for that process, that is, other enzymes that work under the same conditions as EcoRI.

4. Once the EcoRI digestion method is in place, and Benny has shown that he can use EcoRI to cut Lambda and get a reproducible readout (a set of smaller pieces of DNA of reproducible sizes), he will need a method to answer the question that frames his Experimental Project: “What DNA sequences are sufficient for EcoRI to cut DNA?” Because it is now very easy to sequence DNA, Benny decides on a process that asks the following questions:
 - a. What is the sequence of Lambda DNA?
 - b. What size pieces of Lambda DNA are produced by cutting with EcoRI?
 - c. For each piece of Lambda cut by EcoRI, what are the sequences of the severed ends?
 - d. Does Lambda contain any repeated DNA sequence that, when cut, would result in the sizes obtained using EcoRI, and which contains the DNA sequences determined in Step c?
 - e. If the answer to Step d is yes, what is that sequence?¹⁰
 - f. Does the sequence obtained in Step e, when inserted into Theta, allow EcoRI to digest Theta?
 - g. Can we now say that the question “What DNA sequences are sufficient for EcoRI to cut DNA?” is answered? In other words, does the discovered sequence account for all of the cuts made by EcoRI in Lambda?
 - h. If the answer to Step g is no, what are the remaining cuts not accounted for by the discovered sequence?

ESTABLISHING THE SYSTEM

Before Benny can proceed to ask these questions, he must make sure that he has a functional system. If he had a defunct sample of EcoRI (if someone had accidentally exposed his EcoRI to 65°C or had contaminated the sample with a protease), he

¹⁰If the answer to this question is no, then EcoRI may be in the rare class of restriction enzymes that cut at more than one DNA site. In that case, Benny will have to change his methodology to account for this possibility. A complete listing of the experimental methodology for this project would include the “no” case for this question.

would not get very far. If, even under the most careful conditions, the activity of EcoRI degraded over time and needed to be prepared “fresh” each time, Benny would need to know this. Furthermore, even if he has an active enzyme, he may not have a workable system. For example, if he makes a mistake in preparing the buffer required—if he accidentally puts in 100 mM potassium cyanide instead of 100 mM sodium chloride—he would not obtain any data because EcoRI would not function.

In addition to making sure that he actually has active EcoRI, Benny must make sure he can answer the experimental questions listed in his methodology. We list again each of the questions Benny wants to answer and this time include what he will need to validate in order to perform these experiments. Next, we actually map out the experiments Benny will do to answer each question, including the relevant controls. Note that in each case, we do not offer a hypothesis. Somehow, however, data are gathered and a model built:

a. What is the sequence of Lambda DNA?

To answer this question, Benny must be able to sequence DNA. To make sure that he is accurately sequencing Lambda, he will need to sequence a piece of DNA whose sequence is known, as a positive control. To make sure that his buffer is not contaminated with some random piece of DNA and that he is truly sequencing Lambda¹¹ and not this contaminating DNA, he will need to try sequencing the buffer alone as a negative control. So he ends up attempting to sequence not just Lambda, but also a piece of known DNA as a positive control and his buffer as a negative control. In this way, Benny is establishing that he can sequence DNA (because he should be able to get the known sequence for the positive control) and that the sequence he gets is actually Lambda and not some contaminant. One might notice that by using the correct controls, Benny is establishing a system for answering the question that frames his experimental project and, in the process, establishing and validating subsystems for each of the individual experimental questions. (In other words, one cannot answer the experimental question “What is the sequence of Lambda DNA?” unless it has been validated that one can sequence DNA.)

To incorporate other experimental guidelines that have been discussed (or that will be discussed in subsequent chapters), such as the need to both repeat experiments and address questions in more than one way, Benny will sequence both strands of Lambda. Without going into unnecessary detail, sequencing methodology results in multiple, overlapping readouts of DNA sequence, which allows the sequence to be confirmed multiple times. The sequence of the second DNA strand can also be established and used to check the accuracy of the first strand

¹¹Again, various controls are being introduced in these examples ahead of the chapters that actually list out and formalize these controls. This may seem backward, but the idea is to give the reader a flavor for how controls are used, and why they are necessary, so that by the time we get to the descriptions, there will be an appreciation as to why they are so important.

sequence.¹² This is a powerful check, because, if correct, the sequence obtained from the second DNA strand will result in a sequence that is absolutely complementary to the first DNA strand; if it does not, then the sequence in question must be reanalyzed. Recall the procedural checklist for conducting experiments, as modified in Chapter 6:

1. Decide on an experimental project.
2. Ask a broad question to frame the experimental project.
3. Ask a subset question to frame a specific experiment to start deriving data that will be used to answer the question in Step 2.
4. Answer the subset question.
5. Determine whether the answer is accurate by asking whether the derived answer represents the answer to the same question when it is asked again.
6. Use the answer to build the model.
7. Ask a new subset question.

The question “What is the sequence of Lambda DNA?” is an example of Step 3: “Ask a subset question to frame a specific experiment.” The repetitive process of finding the sequence fulfills the requirement that the same answer be achieved again, when the question is repeated. The answer to the question “What is the sequence of Lambda DNA?” will eventually be necessary to build a model in response to the question framing the experimental project. This will be illustrated as we progress with the example. For now, we can summarize the experiment that will be done to answer the question “What is the sequence of Lambda DNA” in the following way:

1. Sequence both strands of Lambda DNA using standard sequencing methodology that calls for each base to be sequenced several times. Validate data by checking that the sequences for the two DNA strands are exactly complementary.
2. As a negative control, make sure that no sequence data are obtained using the buffer alone.
3. As a positive control, use a piece of DNA of known sequence and make sure that the methods confirm the sequence of the positive control DNA. The same buffer and materials should be used to sequence the positive control DNA and the Lambda DNA (and the negative control).

¹²Because of the specific base pairing that occurs in DNA (A pairs with T and C pairs with G), the first strand is always “complementary” to the second strand, for example, 5' ATGTGA 3' would pair with the complementary sequence 3' TACTAC 5'.

The next question Benny wanted to answer as part of his experimental project was:

b. What size pieces of Lambda DNA are produced by cutting with EcoRI?

To answer this question, Benny must do two things:

1. Cut Lambda DNA with EcoRI.
2. Separate the pieces by size and figure out what the different sizes are.

By separating the resultant pieces of Lambda and visualizing them, Benny will have shown that he actually cut the DNA with EcoRI, so the two things Benny must do to answer this experimental question are linked. We previously discussed what Benny would need to do to show that he can cut Lambda DNA with EcoRI. Here is the summary of that experiment:

1. Incubate Lambda DNA under the specified conditions (as established by Marshall) with EcoRI.
2. As a negative control, to prove it is the EcoRI that affects the digestion, incubate another sample of Lambda DNA without EcoRI. This negative control will also show that the Lambda DNA does not separate into small pieces in the absence of EcoRI.
3. As a positive control, to show that the conditions are appropriate for a restriction enzyme digestion to take place, use an enzyme other than EcoRI to digest Lambda. Optimally, this other enzyme should work under the same conditions and also be capable of digesting the Lambda DNA; however, the resultant pieces should be distinguishable in size from the results of EcoRI digestion.

Next, once EcoRI digestion is complete, Benny must prove that he can separate DNA by size using a procedure called "gel electrophoresis." He also must be able to identify the sizes into which EcoRI cuts the Lambda DNA. Therefore, he will need to obtain some DNAs of known sizes (a "size standard") with which to measure his cut Lambda DNA. Benny may have to repeat this experiment several times, perhaps with different size standards in order to obtain as precise a readout as possible. We can summarize this part of the experiment to establish the sizes of EcoRI-cut Lambda DNA fragments in the following way:

1. Perform gel electrophoresis on the Lambda DNA cut by EcoRI.
2. As a negative control, perform gel electrophoresis on the Lambda DNA incubated without any EcoRI.
3. As a positive control, perform gel electrophoresis on the Lambda DNA incubated with the other enzyme known to cut Lambda.
4. As a control to measure the size of the resultant pieces of DNA after EcoRI digestion, perform gel electrophoresis on "size standards," pieces of DNA of predetermined size.

Note that in this particular experiment, there is no “built-in” repeat, as was the case for the sequencing experiment. So, to satisfy the requirement for repetition, Benny will need to do this experiment a few times. It would also be helpful if Benny uses multiple size standards and runs all of his samples on the same gel, so he can more accurately compare the sizes of the DNA fragments. (Fig. 1).

After obtaining experimental data to answer this particular question, Benny can begin to put together a model for how EcoRI works. For one thing, he will now have shown that EcoRI cuts Lambda DNA, resulting in a few discrete pieces of DNA of reproducible size. These data will help to confirm what Marshall told Benny about EcoRI’s properties. The fragments themselves give Benny the material he will need to answer his main experimental question, as we will see. But for now, let’s progress to the next experimental question Benny wanted to answer:

c. For each piece of Lambda cut by EcoRI, what are the sequences of the severed ends?

To answer this question, Benny must be able to sequence the ends of the DNA fragments cut by EcoRI. This experiment would be validated in much the same way as the experiment in Step a. When Benny obtains these data, he will learn whether EcoRI results in a discrete and reproducible cut. Let’s say that he performs the experiment and finds that EcoRI leaves “sticky ends,” as defined earlier. He finds that each and every one of the ends has the sequence 5’ AATT 3’. Therefore, with these data, Benny will finally have evidence consistent with the idea that EcoRI is a “true” restriction enzyme, because it appears to be cutting at a consistent site. Note that the sequence is palindromic.



Now, Benny is in a situation that commonly occurs in science. He has data that allow him to begin building a model, yet he does not know where he is in terms of answering his main question. In this case, he does not know whether the end sequence left by cutting with EcoRI defines the full “cut site” of the enzyme. It may be that AATT is the full extent of the site. Alternatively, these bases may represent only a portion of a larger recognition sequence. The purpose of sequencing the EcoRI-cut ends was to allow Benny to orient himself in the sequence; in other words, he can now use a computer to find the locations where AATT occurs in the Lambda genome and determine whether EcoRI-induced digestion of Lambda at those locations would predict the fragment sizes he achieved experimentally or whether he must look for additional sequences around the “AATT” that are also required to explain the data.

Formally, Benny now knows the sizes of the fragments produced by cutting with EcoRI and he can predict that each of the EcoRI digestions occur when there is at least the sequence AATT; in other words, this sequence is apparently necessary for an

EcoRI cut site, because it is found at each end left by EcoRI digestion, but Benny does not yet know whether it is sufficient for EcoRI to cut.

**“JUMPING THE GUN” AND A SETTING WHERE
A HYPOTHESIS IS FEASIBLE**

Often in experimental science, a scientist gathers a little bit of data and then over-interprets it. So rather than using the data to begin constructing a model, these same scientists think they have “the whole story” when this may not in fact be the case. Let’s say that in the present example, Benny takes the data that he has gathered so far—that EcoRI digestion of Lambda DNA results in DNA fragments with an end structure that is 5’ AATT 3’—and concludes that he has answered the main question that “AATT is the sequence that is sufficient to allow EcoRI to cut DNA.”

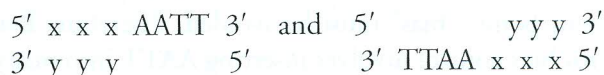
This statement may be correct, but then again, it may not be. One thing is clear—this statement is not a response to the question Benny asked, which was simply:

c. For each piece of Lambda cut by EcoRI, what are the sequences of the severed ends?

This question has been answered. The answer is that the sequence at the severed ends of EcoRI-cut DNA is AATT, and Benny can therefore conclude that digestion with EcoRI requires at least the following sequence:



and that EcoRI digests this sequence to produce the following pieces:



Benny has not finished his experimental outline and has therefore not yet searched the Lambda sequence to see whether there is additional DNA around the AATT that could contribute to a palindrome. More importantly, he has not performed an experiment to ask whether adding “AATT” to the Theta DNA sequence would be sufficient to imbue Theta with sensitivity to EcoRI. Most egregiously, because it would be so simple, he has not even checked his Lambda sequence to determine how many times, or where, the sequence AATT occurs. If he at least did this, Benny would have seen whether the predicted digestion of those sites would result in the fragment sizes that he determined experimentally.

Note also that it is at this point where a hypothesis could be usefully made. Now that the number of possible EcoRI recognition sites have been limited to those sites that contain the sequence AATT, it becomes feasible to construct a falsifiable hypothesis. Such hypotheses could include the following:

1. The EcoRI cut site contains AATT.
2. The sequence AATT comprises the EcoRI restriction site.

Whether or not Benny has “jumped the gun,” one can quickly test the second hypothesis by doing an experiment that was previously outlined as part of the Experimental Question. In the “question/answer framework,” the experiment would be performed in response to the following question:

Is the presence of AATT sufficient to allow EcoRI to cut DNA?

This question is similar to the question in Benny’s Experimental Design, which was

f. Does the sequence obtained in Step e, when inserted into Theta, now allow EcoRI to digest Theta?

In Benny’s experimental design, however, he was supposed to check the Lambda sequence first to determine whether all of the AATT sites correlated with the DNA fragment sizes he obtained experimentally. Even though the outlined approach would have saved Benny a significant amount of time, we will see that Benny will still obtain the right answer and hopefully learn a useful lesson about “jumping the gun” when he follows his present course. So, Benny is now faced with the following question:

Is the presence of AATT sufficient to allow EcoRI to cut DNA?

To answer this question, Benny must be able to insert the sequence AATT into the piece of Theta DNA that cannot be digested by EcoRI. In addition, and this is a very important point, “bias” must be avoided. Therefore, Benny needs to do a “bias control.” This bias control involves inserting AATT into many different locales in the Theta genome; we will shortly see how valuable this control is. For now, it should just be noted that in order to test “fairly” whether AATT is sufficient to allow EcoRI to cut DNA, it must be inserted at different locales in the Theta DNA, which do not have any common elements.

Benny uses polymerase chain reaction (PCR)¹³ to insert AATT into the Lambda genome. To avoid “bias,” he inserts the sequence between each possible DNA “neighbor combination,” at the *, as illustrated in the following table, to give the following sequences.

¹³Polymerase chain reaction is a technique so useful that it was sufficient to merit the Nobel Prize in Chemistry, awarded in 1993.

Now
sites
esis.

and
the
would

was

bda
DNA
ach
will
g the
ing

into
is is a
bias
in the
d just
f to
not

oda
eigh-
wing

rite in

Tube	Insertion site	Resulting sequence
1	A [*] A	AAATTA
2	A [*] C	AAATTC
3	A [*] G	AAATTG
4	A [*] T	AAATTT
5	C [*] A	CAATTA
6	C [*] C	CAATTC
7	C [*] G	CAATTG
8	C [*] T	CAATTT
9	G [*] A	GAATTA
10	G [*] C	GAATTC
11	G [*] G	GAATTG
12	G [*] T	GAATTT
13	T [*] A	TAATTA
14	T [*] C	TAATTC
15	T [*] G	TAATTG
16	T [*] T	TAATTT

Realize that each sequence is surrounded by the rest of Theta. So, for example, TAATTT may be surrounded by any other sequence. Benny then digests the resulting modified Theta DNAs in each tube, in addition to unmodified Theta as a negative control and Lambda as a positive control with EcoRI. As a control for contamination with some other enzyme, he has a tube of Theta DNA with no EcoRI added. Finally, to ensure that Theta DNA is not contaminated with a restriction enzyme inhibitor, he uses another enzyme, HindIII, which he knows can cut unmodified Theta DNA.¹⁴ Benny performs the digestions and uses gel electrophoresis to determine whether the Theta DNA was digested. He obtains the following data:

Tube	Insertion site	Resulting sequence	EcoRI cuts?
1	A [*] A	AAATTA	No
2	A [*] C	AAATTC	No
3	A [*] G	AAATTG	No
4	A [*] T	AAATTT	No
5	C [*] A	CAATTA	No
6	C [*] C	CAATTC	No
7	C [*] G	CAATTG	No
8	C [*] T	CAATTT	No
9	G [*] A	GAATTA	No
10	G [*] C	GAATTC	Yes
11	G [*] G	GAATTG	No
12	G [*] T	GAATTT	No
13	T [*] A	TAATTA	No
14	T [*] C	TAATTC	No
15	T [*] G	TAATTG	No
16	T [*] T	TAATTT	No
17	Unmodified Theta		No
18	Lambda DNA		Yes
19	Unmodified Theta, without EcoRI		No
20	Unmodified Theta, with HindIII		No

¹⁴As noted previously, each of these types of controls will get their own chapter later on.

Benny repeats this experiment three times. Each time, he gets the same results. How are these data interpreted? Recall the possible frameworks for this experiment. First, there was the hypothesis “The sequence AATT comprises the EcoRI restriction site.” Then there was the question

Is the presence of AATT sufficient to allow EcoRI to cut DNA?

This is what Benny is obligated to do, for each framework: The hypothesis that “The sequence AATT comprises the EcoRI restriction site” is *falsified*, because in 15 of 16 cases, the presence of AATT did not cause Theta to be cut. Therefore, AATT does not comprise the EcoRI restriction site. For the question “Is the presence of AATT sufficient to allow EcoRI to cut DNA?” the answer is no, because in 15 of 16 cases, the presence of AATT was not sufficient to cause EcoRI to cut Theta. At this point, Benny has a falsified hypothesis and a question that resulted in a “no” answer. Yet Benny is overjoyed. He is elated, in fact. Why? Because in one case, tube 10, the sequence that Benny added in to Theta, GAATTC, was sufficient to allow EcoRI to cut the Theta DNA. Benny figures that if he just changes his hypothesis to “The sequence GAATTC comprises the EcoRI restriction site,” he will have proven it to be true. Alternatively, if he reframes his question as “Is the presence of GAATTC sufficient to allow EcoRI to cut DNA?” then the answer will probably be “yes.”

Benny goes to Marshall with this news, brimming with happiness, and tells Marshall that he solved the problem and that GAATTC is the EcoRI cut site. He goes through all of his data with Marshall and shows him the evidence revealing that adding the sequence AATT between a G and C in Theta caused EcoRI to cut the Theta DNA, whereas AATT cut in no other situation.

Marshall tells Benny that he is not able to conclude that “GAATTC” is the EcoRI cut site. The main thing that Benny has done is to prove that AATT is not sufficient to cut DNA and that AATT is not sufficient to induce EcoRI to cut Theta. As for the data from “tube 10,” Marshall says the following: “You have increased your Inductive Space.” What Marshall means is that Benny has learned something from his experiment, and he can use this to ask a new question or make a new hypothesis. Let’s stick to the Question/Answer methodology and frame a possible new question for Benny:

Is the presence of GAATTC sufficient to allow EcoRI to cut DNA?

Remember that Benny considered reframing his previous question to this, in a retrospective matter. We should just state categorically the following: **You cannot go back and alter a hypothesis or a question retrospectively, that is, you cannot change the framework for an experiment that has already been performed.**

However, you can use information gathered to do a new experiment under a new hypothesis or question. Note that the question “Is the presence of GAATTC sufficient to allow EcoRI to cut DNA?” is identical in structure to the previous question “Is the

presence of AATT sufficient to allow EcoRI to cut DNA?" However, Benny has now added the information from his previous experiment to refine his question. When Benny tells Marshall that he is going to try to answer this question, Marshall patiently reminds Benny that this was not the original plan. Originally, Benny had mapped out this approach (see page 86):

- a. What is the sequence of Lambda DNA?
- b. What size pieces of Lambda DNA are produced when cutting with EcoRI?
- c. For each piece of Lambda cut by EcoRI, what are the sequences of the severed ends?
- d. Does Lambda contain any repeated DNA sequence that, when cut, would result in the sizes obtained using EcoRI, and which contains the DNA sequences determined in Step c?
- e. If the answer to Step d is yes, what is that sequence?¹⁵
- f. Does the sequence obtained in Step e, when inserted into Theta, allow EcoRI to digest Theta?
- g. Can we now say that our question "What DNA sequences are sufficient for EcoRI to cut DNA?" is answered? In other words, does the discovered sequence account for all of the cuts made by EcoRI in Lambda?
- h. If the answer to Step g is no, what are the remaining cuts not accounted for by the discovered sequence?

Benny got so excited when he discovered that the answer to Step c was AATT that he changed his approach. Now he can keep going along his new approach, and he may find the correct answer, but Marshall wants Benny to see what will happen if he follows his methodology as originally outlined.

Benny decides to listen to Marshall, but he also wants to ask his new question. Marshall points out that the next two steps require a 1-second computer search. Finally persuaded, Benny returns to his original experimental plan and rereads it. He realizes that he has stopped short of the fourth experiment, Step d. Here was the question in Step d and the following question in Step e:

- d. Does Lambda contain any repeated DNA sequence that, when cut, would result in the sizes obtained using EcoRI, and which contains the DNA sequences determined in Step c?
- e. If the answer to Step d is yes, what is that sequence?

¹⁵If the answer to this question is "no," then EcoRI may be in the rare class of restriction enzymes that cut at more than one DNA site. In that case, Benny will have to change his methodology to account for this possibility. A complete listing of the experimental methodology for this project would include the "no" case for this question.

Benny goes back to the sequence for Lambda, and identifies all of the positions for AATT.¹⁶ He then asks the computer whether some combination of these AATT sites, when cut, would result in the fragment sizes he obtained by cutting Lambda, and whether that combination has a common sequence surrounding AATT. The answer turns out to be “yes.”

When Benny then goes to Step e and asks “If the answer to Step d is yes, what is that sequence?” the result is the following:

GAATTC

In other words, if Benny simply takes the Lambda sequence and searches for the AATT ends he found when he digested Lambda with EcoRI, and then he asks if there is any sequence in common with AATT that results in the fragment sizes he obtained, he finds that the computer predicts that digestion of Lambda at “GAATTC” would result in the fragment sizes he in fact achieved.

Benny is now a little chagrined, because he realizes that if he had not “jumped the gun,” he would have been able to deduce the “correct answer.” He goes back to Marshall and tells Marshall that he was right and that following the protocol would have provided the answer without doing the extra experiment. When Marshall asks Benny what he means, Benny explains that it turns out that GAATTC exists only a few times in Lambda, and if Lambda were cut at those sites, then one would obtain the fragment sizes that Benny in fact obtained experimentally when he cut Lambda with EcoRI.

Marshall patiently explains that Benny still has not “proven” anything, but rather has simply added even more to his inductive space. The next question in the Experimental Design is the following question:

f. Does the sequence obtained in Step e, when inserted into Theta, allow EcoRI to digest Theta?

Or, specific to this example, the question can be restated in this way:

Is the presence of GAATTC sufficient to allow EcoRI to cut DNA?

¹⁶To answer this question, Benny requires a simple computer program. Once Benny has determined the sizes of DNA cut by EcoRI, using gel electrophoresis, and the end sequences of each of his DNA pieces, he will input these data into his computer. The computer will run an algorithm, placing the various DNA lengths at each potential interval in the sequence, and asking if there exists the sequenced DNA sequence at these positions that would account for the DNA sizes he obtains. Once the computer finds the end-sequenced DNA, it will determine if there is surrounding DNA that comprises a “repeated sequence” that might account for EcoRI’s cut site.

For example, say Lambda is 10,000 bp long, and EcoRI digestion cuts Lambda in three pieces. One piece is 2000 bp long, one piece is 3000 bp long, and one piece is 5000 bp long. After determining this, the computer program will be asked if there is a sequence in Lambda that is a repeated sequence containing the DNA sequence AATT that will result in these fragment sizes. The computer program then provides a readout. In the present case, the program results in the sequence GAATTC.

Note that this question is identical to the one Benny would have asked anyway, building up from AATT to the experimental data he achieved from “tube 10” of his earlier experiment. This illustrates the following:

1. There is more than one way to get to the “right answer.”
2. Both methods require that the new question be asked; it is not enough to have suggestive data from a previous experiment that was designed to answer a different question.
3. The final proof from distinct methodologies usually converges on a single, binary question, which is highly restricted.

Marshall now gives Benny the go-ahead to ask his question “Is the presence of GAATTC sufficient to allow EcoRI to cut DNA?” To answer this question, as in the case with the previous question involving just AATT, Benny must be able to insert the sequence GAATTC into the piece of Theta DNA that cannot be digested by EcoRI. In addition, as before, and even now more than in the previous instance, bias must be avoided, especially because by now Benny is “really” convinced that the answer to the main experimental question is GAATTC. As in the previous question, to “fairly” test whether GAATTC is sufficient to allow EcoRI to cut DNA, it must be inserted at different locales in the Theta DNA that do not have any common elements. Benny again uses PCR, this time to insert GAATTC, as opposed to just AATT, into the Theta genome. As he did in the previous case, he inserts GAATTC between each possible DNA “neighbor combination,” at the *, as illustrated here, resulting in the following sequences:

Tube	Insertion site	Resulting sequence
1	A*A	AGAATTCA
2	A*C	AGAATTCC
3	A*G	AGAATTCG
4	A*T	AGAATTCT
5	C*A	CGAATTCA
6	C*C	CGAATTCC
7	C*G	CGAATTCG
8	C*T	CGAATTCT
9	G*A	GGAATTCA
10	G*C	GGAATTCC
11	G*G	GGAATTCG
12	G*T	GGAATTCT
13	T*A	TGAATTCA
14	T*C	TGAATTCC
15	T*G	TGAATTCG
16	T*T	TGAATTCT

Benny then tests each resulting sequence to ask whether EcoRI cuts the Theta DNA when these sequences are inserted into Theta. He has the same negative and positive controls as he had in the previous experiment: As a negative control, he uses Theta DNA that has no new sequence inserted into it, and as a positive control, Benny uses the Lambda DNA that he has already proven can be cut by EcoRI. He also uses the HindIII control and the “no restriction enzyme” control to test for any contaminating restriction enzymes. He performs the experiment, and obtains the following data:

Tube	Insertion site	Resulting sequence	EcoRI cuts?
1	A*A	AGAATTCA	Yes
2	A*C	AGAATTCC	Yes
3	A*G	AGAATTCT	Yes
4	A*T	AGAATTCT	Yes
5	C*A	CGAATTCA	Yes
6	C*C	CGAATTCC	Yes
7	C*G	CGAATTCT	Yes
8	C*T	CGAATTCT	Yes
9	G*A	GGAATTCA	Yes
10	G*C	GGAATTCC	Yes
11	G*G	GGAATTCT	Yes
12	G*T	GGAATTCT	Yes
13	T*A	TGAATTCA	Yes
14	T*C	TGAATTCC	Yes
15	T*G	TGAATTCT	Yes
16	T*T	TGAATTCT	Yes
17	Unmodified Theta		No
18	Lambda DNA		Yes
19	Unmodified Theta, without EcoRI		No
20	Unmodified Theta, with HindIII		N.A. ^a

^aNot available.

Benny is about to bound into Marshall's office yelling “eureka,” but fortunately stops himself, remembering the look that Marshall gave him the last time he ran in there prematurely. This time, Benny sits down at his desk and pulls out his Experimental Design, to make sure he did not forget anything. The first thing he remembers to do is repeat his experiment. He does so—five times, as a matter of fact. Each time, he gets the same result. Next, he pulls out his “Experimental Plan.” He reads it and notices the following questions:

- g. Can we now say that our question “What DNA sequences are sufficient for EcoRI to cut DNA?” is answered? In other words, does the discovered sequence account for all of the cuts made by EcoRI in Lambda?
- h. If the answer to Step g is no, what are the remaining cuts not accounted for by the discovered sequence?

He returns to his computer program and this time does a search using the sequence “GAATTC.” The program lists the positions in Lambda where this

sequence can be found. Benny then does the arithmetic and confirms that if EcoRI were to cut at all of the GAATTC sites, the result would be exactly the fragment sizes he determined experimentally. Benny checks further. There are no more questions in his Experimental Program, so now he refers back to his "Framework Question":

What DNA sequences are sufficient for EcoRI to cut DNA?

He writes underneath the question this answer:

GAATTC

He looks back at the question and sighs, because he realizes that the question asks him to determine whether more than one sequence is sufficient. However, at least for Lambda, he demonstrated that GAATTC accounts for all of the cut sites. So his answer, even to this question, is the same: GAATTC.

GOING ABOUT THE EXPERIMENTAL QUESTION IN MORE THAN ONE WAY, ESTABLISHING NECESSITY AND SUFFICIENCY

At this point, Benny really thinks that he has the answer to his question. He has proved that adding GAATTC to Theta, which normally cannot be digested by EcoRI, was sufficient to cause Theta to be cut. He also proved that the GAATTC sequences that naturally exist in the Lambda sequence are sufficient to account for the cuts induced by EcoRI in Lambda. Yet, Benny wants to avoid looking like a novice by missing something, so he decides to ask a new question.

Is mutation of the GAATTC sites in Lambda sufficient to make those regions immune to digestion by EcoRI?

Note that this question goes to a different issue. It does not ask whether "GAATTC" is sufficient to allow EcoRI to cut DNA; rather, it asks whether GAATTC is "necessary" to allow EcoRI to cut DNA. For example, there may be some combination of bordering sequences, plus a large portion of the GAATTC sequence, that allows EcoRI to cut. This may be unlikely, especially because addition of GAATTC to a new piece of DNA (Theta) lead to its digestion with EcoRI. However, the new experiment has the virtue of approaching EcoRI's specificity in a distinct way; for example, one might wonder whether the naturally occurring Lambda DNA might have evolved some additional properties to make it particularly permissive to EcoRI digestion. Put another way, just because you need GAATTC to get EcoRI to cut Theta does not necessarily prove that the entirety of GAATTC is required for EcoRI to cut Lambda; the whole sequence may just make the process more efficient. Therefore, Benny's experiment with mutant sequences will do more than query "necessity" versus "sufficiency"; it also functions as a control for the Theta DNA to ensure that there was nothing "particular" about that DNA that produced a requirement for the entire GAATTC sequence.

To answer this new question, Benny takes his Lambda DNA sequence, and he mutates, in turn, each naturally occurring EcoRI site at a different position in the GAATTC sequence. As a control, he keeps a sample of wild-type (unmutated) Lambda DNA. As a further control, he takes a piece of Lambda and mutates all of the GAATTC sites at the same time, turning each into GAATTA. Thus, he has the following tubes of DNA:

Tube	Sequence
1	Mutate site 1 to GAATTA, keep the rest GAATTC.
2	Mutate site 2 to GAATAC, keep the rest GAATTC.
3	Mutate site 3 to GAAATC, keep the rest GAATTC.
4	Mutate site 4 to GATTTC, keep the rest GAATTC.
5	Mutate site 5 to GTATTC, keep the rest GAATTC.
6	Mutate site 6 to TAATTC, keep the rest GAATTC.
7	GAATTC (unmutated control)
8	Mutate all sites to GAATTA.

Benny performs the digestions and runs DNA from each tube on an electrophoresis gel. He obtains the following data:

Tube	
1	Fragment sizes demonstrating sites 2–6, but not 1, were digested.
2	Fragment sizes demonstrating sites 1 and 3–6, but not 2, were digested.
3	Fragment sizes demonstrating sites 1, 2, and 4–6, but not 3, were digested.
4	Fragment sizes demonstrating sites 1–3, 5, and 6, but not 4, were digested.
5	Fragment sizes demonstrating sites 1–4 and 6, but not 5, were digested.
6	Fragment sizes demonstrating sites 1–5, but not 6, were digested.
7	All sites were digested.
8	The DNA was not digested.

Benny repeats this experiment five times and each time gets the same results. From these data, Benny concludes that mutating the GAATTC sequence, at any position, prevents EcoRI from cutting the site. Therefore, not only is GAATTC sufficient to cause EcoRI to digest DNA, it is also necessary, in each of its components. It should be noted that not everything must meet the “necessary” and “sufficient” criteria for a particular event to be “important.” For example, it can be demonstrated that “smoking causes cancer,” and yet it can be proven that smoking is not sufficient to cause cancer (because not everyone who smokes gets cancer) and that smoking is not necessary to cause cancer (because an individual can get cancer without smoking). Nevertheless, if it is shown that there is a statistically significant increase in the rate of cancer as a result of smoking, then one can still conclude that “smoking causes cancer.” However, the “necessity” and “sufficiency” tests will help to induce the scientist to find the additional conditions that make smoking more or less likely to cause cancer or to figure out why, in some cases, someone can smoke three packs a day for 50 years and not get cancer. Thus, these concepts are always useful. In the present case,

the scientist is obliged to ask whether the DNA sequence GAATTC is necessary for EcoRI digestion of DNA, because these concepts are implied by the framework question for Benny's experimental project.

MODEL BUILDING, AND USING THE MODEL TO UNDERSTAND WHAT WILL HAPPEN IN THE FUTURE

Benny then recalls the need to use his data to build a model. First, he rewrites his Framework Question: "What DNA sequences are sufficient for EcoRI to cut DNA?" He proposes to answer this question with a model. The following sequence is sufficient for EcoRI to cut DNA:



Benny goes further and adds to his model this statement: EcoRI cuts at the site



He considers adding a statement about the end structure left by EcoRI, but because that was not the point of his project, he leaves that for further confirmatory experiments.

With this model in place for the DNA sequence that is sufficient to allow EcoRI to cut DNA, Benny now wants to verify that the model represents what will happen in a future experiment. He asks the question "How accurately does the model predict where EcoRI will cut?"

To answer this question, Benny decides to perform a new experiment. He acquires several bacteria whose DNA has been sequenced and plans to use his computer program to predict the fragment sizes that would be generated by cutting these DNA sequences with EcoRI. He intends to cut the DNA with EcoRI and to see whether his prediction is accurate (and to what degree it is accurate). He designs the following experiment:

1. To digest five DNA sequences, A–E, that have at least five GAATTC sites in each sequence, for a total of 25 sites with EcoRI. As a positive control, to ensure that the DNA is not contaminated with a restriction enzyme inhibitor, five sequences will be digested with a second restriction enzyme called HindIII. As a negative control, to ensure that the DNA is not contaminated with a distinct restriction enzyme, Benny will have tubes of each DNA, A–E, with no added restriction enzyme.
2. To digest five DNA sequences, F–J, that do not have a GAATTC site in them with EcoRI. As a positive control, these five sequences will be digested with a second

restriction enzyme, HindIII, which he knows cuts these sequences. As a negative control, he will have tubes of each DNA, F–J, with no added enzyme.

Benny performs the experiment and finds that the fragment sizes left by digesting A–E with EcoRI conform to the predicted sizes. Furthermore, he observes that DNAs F–J, which lack the GAATTC site, were left unperturbed by EcoRI while being digested by HindIII. A–E were not cut in the absence of EcoRI or HindIII. Thus, in 25 of 25 events, Benny's model accurately reflected what would happen in the future; he was able to predict accurately where EcoRI would cut.

What if Benny had not had access to sequence data before he determined where EcoRI cuts a particular piece of DNA? Would he still be able to subject the model to the requirement that it be verifiable in future experiments? The answer is yes, simply by repeating his experiment multiple times. In other words, once he had obtained the data on fragment sizes induced by EcoRI digestion, he could then ask whether the same result would happen again. Thus, we see that simple repetition, to gain statistical significance, can be recast as the requirement of "predicting the future," in just the same way as allowing a ball to fall to the ground multiple times to confirm the existence of gravity.

DECLARING THE EXPERIMENTAL QUESTION TO BE ANSWERED

Benny gathers all of his data and the model he has put together, and he presents his findings to Marshall. Marshall goes through everything. For each experiment, Marshall asks whether Benny has repeated his experiments and finds that he has. Marshall then asks to see the primary data—the pictures of the gel electrophoresis experiments—and he checks the fragment sizes produced by EcoRI digestion for himself. His analysis agrees with Benny's. Then Marshall does something that Benny rather resents, but which Marshall explains is an "experimentalist control" (a subject we will discuss further later on). Marshall goes through the literature and finds a new piece of DNA, with known sequence, that Benny did not examine, and Marshall asks another scientist in the lab to digest this new piece of DNA with EcoRI, to see whether the data are consistent with Benny's findings. Marshall calculates the fragment sizes that Benny's model predicts will be left by EcoRI digestion of the GAATTC sites in the new piece of DNA. The second scientist does the experiment and shows Marshall and Benny the data. The fragment sizes left by EcoRI digestion of the new piece of DNA are consistent with Benny's model.

Marshall and Benny now agree that the answer to the question, "What DNA sequences are sufficient for EcoRI to cut DNA?" is "GAATTC," and that EcoRI cuts at GAATTC. They start to map out Benny's first scientific manuscript.



Experimental Repetition

The Process of Acquiring Data to Model Future Outcomes

REPETITIVE TESTING—AS WHEN AN OBSERVER NOTES that the sun rises day after day—is the process by which scientists gain sufficient knowledge to build models that represent reality. Without repetition, it is not known whether the outcome of any particular experiment is representative of the “normal” case. In the case of the question “What color is the sky?” a number of repetitions were performed to assess the minute-by-minute and day-to-day variation that exists for sky color; if only a single time point had been chosen, its predictive value would have been low. The idea that a certain number of data points are necessary to be “confident” of providing an accurate representation of a system’s behavior is formalized in the mathematical discipline of statistics and forms the basis of statistical determinations of “significance.”

In this chapter, different types of repetitions are discussed. It will be shown why each type of sampling is required to generate sufficient data to produce an “accurate” model. And it bears repeating to say that accuracy in a model is determined not by its fidelity in describing the past, but rather how well it represents the future; this is the only way to distinguish between science and science fiction.

DETERMINING THE NUMBER OF MEASUREMENTS NECESSARY TO ACHIEVE STATISTICAL SIGNIFICANCE

Other texts written about experimental design deal almost exclusively with statistics. This volume, instead, starts earlier in the process. It is geared toward helping a scientist establish an experiment that is unbiased¹ in design and in interpretive framework and that has the appropriate controls in place to alert the scientist to any issues in execution. Such preparation will allow statistics to be used to the greatest effect.

¹Or, at least, a framework within which biases may be better recognized.

However, in this chapter, we get closer to statistical issues because it is from statistical calculations that we learn how many data points an experiment must contain. Without this critical number, the experiment will not be sufficiently “powered” to produce a demonstrably predictive result.

Statistics are only as good as the quality of the experimental framework in place for any given experiment. A statistical calculation can give “false” confidence if the data set is not representative of the general condition. For example, if a study is performed to determine whether a weight-loss drug will cause people to eat less and the study sample consists entirely of people who are highly motivated to lose weight, willing to exercise, and change their diets, it would not be surprising if data gained from this study failed to predict the drug’s effectiveness in people who were less motivated.

The number of measurements required for “statistical significance” will be dictated in part by the variability of the system under study. The variability itself can be ascertained only by performing a certain number of repetitions. Thus, there is a requirement for an iterative process: A series of measurements must be used to appraise the variability of the system and to estimate the number of data points needed to obtain statistical significance in a new experiment. And lest the idea of “statistical significance” sound like an unjustified goal, it should be emphasized that this parameter is validated by its ability to predict whether a model can be verified by future experimentation.

For example, at noon the sky is almost always blue or gray; only a few measurements will be necessary to obtain a data set that allows an accurate and predictive model to be built of the noon sky. However, if the sky is measured near sunrise or sunset, then many more measurements will be necessary, even if the data set is of “good quality,” meaning that multiple measurements and controls all confirm that the sky is red at the time points measured. Thus, it could be said that the first goal of performing repetitions is to determine the range of variation present in the system. In doing so, we might also find that we are interested in data other than the “norm,” but it is impossible to make that choice until the norm has been defined.

CATEGORIES OF EXPERIMENTAL REPEATS

Let’s return to our question “What color is the sky?” When designing a set of experiments to address this question, five distinct categories of repetitive measurements can be envisioned:

- It is only from performing multiple measurements of the sky at a single time point that a scientist can determine whether the measurements are accurate—each measurement verifies another.
- It is only by taking multiple measurements of a particular time point over many days that a scientist can determine whether the time point produces a consistent

outcome. Through this type of repetition, one discovers that sunrise and sunset occur at different times throughout the year, and, after taking measurements over several years, one learns that these variations are stable and predictable.

- It is only by measuring sky color at multiple times during the course of one day that a scientist may realize that the sky's color changes throughout the day. A measurement made at midnight does not predict the sky color at noon.
- It is only by taking multiple measurements at different points in the sky that a scientist discovers that the sky color along the horizon may not be predictive of other areas.
- It is only by repeating experiments in toto that a scientist can discern whether the outcomes from multiple sets of experimental measurements are consistent; in this way, sufficient data are gathered to gain confidence in a model. For example, if an experiment is designed to take multiple measurements of sky color every five minutes over the course of one day, then we cannot say that this set of measurements is representative of the general case until the entire experiment has been repeated several times in the same manner.
- A model derived from the experimental data can be said to be "verified" only if the experiment is repeated under the Experimental Question "Is the model accurate?"

This last type of repetition forces the scientist to repeat the experiment after the model is in place as a rigorous way of examining all aspects of the model for its ability to accurately represent what will happen in the future.

ILLUSTRATING THE DIFFERENT CATEGORIES OF EXPERIMENTAL REPEATS WITH A BIOLOGICAL EXAMPLE

Let's say that a scientist—we'll call him Scooter—is interested in investigating liver gene expression in animals on high-fat diets. This project is formulated into an experimental question:

open ended question { **What gene-expression changes are observed in the livers of rats that are on a high-fat diet—sufficient to cause obesity—compared to those that are on a normal diet?**

expression genes liver

Scooter goes through all of the preliminary steps that we previously discussed:

1. The Experimental Project is defined using an open-ended question. The primary goal of Scooter's Principal Investigator, Leorrah, is to understand obesity. Leorrah gave this question to Scooter as a way of determining the changes in the liver that are caused by obesity-inducing diets. Given this over-arching goal, Scooter first

istical
tain.
" to
ce for
a set
ed to
mple
cise,
ailed
ated
ascr-
mire-
: the
ed to
istical
ara-
ature
ure-
ictive
r sun-
ood
sky is
form-
oing
it it is
ber-
ts can
point
mea-
many
tent

defines the broadest question as “What are the effects of obesity on the body?” Within that broad question, he asks “What are the effects of a high-fat diet—sufficient to cause obesity—on the body?” Then he gets to his project: “What gene-expression changes are observed in the livers of rats that are on a high-fat diet—sufficient to cause obesity—compared to those that are on a normal diet?” He constructs this question hierarchy so that he knows to access the “inductive space” made available by the broader questions.

2. The terms “high-fat diet” and “normal diet” are defined: The “high-fat diet” has 70% fat, 15% protein, and 15% carbohydrates, and the “normal diet” has 30% fat, 40% protein, and 30% carbohydrates. A “gene-expression change” is decided to be a gene-expression difference of at least threefold in either direction. An inbred strain of rats that has previously been shown to gain weight on a high-fat diet is chosen for the study.²
3. Steps toward system validation are taken. It is shown that the rats will tolerate both types of diet. Scooter demonstrates that he can isolate RNA from rat livers and measure changes in gene expression using microarrays. As a control, two samples from other rat studies are used as “universal standards” so that the changes seen in the current study can be compared to future studies (as long as the “universal standards” are used in the subsequent studies).

At this point, Scooter has a specific “system validation” question:

How many rat livers should be sampled at a given time point?

To illustrate how this issue is settled, we consider several distinct cases, from a “naive” experimental design to a “sophisticated” experimental design. In addition, it will be shown how resolving this issue of system validation causes Scooter to reconsider his definitions.

A NAIVE EXPERIMENTAL DESIGN

In this example, Scooter fails to consider the possibility that gene expression in a single rat’s liver may not be representative of the gene expression changes in the livers of other rats. Nor does he consider that a result recorded at a single point in time may not be representative of the changes that might be observed in multiple measurements over a period of time. It is a chore to take care of rats, so Scooter decides to conclude the experiment 12 hours after initiating the high-fat diet. When he mentions this to Benny, who is a buddy of his, Benny compliments Scooter on this

²Because the over-arching purpose of the experiment is to understand obesity, Scooter wants to make sure his experiment is relevant to that phenotype. Therefore, he will correlate the diet to fat gain, so that he can be sure to study liver changes on a diet that is sufficient to cause an increase in weight and adiposity.

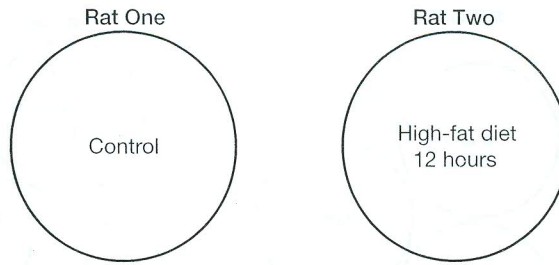


Figure 1. Naive design. Comparison of a single liver from a rat on a high-fat diet for 12 hours to a liver from a control rat on a normal diet.

design and notes that by taking such an early time point, Scooter may discover the key genes that initially trigger the liver's metabolic response to a high-fat diet. Scooter is happy to hear that he can justify his 12-hour time point.

With the experimental design now in place, Scooter compares the liver mRNA from one rat that has been on a high-fat diet for 12 hours to the liver mRNA from one rat that has been maintained on a normal diet (Fig. 1). Note that in this naive experimental setting, there is no mechanism for Scooter to learn the degree of variability in the system. Because there is only a single time point available for each condition, it is impossible to calculate a mean, a median, or a standard deviation for the expression level of any particular gene. Thus, if it is found that the mRNA level of the gene encoding leptin is down-regulated by a factor of 5 in the rat fed a high-fat diet and remains unchanged in the control rat, there is nothing one can do besides accept the results as given. If Scooter dismisses the need for statistics and the need to ascribe a level of "significance" to the results, he might simply say that leptin is down-regulated fivefold when on a high-fat diet. These results would then form the basis for his experimental model.

But our naive scientist knows at least this much: The experimental model must be questioned for its predictive power. Scooter thus repeats the experiment using liver mRNA samples from different rats, one on a high-fat diet and one on a normal diet. In this follow-up experiment, Scooter finds that the expression level of the leptin gene has not changed. He has now performed two experiments and obtained conflicting data, with no rationale that one data set is "more valid" than the other. Absolutely no progress has been made.

3. drugi de u prvom dizajnu ne može izračunati srednji vrijedost i SD

A SLIGHTLY LESS NAIVE STUDY DESIGN

Scooter reads a few paragraphs from a book on experimental design and thus becomes slightly less naive than he was previously. He now realizes that for system validation, he would need to determine the variability of even a single point. So this time, Scooter still takes only a single liver sample from each rat, but he divides each mRNA sample into three batches. In doing so, he can determine

*jebra 1stakura
3 mRNA uzelo*

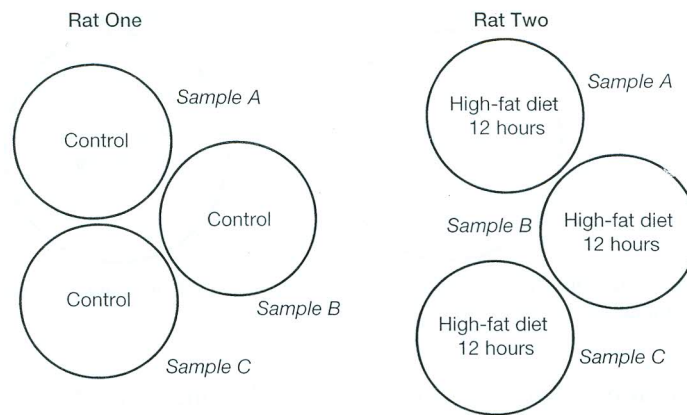


Figure 2. Slightly less naive design than that in Figure 1. Three samples of identical mRNA from two rats, to measure gene changes.

whether his method of measuring gene changes is consistent on identical batches of mRNA (Fig. 2).

In this slightly less naive study, Scooter learns that his data are extremely variable. Even though the mRNA samples came from the same liver, one microarray chip shows no increase in leptin gene expression in rats on the high-fat diet, whereas the other two chips show that it increases fivefold.

Therefore, by performing repeats of a single time point, Scooter learns that his experiment was not trustworthy. He learns that the microarray chip is unreliable in its ability to measure leptin, so he e-mails the manufacturer, requests that a different system be used on future chips, and points out the need for the leptin gene, in particular, to be revalidated. Scooter also discovers that the mRNA sample that showed no change in leptin expression had degraded, so he changes that sample for future studies.

It is by having multiple samples of the same condition that Scooter learns whether his sample is trustworthy and whether his data analysis system is reliable. The use of multiple samples can be thought of as a "system control." These controls encompass both the chip and the mRNA, or, in more general terms, the system controls confirm that the process of measuring the experimental subject is reliable.

After having discarded his first experiment due to the system problems, Scooter performs the experiment again. This time, the data are stable. In the three liver samples from a single rat on a high-fat diet, the expression of leptin was elevated fivefold relative to the three liver samples from a single rat on a normal diet. This time, Scooter is more confident and establishes a model that leptin mRNA increases on a high-fat diet.

To interrogate his model, Scooter repeats the experiment using a different rat fed a high-fat diet and a different control rat fed a normal diet. Once again, the mRNA from each liver is divided into three samples that function as consistency controls. Unfortunately, although the results from the three samples were consistent, Scooter fails to verify his model: The levels of leptin did not change in the rat that was fed a high-fat diet in the second experiment.

AN IMPROVED STUDY, TAKING INTO ACCOUNT VARIABILITY AT A SINGLE STUDY POINT

By now, Scooter realizes that different rats fed high-fat diets for a 12-hour period show different changes in leptin, the gene he has chosen to analyze. When Scooter goes back and examines the thousands of genes on the microarray chips, he realizes that many other genes were similarly variable. In one rat, the expression level of a gene may have decreased 1.5-fold, whereas in a second rat, it may have decreased 10-fold. Scooter therefore realizes that his system is still not properly validated and that he needs to understand how many livers should be analyzed in order to obtain data that hold up to experimental repetition.

Scooter decides to sample liver mRNA from 20 rats on a high-fat diet and to compare it to liver mRNA from 20 rats on a normal diet. Because he has proven that his technique is now consistent and reliable, he might decide to measure each rat only once, but given his prior experience, he still performs three measurements per liver. Therefore, he needs 60 chips for each experimental condition, or 120 chips in total. This is an expensive experiment, but he realizes that he has already wasted a lot of time and material without gaining any knowledge of the system.

Scooter first performs the correct statistics and determines a degree of confidence for the gene changes. He eliminates genes due to the amount of variability in the data; upon calculating the standard error, he finds that the “error bars” are too high for certain genes. In this experiment, 1 of the 300 genes that is shown to change is leptin; it increases, on average, threefold in rats on the high-fat diet.

Next, when the remaining genes from this experiment are analyzed, Scooter observes that the expression of 300 genes are either up- or down-regulated, at least threefold, in rats on a high-fat diet in comparison to control animals and, using the proper statistical test, he calculates a “*P* value” of <0.05 ,³ indicating that the changes in these 300 genes have reached statistical significance.

Because this was the first time that the experiment was performed, a *P* value of 0.05 predicts that 5% of the data will not be reproducible.⁴ Therefore, a repeat experiment using the same experimental design must be conducted, asking the question “Is the behavior of the 300 genes that were shown to be perturbed in the previous experiment consistent in a second experiment?”⁵ Note that this is a “yes/no” or a “binary” question. A more accurate question would probably be phrased “Which of the 300 genes that

³Again, the purpose of this volume is not to discuss the correct statistical test, but rather the preliminary considerations that go into deciding on experimental rationale, which are then subjected to statistical analysis.

⁴After conducting a repeat experiment, the “randomness” factor goes down dramatically; a gene that is shown to be up-regulated threefold with a *P* value of 0.05 in two distinct experiments will have a 0.05×0.05 chance (or 0.0025) of being random.

⁵The reader may wonder why the follow-up question is not simply “Did the experiment repeat?” This phraseology can indeed be used, but sometimes—particularly in microarray experiments—there is a “filtering effect” of repetition, which winnows down the hundreds of changes one may see on any particular chip to those changes that are consistently observed in an experimental condition.

were perturbed in the previous experiment were also perturbed in a second experiment?" This phraseology takes note of one aspect of the model: It can be refined with more data, rather than being accepted or rejected in toto.

One other thing that Scooter can determine from these data is how many livers he "really needs." By taking the data from the 20 livers and performing a subset analysis, he will determine whether data from any smaller set of livers would have given the same result as the set of 20. Scooter looks at random sets of three livers in his data set, and he finds that different sets of three livers produce rather variable data: Some confirm all 300 genes, other sets confirm only 50 genes, and some sets pick out additional sets of genes. However, this problem starts to diminish as more livers are added to the analysis. By the time Scooter gets to 10 livers, he finds that any set of 10 chosen from his 20 would result in almost identical data as the full set. From this analysis, one can learn how large a data set must be used to account for the variability found in the experimental system. In this instance, the data obtained from any 10 livers in the set is as robust as the data from all 20 livers.

Having conducted this analysis, Scooter frames a model for gene induction in liver from rats on a high-fat diet, based on the changes in mRNA levels that were observed for 300 genes. Note, however, that this is premature because the model should not be formulated until the first experiment has been repeated once more. But Scooter does not know this yet.

Scooter now uses his set of 300 genes as a model, and he asks whether this model predicts the gene changes in a new set of animals. When the verification experiment is performed, Scooter is able to confirm 200 of the genes. In other words, the model was predictive for two thirds of the total genes shown to vary in his model. Scooter therefore modifies his model, decreasing the verified gene set by 100.

The question, however, remains: Given that 20 animals were sampled and that the statistics "looked good," why are the findings on 100 genes not reproducible? Another way to address the issue is to note that the formulation of the model was premature; it should not have been attempted before the experiment had been repeated. This was evidenced by the relative lack of consistency in the data. This issue is discussed further in the next section.

TAKING INTO ACCOUNT THAT VARIABILITY IS BIOLOGICALLY RELEVANT TO THE EXPERIMENTAL SYSTEM BEING STUDIED

Scooter is somewhat depressed about his progress. He decides to try to figure out what is going on with his rats, so he takes a walk to the animal room where the rats are housed. The animal room contains a pretty sophisticated setup: There are infrared cameras in place so that the animals can be visualized at all times during the day. These cameras provide a continuous feed to large computer monitors outside of the room, so an observer can watch the rats without actually entering the facility and disturbing them. Furthermore, the rats are housed in so-called "metabolic cages" that

allow the scientist to monitor specific behaviors, such as when and how much the rats eat, their degree of movement, and their respiration. So, to capture a more personal picture of the rats, Scooter pulls a nice comfortable chair over to the computer monitors, takes a seat, pulls out a notebook, a pen, and a timer, and begins to observe the animals. It is 6:00 a.m. when Scooter begins his observations.

One monitor is broadcasting images from a cage of rats being fed a normal diet. The monitor next to it is focused on a cage of rats being fed a high-fat diet. The food is on the top of each cage, and the rats eat ad libitum through the bars of the cage. Scooter thinks about his experiments: The rat livers had been taken 12 hours after the high-fat diet had been initiated. Scooter had gone into the animal rooms at the end of the previous day, at 7:00 p.m., and manually switched the food for the rats on the high-fat diet. At 7:00 a.m. the next day, the experiment was completed.

Now, upon observing the rats, Scooter notes that some animals are eating pretty regularly, whereas others are not. During the first hour of observing rats in the high-fat diet cage, one big fellow steadily chewed away, whereas two other rats ate only once. A fourth rat did not eat at all. By the end of the hour, all of the rats had stopped eating. During the second hour, which was from 7:00 a.m. to 8:00 a.m., only the big fellow ate any food.

Scooter has access to a lot of previously unexamined data on his rats from the "metabolic cage" setup. He goes back to the data on food consumption and activity and learns that rats—being nocturnal creatures—actually do most of their eating while it is dark in the room. By 7:00 a.m., when the study had been completed, the animals were in the process of settling down to sleep after a long night of activity and eating, which is probably why, during the first hour of observation, some animals had completely stopped eating, others were winding down, and only one continued to feed. When Scooter carefully scrutinizes the food-intake data for the evening, he sees that there is much less variability in consumption between the hours of 11:00 p.m. and 5:00 a.m.

Now that he has become sensitized to the need to assess variability in feeding behavior, Scooter downloads all of the data on his rats, performs the proper statistical analyses, and determines a time period in which food intake is least variable. On the basis of these results, he decides that in the future, he should take his 12-hour time point at 4:30 a.m., not 7:00 a.m. Of course, this decision that Scooter just made should itself be tested by experimentation: He should ask whether the early gene responses to the high-fat diet are more stable at 4:30 a.m. or at 7:00 a.m., or even whether the gene-expression changes observed at the different times are biologically relevant. By undergoing this quandary, Scooter is gradually deriving a rationale for the "time-course experiment." He, and any scientist studying any system, must know whether the time point chosen to measure the "experimental case" is representative of the general condition.

Scooter, however, has not yet considered a time-course experiment; he is still attempting to obtain reproducible data at a single time point. In addition, now that he has done a more thorough job of system validation, he repeats his study. He initiates the high-fat diet at 4:30 p.m. and takes his data at the single time point of

u 6 p
sub jid
100
7-86
with pda

prah
piding
ais se
nagujedran
ad 11pm
do 5.00a
4.30am

uzchi
za
exp

NT

4:30 a.m. This time, he finds that the data are more reproducible: After performing the experiment for a second time, again using the 4:30 a.m. time point, 95% of the gene changes seen during the first experiment were documented a second time.

Scooter has now succeeded in producing data that hold up to repetition. Thus, he can say that his model is predictive of how gene expression will change in a similar experimental system in the future. However, one should pause here and ask whether Scooter answered the question that he set out to address. ?

ENSURING THAT THE EXPERIMENT IS DESIGNED TO ADDRESS THE QUESTION

Scooter had justified the 12-hour time point with the rationale that an early measurement would allow for the documentation of changes in gene expression that may involve the adaptation of the liver to the high-fat diet.⁶ When he read some papers on obesity, he noticed that other scientists had shown that gene changes can be more extreme after initially switching diets. However, remember that Scooter's experimental question did not focus on the issue of acute changes upon starting a high-fat diet. The initial question was "What gene-expression changes are observed in the livers of rats that are on a high-fat diet—sufficient to cause obesity—compared to those that are on a normal diet?" The question was not "What gene-expression changes are observed in the livers of rats 12 hours after starting a high-fat diet compared to those that are on a normal diet?"

In checking back to the "definitions" chosen for the terms in the experimental question, one can see now that Scooter did not define "diet" in terms of the length of time on the program. Is the point of the question to understand what is happening in the liver of individuals who have been on a high-fat diet long enough to become obese, or is the point of the question to understand how the liver adapts to a high-fat diet? Put in those terms, it becomes obvious that, although a 12-hour time point might be of interest under the second interpretation, it may not capture the purpose of the question if the first interpretation is what was desired.

This confusion is due to the scientist having established an experimental design before there was a clear understanding of the experimental framework. Scooter will come to realize this issue after performing enough repetitions to understand that gene expression continues to change over time, as the animal first becomes obese and then later becomes insulin-insensitive, before eventually developing full-blown diabetes. He therefore must choose what aspect of the problem he is really interested in and refine the experimental question accordingly. For example, if the purpose of the

⁶ . . . at least that is his claim. However, another scientist might actually have had this idea in the first place—that gene changes early on may provide insight into how the liver adapts to an increase in fat content in the diet. Whatever the rationale, this may not be a justifiable reason to perform the experiment as Scooter had designed it, for reasons that will now be discussed.

prochlost

avalon
realtor
news
recess
prochlost

study is to model what genes are elevated in an obese rat on a high-fat diet, then it may seem especially wrong-headed to study very early changes in the liver. Perhaps Scooter should, instead, wait for the animals on the high-fat diet to become obese—by whatever criteria have been established for obesity—and then conduct the study.

The point of these ruminations is to illustrate the kinds of issues that are critical to a successful experimental program but that usually do not crop up after an experimental design is in place and statistics have been performed to see whether a “significant” result was achieved. As was shown, it is perfectly feasible to obtain a tight, reproducible data set, and yet come to conclusions that are not responsive to the experimental question. In the current case, it may be a non sequitur to use gene changes that occur after 12 hours on a high-fat diet as responsive to a question about gene changes in a diet that is sufficient to induce obesity. There is an unjustified assumption in Scooter’s initial experimental design that the changes one sees at the start of the diet are representative of those one sees after obesity has developed.

This returns us to the issue of variability. To understand how many measurements must be taken, the scientist must conduct multiple measurements and to correlate those to the full spectrum of the subject. It is interesting that a rather simple validation question is so useful in bringing out these issues:

How many repetitions are necessary in order to obtain a representative result?

The crux of the question is the word “representative.” Scientists must determine what it is they want to study. This is now illustrated further below.

Scooter decides to put his rats on a high-fat diet for 3 months in order to determine what gene-expression changes occur over that time period, in comparison to controls. In addition, he will take a series of metabolic and blood chemistry measurements to determine whether any biochemical changes occur in the animals over time. Finally, he will measure the animals’ body mass indices. When he conducts this large experiment, he learns the following information:

1. On a high-fat diet, rats immediately start to gain weight when compared to control rats on a normal diet. After the first few days, the changes in gene expression are stable for some time. In other words, the changes seen after 72 hours on a high-fat diet are very similar to those seen after 1 week on a high-fat diet.
2. After 2 weeks on a high-fat diet, the animals begin to develop insulin insensitivity and elevated glucose levels. At this time point, new changes in gene expression can be seen in the liver.
3. After 8 weeks on a high-fat diet, the animals are obese and diabetic, and liver enzyme levels begin to rise. At this time point, there are, again, new changes in gene-expression patterns.
4. After 3 months on a high-fat diet, fatty deposits in the liver are obvious. There are further changes in gene expression at this time point.

THE COMPETENT STUDY, IN WHICH "MARKERS" ARE USED TO DETERMINE THAT THE SUBJECT OF THE QUESTION IS REALLY BEING ADDRESSED

Scooter started with this experimental question: "What gene-expression changes are observed in the livers of rats that are on a high-fat diet—sufficient to cause obesity—compared to those that are on a normal diet?" By performing sufficient system analyses, he learned that there are a range of changes in gene expression that correspond to different phenotypic "markers" observed in obesity: weight gain, insulin insensitivity, liver enzyme changes, and a fatty liver.

These "markers" are critical to a successful experimental design. The marker that is chosen as the definition for the experimental outcome will allow Scooter and other scientists to repeat the experiment and to establish a firm association between the phenotype of interest and the gene changes that occur in the setting of that phenotype.

For example, Scooter may decide that a key "marker" in this study is obesity, which he defines by the weight gain and insulin insensitivity observed at 2 months, as well as by the underlying gene-expression changes that occur at this stage. One need not then focus on the further changes in gene expression that happen at 3 months; those changes are no longer relevant to the issue. Similarly, if it is decided that an animal must be obese before a gene change is considered relevant, then all of the experiments conducted at the 12-hour time point can be discarded.

The scientist might define different "markers" for the study, and those decisions would not be "wrong." However, having a clear "marker" or criterion for defining the relevant readout allows for a rather simple final experimental design.

THE FINAL EXPERIMENTAL DESIGN

Scooter decides that in his experimental framework the question "What gene-expression changes are observed in the livers of rats that are on a high-fat diet—sufficient to cause obesity—compared to those that are on a normal diet?" refers to the situation where the animals are obese and diabetes is established. He now designs the following experiment:

1. Ten animals will be fed normal diets.
2. Another set of ten rats will be fed high-fat diets.
3. Weights and serum chemistries will be determined every week. When the animals exhibit obesity and insulin insensitivity (which previously occurred by 8 weeks after initiation of the high-fat diet), livers will be taken for analysis.
4. The animals placed in each group will be matched for age, gender, and initial body weight.

When this study is over, Scooter finds 500 genes that are altered in the “obese” livers compared to “control” livers. When he repeats the study, he finds that 90% of the genes are found to change again the second time, meaning that changes in 450 genes were reproducible. Scooter then performs the experiment a third time, under the question “Which of the genes reproducibly change on a high-fat diet—sufficient to induce obesity—as determined by weight gain and insulin insensitivity?” In this third experiment, he finds that 430 of the 450 genes change in the same way. He then constructs a model that includes the 430 genes as “markers” for the gene changes seen when a rat is on a high-fat diet sufficient to cause obesity, as determined by weight gain and the onset of insulin insensitivity.

Scooter can finally verify his model by asking whether it accurately reflects the gene changes seen in a future experiment, when an animal is put on a high-fat diet until it becomes obese and insulin-insensitive. Scooter does this fourth experiment, and using the appropriate statistics, he finds that 425 of the genes again demonstrate the same perturbations. Thus, he has generated a verified model for how gene expression is altered on a high-fat diet, during obesity, as defined in this experimental project.

genes are
insulin

er that
other
between
of that

obesity,
months,

One
en at
decided
all of

decisions
ing the

expres-
sion to
situation
lowering

animals
weeks

initial



The Requirement for the Negative Control

EXPERIMENTAL QUESTIONS OFTEN TAKE THE FORM “What is the effect of X on Y ?” To see the effect of X , one needs to compare the setting in which X is added to the system with the condition in which X is not added to the system. Thus, there is a required additional component to the question “What is the effect of X on Y ?”: “What is the effect of X on Y , in comparison to ‘not X ’?” The purpose of this “control” is to ensure that the conditions under which the experiment is performed vary only for X .

THE NEGATIVE CONTROL AS THE UNPERTURBED CASE

When one asks “What is the effect of caffeine on blood pressure?,” the answer can be determined only by having something with which to compare the effects of caffeine, that is, the setting in which there is no caffeine. If coffee was used as the caffeine source, it would be important to know that the test subjects do not consume other drinks that may perturb blood pressure. To understand the effect of coffee, the scientist will also need a reference point: a set of noncoffee drinkers with whom to compare the coffee drinkers. Without this “negative control,” the scientist would not know whether caffeine had induced any change.

In its simplest form, the “negative control” is the setting in which the experimental subject is not perturbed by the variable under study. This may sound straightforward, but it should be pointed out that the goal of achieving the “unperturbed case” is often difficult. Think again of the experimental question “What is the effect of caffeine on blood pressure?” In a study designed to address this question, subjects in one group would be given caffeine to ingest, whereas another group would not be given caffeine. In gathering people for the two study groups, the “caffeinated” group and the “negative control” group, first consider what would happen if the chosen “negative control” group had other conditions that altered their blood pressure. For example, what if the “negative control” group had a greater number of people who

tend to get anxious around medical personnel? In that case, the noncaffeinated group might end up with a higher-average blood pressure than the caffeinated group due to this undetected variable.

Next, consider what would happen if individuals in the “caffeinated” group contained a greater number of people with hypertension or another medical condition that elevated blood pressure. In this setting, the scientist might conclude that the caffeine resulted in the higher blood pressures, but in reality, it was the prevalence of additional medical conditions in the “caffeinated” group that resulted in this effect. Thus, it should be clear that individuals in each group must be controlled for possible “relevant variables,” meaning additional factors that may influence the experimental output.¹ This issue highlights another setting in which accessing the “inductive space”² is useful; without doing this, one would be more likely to corrupt a study with multiple variables that affect the output.³ With the inductive space in mind, one might take several steps to establish that the unperturbed “negative control” group is matched with the “caffeinated” group for whatever relevant variables that can be identified in advance. For example, the scientist who is designing the study might decide that each group should be screened for preexisting hypertension, and those who have such a condition should be eliminated from the study. In addition, the starting blood pressure and the BMI⁴ should be determined for each individual, and individuals of similar blood pressures and similar BMIs should be distributed equally between the groups.⁵ The groups should be age-matched and controlled for other substances containing caffeine.

Let’s stop at this last point: What if the screening questionnaire simply asked if people had ingested coffee that day, but it failed to ask whether individuals had ingested other caffeinated beverages such as tea or soft drinks? To establish that the groups are not caffeinated before the study, it may be necessary to have all of the individuals come into a controlled area where they do not ingest anything that contains caffeine

¹A relevant variable might be the fact that the technician measuring blood pressure is very frightening. This is opposed to a “dependent variable,” which is what the scientist is trying to measure. In the question “Does caffeine increase blood pressure?,” the dependent variable is blood pressure. Dependent variables are discussed in a later section; they need to be understood and handled distinctly from relevant variables.

²For those who did not read the previous chapters, “accessing the inductive space” means constructing an experimental design using previously discovered information. But you should really go back and read the rest of the book; it is quite fascinating.

³For example, it is a mathematical certainty that if something known to affect blood pressure is left unaccounted for, then the resulting data will be more variable. Realizing this encourages the scientist to identify all of the potential variables that may exist, which in turn encourages the scientist to control for as many variables as possible. This point is examined in more detail in the section on using individual study subjects as their own controls.

⁴Body mass index is a measurement of weight as a function of height and gender.

⁵Group randomization is discussed later. There are many approaches to this idea, but the strategy advised here is that potential study subjects be screened for independent variable criteria and be distributed among the study groups based on those criteria. For example, if there are two 45-year-old women that weigh 60 kg and have blood pressures of 120/85, one of these women would be placed in the “caffeinated” group and the other would be placed in the “control” group.

for several hours before the experiment. All of these steps may be taken to ensure that the unperturbed “negative control” group contains people as similar as possible to the “caffeinated” group, except that they lack caffeine.

One might say that controlling for relevant variables within each group is an attempt to ensure that the comparison between the “ X ” and “not- X ” groups provides a setting where “ X and only X ” is varied. In the present case, it is critical that the two groups do not vary in other factors that may make it impossible to understand how caffeine affects blood pressure.

QUESTIONS THAT CONTAIN MULTIPLE VARIABLES LEAD TO THE CONSTRUCTION OF THE “UNPERTURBED BY X ” NEGATIVE CONTROL

Next consider the more complex framework question “What is the effect of caffeinated coffee on blood pressure?” In this case, the scientist is interested in two things: whether caffeinated coffee changes blood pressure and whether the caffeine in the coffee is the culprit. Therefore, the scientist must be aware that there is more than one variable of interest. This might be made more explicit by stating the question as follows:

Does caffeinated coffee affect blood pressure, and if so, is the caffeine responsible?

To answer questions such as this, an additional set of controls is needed to dissect the variables “coffee” and “caffeine.” One might consider these as “unperturbed by X ” controls, where X is coffee in one case and X is caffeine in another. For example, the scientist could establish a study of the following design.⁶

Study Question: Does caffeinated coffee affect blood pressure, and if so, is the caffeine responsible?

Study Design: Establish six groups, each containing 50 people. Determine the subjects’ starting blood pressures. Eliminate from the study anyone with blood pressure greater than 140/90 or less than 90/60. Match study subjects by gender. BMI should be no greater than 40 and no less than 19. Match groups for average BMI and matched ranges of BMI. Subjects should be aged between 18 and 45. Match groups for average age and similar ranges in age. Study subjects should not ingest caffeinated beverages for 72 hours before the study. Question the study subjects for normal daily caffeine intake, anxiety disorders, and family history of hypertension. Exclude from the study people with anxiety disorders. Measure blood pressure twice each day, every day for 1 month. Take serum and urine samples at study initiation, and every third day during the experiment; save them for future examination. Also take DNA samples for future study. Treat the groups as follows:

⁶It should be noted here that many of the issues that go into this study design are discussed in the other chapters on experimental controls; in this chapter, we are concentrating on the negative control.

Group	Treatment
A	No treatment
B	Water: Four 8-ounce cups per day, as matched with other groups
C	Decaffeinated coffee: Four 8-ounce cups per day
D	Caffeinated water: Four 8-ounce cups per day; caffeine levels set to match the caffeine in the caffeinated coffee
E	Caffeinated coffee: Four 8-ounce cups per day
F	Caffeinated cola: Four 8-ounce cups per day; caffeine levels set to match the caffeine in the caffeinated coffee

The degree of change in the blood pressures of individuals in each of these six groups will be compared to one another. Each of the groups, except for the caffeinated coffee group, establishes an “unperturbed by X ” negative control. Below is the explanation of the “ X ” for each group:

Group	Treatment
A	No treatment: This group is unperturbed by any change, including the fluid change provided by drinking four cups of water per day.
B	Water: This group is unperturbed by the additional ingredients provided by the coffee or the cola, each of which is dissolved in water.
C	Decaffeinated coffee: unperturbed by caffeine, in comparison to the caffeinated coffee
D	Caffeinated water: This group is unperturbed by coffee, in comparison to the caffeinated coffee; it is also unperturbed by the other ingredients in caffeinated cola, in comparison to the caffeinated cola.
E	Caffeinated coffee: This group is the test case.
F	Caffeinated cola: This group provides an assumption control; this is not a negative control.

Now consider how changes in blood pressure are determined in this study. First note that the starting blood pressure for each individual is taken. Remember that the question is “Does caffeinated coffee affect blood pressure, and if so, is the caffeine responsible?” Therefore, to measure a change in blood pressure for each person, starting blood pressure must be measured to determine whether the test condition is responsible for a change.

Next, let’s examine each study group and determine how data obtained from that particular group may affect the construction of a model designed to address the study question. One would not be able to assess whether it was the caffeine in the “caffeinated coffee” as opposed to some other ingredient in coffee that causes a change in blood pressure unless it were possible to compare the caffeinated coffee group with the decaffeinated coffee group, and the water group. Without the just-water group, a blood pressure effect contributed by an ingredient of coffee other than caffeine would be missed. The caffeinated water group allows the scientist to determine whether caffeine is sufficient to induce a change in blood pressure. Let’s dissect the study to illustrate these points. Imagine a setting where the scientist establishes only the following two groups:

Group A	Decaffeinated coffee
Group B	Caffeinated coffee

The scientist reasons that the decaffeinated coffee controls for everything except caffeine, so it provides the ideal negative control to ask whether caffeinated coffee increases blood pressure, and whether the caffeine is responsible.

Let's say that the following results are obtained from such a study:

Decaffeinated coffee	10% increase in blood pressure
Caffeinated coffee	30% increase in blood pressure

Without the additional controls, one might say that the caffeine in caffeinated coffee triples the increase in blood pressure seen with decaffeinated coffee. And one would have difficulty concluding what caused the effect seen with the decaffeinated coffee. The scientist might be tempted to ascribe the 10% effect to some other uncontrolled variable, such as anxiety induced by study conditions, and thus ascribe the caffeine in coffee as the major cause of blood pressure increases induced by caffeinated coffee. Now, let's say that some additional controls are included and the following results are obtained:

Decaffeinated coffee	10% increase in blood pressure
Caffeinated coffee	30% increase in blood pressure
Caffeinated water	10% increase in blood pressure

If these were the results, one might see that caffeinated coffee caused a 30% increase in blood pressure, but caffeine by itself induces only a 10% increase. Therefore, it now seems that there may be a greater-than-additive effect of the caffeine and some other ingredients in the decaffeinated coffee that affects blood pressure. This additional control significantly alters the conclusion that the scientist reached when caffeinated water was excluded from the study; in the first case, caffeine was assigned as the major cause of blood pressure increases. Now, it seems that caffeine does not have the major role, but that it has a significant additional effect when combined with other ingredients in coffee. Let's see what happens when more "unperturbed by X" controls are added:

Decaffeinated coffee	10% increase in blood pressure
Caffeinated coffee	30% increase in blood pressure
Caffeinated water	10% increase in blood pressure
Water	5% increase in blood pressure

With the additional control provided by the just-water group, it now seems that the effect provided by caffeine alone is even less than it previously appeared to be. Simply drinking four additional glasses of water per day increased blood pressure by 5%. Thus, the model as to how caffeine works may be very significantly altered; by itself it has a very small effect, but in the context of coffee, it triples the increase in blood pressure compared to decaffeinated coffee alone. With these data, the scientist might start looking at the other ingredients in coffee that synergize with caffeine to cause such a significant rise in blood pressure. Finally, let's include the rest of the controls:

Decaffeinated coffee	10% increase in blood pressure
Caffeinated coffee	30% increase in blood pressure
Caffeinated water	10% increase in blood pressure
Water	5% increase in blood pressure
Caffeinated cola	10% increase in blood pressure
Nothing	0% increase in blood pressure

With these results, one might still conclude that there is a benefit in switching to decaffeinated coffee, but also that there is something in the decaffeinated coffee that increases blood pressure relative to water. The “no treatment” control demonstrates that there is a “real water” effect that is not caused by some uncontrolled variable such as anxiety. If the “nothing” group had also experienced a 5% increase in blood pressure, then that degree of increase might be attributed to the study conditions.

At this point, it should be clear why “unperturbed by X ” controls are important to good study design. Additional comments should be made about the study design as it was constituted. First, having a sufficient number of people in each group will “power” the study to control for undetermined genetic variations. If every human were a genetic clone, it would be unnecessary to have large cohorts of people in human studies. But because people are so heterogeneous, it is necessary to study many individuals to obtain the “general case.” This point was discussed in Chapter 10. In this chapter, it can be added that in our genetically heterogeneous population, multiple individuals help to establish the “unperturbed by X ” negative control. The use of a large group of people can increase the odds that undetermined relevant variables will be randomized between the groups, and thus controlled for. Second, note that multiple measurements are made for blood pressure, consistent with the chapter on repetitions. System controls and potential dependent variables for this study are covered in a later chapter.

For now, by following this example, we can see that the quest for the “unperturbed case” may not be simple, especially when dealing with complex study subjects such as humans or other animals. Nevertheless, it is quite important to meet this goal insofar as possible. Multiple “unperturbed by X ” controls allow the scientist to dissect out multiple variables that are contributing to the observed outcome. These controls are more effective than merely having a simple “unperturbed” setting. Put another way, the fewer negative controls in place,⁷ the more skeptical one should be about the conclusions drawn.

As noted previously, the indication that data obtained from a study are reliable depends on the ability to repeat the study and reproduce the same data. Whether the mechanistic conclusions drawn from the study are correct can be determined by performing new experiments with more directed questions, as has been illustrated in the earlier examples.

⁷That is, a control that represents an “unperturbed case” for the variable(s) present in the study, which could alter experimental outcomes.

ESTABLISHING THE "UNPERTURBED BY X" NEGATIVE CONTROL IN A TISSUE CULTURE EXPERIMENT

The human caffeine study is fairly complicated, and identifying the "unperturbed case" may be easier in simpler settings. However, care must still be taken to adequately understand the experimental system. Take, for example, the question "Does nerve growth factor (NGF) cause phosphorylation of Akt?" as applied in a setting where a clonal cell line is used for the study.⁸

To explain why the scientist may be interested in this question—NGF is a secreted protein that induces certain cells to become "neuron-like"—the precursor cells differentiate into cells that put out very long processes that resemble axons, and these cells then express protein markers normally found in neurons. The scientist may wonder which of the many potential cell-signaling pathways that exist might be responsible for the cellular differentiation induced by NGF. The protein Akt has been shown in other cellular settings to increase cell survival, but at the time that the scientist performs the experiment, Akt has not been shown to be required for neuronal differentiation. Therefore, the scientist first wants to ask whether NGF stimulates Akt at all and, if so, to determine whether Akt might be required for NGF-induced differentiation of cells into the neuronal phenotype.

Tissue culture—where cell lines⁹ are immortalized and maintained in controlled settings in laboratory incubators—provides a very common and tractable experimental system. A particular immortalized cell line may be studied worldwide in hundreds of different labs. The availability of such material allows many scientists to study highly similar cells in similar settings, providing scientists with a means of checking and building on each others' data. The data that are obtained by using these cells can then be applied to whole animals, and it can be determined whether the tissue culture results are relevant to the intact animal. The reason for initially performing experiments in the tissue culture setting is that the clonal cell allows for better controls of relevant variables and for experiments to be more easily conducted than animal studies. This does not mean that cell culture data always translate well to the animal, but

⁸NGF is a secreted protein that induces phosphorylation of certain cellular proteins in cells expressing the "receptor" to NGF. The receptor for NGF is called TrkA, which is a protein that can mediate the phosphorylation of other proteins. Sometimes, phosphorylation results in the binding of two proteins together, and sometimes it results in a cellular activity, such as an increase in synthesis of new proteins. Phosphorylation of certain proteins called transcription factors may lead to gene activation. Phosphorylation of another type of protein might cause it to become degraded, thereby modifying other cell-signaling pathways. But in order for TrkA to initiate these processes, it must first be bound by NGF. NGF brings two TrkA molecules together, and that process, called "dimerization," activates the system—one TrkA molecule phosphorylates the other in the dimer, which gets the ball rolling. For the purpose of this example, it can be stipulated that if TrkA is not present on a cell, NGF will have no effect because it has no other means of initiating cell signaling other than binding its receptor. Finally, Akt is a protein that is activated upon phosphorylation, which then, in turn, participates in the sort of cascade described here, inducing the phosphorylation of still other proteins.

⁹A "cell line" refers to a population of cells obtained from a single clonal ancestor.

it does allow the scientist to see how the cell behaves on its own. This is valuable in itself; if the cell behaves differently within the animal, there may be some additional control mechanisms in the native context that do not exist in the setting of the isolated cell, and these additional mechanisms can then be determined by comparing the isolated cell to the cell within its normal context.

The question “Does NGF cause phosphorylation of Akt?” may at first seem very straightforward to control for: Take cells, treat some with NGF, and, as a negative control, leave other cells untreated. However, even this case is not quite that simple. A careful analysis will demonstrate the same need for defining terms and for system validation as has been demonstrated throughout this volume in other examples.

It is worth reexamining the issue of system validation, because it will help to illustrate that even an appropriate negative control has no meaning unless it is presented in a validated context. For example, without system validation, imagine what would happen if the scientist used NGF on cells that did not contain the NGF receptor, TrkA, to address the question “Does NGF cause phosphorylation of Akt?” In such an experimental setting, the answer to the question would of course be “no,” because such cells are unable to respond at all to NGF.¹⁰ This instance demonstrates again why defining terms and determining the inductive space that is relevant for the question are important. By asking “Does NGF cause phosphorylation of Akt?” the scientist means to ask the question “Does NGF cause phosphorylation of Akt in those cells that are capable of responding to NGF?” or “Does NGF cause phosphorylation of Akt in those cells that contain the NGF receptor TrkA?” If the scientist ignored the issue of establishing that the cells to be used in the experiment expressed TrkA, the answer would be meaningless. It would be as useful as if the question “Does caffeine cause an increase in blood pressure?” was addressed using dead people or those who were on antihypertensive medications. Experimental questions—and hypotheses—all have premises, and it is important to understand what these are and to use them in system validation. Without that as a preliminary step, the negative control is meaningless: It is not possible to show an “X versus not-X” effect if the system is incapable of responding to X in the first place.

To establish the system, our scientist therefore obtains cells that demonstrably possess the NGF receptor, TrkA. In the next chapter, on positive controls, this example will be pushed further to see how phosphorylation of TrkA is used to determine how much NGF should be given in order to ask whether Akt is phosphorylated upon NGF stimulation. However, for now, we can say that the “unperturbed by X” case is “no NGF” in a system capable of responding to the growth factor.

One further point must be made about the question “Does NGF cause phosphorylation of Akt?” One must establish an experimental system where an increase in phosphorylation of Akt can be determined. If Akt is already phosphorylated, even in the absence of NGF, then one would not be able to see whether NGF was having an

¹⁰There may be an additional receptor, the LNGFR, that mediates some responses to NGF, but for the purpose of this example, we stipulate that TrkA is the only receptor. The reader should know, however, that the actual situation may be more complicated.

INDUCTIVE
SPACE

Neg.
control

positive

increase
phosphorylation
Akt

effect. The “unperturbed by *X*” control thus means “unperturbed by things that impact Akt phosphorylation.” In other words, there must be a negative control for Akt phosphorylation; the requirement of measuring an increase in Akt demands a system in which Akt phosphorylation is normally low.

In a “normal” tissue culture setting, cells are maintained in a fluid medium that contains growth factors to promote cellular survival. It can be demonstrated that some of these other growth factors cause Akt to become phosphorylated. Therefore, to measure the effects of NGF, the cells must be “starved” for these other growth factors. As in the caffeine example, where people were deprived of coffee and other caffeinated beverages in order that the effect of caffeine could be properly determined, the cells in this case must be deprived of other factors that induce Akt phosphorylation. The study design may look something like this:

Study Question: Does NGF cause phosphorylation of Akt?

Study Design: After determining that a cell line called PC12 cells contains the NGF receptor, TrkA, place PC12 cells into each of 12 plates.¹¹ In a system-design experiment, determine how long it takes for Akt phosphorylation to diminish to baseline in the PC12 cells treated with media devoid of growth factors. Once this has been determined, take the 12 plates of starved PC12 cells and treat as follows:

Plates 1–3	Nothing
Plates 4–6	NGF, at levels sufficient to activate TrkA (as determined in a system validation experiment), dissolved in a buffer
Plates 7–9	Buffer alone
Plates 10–12	A growth factor known to induce Akt phosphorylation in PC12 cells

In this design, the negative controls are plates 1–3, which provide the “unperturbed by *X*” control, where *X* is NGF. The issue of positive controls is discussed more in the next chapter.

ESTABLISHING THE “UNPERTURBED BY *X*” NEGATIVE CONTROL IN A GENETIC EXPERIMENT

Consider the question “Does deletion of the *BRCA1* gene increase the incidence of breast cancer in mice?” To answer this question, mice in which the *BRCA1* gene has been deleted in breast tissue are compared to mice in which the gene is intact.¹² To

¹¹A “plate of cells” refers to the usual way cells are “cultured” or grown in the lab. Commonly, a particular cell line (meaning many cells that are derived from a single clonal cell) is grown in plastic “dishes” or “plates.” Each plate is surrounded by a plastic wall that is 1–2 cm in height. Usually, the cells adhere to the plastic on the bottom of the dish and are covered with “tissue culture media,” usually a red broth containing glucose, amino acids, salt, and growth factors.

¹²*BRCA1* is required for normal embryogenesis; embryos that are null for the gene do not survive more than 8.5 days after conception. Therefore, to study the gene’s role in breast cancer, “conditional” knockouts must be made, where the gene is deleted only in breast tissue. This is another example of building on prior knowledge to address a question in a prospective setting.

Handwritten notes:
 nes. ...
 name
 of
 defunct
 to do
 class
 to be
 to
 to
 to

eliminate as many relevant variables as possible, the scientist designing the experiment might establish breeding pairs from an inbred strain of mice that is heterozygotic for the *BRCA1* deletion, meaning that these mice can give birth to either “knockout” mice, which are null for *BRCA1* in breast tissue, or “normal” mice, which contain two normal copies of this gene in their breast tissue. Littermates, which by definition are closely related but differ in the presence or absence of *BRCA1*, would provide a system where most of the “relevant variables,” such as other genes that might differ between the “unperturbed” group and the “*BRCA1*-knockout” group, can be controlled for. Using mice that are inbred for many generations means that the scientist is starting with animals that are genetically identical, so that if the *BRCA1* gene is mutated and progeny are derived, the scientist can show that *BRCA1*, and only *BRCA1*, differs between the “*BRCA1*-knockout” group and the “negative control” group.

Let’s say that the scientist compares 20 “*BRCA1*-null” mice with 20 “control” mice that contain *BRCA1* and follows them over time. It is found that the breast cancer rate in the “*BRCA1* null” mice is 30%, and the incidence in the mice that have a normal pair of *BRCA1* genes is 2%. Note that without the “control” mice, there would be no way of determining whether there was a changed rate of breast cancer in a genetic setting that was null for *BRCA1*.

In a “system validation” study, the age at which mice tend to get breast cancer in the setting of the *BRCA1* deletion must be determined. Thus, a pilot study might be done, in which *BRCA1* knockouts are studied at various ages, to get an idea of when breast cancer can be observed. One might object that this inductive knowledge biases the study; in response, consider the case where mice are studied at only 2 weeks of age, and no breast cancer is found in the *BRCA1* nulls. This result is “correct,” but it does not address the issue of whether there is an increased incidence of breast cancer in the life span of the mice, and that is the purpose of the study. This issue was discussed in earlier chapters, showing why repetitive testing is required in order to get “the full picture.” Therefore, to determine in advance how long the animals must be studied for the breast cancer to become manifest is not biasing. It is responsive to the question. This setting is also analogous to the “What color is the sky?” question. The “full picture” would be illustrated only by following normal versus *BRCA1* knockouts over time and assessing their breast cancer rates at relevant intervals. The study could be designed in this way:

Study Question: Does deletion of the *BRCA1* gene increase the incidence of breast cancer in mice?

Study Design: Set up 20 breeding pairs of *BRCA1*^{+/-} mice in an inbred strain of mice. Genotype female offspring. Distribute littermates into three groups:

Group 1	<i>BRCA1</i> ^{+/+} or “ <i>BRCA1</i> -normal” mice
Group 2	<i>BRCA1</i> ^{+/-} or “ <i>BRCA1</i> -heterozygote” mice
Group 3	<i>BRCA1</i> ^{-/-} or “ <i>BRCA1</i> -knockout” mice

For each group, biopsy breast samples at:

- Two weeks
- One month
- Two months
- Six months
- One year
- 18 months
- Two years

Determine the breast cancer incidence for each group, at each time point. Let's examine a possible outcome of this study:

	Breast cancer incidence		
	Normal	<i>BRCA1</i> ^{+/-} (hets)	<i>BRCA1</i> ^{-/-} (knockout)
Two weeks	0%	0%	0%
One month	2%	0%	1%
Two months	2%	3%	2%
Six months	2%	3%	2%
One year	2%	3%	25%
18 months	2%	3%	30%
Two years	2%	5%	30%

From these data, a model can be built: Deletion of *BRCA1* in the breast tissue of mice results in a significantly increased incidence of breast cancer. Furthermore, the mice must be null for the gene; loss of only one copy does not significantly alter the incidence of cancer. It can also be shown that tumors do not appear until 1 year after birth (middle age for a mouse). Thus, for this time point, the earlier time points can be thought of as "unperturbed by 12 months" controls. In other words, one would not know if it actually took a full year for the effect of the gene deletion to become manifest unless the other time points were in place. Given the large leap in cancer incidence observed between 6 months and 1 year, one might decide to include time points at 8 and 10 months in a future study to further determine how the cancer develops. If there really is a dramatic sudden increase at a particular time point, then the scientist may think about surveying other gene changes that occur at that time to determine whether these potentiate the effect seen by deleting *BRCA1*. All of these notions would be the subject of new questions, for future experiments, after the data in this experiment were confirmed by repetition.

THE "NOT-X" CASE, AS APPLIED TO USING AN ANTIBODY TO DETERMINE WHETHER A PROTEIN IS PRESENT IN A TISSUE

For questions of the type, "Does Y contain X?" or "Is YX?," one needs to compare "X" to "not X." The negative control provides a point of reference, to assure that a measurement is occurring.

Consider the scientist, Betty Sue, who wants to determine whether skeletal muscle contains a protein called “nebulin” and where in the muscle the nebulin might be located. The first question can be stated as “Does skeletal muscle contain nebulin?” To answer this question, Betty Sue uses an antibody that she has demonstrated during system validation¹³ to react against the nebulin protein. Betty Sue then uses the anti-nebulin antibody in an immunohistochemistry experiment: Tissue sections of skeletal muscle are placed on slides, and the nebulin antibody is used to stain some of these sections. There are various ways to detect the nebulin antibody. In this instance, let’s keep things simple and say that the antibody is bound with a fluorescent protein, such that muscle structures that bind the nebulin antibody will glow in the dark.

Betty Sue applies the anti-nebulin antibody, and certain structures in muscle, such as the contractile apparatus, now glow in the dark. Can Betty Sue now answer her question and say that “yes, skeletal muscle contains nebulin”? How does the scientist know that the antibody is actually binding nebulin, as opposed to binding some other protein in skeletal muscle? This is the purpose of the negative control in the “Does *Y* contain *X*?” question, to demonstrate that *X* is being measured, as opposed to “not *X*.”

In the present example, what the “not-*X*” control must show is that the antibody binds nebulin, as opposed to “not nebulin,” meaning some muscle substrate other than nebulin. One way to address this is to take advantage of the nebulin peptides that were used to raise the anti-nebulin antibody.¹⁴ The idea is to bind the antibody to the nebulin peptide before putting it on the tissue, such that if the antibody still binds to the muscle, it must be doing so by binding some other protein besides nebulin. Therefore, anti-nebulin antibody already bound to nebulin peptides is a negative control for an unadulterated anti-nebulin antibody; it allows Betty Sue to determine *X* (nebulin) as opposed to “not *X*.”

A more recent innovation that allows for a less complicated “not-*X*” control is the generation of “knockout” mice (as mentioned above). If skeletal muscle tissue sections could be obtained from an animal that does not produce nebulin because the gene encoding it has been deleted, that would provide a more straightforward “no nebulin” control. In this instance, an anti-nebulin antibody could be applied to skeletal muscle from a “nebulin-null” mouse and to muscle from a normal mouse. If the anti-nebulin antibody showed the same pattern of binding in both mice, Betty Sue would know that the antibody was binding to proteins other than nebulin. If, however, Betty Sue showed that the anti-nebulin antibody bound to the contractile apparatus in the “normal” animals, but showed no reactivity in the “nebulin-knockout” animals, then she could say, “yes, there is nebulin in skeletal muscle at the contractile apparatus.” This model would then be subject to verification by a variety of means, which will be discussed further in subsequent chapters.

¹³Negative controls for system validation are discussed further in a separate section.

¹⁴An antibody can be “raised” against a protein by injecting the protein or a portion thereof (a peptide) into an animal (usually a rabbit). The animal’s immune system is usually stimulated by the injected material, resulting in the production of an antibody that can bind the protein or peptide.

THE INTRASYSTEM NEGATIVE CONTROL

For validation of practices within a system of measurement, a negative control is used to test the key system components. By “negative control,” we mean “not X ,” where X is the parameter being measured. Let’s return to the question “Is the sky red?” To answer this question, one needs to be able to compare red to a setting that is not red. Imagine that the scientist’s “red-ometer”¹⁵ was broken and locked in the “positive” position. Without a negative control, that is, something that is not red, there would be no way of knowing that the system was not working. If the scientist pointed the meter only toward something that was red, even the broken meter would give a “positive” readout, and the scientist would be under the false impression that it was working accurately. Thus, the idea that “the whole world is red” could result from reliance on a red-ometer that reported red and only red, irrespective of the true color. So, for the question “Is the sky red?,” an object that was green or blue, for example, must be used as a negative control. The blue object is not “the unperturbed case”; rather, it is the “not- X ” case, where X is red. Here, we see that a negative control may simply provide a point of contrast to ensure that a measurement is occurring.

THE NEGATIVE CONTROL FOR THE “IS Y X ?” QUESTION

Consider the example of a pH meter. This type of instrument must measure a range of conditions, from highly acidic (pH 1) to highly basic (pH 14). As will be discussed in the section on “positive controls,” to validate this system, “standards” of known pH (validated by some independent means) are needed. For example, to demonstrate that the pH meter is capable of measuring pH 12, a “standard” solution of pH 12 would be used. However, another solution would also be needed: One that is not pH 12, to prove that the “standard” really is being measured. Usually, the scientist simply uses water for this task; if the scientist did not have water as a comparison for the pH 12 standard, it would be impossible to prove that placing the probe into the pH 12 fluid altered its state. Thus, we see that the negative control provides a mechanism of identifying whether anything happens in the test case setting. As stated earlier, it is impossible to determine whether X causes Y unless it can be proven that Y happens. In the case of the pH meter, it is impossible to prove whether adding acid or base to a particular solution results in a particular pH, unless there are standards for that pH, and—as a negative control—a condition providing a comparison pH that is other than the standard (and thus not X).

THE INTERSYSTEM NEGATIVE CONTROL

How does the scientist know whether the system of evaluation being used in a particular experimental setting is itself causing the outcome to be altered? For example,

¹⁵This is the instrument used to measure the color red, introduced earlier in Chapter 2.

consider that a system was in place for a technician to measure blood pressure in a colony of monkeys. However, this technician was so beloved by the monkeys that they would relax whenever they saw him, and their blood pressures would decrease appreciably. As a second, perhaps more likely example, consider a tissue culture setting, where in order to add a growth factor to the cells, the plates need to be removed from the incubator. However, by removing the plates in this way, their temperature decreases and their carbon dioxide level is altered. What if these changes are sufficient to alter the outcomes being studied?

One way to determine this is to have an “unperturbed by X ” control, where X is the system being applied. In the case of the monkeys, a different system of measurement may be used that does not involve human contact. In the tissue culture example, a second system may be established in which the growth factor is injected while the plates are still within the incubator. These are examples of “intersystem negative controls” that ensure that the experimental system is not in itself perturbing the outcome.

BLINDED ANALYSIS PRESENTED AS AN “UNPERTURBED BY X ” NEGATIVE CONTROL

In controlling for measurement perturbations that may be induced by the system itself, one should not ignore the issues of the scientist as a potentially biased evaluator, and of the study subject as being affected by the study.

To minimize potential scientist bias, it is helpful to establish criteria for evaluation in advance of the study; furthermore, if the perturbation used for each group is not revealed, the scientist will be unable to apply a separate evaluation filter for one group versus another. If it is known that one group was treated with caffeine and another group was not, the scientist might be encouraged to look for instances of a caffeine effect in the plus-caffeine group and explain away blood pressure changes in the decaffeinated coffee group. However, if the data and statistical methodologies are evaluated and “locked in” before the group treatments are revealed, the scientist will have no choice but to accept the data and evaluation as they are.

In addition to controlling against an evaluator effect, there is also a need to control against a subject perturbation, meaning an “influencing effect” on the subject that may alter the experimental readout. The placebo effect has been well documented elsewhere. If patients are told that they are on a weight-loss drug which will decrease appetite and allow more exercise, they may feel encouraged to change their habits to meet expectations. However, if the patients are blinded as to which group they are in (drug or placebo), then an attempt has been made to control for a potential “influencing effect.” Let’s reexamine the experiment that is designed to determine the effect of caffeinated coffee on blood pressure. Here again are the groups that were established:

Group	Treatment
A	No treatment
B	Water
C	Decaffeinated coffee
D	Caffeinated water
E	Caffeinated coffee
F	Caffeinated cola

In the context of the negative controls, one may have increased appreciation for the water group and the decaffeinated coffee group. Having both water and caffeinated water groups would allow the scientist to test the effect of caffeine on a patient without them being able to distinguish which treatment they were getting, as long as the caffeine is tasteless¹⁶ and as long as the patient is not told what group they are in. Having decaffeinated coffee also serves as a placebo control for the caffeinated coffee. In both cases, the subjects know that they are drinking coffee, yet they can be blinded as to whether it has caffeine.

Missing from the experimental design of the caffeinated coffee study described earlier in this chapter was the statement that the study would be “double-blinded,” meaning neither the investigator nor the study subjects would know what each individual was being given to drink. By constructing the study in this way, there is an attempt to ensure the results are “unperturbed by X ,” where X might be the investigator, the study subject, or some interaction between the two.

SUMMARY

The establishment of the appropriate negative controls is among the most important factors in successful experimental design. The experimental perturbations must be controlled for in such a way that the scientist can understand the role of each variable in the system. The scientist must also be able to prove that a measurement is being made. In subsequent chapters, additional roles for the negative control are discussed, including significance, sufficiency, and necessity; these determinations are made by a combination of negative controls and other types of controls.

¹⁶... and devoid of color or odor.



The Requirement for the Positive Control

TO ANSWER QUESTIONS OF THE TYPE “What is the effect of X on Y ?” one must first prove that the relevant effect on Y can be measured; second, that Y is actually Y ; and third, that X is actually X .

The question “What is the effect of caffeine on blood pressure?” can only be answered in a setting where it is shown that changes in blood pressure are possible and measurable. To repeat a trivial example from the previous chapter: It would not be meaningful to give caffeine to a dead person or to someone on antihypertensive drugs, because in these settings, either there is no blood pressure at all or the ability to respond to blood pressure stimuli is blunted by medication. As an additional example, it is impossible to determine if caffeine causes an increase in blood pressure if the instrument being used gives an incorrect readout when blood pressure increases over the normal level. Finally, it is not possible to determine changes in blood pressure if the individual performing the reading does not know how to use the instrument.

Consider again the question “What is the effect of caffeine on blood pressure?” Because the question does not contain a bias as to whether blood pressure might be increased or decreased, the experimental system should be capable of measuring either case. This means that there should be positive controls established to demonstrate this capability. One source of such controls would be known blood pressure stimulants or depressants.¹ Using these controls, one can first do system validation and then have a mechanism to determine whether the system is still operational at the time the caffeine experiment is performed.

The scientist might object to using positive controls, because the idea that some other compound is “guaranteed” to increase or decrease blood pressure is based on prior studies, which may or may not be accurate. What if the previous

¹It may not be ethical to use a pro-hypertensive agent as a positive control in a human population.

data are wrong? If an incorrect report is used to establish the experimental system, then the scientist might think the instrument or methodology is invalid when the problem is really with the positive control. However, as was discussed in previous chapters, the scientist has a clear-cut way to determine if this sort of problem exists. The proof of the positive control is its ability to perform as predicted, in a statistically significant manner. For example, if the positive control antihypertensive drug is used on 15 patients, and none of them experience a decrease in blood pressure, then the positive control is not valid (assuming that the relevant other system validation has been done to show that the instrumentation and other aspects of the experiment are operational). However, if a statistically significant number do experience a decrease in blood pressure, then the positive control is valid, and the system is capable of measuring decreases in blood pressure. Most importantly, it is thus established that the subjects are capable of undergoing a change in blood pressure.

Note that a positive control cannot be used effectively if other aspects of system validation have not been performed. Had the measuring methodology not been validated, the scientist might wonder if a negative result obtained using the positive control was due to some other aspect in the system. Thus, if there are unvalidated components of the system, positive controls cannot give effective support for the methodology.

THE POSITIVE CONTROL IN THE CAFFEINE STUDY

In Chapter 12, on negative controls, a study was designed to address the question "What is the effect of caffeine on blood pressure?" Consider data obtained from a caffeine study that simply compared a negative control group given water to a "test group" given caffeinated water. The amount of caffeine dissolved in the water given to each subject is the equivalent of that found in the average cup of coffee. Below is the summary data comparing the two groups:

Treatment	% Increase in blood pressure
Water	10
Caffeinated water	10

From these data, one could only conclude that the caffeine had no additional effect on the experimental subjects in comparison to water alone, because there was no difference between the negative control (water) group and the test (caffeinated water) group. However, because there was no tool in place to demonstrate that the system was operational, the scientist who performed the experiment might wonder retrospectively whether the data are "valid" or whether there was a problem in some

aspect of the methodology. The 10% increase in blood pressure caused by the water might add to that impression.²

Consider the same data in the setting where a positive control is included. The positive control is a drug that has been previously validated to increase blood pressure.

Treatment	% Increase in blood pressure
Water	10
Caffeinated water	10
Hypertensive drug	30

The positive control demonstrates that the experimental methodology is functioning; an increase in blood pressure occurred, and this change could be measured. In the context of these data, the lack of a response using caffeinated water now looks much more convincing.

As noted early on in this volume, the scientist who really is convinced that caffeine increases blood pressure may not be satisfied with these data, but it is much more difficult to dispute the evidence in the context of the positive control. Still, the scientist might be tempted to give more caffeine, and this decision would be valid if the dose used in the first experiment was not consistent with some range of doses normally used by people the study intends to model. At least the positive control makes it clearer that the issue may be with the caffeine itself, as opposed to some other aspect of the experimental design.

Let's examine data produced using a positive control in a "dose-response" caffeine experiment, meaning that different test groups are administered increasing levels of caffeine. The scientist does some reading and determines from prior epidemiology studies that most people who drink caffeinated coffee consume in the range of one to four cups per day. Therefore, the scientist chooses this range for the dose-response

²The fact that water alone induced a 10% increase in blood pressure again illustrates the necessity of the negative control (water) group. If the scientist had just tested caffeine in patients—without the water-control group—the conclusion may have been that the amount of caffeine given in this first experiment increased blood pressure. Note also that one can justifiably call the "untreated state" of the caffeine group prior to treatment the negative control. Because the patients who were to be treated with caffeine had their baseline blood pressure measured, their "unperturbed state" is itself a negative control, as opposed to the "perturbed state" when caffeine was given. However, if only that negative control was used and the water control was left out, then the effect induced by the other experimental parameters would have been missed. The water serves as the "unperturbed by X" control discussed in the previous chapter; it gives the patient everything in the caffeine group except the caffeine. Including the water group allowed the scientist to determine that the stresses of the experimental conditions, or the act of drinking a liquid, were sufficient to induce a 10% rise in blood pressure in the negative control group.

experiment. Below are the data produced using such a regime:

Treatment	% Increase in blood pressure
Water	10
Caffeinated water; caffeine equivalent of one cup of coffee	10
Caffeinated water; caffeine equivalent of two cups of coffee	12
Caffeinated water; caffeine equivalent of three cups of coffee	15
Caffeinated water; caffeine equivalent of four cups of coffee	20
Hypertensive drug	30

The scientist does the appropriate statistics on the data and finds that the change induced by drinking three cups of coffee, four cups of coffee, or the hypertensive drug are all significant in comparison to the group that just consumed water. These experimental results give considerably more information than found in the previous design.

It is now clear that one starts to see a statistically significant effect only after administering at least three-coffee-cup equivalents of caffeine. This finding helps to confirm the previous data, which demonstrated that there was no effect with a one-coffee-cup equivalent. Note too the additional “contextualizing role” the positive control has in this experiment: It demonstrates that four-coffee-cup equivalents of caffeine each day causes a rise in blood pressure, although less than that induced by the hypertensive drug. Thus, the positive control provides a point of comparison by which one can show first that the system is operational and second how the test agent compares to the positive control in the aspect measured by the study. Next let’s consider what might have happened if the data from the caffeine study looked like this:

Treatment	% Increase in blood pressure
Water	10
Caffeine found in one cup of coffee	10
Caffeine found in two cups of coffee	10
Caffeine found in three cups of coffee	10
Caffeine found in four cups of coffee	10
Hypertensive drug	10

With these results, the scientist would know that there is a problem with the study, because the positive control did not result in an increase in blood pressure. Without the positive control, the scientist would have no reason to question the data, and the logical conclusion would have been that even the amount of caffeine found in four cups of coffee was not sufficient to induce a rise in blood pressure. But with the positive control in place, it is now clear that there was an actual problem in the study.

These kinds of data would induce the scientist to go back and check the system to see if it was operational. Part of that exploration would be to double-check the subjects used in the study. One might not be surprised to discover that the scientist in this case had neglected to ask whether the subjects were on antihypertensive medication.

**THE POSITIVE CONTROL ESTABLISHES THAT THE SYSTEM IS
OPERATIONAL, USING A PERTURBATION DISTINCT
FROM THE OBJECT OF THE STUDY**

In Chapter 12, on the negative control, a question with multiple variables was considered: “What is the effect of caffeinated coffee on blood pressure?” This can be stated as a two-tiered question: “Does caffeinated coffee affect blood pressure, and if so, is the caffeine responsible?” To address this question, a set of “unperturbed by *X*” controls that dissect the variables “coffee” and “caffeine” were put in place using the following study groups:

Group	Treatment
A	No treatment
B	Water
C	Noncaffeinated coffee
D	Caffeinated water
E	Caffeinated coffee
F	Caffeinated cola

It should now be clear that what was missing from this study is the positive control. The reader might ask whether the “caffeinated cola” group might function as a positive control. The answer is no. A positive control should demonstrate that the system is capable of detecting the experimental readout (in this case, an increase or decrease in blood pressure). However, it should do so using a perturbation distinct from that used in the study (in this case, something other than caffeine). Obviously, if it had already been established that caffeine induces a change in blood pressure, it would not need to be the object of the study; therefore, because caffeine has not been demonstrated to cause blood pressure changes, the positive control must make use of a validated agent. Below is the study with a positive control in place:

Group	Treatment
A	No treatment
B	Water
C	Noncaffeinated coffee
D	Caffeinated water
E	Caffeinated coffee
F	Caffeinated cola
G	Hypertensive drug

Once again, the hypertensive drug would prove that the system was capable of measuring a change in blood pressure. It would also provide a reference point to see how potent caffeine is in perturbing blood pressure (if it has any effect at all).

POSITIVE CONTROLS CAN TEST MULTIPLE ASPECTS OF A SYSTEM

So far, the only positive control that has been used in the caffeine experiment is a drug that will determine whether the readout for the experiment—a change in blood

pressure—can be assessed. However, there are many other aspects of the experiment that could benefit from a positive control. For example, it would be useful to prove that the experimental subjects were actually consuming, absorbing, and metabolizing the caffeine they were given. Such a positive control would have a few benefits: First, it would allow the investigator to cull those subjects who, for whatever reason, were not drinking their coffee or caffeine and were thus skewing the data inappropriately; second, it would control for other factors that might cause differential absorption and metabolism of the caffeine. This could be used to “normalize the data” and thus compare blood pressure increases in individuals who had been exposed to equivalent levels of the experimental drug. As a silly hypothetical, let’s say that the amount of caffeine usually found in a single cup of coffee turned the subject green and the equivalent amount of caffeine found in four cups of coffee turned the subject red. These colors can be thought of as “markers” for the amount of caffeine the subject has consumed, and measuring these markers would give the scientist a positive control to prove that the subject had in fact consumed a particular amount of the drug. The plausible way that a marker analysis could actually be done in the current study could be to measure caffeine-breakdown products in the subjects’ urine.³ By measuring such metabolites in the urine, the scientist can be assured that the subject had in fact been exposed to the experimental perturbation.

The scientist might think that measuring urine is not a positive control in the traditional sense, but rather an exercise in system validation; however, anything that helps to demonstrate that an aspect of the experimental design is operational, including that the experimental subjects drink their coffee, can be thought of as a positive control. Indeed, the main purpose of the positive control is system validation within the experimental framework. This will be demonstrated further using other examples.

ESTABLISHING POSITIVE CONTROLS IN A TISSUE CULTURE EXPERIMENT

Let’s return to another question that was asked in Chapter 12: “Does nerve growth factor (NGF) cause phosphorylation of Akt?” Recall that this question was addressed using a cell line that expresses the receptor to NGF, TrkA. It can now be pointed out that proving the cell line is “TrkA positive” and that the receptor TrkA is being stimulated upon NGF treatment are positive controls in this experiment.

Just as in the caffeine case, where a marker was needed to demonstrate that the subject is actually metabolizing the caffeine, in the present example, the investigator needs a control showing that the NGF is being effectively delivered to the cell. To understand the positive control in this example, the reader needs to know more about how NGF, which is delivered to the outside of a cell, causes intracellular signaling. The growth factor binds to its receptor, TrkA, which is partially exposed on the sur-

³Caffeine is metabolized in the liver, and the breakdown products paraxanthine, theobromine, and theophylline can be measured.

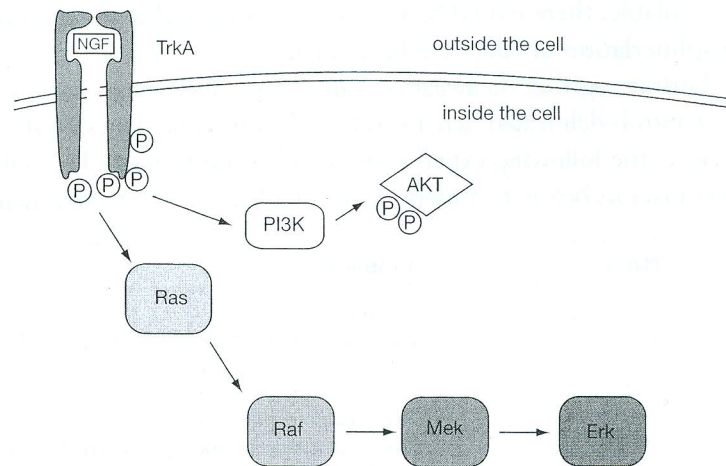


Figure 1. NGF binds to TrkA.

face of the cell, but which also spans through the cellular membrane and has a signaling component within the cell (Fig. 1). The signaling component of TrkA is an enzyme that causes the addition of a phosphate onto the amino acid tyrosine.

NGF actually binds two TrkA molecules and, in so doing, brings the signaling components of each of the two TrkA receptors into close proximity, allowing each to activate the other by adding phosphates to particular tyrosine amino acids (Fig. 1). These phosphorylated tyrosines become binding sites for intracellular enzymes, allowing them to bind and in turn be activated (Fig. 1). Once activated, the intracellular enzymes perturb other proteins in the cell. Such a sequential process of binding, phosphorylation, and activation is referred to as a signaling cascade, as briefly explained in the previous chapter.

To determine whether NGF causes a particular signaling molecule inside the cell, like Akt, to become activated, one would first need to know that sufficient NGF had been delivered to induce TrkA phosphorylation. This is why TrkA phosphorylation is a positive control for the question “Does NGF cause phosphorylation of Akt?” If TrkA is left unstimulated by NGF (e.g., if insufficient NGF is given or if the NGF used is in some way defective), then the question cannot be answered.

Two other positive controls are necessary to answer the question “Does NGF cause phosphorylation of Akt?” First, it must be shown that it is possible to measure phosphorylation of Akt. And, in a related control, it must be shown that Akt can be activated in the particular cell line being studied, using a positive control growth factor that has already been shown to be capable of activating Akt. Let’s say that a scientist, Carlita, consults the literature and finds out that a protein growth factor called insulin-like growth factor-1 (IGF-1) induces the phosphorylation of Akt in the relevant cell type. This cell expresses the IGF-1 receptor on its surface, in addition to the TrkA receptor. Using IGF-1 as a positive control, Carlita can stimulate the cells, test antibodies to phosphorylated Akt, and thereby demonstrate that activation and subsequent phosphorylation of Akt can be measured. If the IGF-1 positive control were

or lysed, using a detergent buffer. The detergent dissolves the cell membranes, allowing the proteins in the cells, including Akt, to escape into the buffer. The proteins in this “cellular lysate” can now be analyzed using specific antibodies that bind to particular proteins. Now that Carlita has her protein lysates, she lists the questions she needs to answer in the biochemistry phase of the experiment:

1. Was the IGF-1 receptor phosphorylated upon IGF-1 treatment? This control was not mentioned previously, but in order to ensure that the IGF-1 is actually functioning as a positive control, it is obviously important to prove that a sufficient amount of IGF-1 was used in the experiment and that the IGF-1 is patent.⁵
2. Was the TrkA receptor phosphorylated upon NGF treatment?
3. Was Akt phosphorylation induced by NGF and IGF-1, in comparison to the negative controls?

Note that before Carlita can answer her ultimate question, she first must go through the positive control experiments. Let’s say that the first two steps were done, and Carlita proved that she gave enough IGF-1 and NGF to induce cell signaling. Now she is ready to do the experiment that addresses the question “Does NGF cause phosphorylation of Akt?” To accomplish this, she uses two antibodies: one that recognizes only the phosphorylated (activated) form of Akt and one that recognizes Akt, whether or not it is activated. By using these two antibodies, Carlita can determine the percentage of Akt that is activated (by comparing the amount of phosphorylated Akt as to the total amount of Akt). Putting the positive control experiments and the Akt experiment together, below are the data Carlita obtained from this experiment:

Plate	Treatment	Akt activation	TrkA activation	IGF-1R activation
1	Nothing	(-)	(-)	(-)
2	Nothing	(-)	(-)	(-)
3	Nothing	(-)	(-)	(-)
4	Buffer	1%	2%	0%
5	Buffer	3%	3%	2%
6	Buffer	2%	0%	1%
7	IGF-1	220%	3%	500%
8	IGF-1	380%	8%	625%
9	IGF-1	340%	4%	400%
10	NGF	410%	745%	4%
11	NGF	290%	333%	4%
12	NGF	320%	530%	5%

⁵Similarly, TrkA and IGF-1 receptor activation is determined by comparing the amount of phosphorylation with the amount of total receptor protein. Thus, if the amount of phosphorylated receptor increases by 10%, and the amount of total protein does not change, the degree of activation is 10%. However, if the amount of phosphorylated receptor increases 10%, but it turns out that 50% more protein is present, then the actual percentage of phosphorylated receptor is substantially lower than in the control.

Looking at the data, Carlita can say the following:

1. The positive control worked. Upon treatment with IGF-1, sufficient to cause activation of its receptor, Akt became activated; the amount of Akt phosphorylation increased between 2- and 3.8-fold.
2. Sufficient NGF was given to cause its receptor, TrkA, to be activated. This proves that the system is “operational”; the cells were able to respond to NGF.
3. NGF induced Akt stimulation as effectively as did IGF-1.

From these data, one could conclude (upon appropriate repetitions of the experiment) that NGF induces phosphorylation of Akt. Now, consider what might have happened if the data looked like this:

Plate	Treatment	Akt activation	TrkA activation	IGF-1R activation
1	Nothing	(-)	(-)	(-)
2	Nothing	(-)	(-)	(-)
3	Nothing	(-)	(-)	(-)
4	Buffer	1%	2%	0%
5	Buffer	3%	3%	2%
6	Buffer	2%	0%	1%
7	IGF-1	220%	3%	500%
8	IGF-1	380%	8%	625%
9	IGF-1	340%	4%	400%
10	NGF	3%	745%	4%
11	NGF	7%	333%	4%
12	NGF	4%	530%	5%

If these were the experimental results, one could conclude that NGF was not capable of inducing Akt, at least under these experimental conditions. Note again the effect on the scientist of getting such “negative” data. If the IGF-1 positive control were not in place, then the data would like this:

Plate	Treatment	Akt activation	TrkA activation	IGF-1R activation
1	Nothing	(-)	(-)	(-)
2	Nothing	(-)	(-)	(-)
3	Nothing	(-)	(-)	(-)
4	Buffer	1%	2%	0%
5	Buffer	3%	3%	2%
6	Buffer	2%	0%	1%
7	NGF	3%	745%	4%
8	NGF	7%	333%	4%
9	NGF	4%	530%	5%

Because there is now no positive control for the activation of Akt, Carlita cannot determine whether her Akt antibody was working. Thus, she might dismiss these data as inconclusive. Such a conclusion would force Carlita to go back and validate the system with the phospho-specific Akt antibody, something she could have avoided if

she had simply had the positive control in place. Next, look at the data if the TrkA control was missing:

Plate	Treatment	Akt activation	IGF-1R activation
1	Nothing	(-)	(-)
2	Nothing	(-)	(-)
3	Nothing	(-)	(-)
4	Buffer	1%	0%
5	Buffer	3%	2%
6	Buffer	2%	1%
7	IGF-1	220%	500%
8	IGF-1	380%	625%
9	IGF-1	340%	400%
10	NGF	3%	4%
11	NGF	7%	4%
12	NGF	4%	5%

If the data looked like this, Carlita might still dismiss the experiment because, in this setting, there is no way to be sure that sufficient NGF was given. Carlita could simply say that “the NGF is probably dead.” It certainly looks dead, because nothing that was measured was influenced by the NGF used in this experiment.

It is much harder to dismiss the data when all of the positive controls are in place. Although it should never be the goal of the scientist to dismiss unwelcome data, humans are but human, and it is therefore best to have all of the relevant controls in place. Next, consider the interpretation of the experiment if the data looked like this:

Plate	Treatment	Akt activation	TrkA activation	IGF-1R activation
1	Nothing	(-)	(-)	(-)
2	Nothing	(-)	(-)	(-)
3	Nothing	(-)	(-)	(-)
4	Buffer	1%	2%	0%
5	Buffer	3%	3%	2%
6	Buffer	2%	0%	1%
7	IGF-1	220%	3%	500%
8	IGF-1	380%	8%	625%
9	IGF-1	340%	4%	400%
10	NGF	50%	745%	4%
11	NGF	40%	333%	4%
12	NGF	32%	530%	5%

Here, it would appear that Akt is in fact stimulated by NGF, but only to about 13% of the levels induced by IGF-1. Rather than going up 2- to 3.8-fold, as seen with IGF-1, the NGF is inducing an increase of only around 40%, even though there is a comparable percentage activation of NGF’s receptor (TrkA) in comparison to the IGF-1 receptor. This result highlights the contextualizing property of the positive control. Compare these data to those that would have been obtained had

cause acti-
orylation

is proves

s of the
what might

1R activation

- (-)
- (-)
- (-)
- 0%
- 2%
- 1%
- 500%
- 625%
- 100%
- 4%
- 4%
- 5%

tr was not
e again the
e control

activation

- (-)
- (-)
- (-)
- 0%
- 2%
- 1%
- 4%
- 4%
- 5%

l a cannot
these data
alidate the
rovided if

the IGF-1 control been missing:

Plate	Treatment	Akt activation	TrkA activation
1	Nothing	(-)	(-)
2	Nothing	(-)	(-)
3	Nothing	(-)	(-)
4	Buffer	1%	2%
5	Buffer	3%	3%
6	Buffer	2%	0%
7	NGF	50%	745%
8	NGF	40%	333%
9	NGF	32%	530%

With these data, an excited Carlita might simply conclude that NGF induces Akt activation. Without the context offered by the IGF-1 positive control, she would have no way of knowing that the amount of Akt activity induced by NGF was rather paltry in comparison to activity induced by another growth factor. Such information might bring into question whether this level of activity is sufficient to induce the entire cascade that normally results from the stimulation of Akt. The basis for such caution is removed when the positive control is eliminated. Finally, consider the conclusion that might be drawn if the data look like this:

Plate	Treatment	Akt activation	TrkA activation	IGF-1R activation
1	Nothing	(-)	(-)	(-)
2	Nothing	(-)	(-)	(-)
3	Nothing	(-)	(-)	(-)
4	Buffer	1%	2%	0%
5	Buffer	3%	3%	2%
6	Buffer	2%	0%	1%
7	IGF-1	220%	3%	500%
8	IGF-1	380%	8%	625%
9	IGF-1	340%	4%	400%
10	NGF	50%	74%	4%
11	NGF	40%	33%	4%
12	NGF	32%	53%	5%

In this case, the amount of Akt stimulation is proportional to the amount of receptor activation. For IGF-1, where its receptor was significantly stimulated, one sees a high level of Akt stimulation. However, when NGF was used, the TrkA receptor was stimulated to only 10% of the levels observed in the IGF-1 receptor upon treatment with IGF-1. So here, one might conclude that even though NGF activates Akt to a much lesser extent, the amount of Akt activation as a proportion of receptor stimulation is the same in comparison to IGF-1. Such a finding might be obtained if too little NGF was used, for example, or if it was not allowed sufficient time to stimulate signaling. Once again, we see that the positive controls give important context to the readout.

ESTABLISHING THE POSITIVE CONTROL IN A GENETIC EXPERIMENT

Now let's return to the genetic question asked in Chapter 12: "Does deletion of the *BRCA1* gene increase the incidence of breast cancer in mice?" Recall that in order to address this question, "conditional knockout" mice were generated, in which the *BRCA1* gene was deleted only in the breast tissue. These "*BRCA1*-knockout" mice were then compared to mice in which the *BRCA1* gene was intact in all tissues. The study, as it was designed in the previous chapter, had these groups:

Group 1	<i>BRCA1</i> ^{+/+} or " <i>BRCA1</i> -normal" mice
Group 2	<i>BRCA1</i> ^{+/-} or " <i>BRCA1</i> -heterozygote" mice
Group 3	<i>BRCA1</i> ^{-/-} or " <i>BRCA1</i> -knockout" mice

Each of those groups was to be biopsied over time, from 2 weeks to 2 years after birth. At every time point, the breast cancer incidence for each group was determined. Here was the first set of data, showing breast cancer incidence, for the study as designed:

Age	Normal	<i>BRCA1</i> ^{+/-} (hets)	<i>BRCA1</i> ^{-/-} (knockout)
Two weeks	0%	0%	0%
One month	2%	0%	1%
Two months	2%	3%	2%
Six months	2%	3%	2%
One year	2%	3%	25%
18 months	2%	3%	30%
Two years	2%	5%	30%

What should now be clear is that there was no positive control for this study. Several such controls would be desirable. First, one might take advantage of a gene or chemical that is already known to cause breast cancer to ensure that the mice are not abnormally resistant to cancer and to prove that the scientist can detect breast cancer when it is present. For example, if it is proven that a chemical called ethylmethane-sulfonate (EMS) induces on average a 25% incidence of breast cancer following 1 year of treatment, this compound can give a "baseline," or context, to the study. Let's see what the data would look like if such a positive control had been used on "normal" mice treated with the cancer-causing EMS:

Age	Normal	<i>BRCA1</i> ^{+/-}	<i>BRCA1</i> ^{-/-}	EMS treated
Two weeks	0%	0%	0%	2%
One month	2%	0%	1%	5%
Two months	2%	3%	2%	8%
Six months	2%	3%	2%	15%
One year	2%	3%	25%	25%
18 months	2%	3%	30%	40%
Two years	2%	5%	30%	80%

ices Akt
ould have
her paltry
n might
e entire
h caution
clusion

activation
(-)
(-)
(-)
0%
2%
1%
0%
5%
0%
4%
4%
5%

mount of
ed, one
A recep-
tor upon
activates
receptor
btained if
to stim-
context

Once again, the positive control provides some context for the experimental treatment. Both the chemical EMS and the loss of the *BRCA1* gene induced a 25% incidence of breast cancer at 1 year. However, for EMS, there was a linear increase in tumors over time. In contrast, there seems to be a “threshold” effect for *BRCA1*; a significant increase in tumors is evident only at year one, when there is a “jump” from the 6-month figure of 2%. The EMS data are especially helpful in making the scientist believe the *BRCA1*-null data from the 6-month time point. Given the severe increase in tumors observed between 6 months and 1 year, one might have legitimately wondered whether there was something “odd” going on at 6 months; perhaps tumors had been missed or were too small to see. However, with the EMS data, the scientist can be confident that whoever measured the tumors at the 6-month time point was able to spot those occurring in the EMS-treated mice. This would be especially convincing if the analysis had been performed “blind,” that is, if the scientist was unaware of which group was being scored. With the positive control in place, the data now make it all the more striking that there is such a jump in the *BRCA1*-null mice. This sort of event raises the question of whether something “special” happens between 6 and 12 months, such as a second genetic event, that renders the mouse sensitive to the loss of *BRCA1*. Such questions become more obvious in the context of the positive control.

A second point that can now be made is that, unlike the EMS group, where almost all of the mice eventually get cancer, only 30% of *BRCA1*-null mice seem to get the disease. Again, this makes one wonder whether *BRCA1* is sufficient to cause tumors or whether some other event is necessary, and whether this second event occurs in only 30% of the mice.

THE POSITIVE CONTROL, AS APPLIED TO USING AN ANTIBODY TO DETERMINE IF A PROTEIN IS PRESENT IN A TISSUE

Remember our scientist Betty Sue, who, in the previous chapter, wanted to determine whether skeletal muscle contained a protein called “nebulin,” and if so, where in the muscle the nebulin was located. What was missing in the last chapter was a positive control to prove that the nebulin antibody was functioning, that is, capable of detecting nebulin. The table below shows how the experiment was left in Chapter 12. Recall that the antibody was used to stain tissue sections obtained from muscle:

Sample	Muscle from nebulin knockout	Muscle from normal mouse
Nebulin antibody	No staining seen	Staining in contractile apparatus
Nebulin antibody plus peptide competition	No staining seen	No staining seen

However, what if the data had looked like this:

Sample	Muscle from nebulin knockout	Muscle from normal mouse
Nebulin antibody	No staining seen	No staining seen
Nebulin antibody plus peptide competition	No staining seen	No staining seen

If this were the result, there would be no way of knowing whether the lack of data was due to some problem with the nebulin-specific antibody. In addition, one would not know whether the method in general was working. For example, something in the way that tissue samples were treated may have prevented the antibody from detecting protein.

There must be two positive controls added to this experiment to demonstrate that the system is operational. First, an antibody should be used that is already proven to detect a protein in skeletal muscle. Prompted to use such a positive control, Betty Sue reads the literature and finds out that there is a huge protein in the contractile apparatus called "titin" and that her next-door lab-mate, Scooter, was the person who developed antibodies to titin. Betty Sue confers with Scooter, who gives her the positive control she needs. This is what the data look like using Scooter's anti-titin antibody (with the appropriate peptide-competition control for the titin antibody as well):

Sample	Muscle from nebulin knockout	Muscle from normal mouse
Nebulin antibody	No staining	No staining
Nebulin antibody plus peptide competition	No staining	No staining
Titin antibody	Contractile apparatus	Contractile apparatus
Titin antibody plus peptide competition	No staining	No staining

With the titin antibody positive control, Betty Sue can be confident that she performed the immunohistochemistry experiment correctly; in other words, she was able to detect a structure (the contractile apparatus) in the muscle tissue using an antibody, and this finding controls for most (although not all) aspects of the experiment. What is left to control for is the nebulin antibody itself. How can Betty Sue be confident that a negative result means that there is no nebulin, as opposed to the possibility that her antibody simply does not work?

The simple way to prove that the anti-nebulin antibody was capable of detecting nebulin would be to produce recombinant nebulin protein (or at least the portion of the nebulin protein against which the antibody was designed to react). This can be readily done⁶ in the laboratory using standard methods. This "recombinant nebulin"

⁶At least a portion of nebulin could be made through standard methods. Nebulin is huge and would thus be difficult to produce.

protein could even be “tagged” with an additional peptide against which the scientist already has a proven antibody. This “epitope-tagged” nebulin would provide an even stronger positive control because now the positive control peptide is physically fused to the protein that Betty Sue wants to detect. Here is how the data would look making use of all of the positive controls (to make things clearer, “no staining” will be designated as [-], and “staining” or reactivity will be designated as [+++]):

Sample	Nebulin knockout	Normal mouse	Epitope-tagged nebulin
Nebulin antibody	(-)	(-)	(-)
Epitope antibody	(-)	(-)	(+++)
Nebulin antibody plus peptide competition	(-)	(-)	(-)
Titin antibody	(+++)	(+++)	(-)
Titin antibody plus peptide competition	(-)	(-)	(-)

With this last positive control in place, Betty Sue finally learns that her nebulin antibody does not work. She cannot even detect recombinant nebulin protein, although she knows that this recombinant protein exists because she put an epitope tag on it and the tag was detected.

POSTSCRIPT

Although it is fair to say that almost every experiment published has in place at least some attempt at a negative control, the positive control is more often ignored (or unreported). As has been shown, the absence of this control can substantially alter the conclusions drawn from experimental data. The positive control not only demonstrates that the system is operational and that the scientist measured the right parameter, but also provides a context for the experiment that may help open up new lines of inquiry.



Method and Reagent Controls

ALTHOUGH IT IS LESS TRUE TODAY than in the past, it is still the case that most scientists are trained to approach problems in a particular way, using one or two favored methodologies. Before they draw a conclusion, some geneticists working on fruit flies or yeast might not go far beyond studying the phenotype induced by mating a particular pair of genetic mutants. In the laboratory of the molecular biologist, scientists are trained to be adept at manipulating DNA. For example, the molecular biologist may induce mutations into a piece of DNA that encodes a particular protein to see how that change perturbs the protein's ability to function. To perform such a study in the context of the protein's "normal environment," the molecular biologist might then need to introduce the mutated DNA into a particular cell or organism by transformation.¹ Because the investigator then needs to study the "transformed" cell, the molecular biologist must now use the skills of the cellular biologist, who is trained to study cells both in tissue culture settings and in the organism. A kindred spirit to the molecular biologist is the pharmacologist; instead of using genetic constructs to perturb a system, the pharmacologist might use a drug, a chemical "reagent," to induce change. The term "reagent" applies to any tool that the scientist might use in an experimental setting to study an effect.² For example, it could be a small molecule, a genetic construct, an antibody, or a detection tool.

Another formerly distinct discipline that is now commonly merged with both molecular and cellular biology is biochemistry. After a mutant form of a gene has been inserted into a cell and studied to determine how the resulting mutant protein alters cell function, the scientist often wants to understand the details of how particular cellular components are perturbed. To do this, certain proteins, lipids, sugars, or nucleic acids may be isolated from the cell to determine their condition. Even though scientific training has evolved to allow for the melding of multiple disciplines, scientists trained in a particular scientific discipline often have a "comfort zone" and tend to apply a rather uniform approach to problem solving.

¹Transformation is the process of inserting a piece of DNA into a cell; it can be done by transfection (without the use of infectious agents such as bacteria or viruses) or infection (using an infectious agent to do the job).

²The dictionary definition of reagent is "any substance used in a chemical reaction," but the term is used somewhat more broadly in biology.

This could be accepted as normal: No one expects a doctor to function as a lawyer as well. And even within the medical profession, surgeons do not step out of their roles to practice psychiatry, nor do psychiatrists ever set foot into an operating theater. People are specialists for a reason; it is difficult to cross disciplines and perform with equal competency. While accepting this reality, we must also recognize that there is a real problem in applying a single methodology to a scientific question, or, to put it another way, there is a strong advantage to be gained from using more than one approach. In all experiments, when a manipulation is performed within a particular system, it is done for a discrete purpose: The scientist introduces the variable X to study the result Y . There is always a risk, however, that in addition to inducing result Y , the manipulation will also cause an unintended and unnoticed effect Z . Even with carefully constructed negative and positive controls of the traditional variety, it can be difficult to determine whether the result of a single methodology or reagent is causing an effect through some unseen additional mechanism. It is only by perturbing the subject through a second mechanism that the scientist puts into place a “not- X ” control, where X in this case is the initial methodology. This type of control is called a “methodology control.” The reason the second perturbation controls for the unforeseen effect of the first method is that the chance of two different methods inducing the same additional effect is remote. Or, one might say that the more discrete ways the scientist finds to perturb the subject and get the same readout, the greater the chances that the readout is being caused by the intended change to the object of the study, and not by some other factor that is affected unintentionally. The rationale for these claims is discussed in detail in the sections that follow.

EXAMPLE OF THE NEED FOR THE METHODOLOGY CONTROL

Let's say that a scientist named François applies a particular drug named Widget to a cell for the purpose of inhibiting a protein called Krakatoa. Unbeknownst to François, Widget inhibits three other proteins, in addition to its actions on Krakatoa. Because François does not know the additional cellular targets of this pharmacologic agent, he will probably assume that the effects he sees with Widget are due to inhibition of Krakatoa alone, particularly since Widget was designed as a Krakatoa inhibitor, without any other activity being demonstrated during the design and system-validation process. The comparison of the negative controls to the Widget-treated subjects will reveal the effects caused by perturbing all four proteins. François, however, will think that he is measuring the effect of inhibiting only Krakatoa. The “normal” system controls are designed simply to prove that Widget does indeed block Krakatoa, and thus the system controls will validate Widget.

Although there are distinct additional system controls that the scientist can include to address whether a perturbation is specific, these vary in their effectiveness, depending on the type of reagent being used. For example, as was pointed out in an earlier

chapter, it is fairly simple to assess whether an antibody is specific. As a positive control, one can use the recombinant form of the protein that the antibody is designed to detect. As a negative control, one can obtain a "protein lysate"³ that does not contain the particular protein being studied. If the antibody detects proteins in the negative control sample, then the scientist will know that the antibody is not specific for the protein of interest. Even without a control sample that is deficient for the protein of interest, it is relatively straightforward to determine whether an antibody binds multiple proteins, and one can thus easily assess specificity. However, it is more difficult to show that a reagent such as an "siRNA" (short interfering RNA) is specific. An siRNA can bind to a particular mRNA and mediate its degradation via a particular RNA-degrading enzyme that recognizes the resulting double-stranded RNA. Even though a particular siRNA is "designed" (by virtue of its sequence) to be specific to an individual mRNA, it might still bind other mRNAs via shorter homology domains. Although it may be difficult to detect such events directly, it is actually quite simple to put in place reagent or methodological controls; these will let the scientist know if a particular reagent is having an effect that is unique to it and distinct from other methods or reagents that "should" have the same effect, since they are all designed to perturb the same target. In the case of siRNA, if a second siRNA were designed and used successfully to block the same gene as the first siRNA even though its activity was distinct from that of the first siRNA, the scientist would be made aware that there was an issue with one of the siRNA molecules that was distinct from its ability to block the intended target. In this setting, because the reagent control failed to validate the result of the original siRNA, the scientist would be deterred from making a conclusion based on either of the siRNAs and would instead wait for further data.

As another example of a setting in which a methodological control would be helpful, take the case of a protein that is inhibited by overexpressing a "dominant-negative" mutant form of a gene. A "dominant-negative" mutation results in the expression of an altered form of the target protein that is capable of displacing the active form, thereby blocking its effects. However, it is possible that by making such a construct, the scientist might cause additional unintended effects. For example, if the endogenous protein, which is now displaced by the dominant-negative mutant, continues to be active but in a different cellular locale, it could induce new effects. Another way a particular mutant form of a protein could have unexpected effects is if it is able to inhibit pathways other than those intended by the scientist, via new binding or cellular localization induced by the mutation.

For all of these possible unintended effects that might come about through the use of a single reagent, it can be demonstrated that using a second reagent, or an entirely distinct methodology, can serve as a valuable control by "unmasking" the previously unseen additional effect.

³A "protein lysate" is a mixture of proteins obtained by lysing a cell or tissue, usually by using a detergent to disrupt cellular membranes.

THE REAGENT CONTROL

Let's say that the Krakatoa protein is composed in part of the following amino acid sequence:

pyecklcllr fsqsglnlrh mrvhgaasmm

An antibody to this Krakatoa protein is raised, using the last eight amino acids of the peptide sequence:

vhgaasmm

François calls this antibody his "GAASM" antibody. Using the resulting GAASM antibody, François first proves that it can detect recombinant, purified Krakatoa and that it is "clean" on a negative control lysate. Below is his "validating experiment":

Antibody	Krakatoa protein	Buffer	Protein lysate without Krakatoa
GAASM	+++ ^a	(-) ^a	(-)
Irrelevant	(-)	(-)	+++

^a+++ means "positive"; the antibody did bind protein in the reaction; (-) means "negative"; the antibody did not bind anything in the reaction.

This is not a bad validation experiment. François has proved that the antibody raised against a Krakatoa peptide does detect Krakatoa, and it does not detect anything in a Krakatoa-free cell lysate that contains thousands of other proteins. Furthermore, as an additional negative control, an "irrelevant antibody," that is, an antibody to some protein that is not Krakatoa, fails to detect Krakatoa, but it does detect protein in the cellular lysate (which includes the target of that antibody). This is the sort of validation experiment that should be performed as a first "screen" in this setting.

To continue the example, François now wants to use his GAASM antibody to see where the Krakatoa protein is allocated within the cell. Such an experiment is called an immunostaining experiment. The scientist makes very thin tissue sections to produce cross sections of individual cells and then applies an antibody to the sections. An antibody-detecting reagent is then applied to reveal the cellular location of the antibody. When François performs his experiment, he sees a few areas in the cell "lighting up" with his GAASM antibody. Given that he had already proved that the antibody recognized Krakatoa but did not detect anything in a lysate containing thousands of others proteins, he therefore concludes that the immunostaining with his GAASM antibody does in fact reveal the location of Krakatoa in the cell. Several months go by and François publishes a paper with his Krakatoa antibody, reporting the cellular localization for Krakatoa.

Eventually, a Krakatoa knockout animal becomes available. This animal has a mutation that causes the *Krakatoa* gene to be inactive; therefore, the animal has no Krakatoa protein in any cell. François gets hold of this knockout and uses it as a new and improved negative control for his GAASM Krakatoa antibody. To his great chagrin, he sees that the

antibody lights up the exact same structures as it did previously. He is thus faced with the unpleasant finding that despite his validation experiment, his antibody reacts with a protein or a molecule other than Krakatoa. How could this have happened? Given the different methods used to prepare cells for an immunostaining experiment, and given that a tissue section has even more structures than a complex protein lysate, the “validated” GAASM antibody was exposed to more structures than in the system control experiment, and it apparently cross-reacted to one of these other structures. This sort of thing happens.

Now that François is faced with this disturbing new result, he is highly motivated to make a new antibody to determine whether he has made some sort of mistake or whether a second reagent will confirm the findings obtained with the first antibody. He therefore goes back to his sequence and raises a second antibody using the following Krakatoa peptide sequence:

pyecklc

Note that this is a completely distinct peptide sequence, but still a component of the Krakatoa protein sequence. François names this second antibody YECKL. It passes the exact same validation experiment as GAASM, but when François performs the immunostaining experiment again, he uses both the YECKL and the GAASM antibodies on sequential tissue sections and compares the staining.

For now, let’s say that François did not have the tissue available from the Krakatoa knockout animal, since that kind of control is often not available. Below are the results François obtains:

Antibody	Reaction
GAASM	Reacts with the chromatin, the endoplasmic reticulum, the cell membrane, and other vesicles.
YECKL	Reacts only with the chromatin.
Irrelevant	Reacts with other structures.

As is evident from these results, two distinct antibody reagents, raised against the same protein, but to different parts of that protein, gave different but overlapping results. The YECKL antibody reacted with a subset of those structures seen with the GAASM antibody. How does François decide which of the results is “real”? The knockout tissue would be the “ultimate negative control” and would provide the answer, but again, this control is often not available. In addition, this chapter is focusing on the power of reagent controls. So in this case, François raises a third antibody to Krakatoa. This third antibody is raised against the peptide shown below, which again is part of the Krakatoa sequence but distinct from the sequences used to raise the other two antibodies:

fsqsgnlnrh

François calls this third antibody “Tiebreaker.” The Tiebreaker antibody passes the same validation control; it reacts against Krakatoa but does not bind to other pro-

teins in a cellular lysate that is Krakatoa-free. Using the Tiebreaker antibody, François again does an immunostaining experiment and obtains these results:

Antibody	Reaction
GAASM	Reacts with the chromatin, the endoplasmic reticulum, the cell membrane, and other vesicles.
YECKL	Reacts only with the chromatin.
Tiebreaker	Reacts only with the chromatin.

Now, armed with these data, François cautiously reports that he had made a mistake previously, but that he can now say that Krakatoa is expressed only in chromatin and does not localize to the other previously reported structures.

François is probably correct in his new interpretation: If a reagent has a “random effect,” the odds that two reagents have the same “random effect” are demonstrably low. In the case of an antibody, there are an almost infinite number of staining patterns that an antibody could have by virtue of binding to some structure other than the one intended; the odds that a second antibody, raised against a distinct peptide, will bind to the same “second structure” are very low. For this to occur, the second structure must have “binding epitopes” common to both of the peptide sequences found in the target protein. However—and this is very important—although the odds may be low, this sort of thing does happen, as in the case where there is a second protein in the same “family” as the one under study that contains many of the same domains as the first protein. This is yet another example where inductive information is important. The scientist might be able to determine in advance whether Krakatoa had “family members” by doing a sequence search, comparing the sequence of Krakatoa to that of every known gene in the genome of the animal being studied. In the event that the scientist found family members related to Krakatoa, it would be possible to align the sequences of all the “Krakatoa-like” proteins with those of Krakatoa itself and thereby ensure that the antibody was raised against a unique part of Krakatoa.

We thus see both the power and the limitations of the reagent control. If it is made in a distinct way, then it should serve as a reliable specificity control to the first reagent (in this case, an antibody). However, if it is made in a way that does not discriminate between the target protein and any unintended proteins that react with the first reagent, the second reagent will not serve as a useful specificity control. On the contrary, it will give the scientist false “validation.”

Just to finish the example, we can say that François did finally test his antibody on the Krakatoa knockout tissue and that the data confirmed his second interpretation that Krakatoa was localized to chromatin: In tissue obtained from the Krakatoa knockout animal, François saw no staining at all with the Tiebreaker antibody or with the YECKL antibody. The Krakatoa knockout tissue also controls nicely for the possibility that the antibodies were binding to family members of the Krakatoa protein, because those family members would still be present in the Krakatoa knockout.

François

NEED FOR METHODOLOGY CONTROLS

The example above demonstrates a setting in which a second antibody is used as a reagent control and thereby indicates the specificity of the antibody being used. However, as indicated above, there are settings in which it is difficult to determine specificity with a second reagent of the same type. In the case of the antibody, a second antibody might still bind to the same set of unintended proteins as those seen with the first antibody, especially if there are many proteins that have peptide sequences similar to those of the protein under study. Thus, using a second antibody as a "specificity control" may be problematic. The only way to mitigate this problem is, as indicated, to be aware of the particular sequence that the antibody is raised against and to determine, in advance if possible, that this sequence is not shared within a family.

As a second example in which a simple reagent control may not be sufficient to determine specificity, let's say that François is using a drug to inhibit Krakatoa and that this drug is designed to fit in a particular "pocket" formed within the surface of the Krakatoa enzyme, within which a chemical agent, ATP (for example), must bind. Using this Krakatoa-inhibiting drug, François observes a particular effect: He sees that cells treated with this drug turn blue and die. Having learned his lesson from the above example, François realizes that he must perform an additional experiment. He obtains an entirely distinct drug that has also been validated as a Krakatoa inhibitor and uses that second reagent-control drug to repeat the experiment. Once again, the cells turn blue and die.

François is just about to finish submitting his paper when he realizes he has another problem—both drugs react with the same domain of Krakatoa. Therefore, even though they may be distinct in chemical makeup, if that "pocket," which exists in Krakatoa, also exists in other proteins, then both drugs may have the same secondary effects. In this case, the reagent control would be analogous to what would have happened in the prior example if François had raised a second antibody to the same peptide as his GAASM antibody. Even though it was a new antibody, it is quite likely that the second GAASM antibody would have the same unintended binding as the first. So too in this case; even though the second drug may have a distinct chemical makeup, if it blocks Krakatoa the same way, it may similarly share secondary effects with the first drug. Therefore, if it is impossible to get a pharmacologic agent that blocks Krakatoa through a different mechanism, then François would be better off using a control other than a pharmacologic reagent control. Similarly, if a protein under study does not contain a demonstrably unique sequence, then a different type of specificity control, other than a second antibody, would be desired for detection of the protein. This second type of specificity control would be a "methodology control."

A methodology control involves verifying a result using a method other than that used in the first experiment. In the first experimental setting, a pharmacological agent (a drug) was used to block Krakatoa; in a second experimental setting, a different method would be used to block Krakatoa. Some examples are given below, but the

key point is that the methodologic control for the drug inhibition of Krakatoa would be the use of anything other than a drug that is still capable of inhibiting the target.

To understand the distinct power of the methodology control, let's continue to discuss the use of a drug to block Krakatoa: In the first experimental setting, the drug that was used has a known mechanism of action. By virtue of its chemical structure, the drug enters into Krakatoa's activity "pocket," where ATP needs to bind, and the drug thereby interferes with ATP binding. As mentioned above, other drugs made against Krakatoa may have this same activity, and therefore, both Krakatoa's first and second pharmacologic inhibitors may have the same spectrum of side effects. Now consider a potential methodological control: a dominant-negative Krakatoa mutant. This Krakatoa inhibitor has a completely different way of blocking Krakatoa. Rather than entering Krakatoa's active site with a small chemical, the dominant-negative Krakatoa protein displaces Krakatoa from its normal place of business, substituting a "dead" protein in its place. Of all the proteins that share Krakatoa's active site, the number that simultaneously shares that characteristic and can also be inhibited by a "dead" dominant-negative form of Krakatoa will be close to zero, if not zero. An example of this is the large family of proteins called "tyrosine kinases." These enzymes work by causing the addition of a phosphate molecule to the amino acid tyrosine on particular substrate proteins. Because the binding pockets in which tyrosine kinases bind to carry out this activity are so similar, many chemicals that block one tyrosine kinase will also block several others. However, it is less usual for a dominant-negative form of a particular tyrosine kinase to inhibit other family members, because the family members bind to so many distinct proteins. Although it is still possible for this to happen, the number of potential "unintended effects" decreases dramatically when using a methodological control to block a particular tyrosine kinase, as opposed to a reagent control in which a second active-site inhibitor is used.

IMPORTANT CAVEAT

Note that in each of the examples given so far, there is an important caveat: The reagent control and the methodology control reduce, but do *not* eliminate, the chance that an experimental result will be caused by an unseen side effect of the reagent. In addition, careful use of a combination of reagents, or distinct methodologies, reduces but does *not* eliminate the chance of a shared side effect complicating a result. Therefore, one can simply point out that if the scientist can reduce the chance of an unseen side effect by using a second reagent or a second methodology, then the chance of an unperceived additional perturbation will be reduced even more by using a third reagent or a third methodology. Furthermore, combining reagent controls and methodological controls provides further additional benefits. The more ways in which a problem is attacked, the more it is possible to dissect away the complications of any particular method or reagent.



Subject Controls

A CONTINUING THEME OF THIS VOLUME has been the emphasis placed on discerning the “representative case” for a particular experimental setting. As explained in previous chapters, it is necessary to demonstrate that results are obtained in a representative setting because the criterion for validating the data-derived experimental model is whether the model accurately represents what will happen in the future. If data used to construct a model are not representative, then the model will likely fail when applied to the representative case. Or, one can say that the unrepresentative data (at best) do not help in constructing a model that may be applied to the usual setting.

In an earlier chapter, an example of this issue was given: a study to determine whether a particular drug could induce weight loss in obese individuals. In this experimental setting, the individuals who comprised the “study subjects” were demonstrably more motivated, and under greater scrutiny, than most other individuals who need to lose weight. Therefore, it was suggested that a result gained in the study group—where people were exercising and watching their diets—would not easily translate to the more usual instance. It should be clear that even if the obesity study on the highly motivated individuals was done well, that is, it was performed in a rigorous fashion—the groups were appropriately randomized for independent variables, it was double-blinded, and so on—it still might not result in a model that accurately predicts how the drug will perform at inducing weight loss on the more “typical” subject. Let’s say for the purpose of the present discussion that the anti-obesity drug was tested again on a more “representative” group. In the second study, individuals were not asked to increase their amount of exercise or to change their diets; the only change was that they were either on the experimental weight-loss agent or on a placebo. The results from the second study differ from those of the first study in many respects. Most tellingly, individuals did not lose weight on the experimental drug as they did in the first study.

**SUBJECTS MAY BE CHOSEN TO REPRESENT A PARTICULAR CASE,
EVEN IF IT IS NOT THE "USUAL" CASE**

The obesity drug example demonstrates the need for the scientist to decide what condition the experimental subjects should represent. If the goal is for the subjects to be representative of a typical patient who might take the drug under study, then that type of patient should be used in the study. However, the scientist may have a different goal. Therefore, when the subjects under study need to be "representative," the scientist must ask, "representative of what?"

For example, for the question "What color is the sky?," it was decided that all times of day were relevant to the question, and thus the sky had to be surveyed at multiple time points, over multiple days, to determine what was "representative." In still another example discussed previously, it was asked whether caffeine could increase blood pressure. Here, it was emphasized that the subjects should not be on antihypertensive medication, because most people are not on such medication, and there was a desire to determine the "representative case." In the obesity example, the scientist must decide what questions need to be answered. Is the question whether the drug is capable of working in a particular type of patient—for example, one who can be motivated to exercise and eat healthy foods—or whether the drug will work on a more typical, less motivated patient? These are very different questions. The first obesity study, in which participants were highly motivated to exercise and watch their diets, demonstrated that the anti-obesity drug could assist people undergoing these additional steps to lose still more weight. This is not an irrelevant point, and it would be of great interest to individuals who fall into a category similar to those in the study group. Even though the second study demonstrated that the anti-obesity drug cannot help an inactive individual to lose weight, that result does not invalidate the first study; rather, it helps to establish the conditions under which the drug will be successful. It should be pointed out that if only the second study were done, the potentially helpful attributes of the pharmacologic agent in the setting where subjects were induced to exercise would remain undiscovered. Therefore, we see that the insistence that subjects always represent the "usual" case may result in the scientist losing valuable information.

What then of the need for an experimental model that predicts the future? Remember that the model is required to represent what will happen. However, the model can be restricted as to the terms of that proposition; in other words, there is no reason one cannot have a model that says "Drug *X* will induce additional weight loss in people who exercise, in comparison to those who exercise but are on a placebo. However, drug *X* will fail to cause weight loss in people who do not exercise." Such a model meets the burden of "representing reality."

In medicine today, there is an increasing push to develop "individualized drugs," that is, drugs that will work on patients with very particular characteristics, even if those characteristics are not the norm. To produce such drugs, one would need to study whether alterations in a particular "marker" that can be used to categorize a sub-

group could be used to determine the setting in which the drug will work. In the obesity example, the marker was the amount of exercise that the subject performed. The combination of the two studies helped to show that the anti-obesity drug's success was correlated with this exercise marker.

FINDING A RESPONSIVE SUBJECT

The initial experimental design for the anti-obesity drug can be recast as a demonstration of where the drug was capable of providing a benefit. In this case, the scientist could have framed the question in the following way: "Under what conditions, if any, will drug *X* induce weight loss?"

Note that this question is purposely biasing. For the question "What color is the sky?" or "Does caffeine increase blood pressure?," the scientist wanted to know the answer that reflects the usual case. However, in the current setting, where a scientist wants to know whether a new drug could provide a benefit, it might not be a concern if this pharmacologic agent fails to help everyone, as long as it can help someone.

There is no scientific problem with such a "bias" as long as the bias is incorporated in the model, so that the purpose of the study is recognized. As another example of this issue, say that a scientist has developed a new anticancer drug. In a large study of patients who have stomach cancer, it is found that the drug provides no statistically significant benefit. Therefore, the drug is initially deemed a failure. However, in a retrospective analysis, the scientist studying the drug notices that individuals who had a mutation in a certain gene, "gene *X*," did seem to have an increase in life span when they were using the drug. As explained in an earlier chapter, such retrospective analyses cannot be used to draw conclusions because it is often possible retrospectively to pick out some correlation that appears to indicate a desired conclusion. However, the finding could be used to develop a new question. Therefore, let's say that the scientist did a new study, asking "Does the anticancer drug increase life span in patients with a mutation in gene *X*?"

The new study is done in the following way: Stomach cancer patients are pre-screened for a mutation in gene *X*. Those who have such a mutation are randomized between a group that receives the potential anticancer drug and a group that receives a placebo. As additional controls, stomach cancer patients who do not have a mutation in gene *X* are randomized between two other groups; again, one group receives the drug and the other receives the placebo. All of the study subjects have the same "stage" of stomach cancer, meaning that they are selected to be equivalent in the severity of the disease. The results of the study are presented as average life span while taking the drug:

	Average life expectancy for stomach cancer patients	
	Gene <i>X</i> mutation	No mutation in gene <i>X</i>
Drug	24 months	6 months
Placebo	6 months	6 months

The data suggest that the drug is capable of increasing the life expectancy of a stomach cancer patient with a mutation in gene *X* severalfold. However, the drug provides no benefit to those patients who do not have this particular genetic marker. The reason that the first study failed to show a benefit for the drug is that mutations in gene *X* are rare: If the study is not designed to focus on patients who have the mutation, then the majority, who do not benefit, will dilute away the effect.

We see from this example that the choice of the subject population under study can dramatically affect the experimental output. In addition, choosing particular, relevant populations to study may provide significant additional information that simple randomization would not provide, even if the population is not the representative case. As more information is learned about a subject (e.g., that people who have stomach cancer can be subdivided according to particular molecular markers), the more the scientist is able to focus on subpopulations for study.

An example of a clinical setting in which a molecular marker determines whether a drug regimen may be successful is breast cancer. A patient who has a tumor that expresses the estrogen receptor is demonstrably responsive to antihormonal therapy. However, patients who have tumors that are negative for the estrogen receptor do not benefit from this drug regimen. Furthermore, a second hormone receptor is also relevant: the progesterone receptor. It has been additionally demonstrated that patients who are positive for both the estrogen and the progesterone receptors do even better on antihormonal treatment than those who are only positive for the estrogen receptor only. Thus, using these markers, the broad set of “breast cancer patients” can be subdivided into new subject categories:

Category	Estrogen receptor	Progesterone receptor
I	(-)	(-)
II	(+)	(-)
III	(+)	(+)

Whereas in the past, breast cancer studies may have focused on the entire set of patients who had this condition, now with the help of the hormonal markers, smaller categories of breast cancer patients can be studied to determine whether they benefit from particular treatments, even if the “larger set” of patients appears to be unaffected. Eventually, one can imagine that these groups may be further subdivided, as additional markers are found to help to predict a therapeutic outcome. For example, if it were demonstrated that a gene *Y* is an additional positive marker, but only in the presence of the progesterone receptor, the following categories would arise:

Category	Estrogen receptor	Progesterone receptor	Gene <i>Y</i>
I	(-)	(-)	Not relevant
II	(+)	(-)	Not relevant
III	(+)	(+)	(-)
IV	(+)	(+)	(+)
V	(-)	(+)	(+)

Note that the “estrogen receptor (–), progesterone receptor (+)” category becomes delineated only after there are data demonstrating that this category could benefit from a particular type of therapy; previous to the finding, it was first necessary to demonstrate that the patient was estrogen-receptor positive (+).

This example demonstrates how increased information about a particular condition may alter the choice in experimental subjects. Focusing on a subject with a particular characteristic can allow the scientist to obtain relevant data that would not have been achieved using the same approach on a different subject subtype or on a more general subject type.

CONTROLLING FOR A PARTICULAR SUBJECT TYPE AND STUDY RANDOMIZATION

In the above example, where an anticancer drug was tested on patients who had a mutation in gene *X*, the placebo-control group also contained stomach cancer patients who had a mutation in gene *X* and whose cancer was at the same clinical stage. Note that the control group is “matched” for relevant characteristics to the study group; both groups have the same stage of stomach cancer, and both have mutations in gene *X*. These characteristics (having a particular stage of stomach cancer and having a mutation in gene *X*) can be thought of as “screening criteria” for being placed in either of the first two groups of patients (one that will receive the anticancer drug, and one that will receive the placebo).

As long as the criteria are determined in advance, and as long as they are represented equally in the study and control groups, the scientist can prescreen individuals for any particular characteristic that is deemed relevant for the study.

RANDOMIZING STUDY SUBJECTS

After the screening process, one might ask how the scientist decides who is to receive the experimental agent and who will be treated with the negative control substance. The answer is that the scientist should not be the one to choose: Study subjects should be “randomized” between these groups.

Randomization refers to a process in which the scientist is not able to pick and choose who goes into a particular study group. For example, if a group of 20 subjects was screened and found to meet the criteria for the study, then each participant would be assigned a random number by a computer. Subjects with certain random numbers would then be chosen to receive either the drug or the placebo, without the person who is making such decisions knowing who has what number. Furthermore, the drug vials should be coded so that the person who distributes the drugs will not know who gets which drug.

This process would mean that both the scientist designing the study and the subject taking part in the experiment are “blinded” to each subject’s group assignment.

This is what is meant by a “double-blind” study. This type of design, where patients can be screened for a particular characteristic and then randomized, can be formalized as follows:

1. Study subjects are selected for particular characteristics.
2. Once selected, the study subjects are randomized between test and control groups.
3. The person administering the drug and the study subject are not aware of which treatment is being administered.

Randomizing the study subjects after they have been selected for particular criteria helps to control against the investigator placing particular subjects in one group versus another. Furthermore, randomization helps to control against the study subjects who are receiving a particular treatment responding in a different fashion from those in a different group.

One might ask why the issue of randomization is included in a chapter on subject controls. This process of blinding both the scientist and the subject is both a subject control and an experimentalist control. As mentioned above, at the point the experimental criteria are established and the experiment is ready to be performed, the less information that the participants and the researcher have as to their group placement, the less chance that one group will either behave differently or be treated differently. Thus, the control subjects will have an improved chance of being “true controls” for the test subjects because they are equivalent for “all but X ,” where X is the agent of perturbation (in the present example, a drug).

MATCHING STUDY SUBJECTS

In some studies, it may be desirable to match study subjects between control groups and treatment groups. This is particularly appropriate in experimental settings where the chance of variation in the experimental output is high.

A typical setting where study subjects may be matched is in animal experiments. Obviously, when the experimental subject is a mouse, there is no chance that the animal will experience a placebo effect. In addition, one can match subjects between groups without the individual who treats the mice knowing which group is being given the experimental agent. As an example, imagine an animal study in which strength is being studied. The scientist wants to answer the following question: “Do anabolic steroids increase strength?” To answer this question, the scientist will take baseline strength measurements and distribute the animals into groups. One group (the test group) will be given anabolic steroids dissolved in a salt solution, whereas the control group will be given just the salt solution.

When the baseline measurements are taken, the scientist discovers that there is a wide variation in strength among the animals, even within a particular gender. Therefore, to ensure that both groups have the same range and have representatives

of the various strengths being measured, the scientist distributes similar animals between groups. Say, for example, that 20 animals have the following strength read-outs, given in arbitrary units: 9, 11, 14, 15, 17, 19, 20, 20.5, 24, 24, 24, 27, 29, 31, 32, 32, 32, 40, 40, and 39. Clearly, there is a great difference in strength, ranging from 9 to 40. Were these animals distributed randomly, it is possible that one group would be comprised of animals with a significantly larger average starting strength than the other group, just by chance. Therefore, the scientist matches the animals as follows:

Steroid group	9,	14,	17,	19,	24,	24,	31,	32,	40,	40
Salt buffer group	11,	15,	20,	20.5,	24,	29,	27,	32,	32,	39

This type of study is more problematic because of the lack of randomization, but it should be clear that the scientist who is administering the experiment can still be blinded as to what each subject is receiving. In addition, closely matched animals could further help to “blind” the scientist as to group-specific differences (compare the case where the animals are matched in size and strength with that where the animals are randomized, but the scientist notices that the animals in the salt group are generally larger).

In this experimental design, matching animals that are highly variable decreases the group sizes that would otherwise be required to achieve statistical significance. This design is acceptable if steps can be taken to prevent scientist bias.

VARIABLES

Genetic “model systems” were developed in an effort to control for the many potential independent variables that the scientist unknowingly encounters when performing genetic studies on heterogeneous populations. These model systems can be traced back to scientists such as Gregor Mendel, who studied strains of peas to analyze particular traits such as shape and color. Mendel thereby demonstrated the concept of a “gene,” meaning a heritable substance that can imbue a subject with a particular trait.

Building on that work, scientists realized many decades ago that it would be quite useful to have strains¹ of mice available that were genetically identical to each other. This would allow a scientist to study the effect of altering a particular gene, while all of the others remained identical and were thus controlled for. These geneticists therefore began a decades-long process of producing clonal mouse strains. Even when

¹According to Jackson laboratories, strains can be termed inbred if they have been mated brother x sister for 20 or more consecutive generations, and individuals of the strain can be traced to a single ancestral pair at the 20th or subsequent generation. At this point, the individuals’ genomes will, on average, have only 0.01 residual heterozygosity (excluding any genetic drift) and can be regarded for most purposes as genetically identical. Inbred strains must be continuously mated brother x sister (or equivalent) thereafter (see http://www.informatics.jax.org/mgihome/nomen/strains.shtml#inbred_strains).

DNA methods were very young, it was already realized that a new trait that cropped up in an individual animal from a clonal mouse strain was necessarily due to a genetic mutation and that this mutation could be identified by following inheritors of the new trait.

In addition to mice, it was realized that other genetic systems might have relevance to the human condition, and therefore fruit flies, worms, zebra fish, and even the lowly single-celled brewer's yeast came to be used as genetic models. Of course, it is not always necessary that a finding in a particular animal be relevant to the way biology works in the human, but from a practical point of view, this is often why such studies are performed.

Even more sophisticated genetic animal subjects can now be generated in the laboratory. For example, to understand the role of a particular gene, mice are often developed that are genetically null for that gene. The genetically altered animals can be called a particular "line," that is, a subset of mice, usually descendants from one or two sets of "founder parents," that is derived from an inbred strain. As explained in an earlier chapter, these genetically null animals are called "knockouts" because the gene of interest has been removed or "knocked out" of the genome. Using such mice, one can control to an even greater degree for the other genetic variations that may occur in more typical animal populations, because now one is starting with a clonal strain and introducing the mutation at a predetermined site, as opposed to screening for mutations that may crop up accidentally and could therefore be caused by multiple alterations.

To study the role of an insulin receptor (IR), for example, that particular gene might be deleted in a particular fertilized oocyte, and from that egg, a mouse could be generated that has one copy (or "allele"²) of the IR removed.³ This "heterozygotic" (IR^{+/-})⁴ mouse is then crossed to a littermate⁵ to produce an animal that has both of its IR gene copies deleted. These are now "insulin receptor knockouts" (IR^{-/-}). It is worth emphasizing again that if the deletion of the IR was performed in a clonal animal strain, then all of the genetic material other than the IR is identical, when comparing strain members that are IR^{-/-} with those that are IR^{+/-}. Therefore, when a scientist studies the IR^{-/-} animals, the negative controls can be any of the "wild-type" or "normal" IR^{+/-} animals in the same strain. However, to produce an even "stronger" negative control, scientists often use the heterozygous IR^{+/-} mice to generate the

²An allele is a gene, which can be thought of as a piece of DNA sufficient to encode a particular trait.

³In almost every eukaryotic cell, there are two copies of each chromosome and therefore two copies of each gene. Exceptions include meiotic germ cells, which have only one copy of every chromosome. In addition, in every male somatic (body) cell, there is a single copy of each of the two sex chromosomes—the X and Y chromosomes (females have two Xs and no Ys; however, one of their X chromosomes is rendered inactive, meaning that, even in females, only one X chromosome is "on").

⁴This nomenclature refers to the two copies of each gene. Thus, IR^{-/-} means that the animal is null or (-) for each of the two copies. In contrast, the normal animal is positive for both copies of the IR and is thus annotated IR^{+/-}.

⁵A littermate is a sibling.

knockouts, because mating these animals will result in offspring that are either “knockouts” ($IR^{-/-}$), “normal” ($IR^{+/+}$), or “heterozygotes” ($IR^{+/-}$). Thus, from one mating, the scientist can generate “normal” negative control siblings to the knockout animals and thus obtain age- and condition-matched controls.

GENERALIZING FINDINGS MADE USING GENETIC CLONES

Let's say that a study is done on such $IR^{-/-}$ mice to understand the role of the IR in maintaining glucose levels. A group of scientists study the mice for years, under a variety of conditions, and thereby generate a large data-based model as to the role of the IR in disease settings such as diabetes.

After the studies are completed, one may inquire as to the relevance of the animal model studies to humans or even to other strains of mice. Although most if not all of the findings may eventually be found to hold in genetically diverse human populations, it is formally possible that some findings will be found to be specific to the particular mouse strain that was studied.

From this example, we see that in addition to the “advantage” provided by having a genetically identical mouse strain to conduct studies, we may also have the “disadvantage” of missing effects induced by particular genetic variations that are capable of altering results for the gene under study. In other words, those variations that are significant at the population level are not represented in our genetically identical mouse strains. We therefore learn from the genetically identical subject—in this case, the clonal mouse strain—that the results obtained from this animal may teach the scientist what a particular gene can do in a particular setting, but it still needs to be demonstrated that the results hold in the general case. This point should not, however, be used to minimize the great utility of using clonal animals to study particular traits. Without such resources, it would be extremely difficult to study the effect of perturbing a particular genetic locus.

Perhaps the most straightforward way to ask whether there is something idiosyncratic about the results obtained from a particular clonal animal strain is to breed a particular knockout animal to another strain of mice and then to backcross the resultant heterozygotic animals to the second strain until the knockout can be said to be within the “genetic background”⁶ of the second strain of animals. Such a “genetic cross” is often done to generalize a finding; one asks whether the same phenotype found in one mouse strain is also found when the same mutation is crossed into a second genetic background.

⁶Genetic background refers to the genetic makeup of an animal. Different clonal strains of animals have regions of DNA that differ. In different backgrounds, some genes behave differently. For example, the action of a gene may be compensated for by a second gene in one strain (and thus one background); however, in a second strain, this second gene may not be expressed to levels high enough to compensate for the action of the gene.

NONHUMAN ANIMALS AS MODELS FOR THE HUMAN

Of course, there is an even more important issue about results obtained using mice, especially if mice are being studied as a surrogate for humans: One must always validate whether a finding on a surrogate for the human actually holds for the human case. The reader might wonder, given this requirement, why all studies are not done on humans. This is because the scientist cannot ethically or practically do the sort of studies in humans that can be done in mice⁷ or, especially, organisms such as fruit flies and yeast. Furthermore, these other eukaryotes usually do provide an accurate surrogate, meaning that the function of a gene in a mouse very often turns out to hold in the human case.⁸ The point being made here is that although the animal model is valuable in understanding a particular gene, the applicability of a particular finding made in an animal subject must still be demonstrated for the human subject.

HUMANS AS "MODELS" FOR THE HUMAN

Just as one might object to applying a mouse finding to the human without experimental validation, an objection can also be made to generalizing a finding made on a particular subset of human subjects. In this case too, one must be careful to validate such a claim. This is an additional reason why experimental repetition is necessary: When experiments are repeated, one can say that the second set of experimental subjects controls for the first, and the repetition answers whether the findings in the first instance are cross-applicable to the second. However, there are additional specific instances when caution should be used when generalizing a result; this caution emphasizes the need to vary subject choices in particular settings. For example, one should be cautious of applying data gathered by studying male subjects to females. It is not difficult to find experimental settings where data gathered from men will not hold for women. As an additional example, it has been shown that the elderly are often more sensitive to pharmacologic agents than the young. Therefore, it is often necessary for physicians to recalibrate suggested drug dosages when treating elderly

⁷The ethics of animal studies is beyond the focus of this book, but it should be mentioned that the scientist is ethically constrained against performing certain studies even in a model such as a rodent. The key questions the scientist must ask when performing an animal study include the following: "Is it necessary to perform the study on an animal?," "Did I take steps to minimize pain?," and "Will the study likely produce useful knowledge?" Of course, there are those who think that nothing can justify experimentation on a mouse. I respectfully disagree, but the arguments that justify this disagreement are, again, beyond the current focus. The salient point is that ethics further constrains the scientist's choices of experimental subject.

⁸There are exceptions, such as particular gene families that diverge rather sharply between mouse and human. In addition, it might be pointed out by simply looking at the mouse and the human side by side that there must be variations in body-patterning genes sufficient to result in the rather obvious differences between the two species. In a single-celled organism such as yeast, the differences between that system and the human are obviously much more numerous; but one might be more surprised by the many cellular systems that are in common between these notably divergent species.

patients. Even differences such as ethnic-group variations can affect an experimental outcome. Only recently have studies been geared to those instances where a particular ethnic group might have a greater disease prevalence or susceptibility.

Obviously, if it can be demonstrated that traits such as gender, age, and ethnic group can affect study outcomes, then these traits constitute “independent variables” that must be taken into account when choosing and randomizing subjects.

PERFORMING GENETIC SCREENS ON HUMAN POPULATIONS

Building on the great progress that has been made using genetically identical strains of mice and other animals (such as fruit flies, worms, and even yeast), the scientist might wonder if there are human populations that are sufficiently uniform to be useful “subjects” in genetic surveys. Again, the reason that one wants rather homogeneous populations is so that most of the genetic material (other than the gene being studied) is identical. In these human populations, a particular genetic trait—such as susceptibility to diabetes—may be due to a mutation in a particular gene, which can be found by comparing the DNA from the afflicted individuals with that of their healthy compatriots, whose DNA is otherwise quite similar.

This type of “genomic” study is done with increasing frequency. The goal is to find “SNPs,” or “single nuclear polymorphisms,” which are changes in a single component of the DNA that could account for a phenotypic alteration. As one might suspect, it is possible to find many such changes when comparing even close populations, but it is rather difficult to prove that such a change is “causal” for a phenotype, meaning that the genetic change is the reason for the phenotype or characteristic. This, again, is why it is useful to study genetically similar populations; the closer one individual is to the next, the better the scientist can follow a particular SNP to determine whether it truly correlates with a characteristic. Once the correlation is established, the scientist can go further and determine whether the alteration is in fact causal. For example, a similar mutation can be inserted into a mouse genome, and a study can be made to see what happens, by comparing mice with the mutated gene to their littermates that do not have the alteration but are otherwise genetically identical.

FINDING GENETICALLY INDEPENDENT VARIABLES

The scientist may also tackle the issue of independent variables among subjects by asking the question directly: “Are there any other gene alterations that can similarly alter the phenotype caused by mutating gene *X*?” One way to find these other genes would be to do what is called a “genetic complementation” screen. Gene *A* “complements” gene *B* if the expression of gene *A* can substitute for the usual effect of gene *B* or ameliorate the phenotype caused by inhibiting gene *B*. For example, in a particular complementation screen for the IR, one might mutagenize the IR-null animals to

g mice,
ys vali-
an case.
one on
of stud-
lies and
rogate,
in the
valuable
e in an

experi-
le on a
validate
ecessary:
al sub-
the first
specific
aution
ple, one
nales. It
ill not
erly are
is often
elderly

the sci-
ent. The
is it nec-
the study
experi-
ent are,
choices

use and
by side
ferences
at system
y cellu-

determine whether an alteration in a second gene might help to ameliorate the phenotype caused by deleting the IR. In a second type of genetic screen, the same sort of mutagenesis might be done to determine whether the loss of a second gene could worsen the phenotype induced by mutating the IR. All of the genes discovered by such a screen may contribute to disease settings such as diabetes, in which the IR is compromised. The expression of these "IR-relevant" genes could then be screened for, or studied in additional experimental subjects.

THE POSSIBILITY OF CONTROLLING AWAY A SUBJECT EFFECT, AND WHETHER THIS RISK IS WORTHWHILE

Sometimes, the effort to produce an "ideal" experimental setting, in which there is no variation other than that intended, can mask an important effect. Let's say that a scientist is studying the effect of an antidepressant drug in humans. To make sure that the subjects are assessed in a controlled environment, all of the individuals in the study are confined to a hospital for the duration of the treatment period. In this experimental design, the scientist finds no difference in depressed individuals receiving the drug when compared to depressed patients who received a placebo. However, had the study been done on subjects allowed to go about their normal routines, it would have been discovered that those who spend at least 1 hour per day in direct sunlight experience a "complete cure" when using the drug. The synergistic effect between sunlight-induced vitamin D production and the drug was missed by the scientist who was trying to be scrupulous in controlling for independent variables.

This example helps to demonstrate another instance where the "representative case" is important. If the goal of a drug is to help people going about their normal lives, then it may, in some instances, be worth the necessary increase in study numbers, and the additional effort in analyzing data, to allow the subjects to go about their regular routines. If the study population is large enough, the inevitable variations that will crop up within individuals may still be controlled for, simply because, as group size increases, it becomes more likely that by random chance each group will include representations of all the relevant variables, and it will be possible to detect the "unforeseen" but still relevant effect.

Although it is easy enough to make note of the caveat, one can still argue that science is done most fruitfully in settings where one can limit unseen variability. Therefore, it may be argued that it is worth missing the occasional fortuitous finding to better guarantee that a study eliminates unfortunate variations that can cause an incorrect conclusion.

CELLS AS SURROGATES FOR ANIMALS: CREATING THE NEED FOR REDUCTIONISM CONTROLS

Many of the experimental examples used so far had humans or nonhuman animals for subjects. Yet in the laboratory setting, the subjects are often much smaller. Rather

than study a question in a complex multicellular organism, single cells can be used as subject populations. It is not difficult to find the reason for this choice: Genetically identical cell lines can be generated easily and grown under highly regulated conditions. Furthermore, although it is extremely difficult to study even a few thousand people or animals, in a tissue culture experiment, millions of cells can be examined. If different perturbations are desired, populations of cells can easily be subjected to the desired variables, with control populations left untreated.⁹ For all of these reasons, it is much easier to generate highly reproducible data on a cell line than on a population of genetically diverse humans.

As is often the case, however, the advantage to using cells as test subjects is precisely the disadvantage. Cells are comparatively simple, and the fact that they are left unperturbed by all of the stimuli found in the body is the reason that a finding on a cell in culture may not hold once the cell is returned to its "natural" context within an organism. Therefore, the scientist must be careful about extrapolating conclusions obtained from cell lines to more complicated organisms. Instead of making such an extrapolation, the finding on a cell can be the justification for asking whether the same finding holds in an animal.

What kind of "reductionism controls" might there be? When doing an experiment, remember that the desire is usually to study the representative case. Therefore, the scientist might ask whether the data obtained in one cell line may hold for a second, similar line. If there are ten different fibroblast lines, for example, these could all easily be used to at least determine whether the observed effect is seen generally in cells of a similar lineage. Next, the scientist can determine whether the effect holds for cells of multiple lineages (if such a question is relevant for the particular finding) and also whether the effect still holds under multiple conditions.

As an example of the need to be aware of the study conditions, an interesting line of study has concerned genes that are sensitive to oxygen levels. For decades, cell lines have been kept at oxygen concentrations that exist in normal air. However, someone finally realized that oxygen levels in most cellular contexts in the body are much lower. Therefore, cells were put in special chambers to limit oxygen concentrations, and it was discovered that an oxygen-sensing system became activated. This system is essentially always "off" in the usual cell culture system. Thus, if the results could be perturbed by this pathway, then data produced in tissue culture would not necessarily translate to the way things work in the animal. This cautionary tale demonstrates that if the cell is being examined out of a desire to model what happens in the complex organism, the "modeling feature" must be validated by proving that things still work the same when in the animal.

⁹When the animal is studied, millions of cells are being surveyed in this case as well. The challenge is that it is difficult to investigate each cell individually when it is housed in an animal. However, in tissue culture, populations can indeed be studied cell by cell if desired, for example, by using a device called a fluorescent-activated cell sorter (FACS) or through other manipulations.

**IN VITRO MOLECULAR SYSTEMS AS SURROGATES FOR CELLS:
CREATING A FURTHER NEED FOR REDUCTIONISM CONTROLS**

A tremendous amount of information can often be learned about the way a particular molecule works by studying it in isolation. Sometimes, this isolation is close to "perfect," as when one takes a protein and purifies away other contaminating proteins as much as possible, to produce crystals. The crystallized protein can then be analyzed by diffraction studies to determine the molecular structure at a resolution that is measured in angstroms. As a second example, an enzyme system can be reconstructed in a test tube so that certain measurements can be made. The scientist may be interested in the rate of enzyme activity, how the substrate concentration affects the enzyme's activity, or whether the enzyme is processive.

As in the example of the isolated cell, the benefit of studying molecules in isolation should not cause the scientist to forget that these findings must be validated in the cellular context if the point of the studies is to model how the molecule behaves in its "regular" context. It is even more difficult to establish the representative case in the test tube. Often, the goal for these types of experiments, which are called "in vitro" or "test tube" experiments, is to put the molecule through its paces, in a variety of settings, to determine its capabilities. Once the scientist knows the range of things the molecule can do, then the scientist is better equipped to ask more relevant questions about the molecule's function in its native context, under various physiologic conditions.

Recall Chapter 10, where the example was given of a restriction enzyme studied in isolation. Eventually, the scientist wanted to return to the bacteria in which the enzyme normally functions and ask whether the data that had been gathered for the molecule in isolation held when it was residing in its usual locale. However, as noted earlier, sometimes the goal is not to model the native setting. Restriction enzymes are great tools for cutting DNA, and they have been made to function well at this job in test tube settings. The scientist who uses DNA for this purpose need not care that the same restriction enzyme may have a different activity in a particular bacterium. Thus, reductionism controls (and objections to particular experiments as modeling systems) need to be appropriate for the particular question being asked.



Assumption Controls

TAKE THE CASE OF A TRIAL TO DETERMINE whether aspirin helps to alleviate the pain caused by headaches. In this study, a certain dosage of aspirin is used, and it is found that those receiving the aspirin report rapid relief of their pain in comparison to those receiving a placebo. This study is then validated through repetition on a second group of individuals. Next, other groups use the aspirin: those that do not fit the entry criteria of the first two studies. Again, it is found that aspirin helps to alleviate pain even in these other groups. Eventually, someone, somewhere comes along who has used aspirin in the past and has experienced a beneficial effect, but who now has a real whopper of a headache. Let's go a little further and say that this individual was even clever enough to have experimented a little in the past, and used gradually more aspirin in certain settings of severe headaches, without ill effect. This person now reasons that "if some aspirin is good, more is better" and then takes ten times the recommended dosage for an extremely bad headache. The dose causes a gastric bleed, resulting in the individual's death.

The person in this case died of an incorrect assumption—that efficacy or benefit is linear. In pharmacology, it is extremely common to find that many beneficial drugs begin to develop unwelcome, and even potentially fatal, effects when used inappropriately. This simple fact helps to illustrate the problem of cross-applying an outcome in an inappropriate way, by making an assumption that a model derived from one system, under particular conditions, will hold true in a different system. An "assumption control" is the effort to turn such unwarranted assumptions into questions, which can then be subjected to experimentation.

CONTROLS TO ELIMINATE ASSUMPTIONS CONTAINED WITHIN THE EXPERIMENTAL QUESTION

Recall the question asked in earlier chapters: "What is the effect of caffeine on blood pressure?" Certain experiments designed to answer this question used caffeinated coffee as the means of delivering caffeine to the experimental subjects. Inherent in this approach is the assumption that the other ingredients in coffee will not perturb the effects of caffeine alone. However, this idea need not be left as an assumption—the scientist included controls to demonstrate whether this is the case. One type of

control that gets at this issue is the “all but X ” negative control introduced earlier. For the negative control, the scientist would use decaffeinated coffee to show that if blood pressure increased in the subjects drinking caffeinated coffee, but did not go up in those drinking decaffeinated coffee, it could be reasonably said that the caffeine was the causative agent. However, the “all but X ” negative control is not the only type of assumption control the scientist may do, and one can go even further and say it is not the only type of control that the scientist should do. In the present example, the scientist would also want to directly test caffeine alone to determine whether it was sufficient to increase blood pressure.

The reader might ask why caffeine was not tested directly in the first place. However, there are many situations where the scientist cannot do this type of direct examination; the pure reagent may be extremely expensive, for example, or difficult to obtain. In the case of caffeine, it is much easier to survey large numbers of people if they can be studied while they are going about their usual routine of drinking either caffeinated or decaffeinated coffee. Therefore, the experimentalist might reasonably posit that the advantage of studying potentially tens of thousands of people far outweighs the disadvantages of using a surrogate reagent, especially since the “all but X ” negative control is used to control for the other ingredients in the coffee. Although this reasoning has merit, the assumption control of using pure caffeine in at least a small group of people serves as a valuable additional piece of evidence. For example, imagine how the scientist would react if it is discovered that the 100 individuals given pure caffeine failed to respond in the same way as those who drank caffeinated coffee. If such a result were obtained, the scientist would be forced to abandon the conclusion that caffeine increases blood pressure. On the other hand, if the “pure caffeine” group responded similarly to the caffeinated coffee group, then the scientist would have an important piece of validating datum, demonstrating that “ X and only X ” is sufficient to cause the effect.

Lest the reader think that the pure caffeine group is just a positive control, it needs to be emphasized that this is not the case. Remember that a positive control is an agent distinct from that being tested, which is used to demonstrate that the measuring system is functional. In the present case, the agent being measured is caffeine, even in the instance that caffeinated coffee is administered. The scientist does not know in advance of the study whether or not caffeine causes an increase in blood pressure—that is why the question is asked. So that is the reason caffeine by itself cannot be considered a positive control. It is something different—it is an assumption control.

ASSUMPTION CONTROLS TO AVOID INAPPROPRIATE DEDUCTIONS

Imagine that the scientist were studying a cell signaling pathway. For this example, it has been previously discovered that the following proteins signal to each other in a

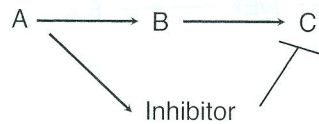


Figure 1. Schematic model of a simple signaling network. This network, in which A might either inhibit or activate C, demonstrates the need for assumption controls.

linear fashion: PI3K, Akt, mTOR, p70S6K.¹ By this, we mean that PI3K activity results in Akt being active; Akt activation then leads to mTOR being induced, and mTOR activity is required in order for p70S6K to be turned on. This pathway can be symbolized as PI3K/Akt/mTOR/p70S6K.

Next, let's say that a scientist discovered that activation of mTOR results in stimulation of another protein, Protein X. Given that both PI3K and Akt can cause mTOR to be active, the scientist might put Protein X in this pathway:

PI3K/Akt/mTOR/ProteinX

Although this may seem to be logical, the scientist must actually demonstrate that the assumption is accurate, by asking whether activation of PI3K or Akt leads to activation of Protein X. Why is it necessary to do these assumption controls? If this were a mathematical equation, or an exercise in deductive logic, one could say that if A leads to B, and if B leads to C, then A leads to C. This example of deduction is completely valid mathematically. However, biology is not math, and there are many examples in biology where A leads to B, and B leads to C, but A does not lead to C. Just to give an example, it could be that A activates an inhibitor of C in addition to activating B (Fig. 1). Therefore, in settings where A is the point of initiation in the pathway, B gets activated, but C does not. However, if B is activated by a different route, then B is free to activate C (Fig. 1). It is because of these types of complexities that assumption controls are required in many aspects of biology. Signaling “pathways” tend to be more than linear systems—they are more accurately thought of as signaling networks, because there are often cross- and counter-regulatory mechanisms at work. As an example of this point, examine the signaling diagram in Figure 2. This is apparently the way cell signaling works in skeletal muscle fibers; in skeletal muscle, as opposed to some other tissues, there is an inhibitory step in which Akt inhibits Raf. This stops Ras from resulting in Erk activation, despite the Ras/Raf/Mek/Erk pathway being operational, because in addition to activating the Ras/Raf/Mek/Erk, in skeletal muscle, Ras also activates the PI3K/Akt pathway, which can turn Erk off by inhibiting Raf (Fig. 2). This sort of cross-regulation helps to show why an “assumption control” is so important in cell signaling experiments; if the scientist knew only that Ras activated, it would be natural to assume that Ras activation would infer that Erk is also activated. However, performing assumption controls allows one to learn that in skeletal muscle Akt inhibits Raf.

¹It could have easily been said that “A leads to B, which leads to C, which is necessary for D.” In this example, the details of the proteins involved do not matter. An actual signaling pathway is used just to make the example clearer for those familiar with these proteins.

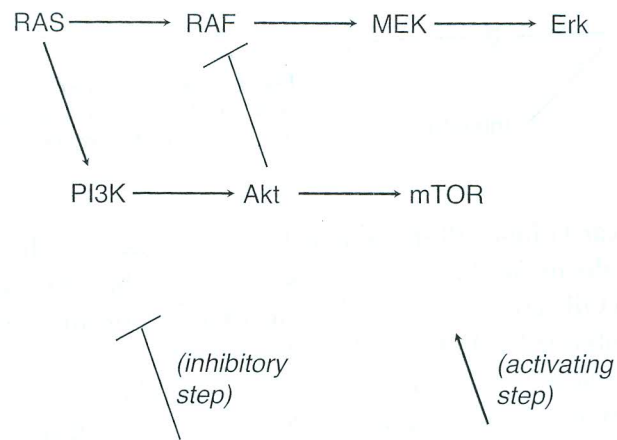


Figure 2. Signaling diagram illustrating part of the RAS network in skeletal muscle.

ASSUMPTION CONTROLS TO DETERMINE WHETHER THE EXPERIMENTAL SETTING IS REPRESENTATIVE: TISSUE SPECIFICITY AS AN EXAMPLE OF THE NEED TO ESTABLISH LIMITS TO CONCLUSIONS

The above example, in which Akt inhibits Raf, but only in skeletal muscle and a few other tissues, demonstrates the need for another type of assumption control, which can be broadly thought of as controlling for the “experimental setting.”

For the case in which cell signaling is being studied, the scientist might ask whether a particular cell is representative of other cells. This concept was mentioned in Chapter 15 on subject controls. In that chapter, the issue that was emphasized was that the scientist must determine whether a particular cell is representative, first of other cells of the same type. The point there was that the scientist wanted to know whether a particular cell was emblematic of the “representative case.” However, in this chapter, the issue can be repositioned by pointing out that even if a result in a particular cell is representative of that cell type, this still does not prove it is representative of other cell types, and that is why other tissues may need to be examined, depending on the breadth of the question.²

The same sort of subject/assumption issue was mentioned briefly when discussing studies on particular groups of people. One can label the need to study each gender individually as an assumption control, to ask whether females and males can be treated similarly in designing an experiment to answer a particular question; even some-

²The issue here is whether a cell, such as a fibroblast, that is under study is being treated as “emblematic” of all cells. In early cell culture experiments, the vast majority of published experiments seemed to be done on particular fibroblast cell lines simply because these cells were available and easy to handle. It was only later, when other cell types were studied, that it became more obvious that not every cell responds in the same way to a particular stimulus, nor does every cell type have the identical signaling networks in place.

thing that might seem to be gender- and race-neutral, such as caffeine—which may be found to cause a blood pressure increase in Caucasian males—would need to be validated as doing so in women or in non-Caucasians. However, again, the assumption control is establishing perhaps more than the “representative case”—it is used to determine answers to other fundamental questions:

1. In what settings does the effect hold true?
 - a. Are there other forces at work that would block the effect?
 - b. Does the effect apply to everyone?
2. Is the experimental setting an accurate representation of the setting in the question?

The second question goes to the issue discussed above, where a surrogate agent, coffee, is used to answer a question about caffeine. The scientist must determine what assumptions are being made in the experimental design, and whether they can be controlled for, in order to make sure the question is being answered.

REDUCTIONISM CONTROLS AS ASSUMPTION CONTROLS

The Selected Group of People as a Reductionist Model

Perhaps the most common practice in laboratory biology is the use of “reductionist” systems as models for what is taking place in more complex systems. The “reductionism” issue was introduced under “subject” controls for the purpose of emphasizing that the scientist needed to validate the model system “representative” of the complex system in terms of the usual need to determine whether one is studying the “representative case.” However, the scientist must go a lot further in the validation process of the reductionist system when designing assumption controls. Perhaps it would be helpful to examine different types of reductionist models and outline potential assumption controls that would help in the system validation process.

First, let’s return to the experimental system in which a group of humans are studied. In this type of system, the study cohort is first determined to be “representative” simply by asking whether the model derived from evaluating one group is predictive of what will happen in a second group of study subjects. Remember though that for the purpose of doing an experimental repeat, the second group of subjects must be selected under the same criteria as the first group. However, after the scientist has done the appropriate experimental repetitions, and thereby validated the model as “accurate” when applied to a particular group of individuals, there eventually comes the desire to know whether a model is accurate when applied beyond the particular experimental system in which it was first derived, to other groups of individuals.

In the case of the human study, the scientist might like to know whether an anti-cancer drug had beneficial effects even in individuals who did not have a particular

EXPERIMENTAL SAMPLE OF

le and a few
rol, which

ask whether
Chapter 15
the scientist
cells of the
a particu-
er, the issue
cell is repre-
cell types,
breadth of

discussing
ch gender
an be treat-
ren some-

as “emblem-
seemed to
to handle.
not every cell
tial signaling

characteristic used to determine entry into the experimental regime. For example, subjects might have been excluded from the study who were obviously close to death, because the severity of the disease might overwhelm the ability to determine any potential therapeutic benefit of the drug. However, once it was established that the drug did have some efficacy in the less-sick patient, it would obviously be of great interest to know whether there would be any utility in the desperately ill individual. Many people might take the posture that “there was nothing to lose” in such a cross-application and that the demonstrated efficacy provides a logical rationale to try the drug in the more advanced cases. However, there are a lot of reasons that such an approach could do more harm than good, and advance the already terminal patient to their demise. The key point here is that there are many differences in the advanced cancer patient in comparison to the earlier-stage patient, both in terms of how the individual is responding to the disease and how the tumor itself is operating. These differences between cancer stages can be thought of under the idea of “representation.” The scientist is asking for an investigation as to how the subject differs under various conditions and whether these differences matter in terms of the application of the model. Therefore, the assumption control needed in this example is a group of subjects with end-stage cancer. In a study that contains this group, the scientist would be able to actually demonstrate whether it is valid to say that the drug that works on early-stage cancer also works on end-stage cancer.

THE SELECTED DRUG DOSE AS AN ILLUSTRATION OF THE ASSUMPTION DILEMMA

Let's approach the issue from another angle and revisit the example used in the beginning of this chapter: The case where a trial was done to determine whether aspirin helps to alleviate the pain caused by headaches. The patient who died of an aspirin overdose, after having taken it previously at levels that were beneficial, illustrated the problem of misapplying data from one condition to a different—untested—condition.

The aspirin example helps to illustrate the scope of the problem when trying to perform an assumption control. The drug dose is just one of a large array of variables that could be altered in an experimental protocol. When applying the results of a study to other conditions and then asking for a new study to query the assumptions that were made in the initial experiment, one is literally asking the scientist to question every “simplification” used in carving out the experimental system, in order to see whether the model still holds true under other conditions. The irony of this requirement is that the scientist may lose one of the “benefits” of simplification—the ability to perform the study in a restricted environment, under particular conditions.

Another example of the assumption problem can be illustrated using the issue of “drug interactions.” It may be true that drug A is beneficial in treating cancer, and drug B is beneficial as well, but the simultaneous use of drug A and drug B results in less benefit than when either drug A or drug B is used individually. This possibility demonstrates

the problem with the assumption that “if two things are good individually, they will go well together,” or, even more generally, that “if something is true under condition X , it will also be true under condition Y .” That is often not the case. In the case of drug interactions, it is particularly difficult to learn early on whether drug A will work under settings where other drugs are used, because study designs often mandate that a drug be studied in the situation where no other drugs are given, simply because one does not want the complicating effects of multiple variables. Although it is a valid experimental design to exclude other drugs from a study, the physician may be left to wonder whether the newly approved drug can be used in combination with past treatment regimes. Whether drugs work in combination then usually becomes the subject of new study, because scientists and doctors have come to realize the folly of simply assuming that such treatments can be mixed without subjecting the procedure to experimentation.

It is obviously very difficult to list all of the potential variables that may affect whether a particular finding X may continue to hold true. Therefore, a convenient prescription cannot be provided as to the kind of other settings in which one must additionally test whether a finding X continues to hold. However, a prescription can be provided to determine how one discovers where there may be a relevant variable Y that affects the finding X : those settings where X no longer holds true. In other words, if the scientist has “validated” by continued experimentation that aspirin induces pain relief, but in a new setting no longer finds this to be the case, this new setting provides a basis for further experimentation, to discover what there is about the different condition that interferes with aspirin’s function. Therefore, one can continue to resort to the requirement that a model holds true if it predicts the future, to say that if a model fails to predict the future under a particular condition, then that setting will help to teach the scientist where an inappropriate assumption was made.

NONHUMAN ANIMAL MODELS AND THE NEED FOR ASSUMPTION CONTROLS TO ESTABLISH RELEVANCE FOR THE HUMAN

Next, let’s consider the experimental system in which animals, such as mice or rats (or really any nonhuman animal), are studied as “model” systems for the human. One might wonder why the issue is framed in this way. Why can’t studies on mice be done simply to see how mice react to a particular perturbation? Although such a study is quite possible, and perfectly valid, one does not then have the “assumption” issue discussed in this chapter. Studying mice to establish the effect of Perturbation Q on mice is exactly the same as studying humans to see the effect of a treatment on humans; there is no additional level of complexity beyond the issue of whether the group of mice chosen is representative of mice in general. However, once a study is initiated on mice as a preliminary step to performing a study on humans, then one must perform assumption controls to make sure that the mouse constitutes a legitimate “test system” for the human.

There is no shortage of examples of particular proteins that differ substantially between rodent and human. There is a whole class of ligands, called “interleukins,” in

which the sequences differ significantly between the species, such that a human interleukin may not stimulate the cognate rat or mouse receptor. Furthermore, in this class of molecules, there is at least some evidence that function may differ between the species. There also are G-protein-coupled receptors that exist in the mouse but not the human, and *visa versa*, and some of those that are in common between the two species have different functions. As a third example of species differences, there is a continuing debate as to whether humans have any residual preservation of the "pheromone system," in which animals secrete a chemical for the purpose of attracting a mate. Therefore, it should be clear why study of any of these systems in the mouse might be questioned for its validity when applied to the human.

The ready response to these objections is simple: The issue should be subjected to experimental validation. The great advantage of using mice as an experimental model is that studies can be performed on mice that are simply impossible to do on the human. For example, mice go through their entire life span in 24 months; therefore, it is possible to do relatively "long-term" studies more quickly than they could be done on humans. Furthermore, given their size, the availability of genetically inbred strains, and the ability to control their conditions, the scientist can study dozens of highly similar individuals in close detail, under extremely controlled conditions. It is for these reasons that one can make much faster progress on the rodent system than when studying the human. Whether this progress constitutes an advance in understanding how the human works can then be determined, without having to do all of the dozens of lead-up experiments that led to the understanding in the rodent. For example, it might take a couple of years to figure out that blockade of a particular receptor cures solid tumors in mice, whereas blockade of other receptors does not have this effect. The successful candidate cancer target can then be studied in the human in a much more efficient fashion.

Of course, one may object that if the mouse is not predictive of the human situation, then all of the advance work on the rodent system was a distraction and actually slowed down progress in understanding human disease; however, there has been enough success with the approach at least to say that it does achieve benefits for the human. And as for losing out by failing to study the human directly, as touched on earlier it is simply not possible to do human experimentation on the basic level that is required to get the data one can obtain from rodents, so nothing is really lost by at least making the attempt in the animal system. The point here is that there is no need to "assume" that the mouse is relevant to the human. Rather, the scientist must validate its relevance by performing subsequent trials on humans.

ISOLATED CELLS AND THE USE OF ASSUMPTION CONTROLS TO ESTABLISH RELEVANCE FOR THE WHOLE ORGANISM

As noted in Chapter 15, on subject controls, the next step in the reductionist march to simplicity is to move away from test organisms completely and to study cells in isolation. One potential advantage to taking this step is that studies can be done on

can inter-
this class
een the
but not
the two
ere is a
of the
of attract-
s in the

jected to
model is
human.
t is possi-
lone on
ains, and
ly similar
reasons
aying the
re human
p exper-
a couple
s in mice,
undicate
on.

man situ-
nd actu-
as been
ts for the
ched on
level that
lost by at
no need
ust vali-

TO

narch to
y cells in
lone on

human cells in addition to mouse or rat or any other kind of animal cell, and this cross-comparison at the cellular level can help to validate whether the various animals might serve as surrogates for one another, and under what conditions. For example, the scientist can determine whether interleukin signaling in a mouse cell works in the same way for a particular interleukin/receptor pair as it does in a human cell. As explained above, these signaling systems are not well conserved between the two species, so it is advantageous to be able to compare them directly in some way; having the cells and the molecules available allows the scientist to do this. Therefore, one can take advantage of the availability of cell lines to do some "assumption control work," as mentioned in earlier sections, such as demonstrating cross-reactivity of protein orthologs between species.

As with the case of doing work in rodent or yeast systems to at least determine how these systems function, tissue culture experiments can be justified without further need for assumption controls if the point is to determine how the isolated cell functions. Such knowledge is obviously useful on its own terms, and is also essential in order to compare how the isolated cell functions with its operation in the body. If the cell changes in the context of its natural environment, the alterations would then give the scientist a clue as to cell-autonomous control mechanisms, for example, influences from other cells or from circulating growth factors.

It is demonstrable that basic cellular functions work the same way in an isolated cell as they do when the cell is contained within a body—functions such as the mechanism of cell division, how mitochondria are maintained, how proteins are synthesized and degraded, how proteins are secreted, the mechanism for fatty acid metabolism, the mechanism by which cells move, or change shape, etc. What differs is how a cell is perturbed by being contacted by certain other cell types or by being stimulated with different hormones delivered by the blood system; these sort of complex interactions are lost when studying the isolated cell. Thus, the limits of cell culture as a model system for the full organism are fairly clear and again can be readily established by asking questions of the cell within the context of the organism, using models derived from experiments on the isolated cell.

An example of the advantages and potential limitations of isolated cell work is drug sensitivity. Bacteria can be tested in culture for their sensitivity or resistance to antibiotics. This is a standard technique used in hospitals to determine which drugs should be used to treat a particular infection. The assay works quite reliably. Antibiotics that are determined to kill a bacterial strain in isolation almost always work in the context of the patient. In contrast, studies performed on cancer cells that are isolated from the body, using chemotherapeutic agents, are seemingly less reliable. The problem is that the cancer cell may change genetically when placed into the tissue culture setting. In addition, the ease at getting a drug to a tumor cell in culture is not comparable to the difficulty of reaching every cancer cell in a tumor. Finally, it may be that tumors are composed of genetically diverse cells and therefore one has the "representation" problem. The cells that are killed in culture may not be emblematic of all the cells in the tumor.

The scientist thus learns of the potential advantages and limitations of using tissue culture systems by testing whether the data gained regarding these cells will predict what will happen when the model derived is applied to the organism. This process is where the assumption control is performed. One asks the question “Does the antibiotic that killed the bacteria on a plate also kill the bacteria when in the body?” One asks the question “Does the drug that kills the cancer cells in culture also cure the tumor in the animal?”

ISOLATED MOLECULES AND THE USE OF ASSUMPTION CONTROLS TO ESTABLISH RELEVANCE FOR THE WHOLE CELL

What happens when the scientist strips a molecule away from its cellular context and studies it in isolation? Again, this depends on the question. If the scientist is simply studying the structure of the molecule, or determining how it behaves in isolation, then no additional assumption needs to be controlled for. However, when performing experiments to understand how the molecule behaves in another context, such as in the cell, the scientist must prove that the data gained regarding the molecule in isolation are relevant to the molecule in the context of interest.

Structural studies are almost always done on proteins in isolation, although occasionally a protein may be crystallized in complex with a particular partner. In addition, there are instances when a protein's structure can be determined in both its “active” and “inactive” conformations. These more complicated studies demonstrate that proteins can change shape depending on their context—an “activated” protein can be structurally very different from an “inactive” protein. This finding demonstrates the need to prove that the context used is relevant. For example, if an enzyme is being studied so that a drug can be developed to insert into its active site, and inhibit its function, then the scientist would need to crystallize the enzyme in its active conformation. In addition, a drug that was found to bind the enzyme in that conformation would then need to be tested in a cellular context to ensure that it continued to bind the enzyme. It is possible that in the cell, the protein is bound by other proteins, such that access to the drug target is impeded.

Note that it is not necessarily the case that “the more reductionist the model, the greater the need for assumption controls.” Rather, it is the case that whatever the model—from a human group to an isolated molecule—assumption controls must be performed to ensure that the model is appropriate for the question being asked.

CONTROLLING FOR THE “META” ASSUMPTION THAT THE EXPERIMENTAL MODEL IS WORKING THE WAY THE SCIENTIST THINKS IT IS WORKING

One other assumption control should be mentioned. This may be the least-practiced control in day-to-day laboratory science and yet one of the most important controls

of using
will pre-
a. This
n "Does
n in the
re also

there is. This control can make the difference between a successful scientific program and a failure. It can help the scientist avoid publishing artifacts of inappropriate experimental designs. Here is the control: The scientist must take an experimental model and subject it to verification using an entirely different methodology than that used to derive the model.

This requirement—that the scientist seek to achieve an answer by asking questions in more than one way—controls against the assumption that the first methodology operated correctly. Or, put another way, the odds are quite small that two entirely different methods will achieve the same artifact.

This issue was touched on in the "experimental repeats" chapter, but it was also pointed out that the control is really an assumption control. Whether talking about using different types of reagents, different experimentalists, or distinct types of equipment, doing things in more than one way will control for the fundamental assumption every scientist makes when reaching a conclusion: The data derived from the experimental system have been validated.

TO

xt and
s simply
solation,
rform-
xt, such
olecule in

gh occa-
In addi-
oth its
onstrate
protein
emon-
enzyme
d inhib-
e con-
nforma-
inued to
oteins,

odel, the
er the
must be
red.

**ENTAL
ING**

practiced
ontrols



Experimentalist Controls

Establishing a Claim to an Objective Perspective

SCIENTISTS MAY NOT LIKE TO BE TOLD that they are just another variable that must be controlled for. This is, however, the case. The scientist does not stand apart from the experimental system. In listing out what constitutes an “objective truth” Robert Nozick¹ noted that the truth must be accessible from different angles (observed in more than one way), intersubjective (observable by more than one person), and independent (if p is an independent truth, it is truly independent of the observer’s belief system, hopes, dreams, desires, etc). We have already touched on some of these strictures, such as the need to establish a finding in more than one way, and the issue of relativism and observers’ belief systems (which is dealt with a little more in the final chapter of this book). In this chapter, more is said about intersubjectivity and independence.

Intersubjectivity—as defined in this volume—does not mean that something can be true only if multiple observers can testify to it being true. For example, it may be easy to find dozens of people willing to say that the Earth is flat or that black is white. Rather, intersubjectivity relies on the concept of “objectivity,” as when one strips away an order to testify in a particular manner or an instilled “need” to say $2 + 2 = 5$ and find, thereby, that $2 + 2 = 4$. Intersubjectivity is what happens when the sun rises and everyone within view perceives this to have happened, or when any individual drops an apple, and each person having done so finds that it falls to the ground. In this last observation, the force on the apple acts without the observer adding anything to the apple; gravity, in its essence, does not require any “help” from the observer or a “belief” on the part of the observer in order to pull the apple down. This is what is meant by an objective fact—an essential character, divorced from any added property

¹*Invariances, the Structure of the Objective World*, by Robert Nozick (2001), Belknap Press of Harvard University Press, Cambridge, Massachusetts, ISBN 0-674-00631-3.

or belief. As noted repeatedly in this volume, one can demonstrate something to be objectively “true” by demonstrating that the Model of that fact is verifiable, in that it represents what will happen in the future.

Cannot a fact be “true” if only one person perceives it to be true? What of the situation when Newton first defined the “law of inertia,” was this law something other than the truth until it was verified by others? Clearly not; at no point is there a claim that any person or group of people need to perceive or believe in something in order for it to be true (the universe will still exist even if human beings manage to incinerate the earth).² What is meant here is that for a phenomenon to be observed in an objective manner, that phenomenon must be observable by multiple reporters. This is what is meant by intersubjectivity. The reader may still balk against this requirement. After all, it is still theoretically possible for some event to occur, in full view of millions, and yet be observed by only a single person, because, thanks to some additional perception, this individual alone is capable of seeing the event. The rub in that example is the “additional perception.” The requirement for intersubjectivity, as defined here, is that if a person is in the position to conduct the observation objectively, the event must be observable. If a particular event thus requires some “extra perception” that only one person has, then the other people cannot confirm or deny the event. Therefore, the need for intersubjectivity cannot be used to “deny” that the thing happened; it is just that, without it, the report cannot be verified. As an example: What if only one person has access to a telescope? This person now has the “additional perception” required to observe celestial objects that other people cannot see. The requirement for intersubjectivity is that were others allowed to look through the telescope, they too would report these celestial objects. Lest this still fails to satisfy the reader, recall that the point of the scientific program, as defined here, is to establish models for how reality functions, in that they represent what will happen. If one person’s report of reality does not agree with anyone else’s report, that individual cannot produce a model that “works” for anyone else.

ESTABLISHING OBJECTIVITY

Much of this volume has been about establishing which steps are necessary to make an objective observation and to decrease the chances of bias and misinterpretation. Here, we list what must be done to the experimentalist to increase the chance that the observation and interpretation of the data will be objective. The goal of producing an “objective” measurement can be thought of as that of establishing an “objective truth.” In these examples, the effort is aimed at establishing that the finding has “independence,” i.e., that it does not require a particular belief system or subjective

²The reader may object that this is not a testable statement. That is not true. One can ask the question “Will the universe continue to exist even after the earth is incinerated?” and even design an experiment to address this question. It is the case, however, that this particular experiment can be done only once.

judgment in order to be upheld. The distinction would be as for establishing that a painting exists (a demonstrably objective truth) versus determining whether the painting is beautiful (a subjective judgment, which requires something to be brought to the painting by the observer).

THE OPEN-ENDED QUESTION STRUCTURE AS A CONTROL FOR THE SCIENTIST'S NEED FOR A PARTICULAR ANSWER

The advice given repeatedly in this volume—to make use of the open-ended question as a frame of reference—is in large part given to insulate the scientist from a subjective “need” to get a particular result. As pointed out early on, if an experiment is framed with an open-ended question, any answer may be deemed responsive. This answer can then take on the role of a proven and objective fact, but only if it results in a model that demonstrably represents reality, by successfully representing a future outcome. Therefore by “demonstrably” we mean that the model accurately represents what will happen when the experiment is done again.

Forcing the scientist to establish the experimental framework in the form of a question emphasizes the scientist’s ignorance as to the outcome of the experiment. It is important to reinforce to experimentalists that they are operating in a state of ignorance. This helps scientists to realize that any intuition or guess that they might have as to the way things are is not yet supported by data. Furthermore, even ideas that are logically induced from objective facts fail to meet the criteria for objective truth until they have been subjected to verification. In this way, even the most brilliant theory must be properly labeled as “unproven” until such evidence is gathered, and the theory is found to be true when evaluated from many angles, by many people, independent of biases or beliefs—in short, when operating from an objective perspective.

Brilliant scientists may object to all of this as robbing them of the “credit” they should rightly get for accurately intuiting how something works, even before doing the experiments to demonstrate the fact. However, such individuals should recognize that with the desire for such fame and glory, is the instilled need to achieve that credit, and the only way to do so is to prove that the particular idea is “true,” even if that idea does not fit the experimental data. As an experimentalist, this is not a good position to be in. The simple mathematics of the situation is that the chance of a theory based on no data turning out to be true in all of its parts is actually pretty small.³

What if a theory is produced that cannot be subjected to experimentation? What if this is the most brilliant theory ever conceived, it explains everything that has ever happened, yet it cannot be subjected to verification, because it is impossible to establish a system in which it can be used to predict an event? Such a situation is probably beyond the scope of this book, since it is beyond any prescription given here for performing “exper-

³Also, there is no reason that the scientist would not experience fame and glory for successfully answering a difficult question. In fact, this occurs quite frequently.

imental science.” However, given the requirement stated earlier, that a fact is distinguished from fiction only by its ability to accurately represent a future event and that the judgment as to whether the model succeeds in representing that future event must be reached in an objectively verifiable manner, one must note that such a theory could not be labeled as objective “truth” or “reality.” Take, for example, the theory that we are all operating within a computer simulation—a reality invented by a child with Attention Deficit Disorder—and the reason that random and contradictory events might occur is that this child is not paying attention to the details; yet the basic structure of the simulation forces the child to obey certain “rules” such as gravity, etc. Such a theory could be said to provide an explanation for everything that has and will happen, but since it is not subject to verification by any of the mentioned criteria, it can be labeled not “true” or not “demonstrably true,” and therefore indistinguishable from fiction.

Is it fair to call the formulation of a question (as opposed to a hypothesis) a “control”? One might say that if a question exists at the same time as the hypothesis, then the different formulation of the two entities may control for the possibility that the hypothesis instills a filter for what is deemed “acceptable” or “positive” data. If such a filter were to be put in place in the setting of the hypothesis, but not in the setting of the question, then the question would function in the same way as other types of “controls”: It would help to show the scientist how the experimental system is operating.

What of the main purpose of the hypothesis as a way to avoid inductive reasoning, and to avoid the difficult problem of verification, or the claim that what has happened in the past allows for a rational prediction of the future? This “advantage” of the hypothesis is simply not consistent with the way experimental science operates. What every empirical scientist does is take results and use them as a way of understanding how things work in order to understand how things will work in the future; this is the process of inductive reasoning. If one cannot rely on something behaving the same way again, there is no workable model for that thing. The entire history of scientific and technological progress is a narrative of individuals building on accumulated knowledge. This makes the use of the Critical Rationalist framework by an experimental scientist particularly ironic, since Popper⁴ explicitly rejects the experimental program as being able to produce a model that represents the way something will operate in the future. Probably the only reason that the Hypothesis has been invoked so consistently is that few people actually practice Critical Rationalism as intended: It is demonstrable that scientists do not seek to deny hypotheses, they in fact seek to verify models. After practicing science for some time, the impression is that people who feel required to use Critical Rationalism to construct the Hypothesis after the experiment is done and the results are gathered. This practice may be undertaken simply to meet the perceived need for a hypothesis to be in place and in recognition of the fact that constructing such a thing before the data

⁴Karl Popper, previously referenced in Chapter 1.

distin-
 hat the
 ist be
 id not
 are all
 ntion
 occur is
 simula-
 id be
 it is not
 or not
 ontrol"?
 the dif-
 othesis
 were to
 uestion,
 would

reason-
 at has
 antage"
 e oper-
 ay of
 in the
 nothing
 entire
 uilding
 frame-
 rejects
 ne way
 othesis
 ritical
 to deny
 ne time,
 o con-
 ed. This
 sis to be
 e data

are available is almost a guarantee that the hypothesis will be falsified, leaving no mechanism through which to demonstrate progress. One can only note that if this impression is correct, then practicing scientists are in a poor state of affairs; the scientist should not be in a position of having to "game the system" in order to conduct the most basic step in the scientific process, establishing a clear and useful framework for experimentation.

The practical need for experimental verification of a model for reality, or for an "objective truth," is a further justification for the question structure, because the combination of the hypothesis as a statement of fact and the need to engage in verification dangerously positions the scientist as trying to prove a statement true, which was made in the absence of data. In addition, as has been pointed out by many philosophers, Popper simply fails in his bid to void induction, since accumulated knowledge is required for the formulation of the hypothesis, and since multiple "successful" tests of the hypothesis necessarily make the hypothesis more secure.⁵

ESTABLISHING PRESET CRITERIA FOR EVALUATION

Another way to control against the "need" to interpret data in a particular way is to establish, in advance of an experiment, the criteria for evaluation. For example, if the question is "Does caffeine cause an increase in blood pressure?," the scientist should first determine what would be deemed an "increase." Does a change in blood pressure from 110/70 to 110/71 merit a "yes" to the question, even if the sample size renders this change statistically significant? If such issues are defined prior to experimentation, then it will not be necessary to go through the evaluation after the data are gathered, when one's pet theory is in danger of being disproved. Another example given earlier concerned a study of "inflammation." There are multiple criteria for what constitutes inflammation; therefore, the term must be defined in the context of the experiment. A measuring system must also be set up to avoid having to redefine "inflammation" postfacto, to fit the data.

The establishment of preset criteria can be seen as an experimentalist control in the same way that the use of the question acts as a control; if it is possible to demonstrate that criteria may be altered to fit the data, then there is clearly a need for preexisting standards.

BLINDED ANALYSES

In Chapter 15 on Subject Controls, it was explained that blinding the subjects to their status in the experimental design prevents them from being influenced by that status. A

⁵See footnote 1 for the reference to Nozick's 2001 book. This last point was made by Nozick in the same paragraph that he labels Popper as "incoherent."

“double-blind” experiment, in which the scientist is also ignorant of the subjects’ status (which was also touched upon), is used to remove potential bias on the part of the scientist (also previously explained). For example, consider the case in which a scientist is trying to answer the question, “What is the effect of Drug J on cell size?” To answer this question, the scientist has an assistant grow cells in tissue culture. The assistant treats a couple of plates with an “all but *X*” negative control—the buffer used to solubilize drug J—and then doses the remaining plates with various concentrations of the drug. The assistant gives each plate a code name, so that the scientist will not know which plate of cells received the drug and which are the negative control plates. At the end of the experiment, the scientist feeds the cells into an instrument capable of analyzing cell size and assigns a size range for each plate of cells. The scientist thereby determines in advance of knowing which plate received a particular treatment whether there is any deviation in size in any of the cell populations when compared to each other. After the data are established and the results tallied, the assistant hands the scientist the code, so that the scientist can discover whether drug J caused an increase in cell size.

DIFFERENT EVALUATORS

Just as the use of a second reagent or measuring system can better establish whether an effect actually occurs (as opposed to the effect being an “artifact” of a particular system), so too does the presence of a second or third scientist to control for any idiosyncrasy in evaluation by a single scientist. The second and third scientists also help to establish a system of “intersubjectivity,” in which more than one individual can seek to conduct an objective observation of an effect.

Note that the way science is often conducted in academia is to put a single graduate student or postdoctoral fellow in sole charge of an experimental project. The thought is that a scientist-in-training should conduct every phase of the study to ensure learning every step of the project. In addition, when dealing with Ph.D. candidates, one wants to make sure that they have done the work that forms the basis for their thesis. Were two graduate students working together, it would almost always be the case that one of the students did more work than the other, and thus there might be a feeling that the slacker is less deserving of credit for the project.

Although these are valid points, there is potential danger to the experimental program in putting a single individual in charge of all steps in design and evaluation of experiments. Even if great care is taken to shield oneself from bias, as in “blinding” oneself as part of the experimental design, at the end of the project confirmation of the observation by another scientist will still be required to control against both bias and any potential idiosyncrasy in the observation or reporting method, and to establish the “intersubjectivity” of the model.

The reader might object that it is just as likely for two people to conduct a biased program as one, but this can be demonstrated to be false; even if a particular bias is common, the chance that two people will have that same bias will always be less than that of one person, unless the bias is universally held.

It is therefore strongly recommended that projects be carried out by more than one person and that these people be encouraged to work together in a team setting. If different people are performing distinct aspects of the same project, each individual will serve as a "check" on each other's methodology and questions or issues that come up can be resolved.

The Principal Investigator must also be closely involved in the Experimental Projects that go on in the laboratory. Although the goal of any particular lab worker and the Principal Investigator might be to produce a result that is publishable, the Principal Investigator should also have as a primary goal the wish to maintain an operation that is credible. Any slip up, such as the publication of demonstrably false data, could threaten the Principal Investigator's status. Such a threat is healthy if it induces the Principal Investigator to instill checks on the experimental program, such as independent observations of the data in an objective manner. Of course, there are well-known instances where the Principal Investigator was the individual instilling bias in the interpretation of results, and it is difficult to insist that lab members function as "controls" in such a setting. Perhaps the best way to "control" for this event is to encourage claimed "objective truths" to be verified in more than laboratory. This will be discussed further, but for now it can be pointed out that if a team of scientists is aware that their data will be subjected to independent verification, as happens almost universally in experimental science, then they may be motivated to produce models that will survive such scrutiny and thus be as free from idiosyncratic bias as is feasible.

OBJECTIVE EVALUATORS—COMPUTERS

In the attempt to obtain an "independent" and "objective" evaluation of data, non-human evaluation systems may be used to perform analyses. Unless one is actively seeking to mislead, as when a computer is actually programmed to treat some groups differently than others, this type of analysis system could be used at least as a defense against subtle or inadvertent biases, such as seeking out a particular type of event in a group where one wants to see that event.

Of course, it is fair to program in certain types of bias, as long as they are applied equally to all groups. For example, it is acceptable to filter data for an "effect," such as when an anticancer drug cures cancer only in certain individuals. Thus, one can see whether there is a particular group of people who had an effect from the treatment and compare that finding to the control group, to see whether there is a similar effect in the control. In this type of analysis, there is a "bias" for an "effect," but this is acceptable since the same bias is applied equally to all groups.

OUTSIDE REPETITION AS THE FINAL ARBITER FOR INDEPENDENCE AND INTERSUBJECTIVITY

In psychiatry, there is a syndrome called a "folie à deux," in which two people share a psychosis, or a belief that is not verifiable by others. This condition—in which it can

be said that something abnormal is going on because the belief structure in which the twosome engaged is not shared by others—implies the last step in the verification process, where someone, other than the individual conducting the experimental program, enters the process by attempting to verify the derived model independently from the scientist who established it. As before, one cannot say that something is an “objective truth” simply because someone else says it is, and, similarly, one cannot deny a finding simply because someone else failed to verify it. However, it can be said that in order for something to be an “objective truth,” it must be verifiable by others positioned to objectively confirm the finding. In practice, there is often a back-and-forth, and some instances of a model are confirmed, whereas others are denied, as different labs enter into the investigative process and seek to apply the resulting model to their own systems.

Occasionally, the investigator will come across a truth such as “objects fall when you release them,” and this proposed “truth” will be easily verified by all, in just about every setting.⁶ In other instances, the investigator will come across a model such as “aspirin is an effective pain-reliever,” and in the confirmation process, limitations will be added to that model, as instances are found in which aspirin is not an effective pain reliever, and other settings are found in which aspirin causes more harm than good. Eventually, one will have a model of how aspirin functions that will allow a verifiably accurate prediction for how it will function in the future. The use of experimentalist controls will help in the development of the accurate model.

⁶ . . . with the appropriate caveats developed for leaves being blown about on a windy day, and why this happens.



A Description of Biological Empiricism

MATHEMATICIANS AND PHILOSOPHERS LIKE THEIR TRUTHS to be black and white. Something is either an “objective truth” or it is not. A hypothesis is either falsified or it is not. A theorem is either proved or disproved. However, the reality is that things are a little messier in biology, because one is dealing with systems that have thousands of variables. It is therefore quite difficult to state something neatly as an immutable “fact” and expect the statement to hold true in every instance. For example, there are people who can, and actually do, eat cheeseburgers every day without a perceptible increase in their cholesterol levels. It is also not hard to find individuals who smoke two packs of cigarettes every day for 50 years without getting lung cancer. Some people, who eat correctly and exercise every day, nevertheless drop dead of a heart attack at the age of 35. The existence of such individuals is part of the reason for the development of statistics, to provide a process to determine whether there is a “real” difference between two studied groups, as demonstrated by showing that the difference, whatever it is, exists to some specified degree when the experiment is repeated over and over again—even if the effect is not absolute in its penetrance.

The complexity of biological systems constitutes another reason to value the malleable model over the rather flinty hypothesis, which is either rejected or accepted in a binary fashion. If a model is simply required to represent the data set, one can never go wrong with it; the scientist can go through a process of iterative editing, to alter the model, as new data comes along, establishing limits and exceptions to previous interpretations. Certainly, to go back to the drawing board every time a falsified hypothesis was encountered would slow progress. And even the lack of falsification does not allow one to attach any additional impetus to a follow-on hypothesis, given the proscription against inductive reasoning, which is a component of Critical Rationalism.

Imagine a setting where a particular chemical is found to cause long life when given to most people. However, certain individuals who have a particular genetic marker die immediately if they are exposed to this same chemical. This combination

of facts results in a model that mirrors the reality that the chemical results in prolonged life span unless an individual is of a particular genetic subtype, and in that instance, it results in rapid death. To this model there may be additional limitations and caveats. For example, it may be found that even those who would die if fed the chemical are protected when they drink some potion, because a protein in this potion binds the chemical and prevents it from doing harm. One can go on for considerable lengths, attaching particular context to each biological finding. After this biological system is characterized, if the scientist then tries to formulate a statement of "truth" about the chemical that does something other than simply describe how it acts in various settings, and establishes the various underlying mechanistic causalities, the effort would be difficult to accomplish and of questionable utility.

With experience in the process of experimental biology comes a gradual understanding of extremely complex systems by first reducing them to simple components and then gradually returning them to their normal environment. The scientist might find that different environments result in different findings, and these contexts would therefore be incorporated in the model. It is demonstrable that an understanding of how the system works has been developed when one produces a verifiable model, a model that represents how the system will work in the future, in a statistically significant manner. This last caveat is important: The scientist will rarely succeed in predicting a system to 100% fidelity, and if such a standard were required, almost no progress in biology would be made, simply because it is not currently possible to track all of the relevant variables.

ASSIGNING CAUSALITY AND ASSESSING THE REQUIREMENTS FOR NECESSITY AND SUFFICIENCY

Given the challenges in establishing an accurate description of a biological system, it may be surprising that any progress at all can be made in demonstrating mechanistic causality, as when one says that A causes B in a biological context. However, it is possible to determine causal links in biological systems. This is especially the case when A can be shown to be either necessary (required) for B and/or sufficient to result in B. In such an instance, the absence of A would result in B being inhibited, or not in effect, whereas the addition of A would result in B. This sort of empirical finding allows one to state as an "objective truth" that A causes B, and one can agree that this is an "objective truth" when the finding is verified along the lines discussed previously.

It should be noted, however, that there are many instances in biological systems where one can show that A causes B and yet find that A is neither necessary nor sufficient for B. For example, the scientist might add A and induce B and yet find settings where B exists without A. In addition, there may be settings where one requires C in order for A to cause B. Consider, for example, the finding that eating fatty foods (A) increases the chance of heart disease (B). However, this may be true only in settings where an individual has high cholesterol levels (C). It may also be

shown that an individual can have a heart attack without eating fatty foods. Therefore, eating fatty foods may be neither necessary nor sufficient to result in heart attacks, and yet the consumption of fatty foods may demonstrably cause heart attacks in those with high cholesterol levels. Thus, a causal link may exist only in a particular context, and the demand for “necessity” and “sufficiency” in order to assign causality may be an inappropriately simplistic demand given the complexity of the biological system.

Another potential reason to resist as an absolute requirement demands for establishing necessity and sufficiency is the demonstration of biological redundancy. There are many biological settings where more than one protein can contribute to a particular effect. For example, imagine an instance where two proteins, Protein A and Protein B, each can comprise the nuclear membrane. Therefore, if either Protein A or Protein B is deleted individually, the nuclear membrane is maintained. One can thus say that neither Protein A nor Protein B is required to have a nuclear membrane. However, if both Protein A and Protein B are simultaneously removed from the system, then the nuclear membrane is completely disrupted. Furthermore, adding Protein A or Protein B back to the disrupted state allows the nuclear membrane to be reconstructed. In this complex system, one can say that neither Protein A nor Protein B is required or sufficient for the existence of a nuclear membrane, but in the setting where the partner protein is absent, both Protein A and Protein B are necessary and sufficient for the structure. On the one hand, this type of complexity might illustrate a problem in establishing requirements for causality such as necessity and sufficiency. On the other hand, one might respond that causality could not be determined for either protein before the partner protein was eliminated from the system. Therefore, it is still valid to insist on necessity or sufficiency to be shown, at least in some context. The problem with even this last stricture is that there can be additional shades of gray.

Let's take the case of a scientist who is studying the dynamics in a particular office. Every day, a clerk named Jebediah comes in at 7 a.m. and turns on the light. This happens without fail; summer, autumn, winter, and spring, Jebediah can be found in the office at 7 a.m. Before he arrives, the office is dark. After he enters, all of the lights in the office are turned on, in the same sequential fashion. As the day goes forward, other office members arrive, and each behaves in a particular way; for example, each day at 7:20 a.m., Betty Sue almost always comes through the doors, heads to the small kitchen area, and starts to brew the first pot of coffee for the day. Although there are some tasks that are done by multiple people, such as copying documents, using the fax, or getting the mail, the office seems to hum along on a pretty set routine, with each individual doing some particular subset of jobs. After having observed this situation for a period of time, the scientist establishes a model: “Jebediah turns on the lights at 7 a.m.” The scientist next queries the model by asking whether it represents the situation in the future and finds that indeed every time it is checked the model is verified. Finally, the scientist decides to do an experiment: Jebediah is sent an e-mail telling him that he will get \$100 if he shows up across town at 7 a.m. the next morning. So as not to be

a bad guy, the scientist actually puts \$100 in an envelope and places it just where the e-mail said it would be. The scientist then settles down in front of the bank of closed-circuit monitors to see what will happen on the following day. The next morning arrives, 7 a.m. comes and goes, and, sure enough, the lights do not come on. The office remains dark. This state of affairs continues until 7:20 a.m., when Betty Sue arrives. She opens the door and hesitates for a moment, seeing that all of the lights are off. Maybe a minute passes, maybe less. Betty Sue then walks into the office, turns to her right, and turns on the light. She then proceeds to turn on the rest of the lights in the suite, although in a different order than was Jebediah's routine. The scientist surveying this scene is now in a quandry: The model states that "Jebediah turns on the lights," but in the absence of Jebediah, the scientist now finds that Betty Sue can complement Jebediah's activity, and that Jeb is not necessary for the lights to go on, although he may still be required for the lights to be on between 7 a.m. and 7:20 a.m. Is there some altered truth-level to the statement "Jebediah turns on the lights" when one finds that Betty Sue also turns on the lights?

Here are some things to think about: Does the fact that Jebediah is not the only one who can do what he does alter the fact that under "normal" or "usual" circumstances he is in fact the one who does what he does? In an exploration of "how things work" is it fair to say that because things work differently under a perturbed condition that this somehow invalidates a finding as to how they work under an unperturbed condition?

The example illustrates that the perturbed state will reveal a "normal function" of a thing only if there are no other agents that can complement for that thing. In other words, if only one person is capable of performing a particular task in the office, this will be revealed by removing that person from the mix; however, if multiple individuals can perform the task, and yet only one person has been assigned the task, one would not easily figure this out by removing the person from the setting.

At the end of the day, the scientist may simply note that the demand for "absolute" causality, as when one requires that only one thing should be responsible for some other thing, may be naive in the context of a biological system, and one should instead be satisfied with showing how things actually work, and to assess the degree to which A contributes to B, even if A is not sufficient to cause B, or absolutely required for B to be in effect.

SETTINGS WHERE NECESSITY AND SUFFICIENCY ARE REQUIRED

So, there are many biological settings in which an agent, a protein, for example, may contribute to a particular subject of study, such as a phenotype, but that agent may be neither necessary nor sufficient to induce that phenotype. However, it should be pointed out that there are instances where there is a particular "need" to find those gene products that are absolutely required for a particular function. For example, if the aim is to cure cancer, one cannot be satisfied with blocking a step that "normally" contributes to tumor growth if there are multiple other mechanisms that can com-

plement in the absence of the targeted mechanism. And it should be pointed out that one can in fact find single genes that are absolutely required, even for critical biological processes. For example, there are dozens of genes that are absolutely critical for an organism's development; this can be demonstrated by removing any one of these genes and finding that in its absence, a fertilized mouse egg will not develop into a mouse. Therefore, it should not be dismissed as irrelevant or unimportant that the finding that a gene, for example, is absolutely necessary and sufficient for a particular process. One should simply be aware that this requirement may cause one to miss settings in which a set of genes, as opposed to a single entity, may be necessary and sufficient for that process, whereas any one in isolation is dispensable.

Consider again the goal of curing cancer. It is demonstrable in many tumor settings that a single chemotherapeutic agent is ineffective in treating the disease, and yet a combination of several drugs can be beneficial. This helps to show that only by perturbing multiple targets can a tumor be slowed down, and if only one target is hit, the tumor is capable of using alternative mechanisms to spread. Another example where multiple targets must be perturbed to obtain a desired biological effect is in anti-HIV therapy. It was only when "triple therapy," the use of a trio of drugs, was used that significant anti-HIV activity was seen in the clinic. The early use of a single drug, AZT, had minimal effect. These examples help to show that to characterize a system in a useful manner, one must accept the biological complexity of that system. Systems are composed of multiple variables, each of which represents only one aspect when studied in isolation.

DIFFERENT TYPES OF BIOLOGICAL QUESTIONS

The example of wanting to know what needs to be done to a biological system in order to induce a "desired effect" helps to show that biological questions can go beyond the straightforward investigation: "How do things normally work?" It may be shown, for example, that a drug can cause weight loss by stimulating a particular receptor, or biological system, and yet this receptor or system may not have a significant role in the human's normal or regular process for regulating food intake. This finding demonstrates that particular questions can be divorced from a study of "normal" processes; one can simply ask "What happens if A is added to B?"

Empiricism is not limited to any particular type of investigation, rather it is a process of investigation that may allow the scientist to establish something as an "objective fact," or at least as a model of how things can work, by verification of a resultant model.

Reference to David Hume might help the scientist to list the sort of questions one might ask, since he enumerated different types of philosophical relations.¹ Hume

¹A *Treatise on Human Nature*, David Hume, Barnes & Noble New York (2005). Originally published in 1739. ISBN 0-7607-7172-3. Page 57; "Of Knowledge and Probability."

came up with four relations that he thought could be objects of knowledge and certainty: “resemblance, contrariety, degrees in quality, and proportions in quantity or number.” Interestingly, he left out “causation,” perhaps recognizing some of the difficulties in establishing causation to the level of “certainty.” However, we must add “causation” to our list of questions that the biologist engages in, because it is what is usually meant when one asks for a “mechanism” in relation to a perceived effect. To this list, one might also add “relations of condition” when one asks whether something affects something else (e.g., whether gene *X* complements for gene *Y* or is dominant in the setting of gene *Y* activity), although this could come under the term “relations of time and place,” which Hume additionally lists as a type of philosophical relation. The reader might wonder about the difference between “causation” and “relations of condition.” Causation refers to a particular type of relation, as when A perturbs the environment to result in B, or when A directly acts on B to change it in a particular way—as when drinking alcohol is found to cause impaired judgment. “Relations of condition” might refer to what can happen under a particular circumstance, as demonstrated that when Jimmy goes to the bar with his buddies, he drinks a lot more than when he goes to the bar with his wife.

One might also add a new term that could be called “solvency,” as when a particular problem needs to be solved, and there is an effort to see what sort of thing might accomplish this. It may be shown that a particular substance or strategy can be used to solve a problem, even if the substance is not accessed by any of the other listed relations.

PROBLEM SOLVING VERSUS UNDERSTANDING THE PROBLEM'S CAUSE

Consider the scientist who is seeking to cure a disease. It may be of interest to ask whether, in that setting, the effort to understand the mechanism underlying the disease may be the only way, or even the best way, to go about finding a cure. As a metaphor, think of a man who is crossing a busy street, where cars are whizzing by at very high speeds. As he crosses, the man sees a sports car heading toward him at 150 km per hour. To get safely across the road, should the man stop and work out why this car is bearing down on him at such a pace or should he take immediate action to get out of the way? This allegory might help to reinforce the need to remember what question is being asked. If the desire of the scientist is to understand how things work, then it is appropriate to ask that question. However, if the goal is to ask how things might work under a particular circumstance, or to understand how to stop something from working in a particular way, avoid the effects of a particular thing, or inhibit or overcome a particular occurrence, then it should be realized that these issues require that different types of questions be asked, not straightforward questions of causality, and they need to be addressed in a distinct manner. Because the goal of understanding causality is so entrenched in scientific culture, one often finds scientists who have lost sight of their experimental framework and who ask questions about mechanisms

of particular components in their system that are not responsive to their framework question.

The reader might object that understanding how a cancer cell works is critical to preventing its spread. But one might tell that to the surgeon, who still has the best track record for curing solid tumors, and whose methodology requires only a determination of what constitutes the boundaries of the tumor, so that it can be removed. This discussion should not be used as an excuse to remain ignorant about mechanism; the point is to ensure that there is a clear question framing the experimental project, so that the correct experimental approach can be established.

ACCEPTING NONUNIVERSAL TRUTHS

Perhaps one of the biggest objections to empiricism is that any set of experimental observations one might accumulate is limited and cannot be shown to encompass multiple aspects of the question. Therefore, it is not rational to say that one can answer the question completely or represent how the system operates in a comprehensive fashion. Nor is it possible to say with certainty that the system will continue to operate in the same way in the future. In addition, if one does not have a way of measuring a particular aspect of the subject, that aspect cannot be assessed. One way this problem has been represented in the past was to compare empiricism to a group of blind people examining an elephant. Each individual has a very different "picture" of the elephant, given the limited component of the animal they can access. The result is that nobody has an accurate model of the subject.

Perhaps we should consider the elephant metaphor a little more closely. If a subject is really elephantine in character, different experiments may be found to be nonreproducible, because each time the subject is accessed, a different portion is seen, and the various portions will not be representative of one another. The scientist may thereby be alerted to the fact that the full scope of the problem is not being dealt with. Alternatively, if the experiment is designed in a rigorous way, the same portion of the elephant will be accessed again and again, and one might thereby get an accurate description of that portion and a correct understanding as to how that part of the elephant operates. There is nothing wrong with limiting a study to an elephant's trunk, for example, as long as the model incorporates the limitation appropriately. The objection that one will not get a "comprehensive" picture assumes that comprehensiveness is necessary, but one can make progress by analyzing subsets of a problem. It is not necessary to see everything, just to describe accurately that which has been seen.

As noted earlier, the most damning criticism of empiricism and "rationalism" is the objection that the future case might be different. But this objection is resolved by establishing the predictive power of the model. Therefore, in response to the person who says that empiricism is flawed, one need not have some fanciful philosophical reply. One need only tell that person to release a ball and say in advance that the ball will fall to the ground. When the objector does this and finds that the ball does indeed

fall to the ground, and yet objects that this does not further justify either empiricism or inductive reasoning, the response is to say "drop the ball again; it will fall to the ground." At some point, the repetitive answers may cause the critic to get a headache. When this happens, one can say "Take an aspirin; it will relieve the headache." When the critic does this and finds that the aspirin does indeed relieve the headache, one can again wonder at the criticism of empiricism and inductive reasoning.

We therefore see that the justification for empirical study, and therefore experimental biology, is that it can result in models of the way systems work, as demonstrated by the models' capacity to show how systems *will* work. To assess the value of such an ability, simply take a quick glance around and see what has arisen from an understanding of how things work and how they will continue to work. From this examination comes a validation of the empirical enterprise.



A Short Synopsis

AFTER HAVING READ THIS WHOLE BOOK, scientists will still have to go into the laboratory and get their projects done. Is it possible to reduce this entire volume to a quick checklist, or bit of advice, so that each time scientists map out an experiment, they will be able to refer to that list to help them assess the experimental design? Let's have a go, with the caveat that it will not make sense unless the rest of the book has already been understood, and one can reference the concepts underlying the various points. This is not as short a list as one might like, but hopefully it will be helpful.

The following lists things to consider when designing an experimental project and subsequent experiments:

1. *Framework.* What is the question that you want to answer?
2. *Inductive space.* Decide what aspects of prior knowledge relate to your question. Read the literature.
3. *System.* What tools will you use to answer your question?
4. *System controls.* How do you know your system works? How do you know your system can provide the type of data you require? Is the chosen system well matched to your question, or might a different system be better?
5. *Experiment.* What are you going to do to answer the question? Be sure that measurements are taken multiple times and that you are measuring the effect in a representative fashion. Study the effect over time and over a range of experimental conditions. Do a dose-response with any experimental agent. Try to determine the "representative case" for the subject if that is relevant to the question. Consult a statistician and discuss the mode of analysis for your data and how many data points are required.
6. *Establish criteria* for the effect in advance of the experiment.
7. *Negative controls.* What negative controls are required? Is there an "all but X" control available?

8. *System-positive controls.* How do you know that the system is still operational? What positive controls are necessary to prove that the thing you want to measure was actually measurable within the context of the experiment?
9. *Effect-positive controls.* How do you know that the effect you want to measure can be produced in your system? What positive controls are necessary to produce the effect?
10. *Assumption controls.* If X is being measured, can something else be measured that has been shown to occur when X occurs? If you think Y has happened as a result of X , is there something else you can measure that has been shown to also happen when Y happens?
11. *Do the experiment.*
12. *Experimentalist controls.* Analyze the data in a blinded fashion.
13. *Repetition.* Repeat the experiment using the same criteria and methodology.
14. *Model building.* What is the answer to the question?
15. *Model check.* Is the answer responsive to the question?
16. *Prediction.* Does the model predict what will happen again? Repeat the experiment.
17. *Extension.* Does the model hold in different circumstances?
18. *Change the system.* Approach the question in another way.
19. *Change the scientist.* See if others can reproduce the effect.
20. *Present the data.* What do others think of the result? What do others think of the interpretation of the result?
21. *Predict the future.* See if the model continues to represent what will happen under various circumstances.
22. *Amend the model* as instances are found in which it is not predictive. Limit your claims as you discover limits to your claims.
23. *If the system is reductionist* in nature, realize that and apply the model in a nonreductionist or less-reductionist setting.
24. *A model that is limited but verifiable is superior* to a model that is comprehensive but not predictive.
25. *If an idea cannot be subjected* to experimentation and verification, it is not relevant to your project.

The reader should now be better prepared to design an experiment. If this list seems intimidating, or too much to take into account in a single experiment, rest

rational?
measure

measure
produce

ired that
a result
also hap-

ogy.

e experi-

rk of the

happen

mit your

nonre-

rehe-

ot rele-

f this list
at, rest

assured that the relative importance of the issues discussed here will become manifest after the first experiment has been done. On examining the initial data set, it usually becomes clear that a particularly important control is missing, which would have illuminated the data or would have shown that the system is not working as expected. It is to be expected that mistakes will be made. Almost always, experimentalists initiate experiments before having worked out every aspect of the experimental design; this is acceptable as long as the scientist is open to redesigning the experiment and revalidating the system as more knowledge is gained. Once some progress has been made, it will become clear at what point a new model, which did not exist before experimental data were gathered, can be formulated and, after the appropriate repeat experiments are done, validated.

When you have built your first model and found that it has predictive power, you will have successfully entered the ongoing process of empiricism and discovery. This is a club that all scientists can try to enter. However, membership is limited to those who are willing to give the data the final word. Good luck with your experiments.

Index

- A**
- Akt
assumption controls in signaling study, 173
nerve growth factor effects, 124–125, 138–144
- Animal models
assumption controls in clinical relevance establishment, 177–178
subject controls in genetic model systems
clinical value of animal models, 166
generalizing findings, 165
independent variable control, 163–164
- Antibody detection of protein
negative control, 128
positive control, 146–148
- Assumption controls
animal models in clinical relevance establishment, 177–178
avoidance of inappropriate deductions, 172–173
cell studies in whole organism relevance establishment, 178–180
drug dosing, 171, 176–177
experimental question formulation, 171–172
importance, 171
in vitro molecular systems in cell relevance establishment, 180
meta assumption of experimental model, 180–181
reductionism controls, 175–176
representative sample determination, 174–175
tissue specificity example, 174–175
- B**
- Bacon Francis, 3
- Biological empiricism. *See also* Necessity; Sufficiency
nonuniversal truth acceptance, 197–198
overview, 191–192
problem solving versus problem's cause, 196–197
question types, 195–196
- Blinding
experimentalist controls, 187–188
“unperturbed by *X*” negative control, 130–131
- BRCA1*, knockout mouse experiments
negative control, 125–127
positive control, 145–146
- C**
- Caffeine
blood pressure experiment controls, 119–122, 131, 133–138, 171–172
cancer risk analysis, 10–12
- Causality, assigning, 192–194
- Cell culture studies
subject controls, 168–169
“unperturbed by *X*” negative control, 123–125
whole organism relevance establishment, 178–180
- Computers, objective evaluators, 189
- Conclusion
hypothesis grammatical structure similarity, 7
model building. *See* Model building
open-ended question grammatical structure dissimilarity, 26–27
- Controls
assumptions. *See* Assumption controls
experimentalist controls. *See* Experimentalist controls

Controls (*continued*)

- methodology. *See* Methodology controls
- negative. *See* Negative control
- positive. *See* Positive control
- reagents. *See* Reagent controls
- subjects. *See* Subject controls
- system validation and accumulation, 62

Critical Rationalism

- bias, 69–70
- inductive reasoning limitations, 16–17
- overview, 2, 7, 15
- problem/question framework comparison, 24–25, 186

D

Data analysis, example in experiment interpretation, 64–66

Descartes, René, 3, 71

Drug dosing, assumption controls, 171, 176–177

E

EcoRI, restriction site determination

- experimental design
- hypothesis application, 91–99
- methodology development, 85–86
- model building, 101–102
- necessity and sufficiency establishment, 99–101
- open-ended question formulation, 76–81
- prior knowledge utilization, 81–84
- restriction enzyme function, 75, 81–82
- system establishment, 86–91
- term defining, 84–85

Empiricism. *See* Biological empiricism

Epistemology, definition, 1

Estrogen receptor, subject controls in breast cancer studies, 160

Experimental repeat

- biological relevance of variability, 110–112
- categories, 104–106
- experimental design example, 105–116
- perturbation of experimental answer, 61
- representative result considerations, 112–113
- scope of single experiment for defining, 61–64
- statistical significance, 103–104
- time course comparison, 59–60

variability at single study point, 109–110

Experimentalist controls

- blinded analysis, 187–188
- computers as objective evaluators, 189
- different evaluators, 188
- intersubjectivity, 183–184
- objectivity establishment, 184–185
- open-ended question as experimentalist control, 185–188
- outside repetition as arbiter for independence and intersubjectivity, 189–190

F

Future, prediction with models, 67–73, 200

H

Human Genome Project, hypothesis

independence, 18–20

Hume, David, 2, 16, 69, 195–196

Hypothesis

- binary distinction between positive and negative data types as filter, 10–14
- conclusion structure similarity, 7
- confirmation as reward, 14–15
- positive result measurement requirements, 8–10
- prior knowledge requirements, 19–20
- psychological importance of proving, 15
- scientific settings where not needed, 17–20
- testing in problem/question framework, 24

I

Inductive reasoning

- Critical Rationalism limitations, 16
- problem solving. *See* Problem/question framework

Insulin receptor, manipulation in mice,

164–165, 167–168

Intersubjectivity, 183–184, 189–190

K

Kant, Immanuel, 2, 16

M

Markers, selection, 114

The Matrix, thought experiment, 73

Methodology controls

caveats, 156

examples, 150–151, 155–156
 overview, 149–150

Model building
 background information necessity, 38–39
 conclusion transformation to unified
 model, 46
 descriptive questions, 39–41
 example, 101–102
 experimental question access of inductive
 space, 37–38, 43, 46–47
 open-ended question in origins, 35–37
 protein structure and function probing,
 43–46
 small inductive space contributions, 48–49
 unknown subject versus alien subject,
 41–42

Model validation, future prediction, 67–73

Molecular systems
 cell relevance establishment, 180
 subject controls, 170

mTOR, assumption controls in signaling
 study, 173

MuRF1, model building of function, 36–49

N

Nebulin, antibody detection
 negative control, 128
 positive control, 146–148

Necessity
 assessing of requirements, 192–194
 biological empiricism overview, 191–192
 settings where required, 194–195

Negative control
 antibody detection of protein, 128
 intersystem negative control, 129–130
 intrasystem negative control, 129
 “Is YX ?” question, 129
 “not- X ” case, 127–128
 “unperturbed by X ” negative control
 blinded analysis presented as, 130–131
 genetic experiment, 125–127
 questions with multiple variables,
 119–122
 tissue culture experiment, 123–124
 unperturbed case, 117–119

Nerve growth factor (NGF)
 negative controls in tissue culture experi-
 ment, 123–125
 positive controls

biochemistry experiment, 140–144
 tissue culture experiment, 138–140
 receptors, 138–139, 141

NGF. *See* Nerve growth factor

Nonuniversal truths, 197–198

Nozick, Robert, 183

Nuclear factor- κ B, hypothesis example, 9–10

O

Objectivity. *See* Experimentalist controls

Open-ended question. *See* Problem/question
 framework

P

Perception, model validation considerations,
 71–72

Popper, Karl, 2–3, 6, 16, 63, 186

Positive control
 antibody detection of protein, 146–148
 biochemistry experiment, 141–144
 caffeine effects on blood pressure experi-
 mental design, 133–134–138
 genetic experiment, 145–146
 importance, 148
 system validation, 137
 testing multiple aspects of system, 137–138
 tissue culture experiment, 138–140

Problem solving, problem's cause, 196–197

Problem/question framework
 acceptable answer features, 33–34
 answer bias prevention, 30
 close-ended question formulation, 30
 Critical Rationalism comparison, 24–25
 discrete questions, 31–32
 hypothesis testing, 24
 model building. *See* Model building
 open-ended question
 conclusion structure comparison versus
 binary question, 26–27
 derivation, 25–26
 experimentalist controls, 185–188
 result requirement independence, 27–28
 scope of problem defining, 29
 subset independence, 29
 success of outcomes, 30
 technology development, 30
 term defining in experimental design,
 57–58
 validation of sequential queries, 21–23

Progesterone receptor, subject controls in breast cancer studies, 160

Q

Question framework. *See* Problem/question framework

R

Randomization, study subjects, 161–162

Reagent controls

caveats, 156

example, 152–154

overview, 149–150

Reductionism controls

assumption controls, 175–176

subject controls

cell studies, 168–169

in vitro molecular systems, 170

Repetition. *See* Experimental repeat

Representative conditions, experimental design, 58–59

Restriction enzyme. *See* EcoRI

Russell, Bertrand, 2, 16

S

Short interfering RNA (siRNA), methodology controls in experimental design, 151

Single nucleotide polymorphisms (SNPs), genetic screens of human population, 167

siRNA. *See* Short interfering RNA

SNPs. *See* Single nucleotide polymorphisms

Statistics

experimental design training in graduate school, 1

repeats required for statistical significance, 103–104

time course versus experimental repeat, 60

Subject controls

controlling away of subjects effects, 168

genetic model systems

clinical value of animal models, 166

generalizing findings, 165

independent variable controls, 163–164

genetic screens of human population, 167

genetically independent variable finding, 167–168

matching subjects between control groups, 162–163

randomizing study subjects, 161–162

reductionism controls

cell studies, 168–169

in vitro molecular systems, 170

responsive subject finding, 159–160

subject group independent variables, 166–167

subject selection factors, 158–159

Sufficiency

assessing of requirements, 192–194

biological empiricism overview, 191–192

settings where required, 194–195

Synopsis, experimental design, 199–201

T

Term defining

example, 84–85

experimental design, 57–58

Time course

experimental repeat comparison, 59–60

perturbation of experimental answer, 61

Tissue culture. *See* Cell culture studies

V

Variability

accounting for at single study point, 109–110

biological relevance, 110–112

Venn diagram

discrete questions in problem/question framework, 31–32

experimental program illustration, 5

Venter, J. Craig, 18