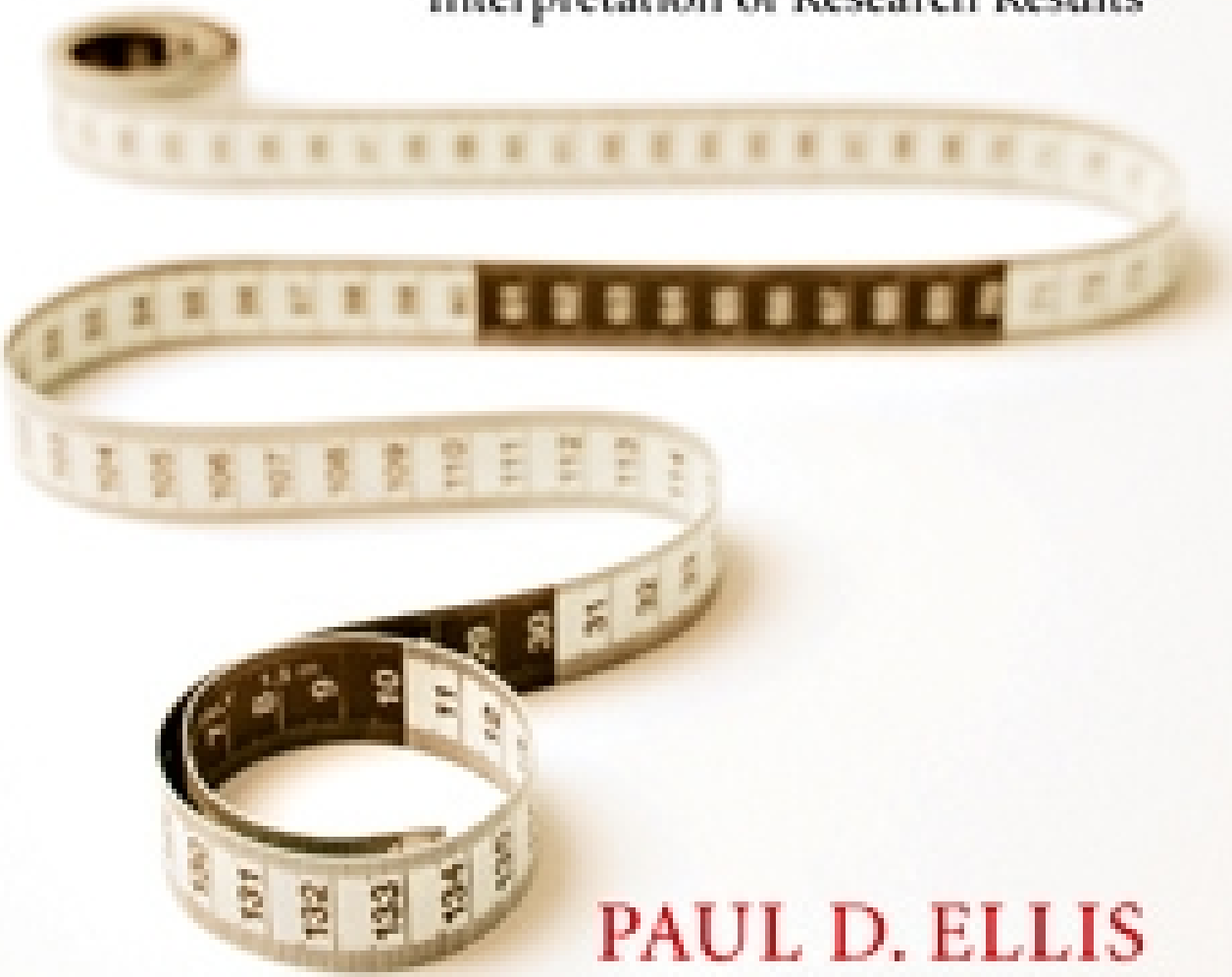
The Essential Guide to EFFECT SIZES

Statistical Power, Meta-Analysis, and the Interpretation of Research Results



PAUL D. ELLIS

CAMBRIDGE

CAMBRIDGE www.cambridge.org/9780521194235

The Essential Guide to Effect Sizes

This succinct and jargon-free introduction to effect sizes gives students and researchers the tools they need to interpret the practical significance of their research results. Using a class-tested approach that includes numerous examples and step-by-step exercises, it introduces and explains three of the most important issues relating to the assessment of practical significance: the reporting and interpretation of effect sizes (Part I), the analysis of statistical power (Part II), and the meta-analytic pooling of effect size estimates drawn from different studies (Part III). The book concludes with a handy list of recommendations for those actively engaged in or currently preparing research projects.

PAUL D. ELLIS is a professor in the Department of Management and Marketing at Hong Kong Polytechnic University, where he has taught research methods for fifteen years. His research interests include trade and investment issues, marketing and economic development, international entrepreneurship, and economic geography. Professor Ellis has been ranked as one of the world's most prolific scholars in the field of international business.

The Essential Guide to Effect Sizes

Statistical Power, Meta-Analysis, and the Interpretation of Research Results

Paul D. Ellis

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521142465

First published 2010

Printed in the United Kingdom at the University Press, Cambridge

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

Library of Congress Cataloguing in Publication data
Ellis, Paul D., 1969–
The essential guide to effect sizes : statistical power, meta-analysis, and the
interpretation of research results / Paul D. Ellis.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-521-19423-5 (hardback)
1. Research – Statistical methods. 2. Sampling (Statistics) I. Title.
Q180.55.S7E45 2010
507.2 – dc22 2010007120

ISBN 978-0-521-19423-5 Hardback
ISBN 978-0-521-14246-5 Paperback

Contents

Figures

Tables

Boxes

Introduction

The primary purpose of research is to estimate the magnitude and direction of effects which exist "out there" in the real world. An effect may be the result of a treatment, a trial, a decision, a strategy, a catastrophe, a collision, an innovation, an invention, an intervention, an election, an evolution, a revolution, a mutiny, an incident, an insurgency, an invasion, an act of terrorism, an outbreak, an operation, a habit, a ritual, a riot, a program, a performance, a disaster, an accident, a mutation, an explosion, an implosion, or a fluke.

I am sometimes asked, what do researchers do? The short answer is that we estimate the size of effects. No matter what phenomenon we have chosen to study we essentially spend our careers thinking up new and better ways to estimate effect magnitudes. But although we are in the business of producing estimates, ultimately our objective is a better understanding of actual effects. And this is why it is essential that we interpret not only the statistical significance of our results but their practical, or real-world, significance as well. Statistical significance reflects the improbability of our findings, but practical significance is concerned with meaning. The question we should ask is, what do my results say about effects themselves?

Interpreting the practical significance of our results requires skills that are not normally taught in graduate-level Research Methods and Statistics courses. These skills include estimating the magnitude of observed effects, gauging the power of the statistical tests used to detect effects, and pooling effect size estimates drawn from different studies. I surveyed the indexes of thirty statistics and research methods textbooks with publication dates ranging from 2000 to 2009. The majority of these texts had no entries for "effect size" (87%), "practical significance" (90%), "statistical power" (53%), or variations on these terms. On the few occasions where material was included, it was either superficial (usually just one paragraph) or mathematical (e.g., graphs and equations). Conspicuous by their absence were plain English guidelines explaining how to interpret effect sizes, distinguish practical from statistical significance, gauge the power of published research, design studies with sufficient power to detect sought-after effects, boost statistical power, pool effect size estimates from related studies, and correct those estimates to compensate for study-specific features. This book is the

beginnings of an attempt to fill a considerable gap in the education of the social science researcher.

This book addresses three questions that researchers routinely ask:

1. How do I interpret the practical or everyday significance of my research results?
2. Does my study have sufficient power to find what I am seeking?
3. How do I draw conclusions from past studies reporting disparate results?

The first question is concerned with meaning and implies the reporting and interpretation of effect sizes. Within the social science disciplines there is a growing recognition of the need to report effect sizes along with the results of tests of statistical significance. As with other aspects of statistical reform, psychology leads the way with no less than twenty-three disciplinary journals now insisting that authors report effect sizes (Fidler et al. 2004). So far these editorial mandates have had only a minimal effect on practice. In a recent survey Osborne (2008b) found less than 17% of studies in educational psychology research reported effect sizes. In a survey of human resource development research, less than 6% of quantitative studies were found to interpret effect sizes (Callahan and Reio 2006). In their survey of eleven years' worth of research in the field of play therapy, Armstrong and Henson (2004) found only 5% of articles reported an effect size. It is likely that the numbers are even lower in other disciplines. I had a research assistant scan the style guides and Instructions for Contributors for forty business journals to see whether any called for effect size reporting or the analysis of the statistical power of significance tests. None did.[1]

The editorial push for effect size reporting is undeniably a good thing. If history is anything to go by, statistical reforms adopted in psychology will eventually spread to other social science disciplines.[2] This means that researchers will have to change the way they interpret their results. No longer will it be acceptable to infer meaning solely on the basis of *p* values. By giving greater attention to effect sizes we will reduce a potent source of bias, namely the availability bias or the underrepresentation of sound but statistically nonsignificant results. It is conceivable that some results will be judged to be important even if they happen to be outside the bounds of statistical significance. (An example is provided in Chapter 1.) The skills for gauging and interpreting effect sizes are covered in Part I of this book.

The second question is one that ought to be asked before any study begins but seldom is. Statistical power describes the probability that a study will detect an effect when there is a genuine effect to be detected. Surveys measuring the statistical power of published research routinely find that most studies lack the power to detect sought-after effects. This shortcoming is endemic to the social sciences where effect sizes tend to be small. In the management domain the proportion of studies sufficiently

[1] However, the *Journal of Consumer Research* website had a link to an editorial which did call for the estimation of effect sizes (see Iacobucci 2005).

[2] The nonpsychologist may be surprised at the impact psychology has had on statistical practices within the social sciences. But as Scarr (1997: 16) notes, "psychology's greatest contribution is methodology." Methodology, as Scarr defines the term, means measurement and statistical rules that "define a realm of discourse about what is 'true'."

empowered to detect small effects has been found to vary between 6% and 9% (Mazen et al. 1987a; Mone et al. 1996). The corresponding figures for research in international business are 4–10% (Brock 2003); for research in accounting, 0–1% (Borkowski et al. 2001; Lindsay 1993); for psychology, 0–2% (Cohen 1962; Rossi 1990; Sedlmeier and Gigerenzer 1989); for communication research, 0–8% (Katzer and Sodt 1973; Chase and Tucker 1975); for counseling research, 0% (Kosciulek and Szymanski 1993); for education research, 4–9% (Christensen and Christensen 1977; Daly and Hexamer 1983); for social work research, 11% (Orme and Combs-Orme 1986); for management information systems research, less than 2% (Baroudi and Orlikowski 1989); and for accounting information systems research, 0% (McSwain 2004). These low numbers lead to different consequences for researchers and journal editors.

For the researcher insufficient power means an increased risk of missing real effects (a Type II error). An underpowered study is a study designed to fail. No matter how well the study is executed, resources will be wasted searching for an effect that cannot easily be found. Statistical significance will be difficult to attain and the odds are good that the researcher will wrongly conclude that there is nothing to be found and so misdirect further research on the topic. Underpowered studies thus cast a shadow of consequence that may hinder progress in an area for years.

For the journal editor low statistical power paradoxically translates to an increased risk of publishing false positives (a Type I error). This happens because publication policies tend to favor studies reporting statistically significant results. For any set of studies reporting effects, there will be a small proportion affected by Type I error. Under ideal levels of statistical power, this proportion will be about one in sixteen. (These numbers are explained in Chapter 4.) But as average power levels fall, the proportion of false positives being reported and published inevitably rises. This happens even when alpha standards for individual studies are rigorously maintained at conventional levels. For this reason some suspect that published results are more often wrong than right (Hunter 1997; Ioannidis 2005).

Awareness of the dangers associated with low statistical power is slowly increasing. A taskforce commissioned by the American Psychological Association recommended that investigators assess the power of their studies prior to data collection (Wilkinson and the Taskforce on Statistical Inference 1999). Now it is not unusual for funding agencies and university grants committees to ask applicants to submit the results of prospective power analyses together with their research proposals. Some journals also require contributors to quantify the possibility that their results are affected by Type II errors, which implies an assessment of their study's statistical power (e.g., Campion 1993). Despite these initiatives, surveys reveal that most investigators remain ignorant of power issues. The proportion of studies that merely mention power has been found to be in the 0–4% range for disciplines from economics and accounting to education and psychology (Baroudi and Orlikowski 1989; Fidler et al. 2004; Lindsay 1993; McCloskey and Ziliak 1996; Osborne 2008b; Sedlmeier and Gigerenzer 1989).

Conscious of the risk of publishing false positives it is likely that a growing number of journal editors will require authors to quantify the statistical power of their studies.

However, the available evidence suggests editorial mandates alone will be insufficient to initiate change (Fidler et al. 2004). Also needed are practical, plain English guidelines. When most of the available texts on power analysis are jam-packed with Greek and complicated algebra it is no wonder that the average researcher still picks sample sizes on the basis of flawed rules of thumb. Analyzing the power inherent within a proposed study is like buying error insurance. It can help ensure that your project will do what you intend it to do. Power analysis is addressed in Part II of this book.

The third question is one which nearly every doctoral student asks and which many professors give up trying to answer! Literature reviews provide the stock foundation for many of our research projects. We review the literature on a topic, see there is no consensus, and use this as a justification for doing yet another study. We then reach our own little conclusion and this gets added to the pile of conclusions that will then be reviewed by whoever comes after us. It's not ideal, but we tell ourselves that this is how knowledge is advanced. However, a better approach is to side-step all the little conclusions and focus instead on the actual effect size estimates that have been reported in previous studies. This pooling of independent effect size estimates is called meta-analysis. Done well, a meta-analysis can provide a precise conclusion regarding the direction and magnitude of an effect even when the underlying data come from dissimilar studies reporting conflicting conclusions. Meta-analysis can also be used to test hypotheses that are too big to be tested at the level of an individual study. Meta-analysis thus serves two important purposes: it provides an accurate distillation of extant knowledge and it signals promising directions for further theoretical development. Not everyone will want to run a meta-analysis, but learning to think meta-analytically is an essential skill for any researcher engaged in replication research or who is simply trying to draw conclusions from past work. The basic principles of meta-analysis are covered in Part III of this book.

The three topics covered in this book loosely describe how scientific knowledge accumulates. Researchers conduct individual studies to generate effect size estimates which will be variable in quality and affected by study-specific artifacts. Meta-analysts will adjust then pool these estimates to generate weighted means which will reflect population effect sizes more accurately than the individual study estimates. Meanwhile power analysts will calculate the statistical power of published studies to gauge the probability that genuine effects were missed. These three activities are co-dependent, like legs on a stool. A well-designed study is normally based on a prospective analysis of statistical power; a good power analysis will ideally be based on a meta-analytically derived mean effect size; and meta-analysis would have nothing to cumulate if there were no individual studies producing effect size estimates. Given these interdependencies it makes sense to discuss these topics together. A working knowledge of how each part relates to the others is essential to good research.

The value of this book lies in drawing together lessons and ideas which are buried in dense texts, encrypted in oblique language, and scattered across diverse disciplines. I have approached this material not as a philosopher of science but as a practicing researcher in need of straightforward answers to practical questions. Having waded

through hundreds of equations and thousands of pages it occurs to me that many of these books were written to impress rather than instruct. In contrast, this book was written to provide answers to how-to questions that can be easily understood by the scholar of average statistical ability. I have deliberately tried to write as short a book as possible and I have kept the use of equations and Greek symbols to a bare minimum. However, for the reader who wishes to dig deeper into the underlying statistical and philosophical issues, I have provided technical and explanatory notes at the end of each chapter. These notes, along with the appendices at the back of the book, will also be of help to doctoral students and teachers of graduate-level methods courses.

Speaking of students, the material in this book has been tested in the classroom. For the past fifteen years I have had the privilege of teaching research methods to smart graduate students. If the examples and exercises in this book are any good it is because my students patiently allowed me to practice on them. I am grateful. I am also indebted to colleagues who provided advice or comments on earlier drafts of this book, including Geoff Cumming, J.J. Hsieh, Huang Xu, Trevor Moores, Herman Aguinis, Godfrey Yeung, Tim Clark, Zhan Ge, and James Wilson. At Cambridge University Press I would like to thank Paula Parish, Jodie Barnes, Phil Good and Viv Church.

Paul D. Ellis
Hong Kong, March 2010

Part I

Effect sizes and the interpretation of results

1 Introduction to effect sizes

The primary product of a research inquiry is one or more measures of effect size, not *p* values.
∼ Jacob Cohen (1990: 1310)

The dreaded question

"So what?"
It was the question every scholar dreads. In this case it came at the end of a PhD proposal presentation. The student had done a decent job outlining his planned project and the early questions from the panel had established his familiarity with the literature. Then one old professor asked the dreaded question.

"So what? Why do this study? What does it mean for the man on the street? You are asking for a three-year holiday from the real world to conduct an academic study. Why should the taxpayer fund this?"

The student was clearly unprepared for these sorts of questions. He referred to the gap in the literature and the need for more research, but the old professor wasn't satisfied. An awkward moment of silence followed. The student shuffled his notes to buy another moment of time. In desperation he speculated about some likely implications for practitioners and policy-makers. It was not a good answer but the old professor backed off. The point had been made. While the student had outlined his methodology and data analysis plan, he had given no thought to the practical significance of his study. The panel approved his proposal with one condition. If he wanted to pass his exam in three years' time he would need to come up with a good answer to the "so what?" question.

Practical versus statistical significance

In most research methods courses students are taught how to test a hypothesis and how to assess the statistical significance of their results. But they are rarely taught how to interpret their results in ways that are meaningful to nonstatisticians. Test results are judged to be significant if certain statistical standards are met. But significance in this context differs from the meaning of significance in everyday language. A

statistically significant result is one that is unlikely to be the result of chance. But a practically significant result is meaningful in the real world. It is quite possible, and unfortunately quite common, for a result to be statistically significant and trivial. It is also possible for a result to be statistically nonsignificant and important. Yet scholars, from PhD candidates to old professors, rarely distinguish between the statistical and the practical significance of their results. Or worse, results that are found to be statistically significant are interpreted as if they were practically meaningful. This happens when a researcher interprets a statistically significant result as being "significant" or "highly significant."[1]

The difference between practical and statistical significance is illustrated in a story told by Kirk (1996). The story is about a researcher who believes that a certain medication will raise the intelligence quotient (IQ) of people suffering from Alzheimer's disease. She administers the medication to a group of six patients and a placebo to a control group of equal size. After some time she tests both groups and then compares their IQ scores using a *t* test. She observes that the average IQ score of the treatment group is 13 points higher than the control group. This result seems in line with her hypothesis. However, her *t* statistic is not statistically significant ($t = 1.61$, $p = .14$), leading her to conclude that there is no support for her hypothesis. But a nonsignificant *t* test does not mean that there is no difference between the two groups. More information is needed. Intuitively, a 13-point difference seems to be a substantive difference; the medication seems to be working. What the *t* test tells us is that we cannot rule out chance as a possible explanation for the difference. Are the results *real*? Possibly, but we cannot say for sure. Does the medication have promise? Almost certainly. Our interpretation of the result depends on our definition of significance. A 13-point gain in IQ seems large enough to warrant further investigation, to conduct a bigger trial. But if we were to make judgments solely on the basis of statistical significance, our conclusion would be that the drug was ineffective and that the observed effect was just a fluke arising from the way the patients were allocated to the groups.

The concept of effect size

Researchers in the social sciences have two audiences: their peers and a much larger group of nonspecialists. Nonspecialists include managers, consultants, educators, social workers, trainers, counselors, politicians, lobbyists, taxpayers and other members of society. With this second group in mind, journal editors, reviewers, and academy presidents are increasingly asking authors to evaluate the practical significance of their results (e.g., Campbell 1982; Cummings 2007; Hambrick 1994; JEP 2003; Kendall 1997; La Greca 2005; Levant 1992; Lustig and Strauser 2004; Shaver 2006, 2008; Thompson 2002a; Wilkinson and the Taskforce on Statistical Inference 1999).[2] This implies an estimation of one or more *effect sizes*. An effect can be the result of a treatment revealed in a comparison between groups (e.g., treated and untreated groups) or it can describe the degree of association between two related variables (e.g., treatment dosage and health). An effect size refers to the magnitude of the result as it occurs, or

would be found, in the population. Although effects can be observed in the artificial setting of a laboratory or sample, effect sizes exist in the real world.

The estimation of effect sizes is essential to the interpretation of a study's results. In the fifth edition of its *Publication Manual*, the American Psychological Association (APA) identifies the "failure to report effect sizes" as one of seven common defects editors observed in submitted manuscripts. To help readers understand the importance of a study's findings, authors are advised that "it is almost always necessary to include some index of effect" (APA 2001: 25). Similarly, in its Standards for Reporting, the American Educational Research Association (AERA) recommends that the reporting of statistical results should be accompanied by an effect size and "a qualitative interpretation of the effect" (AERA 2006: 10).

The best way to measure an effect is to conduct a census of an entire population but this is seldom feasible in practice. Census-based research may not even be desirable if researchers can identify samples that are representative of broader populations and then use inferential statistics to determine whether sample-based observations reflect population-level parameters. In the Alzheimer's example, twelve patients were chosen to represent the population of all Alzheimer's patients. By examining carefully chosen samples, researchers can estimate the magnitude and direction of effects which exist in populations. These estimates are more or less precise depending on the procedures used to make them. Two questions arise from this process; how big is the effect and how precise is the estimate? In a typical statistics or methods course students are taught how to answer the second question. That is, they learn how to gauge the precision (or the degree of error) with which sample-based estimates are made. But the proverbial man on the street is more interested in the first question. What he wants to know is, how big is it? Or, how well does it work? Or, what are the odds?

Suppose you were related to one of the Alzheimer's patients receiving the medication and at the end of the treatment period you noticed a marked improvement in their mental health. You would probably conclude that the treatment had been successful. You would be astonished if the researcher then told you the treatment had not led to any significant improvement. But she and you are looking at two different things. You have observed an effect ("the treatment seems to work") while the researcher is commenting about the precision of a sample-based estimate ("the study result may be attributable to chance"). It is possible that both of you are correct – the results are practically meaningful yet statistically nonsignificant. Practical significance is inferred from the size of the effect while statistical significance is inferred from the precision of the estimate. As we will see in Chapter 3, the statistical significance of any result is affected by both the size of the effect and the size of the sample used to estimate it. The smaller the sample, the less likely a result will be statistically significant regardless of the effect size. Consequently, we can draw no conclusions about the practical significance of a result from tests of statistical significance.

The concept of effect size is the common link running through this book. Questions about practical significance, desired sample sizes, and the interpretation of results obtained from different studies can be answered only with reference to some population

effect size. But what does an effect size look like? Effect sizes are all around us. Consider the following claims which you might find advertised in your daily newspaper: "Enjoy immediate pain relief through acupuncture"; "Change service providers now and save 30%"; "Look 10 years younger with Botox". These claims are all promising measurable results or effects. (Whether they are true or not is a separate question!) Note how both the effects – pain relief, financial savings, wrinkle reduction – and their magnitudes – immediate, 30%, 10 years younger – are expressed in terms that mean something to the average newspaper reader. No understanding of statistical significance is necessary to gauge the merits of each claim. Each effect is being promoted as if it were intrinsically meaningful. (Whether it is or not is up to the newspaper reader to decide.)

Many of our daily decisions are based on some analysis of effect size. We sign up for courses that we believe will enhance our career prospects. We buy homes in neighborhoods where we expect the market will appreciate or which provide access to amenities that make life better. We endure vaccinations and medical tests in the hope of avoiding disease. We cut back on carbohydrates to lose weight. We quit smoking and start running because we want to live longer and better. We recycle and take the bus to work because we want to save the planet.

Any adult human being has had years of experience estimating and interpreting effects of different types and sizes. These two skills – estimation and interpretation – are essential to normal life. And while it is true that a trained researcher should be able to make more precise estimates of effect size, there is no reason to assume that researchers are any better at interpreting the practical or everyday significance of effect sizes. The interpretation of effect magnitudes is a skill fundamental to the human condition. This suggests that the scientist has a two-fold responsibility to society: (1) to conduct rigorous research leading to the reporting of precise effect size estimates in language that facilitates interpretation by others (discussed in this chapter) and (2) to interpret the practical significance or meaning of research results (discussed in the next chapter).

Two families of effects

Effect sizes come in many shapes and sizes. By one reckoning there are more than seventy varieties of effect size (Kirk 2003). Some have familiar-sounding labels such as odds ratios and relative risk, while others have exotic names like Kendall's tau and Goodman–Kruskal's lambda.[3] In everyday use effect magnitudes are expressed in terms of some quantifiable change, such as a change in percentage, a change in the odds, a change in temperature and so forth. The effectiveness of a new traffic light might be measured in terms of the change in the number of accidents. The effectiveness of a new policy might be assessed in terms of the change in the electorate's support for the government. The effectiveness of a new coach might be rated in terms of the team's change in ranking (which is why you should never take a coaching job at a team that just won the championship!). Although these sorts of one-off effects are the stuff of life, scientists are more often interested in making comparisons or in measuring

relationships. Consequently we can group most effect sizes into one of two "families" of effects: differences between groups (also known as the *d* family) and measures of association (also known as the *r* family).

The d family: assessing the differences between groups

Groups can be compared on dichotomous or continuous variables. When we compare groups on dichotomous variables (e.g., success versus failure, treated versus untreated, agreements versus disagreements), comparisons may be based on the probabilities of group members being classified into one of the two categories. Consider a medical experiment that showed that the probability of recovery was p in a treatment group and q in a control group. There are at least three ways to compare these groups:

(i) Consider the difference between the two probabilities ($p - q$).
(ii) Calculate the risk ratio or relative risk (p/q).
(iii) Calculate the odds ratio $(p/(1-p))/(q/(1-q))$.

The **difference between the two probabilities** (or proportions), a.k.a. the **risk difference**, is the easiest way to quantify a dichotomous outcome of whatever treatment or characteristic distinguishes one group from another. But despite its simplicity, there are a number of technical issues that confound interpretation (Fleiss 1994), and it is little used.[4]

The **risk ratio** and the **odds ratio** are closely related but generate different numbers. Both indexes compare the likelihood of an event or outcome occurring in one group in comparison with another, but the former defines likelihood in terms of probabilities while the latter uses odds. Consider the example where students have a choice of enrolling in classes taught by two different teachers:

1. Aristotle is a brilliant but tough teacher who routinely fails 80% of his students.
2. Socrates is considered a "soft touch" who fails only 50% of his students.

Students may prefer Socrates to Aristotle as there is a better chance of passing, but how big is this difference? In short, how big is the Socrates Effect in terms of passing? Alternatively, how big is the Aristotle Effect in terms of failing? Both effects can be quantified using the odds or the risk ratios.

To calculate an odds ratio associated with a particular outcome we would compare the odds of that outcome for each class. An odds ratio of one means that there is no difference between the two groups being compared. In other words, group membership has no effect on the outcome of interest. A ratio less than one means the outcome is less likely in the first group, while a ratio greater than one means it is less likely in the second group. In this case the odds of failing in Aristotle's class are .80 to .20 (or four to one, represented as 4:1), while in Socrates' class the odds of failing are .50 to .50 (or one to one, represented as 1:1). As the odds of failing in Aristotle's class are four times higher than in Socrates' class, the odds ratio is four (4:1/1:1).[5]

To calculate the risk ratio, also known to epidemiologists as **relative risk**, we could compare the probability of failing in both classes. The relative risk of failing in Aristotle's class compared with Socrates' class is .80/.50 or 1.6. Alternatively, the relative risk of failing in Socrates' class is .50/.80 or .62 compared with Aristotle's class. A risk ratio of one would mean there was equal risk of failing in both classes.[6]

In this example, both the odds ratio and the risk ratio show that students are in greater danger of failing in Aristotle's class than in Socrates', but the odds ratio gives a higher score (4) than the risk ratio (1.6). Which number is better? Usually the risk ratio will be preferred as it is easily interpretable and more consistent with the way people think. Also, the odds ratio tends to blow small differences out of all proportion. For example, if Aristotle has ten students and he fails nine instead of the usual eight, the odds ratio for comparing the failure rates of the two classes jumps from four (4:1/1:1) to nine (9:1/1:1). The odds ratio has more than doubled even though the number of failing students has increased only marginally. One way to compensate for this is to report the logarithm of the odds ratio instead. Another example of the difference between the odds and risk ratios is provided in Box 1.1.[7]

Box 1.1 A Titanic confusion about odds ratios and relative risk*

In James Cameron's successful 1997 film *Titanic*, the last hours of the doomed ship are punctuated by acts of class warfare. While first-class passengers are bundled into lifeboats, poor third-class passengers are kept locked below decks. Rich passengers are seen bribing their way to safety while poor passengers are beaten and shot by the ship's officers. This interpretation has been labeled by some as "good Hollywood, but bad history" (Phillips 2007). But Cameron justified his neo-Marxist interpretation of the Titanic's final hours by looking at the numbers of survivors in each class. Probably the best data on Titanic survival rates come from the report prepared by Lord Mersey in 1912 and reproduced by Anesi (1997). According to the Mersey Report there were 2,224 people on the Titanic's maiden voyage, of which 1,513 died. The relevant numbers for first- and third-class passengers are as follows:

	Survived	Died	*Total*
First-class passengers	203	122	*325*
Third-class passengers	178	528	*706*

Clearly more third-class passengers died than first-class passengers. But how big was this class effect? The likelihood of dying can be evaluated using either an odds ratio or a risk ratio. The odds ratio compares the relative odds of dying for passengers in each group:

* The idea of using the survival rates of the Titanic to illustrate the difference between relative risk and odds ratios is adapted from Simon (2001).

- For third-class passengers the odds of dying were almost three to one in favor (528/178 = 2.97).
- For first-class passengers the odds of dying were much lower at one to two in favor (122/203 = 0.60).
- Therefore, the odds ratio is 4.95 (2.97/0.60).

The risk ratio or relative risk compares the probability of dying for passengers in each group:

- For third-class passengers the probability of death was .75 (528/706).
- For first-class passengers the probability of death was .38 (122/325).
- Therefore, the relative risk of death associated with traveling in third class was 1.97 (.75/.38).

In summary, if you happened to be a third-class passenger on the Titanic, the *odds* of dying were nearly five times greater than for first-class passengers, while the *relative risk* of death was nearly twice as high. These numbers seem to support Cameron's view that the lives of poor passengers were valued less than those of the rich.

However, there is another explanation for these numbers. The reason more third-class passengers died in relative terms is because so many of them were men (see table below). Men accounted for nearly two-thirds of third-class passengers but only a little over half of the first-class passengers. The odds of dying for third-class men were still higher than for first-class men, but the odds ratio was only 2.49 (not 4.95), while the relative risk of death was 1.25 (not 1.97). Frankly it didn't matter much which class you were in. If you were an adult male passenger on the Titanic, you were a goner! More than two-thirds of the first-class men died. This was the age of women and children first. A man in first class had less chance of survival than a child in third class. When gender is added to the analysis it is apparent that chivalry, not class warfare, provides the best explanation for the relatively high number of third-class deaths.

	Survived	Died	*Total*
First-class passengers			
– men	57	118	*175*
– women & children	146	4	*150*
Third-class passengers			
– men	75	387	*462*
– women & children	103	141	*244*

When we compare groups on continuous variables (e.g., age, height, IQ) the usual practice is to gauge the difference in the average or mean scores of each group. In the Alzheimer's example, the researcher found that the mean IQ score for the treated

group was 13 points higher than the mean score obtained for the untreated group. Is this a big difference? We can't say unless we also know something about the spread, or standard deviation, of the scores obtained from the patients. If the scores were widely spread, then a 13-point gap between the means would not be that unusual. But if the scores were narrowly spread, a 13-point difference could reflect a substantial difference between the groups.

To calculate the difference between two groups we subtract the mean of one group from the other ($M_1 - M_2$) and divide the result by the standard deviation (SD) of the population from which the groups were sampled. The only tricky part in this calculation is figuring out the population standard deviation. If this number is unknown, some approximate value must be used instead. When he originally developed this index, Cohen (1962) was not clear on how to solve this problem, but there are now at least three solutions. These solutions are referred to as Cohen's *d*, Glass's delta or Δ, and Hedges' *g*. As we can see from the following equations, the only difference between these metrics is the method used for calculating the standard deviation:

$$\text{Cohen's } d = \frac{M_1 - M_2}{SD_{pooled}}$$

$$\text{Glass's } \Delta = \frac{M_1 - M_2}{SD_{control}}$$

$$\text{Hedges' } g = \frac{M_1 - M_2}{SD^*_{pooled}}$$

Choosing among these three equations requires an examination of the standard deviations of each group. If they are roughly the same then it is reasonable to assume they are estimating a common population standard deviation. In this case we can pool the two standard deviations to calculate a **Cohen's *d*** index of effect size. The equation for calculating the pooled standard deviation (SD_{pooled}) for two groups can be found in the notes at the end of this chapter.[8]

If the standard deviations of the two groups differ, then the homogeneity of variance assumption is violated and pooling the standard deviations is not appropriate. In this case we could insert the standard deviation of the control group into our equation and calculate a **Glass's delta** (Glass et al. 1981: 29). The logic here is that the standard deviation of the control group is untainted by the effects of the treatment and will therefore more closely reflect the population standard deviation. The strength of this assumption is directly proportional to the size of the control group. The larger the control group, the more it is likely to resemble the population from which it was drawn.

Another approach, which is recommended if the groups are dissimilar in size, is to weight each group's standard deviation by its sample size. The pooling of weighted standard deviations is used in the calculation of **Hedges' *g*** (Hedges 1981: 110).[9]

These three indexes – Cohen's *d*, Glass's delta and Hedges' *g* – convey information about the size of an effect in terms of standard deviation units. A score of .50 means that

the difference between the two groups is equivalent to one-half of a standard deviation, while a score of 1.0 means the difference is equal to one standard deviation. The bigger the score, the bigger the effect. One advantage of reporting effect sizes in standardized terms is that the results are scale-free, meaning they can be compared across studies. If two studies independently report effects of size $d = .50$, then their effects are identical in size.

The r family: measuring the strength of a relationship

The second family of effect sizes covers various measures of association linking two or more variables. Many of these measures are variations on the correlation coefficient.

The **correlation coefficient** (r) quantifies the strength and direction of a relationship between two variables, say X and Y (Pearson 1905). The variables may be either dichotomous or continuous. Correlations can range from -1 (indicating a perfectly negative linear relationship) to 1 (indicating a perfectly positive linear relationship), while a correlation of 0 indicates that there is no relationship between the variables. The correlation coefficient is probably the best known measure of effect size, although many who use it may not be aware that it is an effect size index. Calculating the correlation coefficient is one of the first skills learned in an undergraduate statistics course. Like Cohen's *d*, the correlation coefficient is a standardized metric. Any effect reported in the form of *r* or one of its derivatives can be compared with any other. Some of the more common measures of association are as follows:

(i) The **Pearson product moment correlation coefficient** (r) is used when both X and Y are continuous (i.e., when both are measured on interval or ratio scales).
(ii) **Spearman's rank correlation** or **rho** (ρ or r_s) is used when both X and Y are measured on a ranked scale.
(iii) An alternative to Spearman's rho is **Kendall's tau** (τ), which measures the strength of association between two sets of ranked data.
(iv) The **point-biserial correlation coefficient** (r_{pb}) is used when X is dichotomous and Y is continuous.
(v) The **phi coefficient** (ϕ) is used when both X and Y are dichotomous, meaning both variables and both outcomes can be arranged on a 2×2 contingency table.[10]
(vi) **Pearson's contingency coefficient *C*** is an adjusted version of phi that is used for tests with more than one degree of freedom (i.e., tables bigger than 2×2).
(vii) **Cramér's V** can be used to measure the strength of association for contingency tables of any size and is generally considered superior to *C*.
(viii) **Goodman and Kruskal's lambda** (λ) is used when both X and Y are measured on nominal (or categorical) scales and measures the percentage improvement in predicting the value of the dependent variable given the value of the independent variable.

In some disciplines the strength of association between two variables is expressed in terms of the **proportion of shared variance**. Proportion of variance (POV) indexes are recognized by their square-designations. For example, the POV equivalent of the correlation r is r^2, which is known as the **coefficient of determination**. If X and Y have a correlation of $-.60$, then the coefficient of determination is .36 (or $-.60 \times -.60$). The POV implication is that 36% of the total variance is shared between the two variables. A slightly more interesting take is to claim that 36% of the variation in Y is accounted for, or explained, by the variation in X. POV indexes range from 0 (no shared variance) to 1 (completed shared variance).

When one variable is considered to be dependent on a set of predictor variables we can compute the **coefficient of multiple determination** (or R^2). This index is usually associated with multiple regression analysis. One limitation of this index is that it is inflated to some degree by variation caused by sampling error which, in turn, is related to the size of the sample and the number of predictors in the model. We can adjust for this extraneous variation by calculating the **adjusted coefficient of multiple determination** (or ${}_{\text{adj}}R^2$). Most software packages generate both R^2 and ${}_{\text{adj}}R^2$ indexes.[11]

Logistic regression is a special form of regression that is used when the dependent variable is dichotomous. The effect size index associated with logistic regression is the **logit coefficient** or the logged odds ratio. As logits are not inherently meaningful, the usual practice when assessing the contribution of individual predictors (the logit coefficients) is to transform the results into more intuitive metrics such as odds, odds ratios, probabilities, and the difference between probabilities (Pampel 2000).

R squareds are common in business journals and are the usual output of econometric analyses. In psychology journals a more common index is the **correlation ratio** or **eta^2** (η^2). Typically associated with one-way analysis of variance (ANOVA), eta^2 reflects the proportion of variation in the dependent variable which is accounted for by membership in the groups defined by the independent variable. As with R^2, eta^2 is an uncorrected or upwardly biased effect size index.[12] There are a number of alternative indexes which correct for this inflation, including **omega squared** (ω^2) and **epsilon squared** (ε^2) (Snyder and Lawson 1993).

Finally, **Cohen's f** and f^2 are used in connection with the F-tests associated with ANOVA and multiple regression (Cohen 1988). In the context of ANOVA Cohen's f is a bit like a bigger version of Cohen's d. While d is the standardized difference between two groups, f is used to measure the dispersion of means among three or more groups. In the context of hierarchical multiple regression involving two sets of predictors A and B, the f^2 index accounts for the incremental effect of adding set B to the basic model (Cohen 1988: 410ff).[13]

Calculating effect sizes

A comprehensive list of the major effect size indexes is provided in Table 1.1. Many of these indexes can be computed using popular statistics programs such as SPSS.

Table 1.1 *Common effect size indexes*

Measures of group differences (the *d* family)		Measures of association (the *r* family)	
(a) Groups compared on dichotomous outcomes		**(a) Correlation indexes**	
RD	The risk difference in probabilities: the difference between the probability of an event or outcome occurring in two groups	r	The Pearson product moment correlation coefficient: used when both variables are measured on an interval or ratio (metric) scale
RR	The risk or rate ratio or relative risk: compares the probability of an event or outcome occurring in one group with the probability of it occurring in another	ρ (or r_s)	Spearman's rho or the rank correlation coefficient: used when both variables are measured on an ordinal or ranked (non-metric) scale
OR	The odds ratio: compares the odds of an event or outcome occurring in one group with the odds of it occurring in another	τ	Kendall's tau: like rho, used when both variables are measured on an ordinal or ranked scale; tau-b is used for square-shaped tables; tau-c is used for rectangular tables
(b) Groups compared on continuous outcomes		r_{pb}	The point-biserial correlation coefficient: used when one variable (the predictor) is measured on a binary scale and the other variable is continuous
d	Cohen's d: the uncorrected standardized mean difference between two groups based on the pooled standard deviation	φ	The phi coefficient: used when variables and effects can be arranged in a 2×2 contingency table
Δ	Glass's delta (or d): the uncorrected standardized mean difference between two groups based on the standard deviation of the control group	C	Pearson's contingency coefficient: used when variables and effects can be arranged in a contingency table of any size
g	Hedges' g: the corrected standardized mean difference between two groups based on the pooled, weighted standard deviation	V	Cramér's V: like C, V is an adjusted version of phi that can be used for tables of any size
PS	Probability of superiority: the probability that a random value from one group will be greater than a random value drawn from another	λ	Goodman and Kruskal's lambda: used when both variables are measured on nominal (or categorical) scales

(*cont.*)

Table 1.1 (*cont.*)

Measures of group differences (the *d* family)	Measures of association (the *r* family)	
	(b) Proportion of variance indexes	
	r^2	The coefficient of determination: used in bivariate regression analysis
	R^2	R squared, or the (uncorrected) coefficient of multiple determination: commonly used in multiple regression analysis
	${}_{\text{adj}}R^2$	Adjusted R squared, or the coefficient of multiple determination adjusted for sample size and the number of predictor variables
	f	Cohen's *f*: quantifies the dispersion of means in three or more groups; commonly used in ANOVA
	f^2	Cohen's *f* squared: an alternative to R^2 in multiple regression analysis and ΔR^2 in hierarchical regression analysis
	η^2	Eta squared or the (uncorrected) correlation ratio: commonly used in ANOVA
	ε^2	Epsilon squared: an unbiased alternative to η^2
	ω^2	Omega squared: an unbiased alternative to η^2
	${R^2}_{\text{C}}$	The squared canonical correlation coefficient: used for canonical correlation analysis

In Table 1.2 the effect sizes associated with some of the more common analytical techniques are listed along with the relevant SPSS procedures for their computation. In addition, many free effect size calculators can be found online by googling the name of the desired index (e.g., "Cohen's *d* calculator" or "relative risk calculator"). One easy-to-use calculator has been developed by Ellis (2009). In this case calculating a Cohen's *d* requires nothing more than entering two group means and their corresponding standard deviations, then clicking "compute." The calculator also generates an *r* equivalent of the *d* effect. A number of other online calculators are listed in the notes found at the end of this chapter.[14]

Table 1.2 *Calculating effect sizes using SPSS*

Analysis	Effect size	SPSS procedure
crosstabulation	phi coefficient (φ)	Analyze, Descriptive Statistics, Crosstabs; Statistics; select Phi
	Pearson's C	Analyze, Descriptive Statistics, Crosstabs; Statistics; select Contingency Coefficient
	Cramér's V	Analyze, Descriptive Statistics, Crosstabs; Statistics; select Cramér's V
	Goodman and Kruskal's lambda (λ)	Analyze, Descriptive Statistics, Crosstabs; Statistics; select Lambda
	Kendall's tau (τ)	Analyze, Descriptive Statistics, Crosstabs, Statistics – select Kendall's tau-b if the table is square-shaped or tau-c if the table is rectangular
t test (independent)	Cohen's d Glass's Δ Hedges g eta^2 (η^2)	Analyze, Compare Means, Independent Samples T Test, then use group means and SDs to calculate d, Δ or g by hand using the equations in the text Analyze, Compare Means, Independent Samples T Test, then calculate $\eta^2 = t^2/(t^2 + N - 1)$
correlational analysis	Pearson correlation (r)	Analyze, Correlate, Bivariate – select Pearson
	partial correlation ($r_{xy.z}$)	Analyze, Correlate, Partial
	point biserial correlation (r_{pb})	Analyze, Correlate, Bivariate – select Pearson (one of the variables should be dichotomous)
	Spearman's rank correlation (ρ)	Analyze, Correlate, Bivariate – select Spearman
multiple regression	R^2	Analyze, Regression, Linear
	$^{adj}R^2$	Analyze, Regression, Linear
	ΔR^2	Analyze, Regression, Linear, enter predictors in blocks, Statistics – select R squared change
	part and partial correlations	Analyze, Regression, Linear, Statistics – select Part and partial correlations
	standardized betas	Analyze, Regression, Linear
logistic regression	logits	Analyze, Regression, Binary Logistic
	odds ratios	As above, then take the antilog of the logit by exponentiating the coefficient (e^b)
	%Δ	As above, then $(e^b - 1) \times 100$ (Pampel 2000: 23)
ANOVA	eta^2 (η^2)	Analyze, Compare Means, ANOVA, then calculate η^2 by dividing the sum of squares between groups by the total sum of squares
	Cohen's f	Analyze, Compare Means, ANOVA, then take the square root of $\eta^2/(1 - \eta^2)$ (Shaughnessy et al. 2009: 434)
ANCOVA	eta^2 (η^2)	Analyze, General Linear Model, Univariate, Options – select Estimates of effect size
MANOVA	partial eta^2 (η^2)	Analyze, General Linear Model, Multivariate, Options – select Estimates of effect size

Reporting effect size indexes – three lessons

It is not uncommon for authors of research papers to report effect sizes without knowing it. This can happen when an author provides a correlation matrix showing the bivariate correlations between the variables of interest or reports test statistics that also happen to be effect size measures (e.g., R^2). But these estimates are seldom interpreted. The normal practice is to pass judgment on hypotheses by looking at the *p* values. The problem with this is that *p* values are confounded indexes that reflect both the size of the effect as it occurs in the population and the statistical power of the test used to detect it. A sufficiently powerful test will almost always generate a statistically significant result irrespective of the effect size. Consequently, effect size estimates need to be interpreted separately from tests of statistical significance.

As we will see in the next chapter the interpretation of research results is sometimes problematic. To facilitate interpretation there are three things researchers need to keep in mind when initially reporting effects. First, clearly identify the type of effect being reported. Second, quantify the degree of precision of the estimate by computing a confidence interval. Third, to maximize opportunities for interpretation, report effects in metrics or terms that can be understood by nonspecialists.

1. Specify the effect size index

It is meaningless to report an effect size without specifying the index or measure used. An effect of size = 0.5 will mean something quite different depending on whether it belongs to the *d* or *r* family of effects. (An *r* of 0.5 is about twice as large as a *d* of 0.5.) Usually the index adopted will reflect the type of effect being measured. If we are interested in assessing the strength of association between two variables, the correlation coefficient *r* or one of its many derivatives will normally be used. If we are comparing groups, then a member of the *d* family may be preferable. (The point biserial correlation is an interesting exception, being a particular type of correlation that is used to compare groups. Although it is counted here as a measure of association, it has a legitimate place in both groups.) The interpretation of *d* and *r* is different, but as both are standardized either one can be transformed into the other using the following equations:[15]

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

Being able to convert one index type into the other makes it is possible to compare effects of different kinds and to draw precise conclusions from studies reporting dissimilar indexes. The full implications of this possibility are explored in Part III of this book in the chapters on meta-analysis.

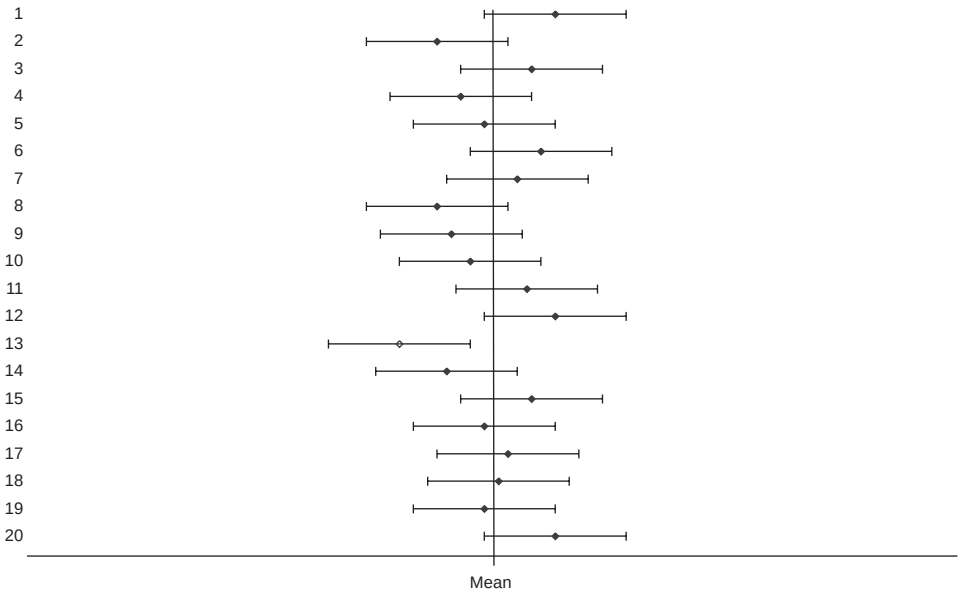


Figure 1.1 Confidence intervals

2. Quantify the precision of the estimate using confidence intervals

In addition to reporting a point estimate of the effect size, researchers should provide a confidence interval quantifying the accuracy of the estimate. A confidence interval is a range of plausible values for the index or parameter being estimated. The "confidence" associated with any interval is proportional to the risk that the interval excludes the true parameter. This risk is known as alpha, or α, and the equation for determining the desired level of confidence or $C = 100(1 - \alpha)\%$. If $\alpha = .05$, then $C = 95\%$. If we are prepared to take a 5% risk that our interval will exclude the true value, we would calculate a 95% confidence interval (or CI_{95}). If we wanted to reduce this risk to 1%, we would calculate a 99% confidence interval (or CI_{99}). The trade-off is that the lower the risk, the wider and less precise the interval. For reasons relating to null hypothesis significance testing and the traditional reliance on $p = .05$, most confidence intervals are set at 95%.

Confidence intervals are relevant whenever an inference is made from a sample to a wider population (Gardner and Altman 2000).[16] Every interval has an associated level of confidence (e.g., 95%, 99%) that represents the proportion of intervals that would contain the parameter if a large number of intervals were estimated. The wrong way to interpret a 95% confidence interval is to conclude that there is a 95% probability that the interval contains the parameter. Figure 1.1 shows why this conclusion can never be drawn. In the figure, the horizontal lines refer to twenty intervals obtained from twenty samples drawn from a single population. In this case the parameter of interest is the population mean represented by the vertical line.

Each sample has provided an independent estimate of this mean and a corresponding confidence interval centered on the estimate. As the figure shows, the individual intervals either include the true population mean or they do not. Interpreting a 95% confidence interval as meaning there is a 95% chance that the interval contains the estimate is a bit like saying you're 95% pregnant (Thompson 2002b). The probability that any given interval contains the parameter is either 0 or 1 but we can't tell which.

Adopting a 95% level of confidence means that in the long run 5% of intervals estimated will exclude the parameter of interest. In Figure 1.1, interval number 13 excludes the mean. It just may be the case that our interval is the unlucky one that misses out. In view of this possibility, a safer way to interpret a 95% confidence interval is to say that we are 95% *confident* that the parameter lies within the upper and lower bounds of the estimated interval.[17]

A confidence interval can also be defined as a point estimate of a parameter (or an effect size) plus or minus a margin of error. Margins of error are often associated with polls reported in the media. For example, a poll showing voter preferences for political candidates will return both a result (the percentage favoring each candidate) and an associated margin of error (which reflects the accuracy of the result and is usually relevant for a confidence interval of 95%). If a poll reports support for a candidate as being 46% with a margin of error of 3%, this means the true percentage of the population that actually favors the candidate is likely to fall between 43% and 49%. What conclusions can we draw from this? If a minimum of 50% is needed to win the election, then the poll suggests this candidate is going to be disappointed on election day. Winning is not beyond the bounds of possibility, but it is well beyond the bounds of probability. Another way to interpret the result would be to say that if we polled the entire population, there would be a 95% chance that the true result would be within the margin of error.

The margin of error describes the precision of the estimate and depends on the sampling error in the estimate as well as the natural variability in the population (Sullivan 2007). Sampling error describes the discrepancy between the values in the population and the values observed in a sample. This error or discrepancy is inversely proportional to the square root of size of the sample. A poll based on 100 voters will have a smaller margin of error than a poll based on just 10.

Confidence intervals are sometimes used to test hypotheses. For example, intervals can be used to test the null hypothesis of no effect. A 95% interval that excludes the null value is equivalent to obtaining a p value $< .05$. While a traditional hypothesis test will lead to a binary outcome (either reject or do not reject the null hypothesis), a confidence interval goes further by providing a range of hypothetical values (e.g., effect sizes) that cannot be ruled out (Smithson 2003). Confidence intervals provide more information than p values and give researchers a better feel for the effects they are trying to estimate. This has implications for the accumulation of results across studies. To illustrate this, Rothman (1986) described ten studies which yielded mixed results. The results of five studies were found to be statistically significant while the remainder were found to

be statistically nonsignificant. However, graphing the confidence intervals for each study revealed the existence of a common effect size that was within the bounds of plausibility in every case (i.e., all ten intervals overlapped the population parameter). While an exclusive focus on *p* values would convey the impression that the body of research was saddled with inconsistent results, the estimation of intervals revealed that the discord in the results was illusory.

Like effect sizes, confidence intervals come highly recommended. In their list of recommendations to the APA, Wilkinson and the Taskforce on Statistical Inference (1999: 599) proposed that interval estimates for effect sizes should be reported "whenever possible" as doing so reveals the stability of results across studies and "helps in constructing plausible regions for population parameters." This recommendation was subsequently adopted in the 5th edition of the APA's *Publication Manual* (APA 2001: 22):

> The reporting of confidence intervals . . . can be an extremely effective way of reporting results. Because confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended.

Similarly, the AERA recommends the use of confidence intervals in its Standards for Reporting (AERA 2006). The rationale is that confidence intervals provide an indication of the uncertainty associated with effect size indexes. In addition, a growing number of journal editors have independently called for the reporting of confidence intervals (see, for example, Bakeman 2001; Campion 1993; Fan and Thompson 2001; La Greca 2005; Neeley 1995).[18]

Yet despite these recommendations, confidence intervals remain relatively rare in social science research. Reviews of published research regularly find that studies reporting confidence intervals are in the extreme minority, usually accounting for less than 2% of quantitative studies (Callahan and Reio 2006; Finch et al. 2001; Kieffer et al. 2001). Possibly part of the reason for this is that although the APA advocated confidence intervals as "the best reporting strategy," no advice was provided on how to construct and interpret intervals.[19]

Confidence intervals can be calculated for descriptive statistics (e.g., means, medians, percentages) and a variety of effect sizes (e.g., the differences between means, relative risk, odds ratios, and regression coefficients). There are essentially two families of confidence interval – central and non-central (Smithson 2003). The difference stems from the type of sampling distribution used (see Box 1.2). Basically central confidence intervals are straightforward to calculate while non-central confidence intervals are computationally tricky. To take the easy ones first, consider the calculation of a confidence interval for a mean that is drawn from a population with a known standard deviation or is calculated from a sample large enough ($N > 150$) that an approximation can be made on the basis of the standard deviation observed in the sample. In either case we can assume that the data are more or less normally distributed according to the familiar bell-shaped curve, permitting us to use the central *t* distribution for critical values used in the calculation.

Box 1.2 Sampling distributions and standard errors

What is a sampling distribution?
Imagine a population with a mean of 100 and a standard deviation of 15. From this population we draw a number of random samples, each of size $N = 50$, to estimate the population mean. Some of the sample means will be a little below the true mean of 100 while others will be above. If we drew a very large set of samples and plotted all their means on a graph, the resulting distribution would be labeled the sampling distribution of the mean for $N = 50$.

What is a standard error?
The standard deviation of a sampling distribution is called the standard error of the mean or the standard error of the proportion or whatever parameter we are trying to estimate. The standard error is very important in the calculation of inferential statistics and confidence intervals as it is an indicator of the uncertainty of a sample-based statistic. Two samples drawn from the same population are unlikely to produce identical parameter estimates. Each estimate is imprecise and the standard error quantifies this imprecision. The smaller the standard error, the more precise is the estimate of the mean and the narrower the confidence interval. For any given sample the standard error can be estimated by dividing the standard deviation of the sample by the square root of the sample size.

The confidence interval for the mean $\overline{X}$ can be expressed as $\overline{X} \pm \text{ME}$ where ME refers to the margin of error. The margin of error is derived from the standard error (*SE*) of the mean which is found by dividing the observed standard deviation (*SD*) by the square root of sample size (*N*). Consider a study where $\overline{X} = 145$, $SD = 70$, and $N = 49$. The standard error in this case is:

$$\begin{aligned} SE &= SD/\sqrt{N} \\ &= 70/\sqrt{49} \\ &= 10 \end{aligned}$$

The width of the margin of error is the *SE* multiplied by $t_{(N-1)C}$, where t is the critical value of the t statistic for $N - 1$ degrees of freedom that corresponds to our chosen level of confidence C.[20] The critical value of t when $C = 95\%$ and $df = N - 1 = 48$ is 2.01. This value can be found by looking up a table showing critical values of the t distribution and finding the value that intersects $df = 48$ and $\alpha = .05$ (or $\alpha/2 = .025$ if only upper tail areas are listed).[21] Knowing the critical t value we can calculate the margin of error as follows:

$$\begin{aligned} ME &= SE \times t_{(N-1)C} \\ &= 2.01 \times 10 \\ &= 20.1 \end{aligned}$$

We can now calculate the lower and upper bounds of the confidence interval by subtracting and adding the margin of error from and to the mean: CI_{95} lower limit = 124.9 (145 – 20.1), upper limit = 165.1 (145 + 20.1). Ideally a confidence interval should be portrayed graphically. There are a couple of ways to do this using Excel. One way is to create a *Stock* chart with raw data coming from three columns corresponding to high and low values of the interval and point estimates. Another way is create a scatter graph by selecting *Scatter* from the *Chart* submenu and linking it to raw data in two columns. The first column corresponds to the interval number and the second column corresponds to the point estimate of the mean. Next, select the data points and choose X or Y error bars under the *Format* menu. Intervals can be given a fixed value, as was done for Figure 1.1, or a unique value under *Custom* corresponding to data in a third or even a fourth column. Additional information, such as a population mean, can be superimposed by using the *Drawing* toolbar.

Formulas can be used to calculate central confidence intervals because the widths are centered on the parameter of interest; they extend the same distance in both directions. However, generic formulas cannot be used to compute non-central confidence intervals (e.g., for Cohen's *d*) because the widths are not pivotal (Thompson 2007a). In the old days before personal computers, these types of confidence intervals were calculated by hand on the basis of approximations that held under certain circumstances. (A review of these methods can be found in Hedges and Olkin (1985: 85–91).) But now this type of analysis is normally done by a computer program that iteratively guesstimates the two boundaries of each interval independently until a desired statistical criterion is approximated (Thompson 2008). Software that can be used to calculate these sorts of confidence intervals is discussed by Smithson (2001), Bryant (2000), Cumming and Finch (2001), and Mendoza and Stafford (2001). Other useful sources relevant to calculating confidence intervals are listed in the notes at the end of this chapter.[22]

3. Report effects in jargon-free language

Earlier we saw how the size of any difference between two groups can be expressed in a standardized form using an index such as Cohen's *d*. Although *d* is probably one of the best known effect size indexes, it remains unfamiliar to the nonspecialist. This limits opportunities for interpretation and raises the risk that alternative plausible explanations for observed effects will not be considered. Fortunately a number of jargon-free metrics are available to the researcher looking to maximize interpretation possibilities. These include the **common language effect size** index (McGraw and Wong 1992), the **probability of superiority** (Grissom 1994), and the **binomial effect size display** (Rosenthal and Rubin 1982).

The first two indexes transform the difference between two groups into a probability – the probability that a random value or score from one group will be greater than a random value or score from the other. Consider height differences between men and women. Men tend to be taller on average and a Cohen's *d* could be calculated to quantify this difference in a standardized form. But knowing that the average male is two standard

Box 1.3 Calculating the common language effect size index

In most of the married couples you know, chances are the man is taller than the woman. But if you were to pick a couple at random, what would be the probability that the man would be taller? Experience suggests that answer must be more than 50% and less than 100%, but could you come up with an exact probability using the following data?

Height (inches)	Mean	Standard deviation	Variance
Males	69.7	2.8	7.84
Females	64.3	2.6	6.76

The common language (CL) statistic converts an effect into a probability. In this height example, which comes from McGraw and Wong (1992), we want to determine the probability of obtaining a male-minus-female height score greater than zero from a normal distribution with a mean of 5.4 inches (the difference between males and females) and a standard deviation equivalent to the square root of the sum of the two variances: $3.82 = \sqrt{(7.84 + 6.76)}$. To determine this probability, it is necessary to convert these raw data to a standardized form using the equation: $z = (0 - 5.4)/3.82 = -1.41$. On a normal distribution, -1.41 corresponds to that point at which the height difference score is 0. To find out the upper tail probability associated with this score, enter this score into a *z* to *p* calculator such as the one provided by Lowry (2008b). The upper tail probability associated with this value is .92. This means that in 92% of couples, the male will be taller than the female.

Another way to quantify the so-called "probability of superiority" (*PS*) would be to calculate the standardized mean difference between the groups and then convert the resulting *d* or Δ to its PS equivalent by looking up a table such as Table 1 in Grissom (1994).

deviation units taller than the average female (a huge difference) may not mean much to the average person. A better way to quantify this difference would be to calculate the probability that a randomly picked male will be taller than a randomly picked female. As it happens, this probability is .92. The calculation devised by McGraw and Wong (1992) to arrive at this value is explained in Box 1.3.[23] A probability of superiority index based on Grissom's (1994) technique would have generated the same result.

Correlations are the bread and butter of effect size analysis. Most students are reasonably comfortable calculating correlations and have no problem understanding that a correlation of -0.7 is actually bigger than a correlation of 0.3. But correlations can be confusing to nonspecialists and squaring the correlation to compute the proportion of shared variance only makes things more confusing. What does it mean to say that a proportion of the variability in Y is accounted for by variation in X? To make matters

Table 1.3 *The binomial effect size display of* $r = .30$

	Success	Failure	Total
Treatment	65	35	100
Control	35	65	100
Total	100	100	200

worse, many interesting correlations in science are small and squaring a small correlation makes it smaller still. Consider the case of aspirin, which has been found to lower the risk of heart attacks (Rosnow and Rosenthal 1989). The benefits of aspirin consumption expressed in correlational form are tiny, just $r = .034$. This means that the proportion of shared variance between aspirin and heart attack risk is just .001 (or .034 × .034). This sounds unimpressive as it leaves 99.9% of the variance unaccounted for. Seemingly less impressive still is the Salk poliomyelitis vaccine which has an effect equivalent to $r = .011$ (Rosnow and Rosenthal 2003). In POV terms the benefits of the polio vaccine are a piddling one-hundredth of 1% (i.e., .011 × .011 or $r^2 = .0001$). Yet no one would argue that vaccinating against polio is not worth the effort.

A more compelling way to convey correlational effects is to present the results in a binomial effect size display (BESD). Developed by Rosenthal and Rubin (1982), the BESD is a 2 × 2 contingency table where the rows correspond to the independent variable and the columns correspond to any dependent variable which can be dichotomized.[24] Creating a BESD for any given correlation is straightforward. Consider a table where rows refer to groups (e.g., treatment and control) and columns refer to outcomes (e.g., success or failure). For any given correlation (r) the success rate for the treatment group is calculated as $(.50 + r/2)$, while the success rate for the control group is calculated as $(.50 - r/2)$. Next, insert values into the other cells so that the row and column totals add up to 100 and voilà!

A stylized example of a BESD is provided in Table 1.3. In this case the correlation $r = .30$ so the value in the success-treatment cell is .65 (or $.50 + .30/2$) and the value in the success-control cell is .35 (or $.50 - .30/2$). The BESD shows that success was observed for nearly two-thirds of people who undertook treatment but only a little over one-third of those in the control group. Looking at these numbers most would agree that the treatment had a fairly noticeable effect. The difference between the two groups is 30 percentage points. This means that those who took the treatment saw an 86% improvement in their success rate (representing the 30 percentage point gain divided by the 35-point baseline). Yet if these results had been expressed in proportion of variance terms, the effectiveness of the treatment would have been rated at just 9%. That is, only 9% of the variance in success is accounted for by the treatment. Someone unfamiliar with this type of index might conclude that the treatment had not been particularly effective. This shows how the interpretation of a result can be influenced by the way in which it is reported.

Table 1.4 *The effects of aspirin on heart attack risk*

	Heart attack	No heart attack	Total
Raw counts			
Aspirin (treatment)	104	10,933	11,037
Placebo (control)	189	10,845	11,034
Total	293	21,778	22,071
BESD ($r = .034$)			
Aspirin	48.3	51.7	100
Placebo	51.7	48.3	100
Total	100	100	200

Source: Rosnow and Rosenthal (1989, Table 2)

Another example of a BESD is provided in Table 1.4. This one was done by Rosnow and Rosenthal (1989) to illustrate the effects of aspirin consumption on heart attack risk. The raw data in the top of the table came from a large-scale study involving 22,071 doctors (Steering Committee of the Physicians' Health Study Research Group 1988). Every other day for five years half the doctors in the study took aspirin while the rest took a placebo. The study data show that of those in the treatment group, 104 suffered a heart attack while the corresponding number in the control group was 189. The difference between the two groups is statistically significant – the benefits of aspirin are no fluke. However, as mentioned earlier, the effects of aspirin appear very small when expressed in terms of shared variability. But when displayed in a BESD, the benefits of aspirin are more impressive. The table shows taking aspirin lowers the risk of a heart attack by more than 3% (i.e., 51.7 – 48.3). In other words, three out of a hundred people will be spared heart attacks if they consume aspirin on a regular basis. To the nonspecialist this is far more meaningful than saying the percentage of variance in heart attacks accounted for by aspirin consumption is one-tenth of 1%.

Summary

An increasing number of editors are either encouraging or mandating effect size reporting in new journal submissions (e.g., Bakeman 2001; Campion 1993; Iacobucci 2005; JEP 2003; La Greca 2005; Lustig and Strauser 2004; Murphy 1997).[25] Quite apart from editorial preferences, there are at least three important reasons for gauging and reporting effect sizes. First, doing so facilitates the interpretation of the practical significance of a study's findings. The interpretation of effects is discussed in Chapter 2. Second, expectations regarding the size of effects can be used to inform decisions about how many subjects or data points are needed in a study. This activity describes power analysis and is covered in Chapters 3 and 4. Third, effect sizes can be used to compare the results of studies done in different settings. The meta-analytic pooling of effect sizes is discussed in Chapters 5 and 6.

Notes

1 Even scholars publishing in top-tier journals routinely confuse statistical with practical significance. In their review of 182 papers published in the 1980s in the *American Economic Review*, McCloskey and Ziliak (1996: 106) found that 70% "did not distinguish statistical significance from economic, policy, or scientific significance." Since then things have got worse. In a follow-up analysis of 137 papers published in the 1990s in the same journal, Ziliak and McCloskey (2004) found that 82% mistook statistical significance for economic significance. Economists are hardly unique in their confusion over significance. An examination of the reporting practices in the *Strategic Management Journal* revealed that no distinction was made between statistical and substantive significance in 90% of the studies reviewed (Seth et al. 2009).

2 This practice can perhaps be traced back to the 1960s when, during his tenure as editor of the *Journal of Experimental Psychology*, Melton (1962: 554) insisted that the researcher had a responsibility to "reveal his effect in such a way that no reasonable man would be in a position to discredit the results by saying they were the product of the way the ball bounced." For Melton this meant interpreting the size of the effect observed in the context of other "previously or concurrently demonstrated effects." Isolated findings, even those that were statistically significant, were typically not considered suitable for publication. A similar stance was taken by Kevin Murphy during his tenure as editor of the *Journal of Applied Psychology*. In one editorial he wrote: "If an author decides not to present an effect size estimate along with the outcome of a significance test, I will ask the author to provide some specific justifications for why effect sizes are not reported. So far, I have not heard a good argument against presenting effect sizes" (Murphy 1997: 4).

Bruce Thompson, a former editor of no less than three different journals, has done more than most to advocate effect size reporting in scholarly journals. In the late 1990s Thompson (1999b, 1999c) noted with dismay that the APA's (1994) "encouragement" of effect size reporting in the 4th edition of its publication manual had not led to any substantial changes to reporting practices. He argued that the APA's policy

> presents a self-canceling mixed message. To present an "encouragement" in the context of strict absolute standards regarding the esoterics of author note placement, pagination, and margins is to send the message, "These myriad requirements count: this encouragement doesn't." (Thompson 1999b: 162)

Possibly in response to the agitation of Thompson and like-minded others (e.g., Kirk 1996; Murphy 1997; Vacha-Haase et al. 2000; Wilkinson and the Taskforce on Statistical Inference 1999), the 5th edition of the APA's (2001) publication manual went beyond encouragement, stating that "it is almost always necessary to include some index of effect size" (p. 25). Now it is increasingly common for editors to insist that authors report and interpret effect sizes. During the 1990s a survey of twenty-eight APA journals identified only five editorials that explicitly called for the reporting of effect sizes (Vacha-Haase et al. 2000). But in a recent poll of psychology editors Cumming et al. (2007) found that a majority now advocate effect size reporting. On his website Thompson (2007b) lists twenty-four educational and psychology journals that require effect size reporting. This list includes a number of prestigious journals such as the *Journal of Applied Psychology*, the *Journal of Educational Psychology* and the *Journal of Consulting and Clinical Psychology*.

As increasing numbers of editors and reviewers become cognizant of the need to report and interpret effect sizes, Bakeman (2001: 5) makes the ominous prediction that "empirical reports that do not consider the strength of the effects they detect will be regarded as inadequate." Inadequate, in this context, means that relevant evidence has been withheld (Grissom and Kim 2005: 5). The reviewing practices of the journal *Anesthesiology* may provide a glimpse into the future of the peer review process. Papers submitted to this journal must initially satisfy a special reviewer

that authors have not confused the results of statistical significance tests with the estimation of effect sizes (Eisenach 2007).

A few editors have gone beyond issuing mandates and have provided notes outlining their expectations regarding effect size reporting (see for example the notes by Bakeman (2001), a former editor of *Infancy*, and Campion (1993) of *Personnel Psychology*). Usually these editorial instructions have been based on the authoritative "Guidelines and Explanations" originally developed by Wilkinson and the Taskforce on Statistical Inference (1999), which itself was partly based on the recommendations developed by Bailar and Mosteller (1988) for the medical field. But for the most part practical guidelines for effect size reporting are lacking. As Grissom and Kim (2005: 56) observed, "effect size methodology is barely out of its infancy."

There have been repeated calls for textbook authors to provide material explaining effect sizes, how to compute them, and how to interpret them (Hyde 2001; Kirk 2001; Vacha-Haase 2001). To date, the vast majority of texts on the subject are full of technical notes, algebra, and enough Greek to confuse a classicist. Teachers and students who would prefer a plain English introduction to this subject will benefit from reading the short papers by Coe (2002), Clark-Carter (2003), Field and Wright (2006), and Vaughn (2007).

For the researcher looking for discipline-specific examples of effect sizes, introductory papers have been written for fields such as education (Coe 2002; Fan 2001), school counseling (Sink and Stroh 2006), management (Breaugh 2003), economics (McCloskey and Ziliak 1996), psychology (Kirk 1996; Rosnow and Rosenthal 2003; Vacha-Haase and Thompson 2004), educational psychology (Olejnik and Algina 2000; Volker 2006), and marketing (Sawyer and Ball 1981; Sawyer and Peter 1983). For the historically minded, Huberty (2002) surveys the evolution of the major effect size indexes, beginning with Francis Galton and his cousin Charles Darwin. His paper charts the emergence of the correlation coefficient (in the 1890s), eta-squared (in the 1930s), *d* and omega-squared (both in the 1960s), and other popular indexes. Rodgers and Nicewander (1988) celebrated the centennial decade of correlation and regression with a paper tracing landmarks in the development of *r*.

3 Using a magisterial mixture of Greek and hieroglyphics, the 5th edition of the *Publication Manual of the American Psychological Association* helpfully suggests authors report effect sizes using any of a number of estimates "including (but not limited to) r^2, η^2, ω^2, R^2, ϕ^2, Cramér's *V*, Kendall's *W*, Cohen's *d* and κ, Goodman–Kruskal's λ and γ . . . and Roy's Θ and the Pillai–Bartlett V" (APA 2001: 25–26).

4 To be fair, Rosnow and Rosenthal (2003, Table 5) provide a hypothetical example of a situation where the risk difference would be superior to both the risk ratio and the odds ratio.

5 This is the same result that would have been obtained had we followed the equation for probabilities above. The odds that an event or outcome will occur can be expressed as the ratio between the probability that it will occur to the probability that it won't: $p/(1-p)$. Conversely, to convert odds into a probability use: $p = odds/(1+ odds)$.

6 We might just as easily discuss the relative risk of *passing* which is 2.5 (.50/.20) in Socrates' class compared with Aristotle's. But as the name suggests, the *risk* ratio is normally used to quantify an outcome, in this case failing, which we wish to avoid.

7 For more on the differences between proportions, relative risk, and odds ratios, see Breaugh (2003), Gliner et al. (2002), Hadzi-Pavlovic (2007), Newcombe (2006), Osborne (2008a), and Simon (2001). Fleiss (1994) provides a good overview of the merits and limitations of four effect size measures for categorical data and an extended treatment can be found in Fleiss et al. (2003).

8 To calculate the pooled standard deviation (SD_{pooled}) for two groups A and B of size *n* and with means $\overline{X}$ we would use the following equation from Cohen (1988: 67):

$$SD_{pooled} = \sqrt{\frac{\sum (X_A - \overline{X}_A)^2 + \sum (X_B - \overline{X}_B)^2}{n_A + n_B - 2}}$$

9 To calculate the weighted and pooled standard deviation ($SD*_{\text{pooled}}$) we would use the following equation from Hedges (1981: 110):

$$SD^*_{pooled} = \sqrt{\frac{(n_A - 1)SD_A^2 + (n_B - 1)SD_B^2}{n_A + n_B - 2}}$$

Hedges' g was also developed to remove a small positive bias affecting the calculation of d (Hedges 1981). An unbiased version of d can be arrived at using the following equation adapted from Hedges and Olkin (1985: 81):

$$g \cong d\left(1 - \frac{3}{4(n_1 + n_2) - 9}\right)$$

However, beware the inconsistent terminology. What is labeled here as g was labeled by Hedges and Olkin as d and vice versa. For these authors writing in the early 1980s, g was the mainstream effect size index developed by Cohen and refined by Glass (hence g for Glass). However, since then g has become synonymous with Hedges' equation (not Glass's) and the reason it is called Hedges' g and not Hedges' h is because it was originally named after Glass – even though it was developed by Larry Hedges. Confused?

10 Both the phi coefficient and the odds ratio can be used to quantify effects when categorical data are displayed on a 2×2 contingency table, so which is better? According to Rosenthal (1996: 47), the odds ratio is superior as it is unaffected by the proportions in each cell. Rosenthal imagines an example where 10 of 100 (10%) young people who receive Intervention A, as compared with 50 of 100 (50%) young people who receive Intervention B, commit a delinquent offense. The phi coefficient for this difference is .436. However, if you increase the number in group A to 200 and reduce the number in group B to 20, while holding the percentage of offenders constant in each case, the phi coefficient falls to .335. This drop suggests that the effectiveness of the intervention is greater in the first situation than in the second, when in reality there has been no change. In contrast, the odds ratio for both situations is 9.0.

11 Some might argue that the coefficient of multiple determination (R^2) is not a particularly useful index as it combines the effects of several predictors. To isolate the individual contribution of each predictor, researchers should also report the relevant semipartial or part correlation coefficient which represents the change in Y when X_1 is changed by one unit while controlling for all the other predictors ($X_2, \ldots X_k$). Although both the part and partial correlations can be calculated using SPSS and other statistical programs, the former is typically used when "apportioning variance" among a set of independent variables (Hair et al. 1998: 190). For a good introduction on how to interpret coefficients in *non*linear regression models, see Shaver (2007).

12 Effect size indexes such as R^2 and η^2 tend to be upwardly biased on account of the principle of mathematical maximization used in the computation of statistics within the general linear model family. This principle means that any variance in the data – whether arising from natural effects in the population or sample-specific quirks – will be considered when estimating effects. Every sample is unique and that uniqueness inhibits replication; a result obtained in a particularly quirky sample is unlikely to be replicated in another. The uniqueness of samples, which is technically described as sampling error, is positively related to the number of variables being measured and negatively related to both the size of the sample and the population effect (Thompson 2002a). The implication is that index-inflation attributable to sampling error is greatest when sample sizes and effects are small and when the number of variables in the model is high (Vacha-Haase and Thompson 2004). Fortunately the sources of sampling error are so well known that we can correct for this inflation and calculate unbiased estimates of effect size (e.g., $_{\text{adj}}R^2$, ω^2). These unbiased or corrected estimates are usually smaller than their uncorrected counterparts and are thought to be closer to population effect sizes (Snyder and Lawson 1993). The difference between biased and unbiased (or corrected and uncorrected) measures is referred to as shrinkage (Vacha-Haase

and Thompson 2004). Shrinkage tends to shrink as sample sizes increase and the number of predictors in the model falls. However, shrinkage tends to be very small if effects are large, irrespective of sample size (e.g., larger R^2s tend to converge with their adjusted counterparts). Should researchers report corrected or uncorrected estimates? Vacha-Haase and Thompson (2004) lean towards the latter. But given Roberts' and Henson's (2002) concern that sometimes estimates are "over-corrected," the prudent path is probably to report both.

13 Good illustrations of how to calculate Cohen's *f* are hard to come by, but three are provided by Shaughnessy et al. (2009: 434), Volker (2006: 667–669), and Grissom and Kimt (2005: 119).

It should be noted that many of these test statistics require that the data being analyzed are normally distributed and that variances are equal for the groups being compared or the variables thought to be associated. When these assumptions are violated, the statistical power of tests falls, making it harder to detect effects. Confidence intervals are also likely to be narrower than they should be. An alternative approach which has recently begun to attract attention is to adopt statistical methods that can be used even when data are nonnormal and heteroscedastic (Erceg-Hurn and Mirosevich 2008; Keselman et al. 2008; Wilcox 2005). Effect sizes associated with these so-called robust statistical methods include robust analogs of the standardized mean difference (Algina et al. 2005) and the probability of superiority or *PS* (Grissom 1994). *PS* is the probability that a randomly sampled score from one group will be larger than a randomly sampled score from a second group. A *PS* score of .50 is equivalent to a *d* of 0. Conversely, a large *d* of .80 is equivalent to a *PS* of .71 (see also Box 1.3).

14 Many free software packages for calculating effect sizes are available online. An easy-to-use Excel spreadsheet along with a manual by Thalheimer and Cook (2002) can be downloaded from www.work-learning.com/effect_size_download.htm. Another Excel-based calculator is provided by Robert Coe of Durham University and can be found at www.cemcentre.org/renderpage.asp?linkID=30325017Calculator.htm. Some of the calculators floating around online are specific to a particular effect size such as relative risk (www.hutchon.net/ConfidRR.htm), Cohen's *d* (Becker 2000), and f^2 (www.danielsoper.com/statcalc/calc13.aspx). Others can be used for a variety of indexes (e.g., Ellis 2009). As these are constantly being updated, the best advice is to google the desired index along with the search terms "online calculator."

15 This is practically true but technically contentious, as explained by McGrath and Meyer (2006). See also Vacha-Haase and Thompson (2004: 477). When converting *d* to *r* in the case of unequal group sizes, use the following equation from Schulze (2004: 31):

$$r = \sqrt{\frac{d^2}{d^2 + \frac{(n_1+n_2)^2 - 2(n_1+n_2)}{n_1 n_2}}}$$

The effect size *r* can also be calculated from the chi-square statistic with one degree of freedom and from the standard normal deviate *z* (Rosenthal and DiMatteo 2001: 71), as follows:

$$r = \sqrt{\frac{x_1^2}{N}}$$

$$r = \frac{z}{\sqrt{N}}$$

16 Researchers select samples to represent populations. Thus, what is true of the sample is inferred to be true of the population. However, this sampling logic needs to be distinguished from the inferential logic used in statistical significance testing where the direction of inference runs from the population to the sample (Cohen 1994; Thompson 1999a).

17 However, even this interpretation is dismissed by some as misleading (e.g., Thompson 2007a). Problems arise because "confidence" means different things to statisticians and nonspecialists. In everyday language to say "I am 95% confident that the interval contains the population parameter" is to claim virtual certainty when in fact the only thing we can be certain of is that the method of estimation will be correct 95% of the time. There is presently no consensus on the best way to interpret a confidence interval, but it is reasonable to convey the general idea that values within the confidence interval are "a good bet" for the parameter of interest (Cumming and Finch 2005).

18 One particularly well-known advocate of confidence intervals is Kenneth Rothman (1986). During his two-year tenure as editor of *Epidemiology*, Rothman refused to publish any paper reporting statistical hypothesis tests and *p* values. His advice to prospective authors was radical: "When writing for *Epidemiology*, you can . . . enhance your prospects if you omit tests of statistical significance" (Rothman 1998). *P* values were shunned because they confound effect size with sample size and say little about the precision of a result. Rothman preferred point and interval estimates. This led to a boom in the reporting of confidence intervals in *Epidemiology*.

19 Possibly another reason why intervals are not reported is because they are sometimes "embarrassingly large" (Cohen 1994: 1002). Imagine the situation where an effect found to be medium-sized is couched within an interval of plausible values ranging from very small to very large. How does a researcher interpret such an imprecise result? This is one of those times where the best way to deal with the problem is to avoid it altogether, meaning that researchers should design studies and set sample sizes with precision targets in mind. This point is taken up in Chapter 3.

20 Sometimes you will see the critical value "$t_{(N-1)C}$" expressed as "t_{CV}," "$t_{(df:\ \alpha/2)}$," or "$t_{N-1}(0.975)$," or even "1.96." What's going on here? The short version is that these are five different ways of saying the same thing. Note that there are two parts to determining the critical value of *t*: (1) the degrees of freedom in the result, or *df*, which are equal to $N - 1$, and (2) the desired level of confidence (C, usually 95%) which is equivalent to $1 - \alpha$ (and α usually $= .05$). To save space, tables listing critical values of the *t* distribution typically list only upper tail areas which account for half of the critical regions covered by alpha. So instead of looking up the critical value for $\alpha = .05$, we would look up the value for $\alpha/2 = .025$, or the 0.975 quantile (although this can be a bit misleading because we are not calculating a 97.5% confidence interval). For large samples ($N > 150$) the *t* distribution begins to resemble the *z* (standard normal) distribution so critical *t* values begin to converge with critical *z* values. The critical upper-tailed *z* value for $\alpha_2 = .05$ is 1.96. (Note that this is the same as the one-tailed value when $\alpha = .025$.) What does this number mean? In the sampling distribution of any mean, 95% of the sample means will lie within 1.96 standard deviations of the population mean.

21 The same result can be achieved using the Excel function: =tinv(probability, degrees of freedom) = tinv(.05, 48).

22 Methods for constructing basic confidence intervals (e.g., relevant for means and differences between means) can be found in most statistics textbooks (see, for example, Sullivan (2007, Chapter 9) or McClave and Sincich (2009, Chapter 7)), as well as in some research methods texts (e.g., Shaughnessy (2009, Chapter 12)). Three good primers on the subject are provided by Altman et al. (2000), Cumming and Finch (2005), and Smithson (2003). For more specialized types of confidence intervals relevant to effect sizes such as odds ratios, bivariate correlations, and regression coefficients, see Algina and Keselman (2003), Cohen et al. (2003), and Grissom and Kim (2005). Technical discussions relating confidence intervals to specific analytical methods have been provided for ANOVA (Bird 2002; Keselman et al. 2008; Steiger 2004) and multiple regression (Algina et al. 2007). The *Educational and Psychological Measurement* journal devoted a special issue to confidence intervals in August 2001. The calculation of noncentral confidence intervals normally requires specialized software such as the Excel-based Exploratory Software for Confidence Intervals (ESCI) developed by Geoff Cumming of La Trobe University. This program can be found at www.latrobe.edu.au/psy/esci/index.html.

23 The example in Box 1.3 illustrates how to calculate the common language effect size when comparing two groups (CL_G). To calculate a common language index from the correlation of two continuous variables (CL_R), see Dunlap (1994).

24 BESDs can be prepared for outcomes that are both dichotomous and continuous. In the first instance percentages are used as opposed to raw counts. In the second instance binary outcomes are computed from the point biserial correlation r_{pb}. In such cases the success rate for the treatment group is computed as $0.50 + r/2$ whereas the success rate for the control group is computed as $0.50 - r/2$. A BESD can also be used where standardized group means have been reported for two groups of equal size by converting d to r using the equation: $r = d/\sqrt{(d^2 + 4)}$. To work with more than two groups or groups of unequal size see Rosenthal et al. (2000). For more on the BESD see Rosenthal and Rubin (1982), Di Paula (2000), and Randolph and Edmondson (2005).

25 Hyde (2001), herself a former journal editor, suggests that one reason why more editors have not called for effect size reporting is because they are old – they learned their statistics thirty years ago when null hypothesis statistical testing was less controversial and research results lived or died according to the $p = .05$ cut-off. But now the statistical world is more "complex and nuanced" and exact p levels are often reported along with estimates of effect size. Hyde argues that this is not controversial but "good scientific practice" (2001: 228).

2 Interpreting effects

Investigators must learn to argue for the significance of their results without reference to inferential statistics. ~ John P. Campbell (1982: 698)

An age-old debate – rugby versus soccer

A few years ago a National IQ Test was conducted during a live TV show in Australia. Questions measuring intelligence were asked on the show and viewers were able to provide answers via a special website. People completing the online questionnaire were also asked to provide some information about themselves such as their preferred football code. When the results of the test were published it was revealed that rugby union fans were, on average, two points smarter than soccer fans. Now two points does not seem to be an especially big difference – it was actually smaller than the gap separating mums from dads – but the difference was big enough to trigger no small amount of gloating from vociferous rugby watchers. As far as these fans were concerned, two percentage points was large enough to substantiate a number of stereotypes regarding the mental capabilities of people who watch soccer.[1]

How large does an effect have to be for it to be important, useful, or meaningful? As the National IQ story shows, the answer to this question depends a lot on who is doing the asking. Rugby fans interpreted a 2-point difference in IQ as meaningful, legitimate, and significant. Soccer fans no doubt interpreted the difference as trivial, meaningless, and insignificant. This highlights the fundamental difficulty of interpretation: effects mean different things to different people. What is a big deal to you may not be a big deal to me and vice versa. The interpretation of effects inevitably involves a value judgment. In the name of objectivity scholars tend to shy away from making these sorts of judgments. But Kirk (2001) argues that researchers, who are intimately familiar with the data, are well placed to comment on the meaning of the effects they observe and, indeed, have an obligation to do so. However, surveys of published research reveal that most authors make no attempt to interpret the practical or real-world significance of their research results (Andersen et al. 2007; McCloskey and Ziliak 1996; Seth et al. 2009). Even when effect sizes and confidence intervals are reported, they usually go uninterpreted (Fidler et al. 2004; Kieffer et al. 2001).

It is not uncommon for social science researchers to interpret results on the basis of tests of statistical significance. For example, a researcher might conclude that a result that is highly statistically significant is bigger or more important than a marginally significant result. Or a nonsignificant result might be interpreted as indicating the absence of an effect. Both conclusions would be wrong and stem from a misunderstanding of what statistical significance testing can and cannot do. Tests of statistical significance are properly used to manage the risk of mistaking random sampling variation for genuine effects.[2] Statistical tests limit, but do not wholly remove, the possibility that sampling error will be misinterpreted as something real. As the power of such tests is affected by several parameters, of which effect size is just one, their results cannot be used to inform conclusions about effect magnitudes (see Box 2.1).

Researchers cannot interpret the meaning of their results without first estimating the size of the effects that they have observed. As we saw in Chapter 1 the estimation of an effect size is distinct from assessments of statistical significance. Although they are related, statistical significance is also affected by the size of the sample. The bigger the sample, the more likely an effect will be judged statistically significant. But just as a $p = .001$ result is not necessarily more important than a $p = .05$ result, neither is a Cohen's d of 1.0 necessarily more interesting or important than a d of 0.2. While large effects are *likely* to be more important than small effects, exceptions abound. Science has many paradigm-busting discoveries that were triggered by small effects, while history famously turns on the hinges of events that seemed inconsequential at the time.

The problem of interpretation

To assess the practical significance of a result it is not enough that we know the size of an effect. Effect magnitudes must be interpreted to extract meaning. If the question asked in the previous chapter was how big is it? then the question being asked here is how big is big? or is the effect big enough to mean something?

Effects by themselves are meaningless unless they can be contextualized against some frame of reference, such as a well-known scale. If you overheard an MBA student bragging about getting a score of 140, you would conclude that they were referring to their IQ and not their GMAT result. An IQ of 140 is high, but a GMAT score of 140 would not be enough to get you admitted to the Timbuktu Technical School of Shoelace Manufacturing. However, the interpretation of results becomes problematic when effects are measured indirectly using arbitrary or unfamiliar scales. Imagine your doctor gave you the following information:

> Research shows that people with your body-mass index and sedentary lifestyle score on average 2 points lower on a cardiac risk assessment test in comparison with active people with a healthy body weight.

Would this prompt you to make drastic changes to your lifestyle? Probably not. Not because the effect reported in the research is trivial but because you have no way of interpreting its meaning. What does "2 points lower" mean? Does it mean you are more

Box 2.1 Distinguishing effect sizes from *p* values

Two studies were done comparing the knowledge of science fiction trivia for two groups of fans, *Star Wars* fans (Jedi-wannabes) and *Star Trek* fans (Trekkies). The mean test scores and standard deviations are presented in the table below.

The results of Study 1 and Study 2 are the same; the average scores and standard deviations were identical in both studies. But the results from the first study were not statistically significant (i.e., $p > .05$). This led the authors of Study 1 to conclude that there was no appreciable difference between the groups in terms of their knowledge of sci-fi trivia. However, the authors of Study 2 reached a different conclusion. They noted that the 5-point difference in mean test scores was genuine and substantial in size, being equivalent to more than one-half of a standard deviation. They concluded that Jedi-wannabes are substantially smarter than Trekkies.

Test scores for knowledge of sci-fi trivia

	N	Mean	*SD*	*t*	*p*	Cohen's *d*
Study 1						
Jedi-wannabes	15	25	9	1.52	>.05	0.56
Trekkies	15	20	9			
Study 2						
Jedi-wannabes	30	25	9	2.15	<.05	0.56
Trekkies	30	20	9			

How could two studies with identical effect sizes lead to radically different conclusions? The answer has to do with the mis-use of statistical significance testing. When interpreting the results of their study, the authors of Study 1 ignored the estimate of effect size and focused on the *p* value. They incorrectly interpreted a nonsignificant result as indicating no meaningful effect. A nonsignificant result is more accurately interpreted as an inconclusive result. There might be no effect, or there might be an effect but the study lacked the statistical power to detect it. Given the result of Study 2 it is tempting to conclude that Study 1's lack of a result was a case of a genuine effect being missed due to insufficient power.

or less healthy than a normal person? Is 2 points a big deal? Should you be worried? Being unfamiliar with the scale, you are unable to draw any conclusion.

Now imagine your doctor said this to you instead:

Research shows that people with your body-mass index and sedentary lifestyle are four times as likely to suffer a serious heart attack within 10 years in comparison with active people with a normal body weight.

Now the doctor has your full attention. This time you are sitting on the edge of your seat, gripped with a resolve to lose weight and start exercising again. Hearing the

research result in terms which are familiar to you, you are better able to extract their meaning and draw conclusions.

Unfortunately the medical field is something of a special case when it comes to reporting results in metrics that are widely understood. Most people have heard of cholesterol, blood pressure, the body-mass index, blood-sugar levels, etc. But in the social sciences many phenomena (e.g., self-esteem, trust, satisfaction, power distance, opportunism, depression) can be observed only indirectly by getting people to circle numbers on an arbitrary scale. A scale is considered arbitrary when there is no obvious connection between a given score and an individual's actual state or when it is not known how a one-unit change on the score reflects change on the underlying dimension (Blanton and Jaccard 2006). Arbitrary scales are useful for gauging effect sizes but make interpretation problematic.

The field of psychology provides a good example of this difficulty. Psychology researchers have a professional imperative to explain their results in terms of their clinical significance to practitioners and patients (Kazdin 1999; Levant 1992; Thompson 2002a). But many effects in psychology are measured using arbitrary scales that have no direct connection with real-world outcomes (Sechrest et al. 1996). Consider a study assessing the effectiveness of a particular treatment on depression. In the study depression is measured before and after the treatment by getting subjects to complete a pencil and paper test. If the "after" scores are better than the "before" scores, and if the difference between the scores is nontrivial and statistically significant, the researcher might conclude that the treatment had been effective. But this conclusion will not be warranted unless the change in test scores corresponds to an actual change in outcomes valued by the patients themselves. From their perspective the effectiveness of the treatment would be better evidenced by measures that reflect their quality of life (e.g., the number of days absent from work or the amount of time spent in bed).

A similar problem afflicts research in education, business, social work, sociology, and indeed any subject that measures variables using arbitrary scales. If Betty scores 60 on an intelligence test while Veronica scores 30, it would appear that Betty is smarter. But how much smarter? When the honest answer is "we don't know," the question becomes "so what?" (Andersen et al. 2007). Or consider the management consultant who promises that his weekend course on time management will lead to an average 10-point improvement on a worker efficiency scale. Is 10 points a big improvement? Is it worth paying for? Unless these results can be translated into well-known metrics, there is no easy way to interpret them and our "Research Emperor" has no clothes (Andersen et al. 2007: 666).

A recent flurry of literature on this topic belies the difficulty scholars have with converting arbitrary metrics into meaningful results (Andersen et al. 2007; Blanton and Jaccard 2006; Embretson 2006; Kazdin 2006). Surveys of reporting practices reveal that most of the time social scientists just ignore the interpretation problem altogether. In their review of research published in the *American Economic Review*, McCloskey and Ziliak (1996: 106) found that 72% of the papers surveyed did not ask, how large is large? That is, they reported an effect size (typically a coefficient) but failed to

interpret it in meaningful ways. In a similar study of research published in the *Strategic Management Journal*, the corresponding proportion of studies lacking interpretation was 78% (Seth et al. 2009). In a survey of research in the field of sports psychology, Andersen et al. (2007) found that while forty-four of fifty-four studies reported effect size indexes, only a handful (14%) interpreted those effects in terms of real-world meaning.

If we are to interpret the practical significance of our research results, nonarbitrary reference points are essential. These reference points may come from the measurement scales themselves (e.g., when measuring a well-known index like return on investment, IQ score, or GMAT performance), but this may not be possible when measuring latent constructs like motivation, satisfaction, and depression. Fortunately, there are at least three other ways to interpret these kinds of effects. These methods could be labeled the three Cs of interpretation – context, contribution, and Cohen.

The importance of context

When it comes to interpreting effects, context matters. Consider the case of seven-year-old Law Ho-ming of Hong Kong who died after being admitted to hospital with the flu in March 2008. In normal circumstances the death of a schoolboy, although tragic for the family concerned, is an inconsequential event in the life of a large city. But in this particular case Law's death prompted the government to shut down all the schools for two weeks. Although Hong Kong's health minister claimed that this was nothing more than a seasonal outbreak of influenza, the decision to keep hundreds of thousands of children at home was justified as a precautionary measure. This was Hong Kong after all, the city that became famous as the incubator of the SARS virus in 2003 and where the risk of avian influenza is considered sufficiently serious that nightly news bulletins report on autopsies done on birds found dead in busy neighborhoods.

In the right context even small effects may be meaningful.[3] This could happen one of four ways. First, and as the story of Law Ho-ming illustrates, small effects can be important if they trigger big consequences, such as shutting down hundreds of schools. This is the "small sparks start big fires" rationale. On July 2, 1997, the Thai government devalued the baht, triggering the Asian financial crisis. On September 14, 2008, the financial services firm Lehman Brothers announced it would file for bankruptcy, an event that some argued was a pivotal moment in the subsequent global financial crisis. In both cases prior conditions provided fuel for a fire that only needed to be ignited.[4]

Small effects can trigger big outcomes, even in the absence of pending crises. There is evidence to show that physical appearance can influence the judgment of voters (Todorov et al. 2005), lenders (Duarte et al. 2009), and juries (Sigall and Ostrove 1975).[5] One particularly startling demonstration of the "big consequences" principle was provided in a classic study by Sudnow (1967). Based on his observations of a hospital emergency ward, Sudnow found that the speed with which people were pronounced dead on arrival was affected by factors such as their age, social background, and perceived moral character. For instance, if the attending physician detected the

smell of alcohol on an unconscious patient, he might announce to other staff that the patient was a drunk. This would lead to a less strenuous effort to revive the patient and a quicker pronunciation of death. Sudnow concluded that if one was anticipating a major heart attack and a trip to the emergency ward, one would do well to keep oneself well dressed and one's breath clean. It could mean the difference between being resuscitated or sent to the morgue!

Second, small effects can be important if they change the perceived probability that larger outcomes *might* occur. A funny heart beat might be benign but prompt a radical change in lifestyle because of the thought that a heart attack might occur. The delivery of missiles to Cuba became an international crisis because of what might have occurred if the Soviet Union and the US had not backed down from the brink of war. If the asteroid Apophis were to collide with a geosynchronous satellite in 2029, this might increase the chances that it will subsequently plow into the Atlantic Ocean, destroying life as we know it (see Box 2.2). In the case of the Hong Kong schoolboy, the authorities interpreted his untimely death as signaling an increased risk of an influenza outbreak. No outbreak occurred, but the thought that it might occur compelled the government to interpret the death as an event warranting special attention.

Box 2.2 When small effects are important

Apophis and the end of life on earth
NASA's Near-Earth Object program office has calculated that the 300m wide asteroid Apophis will pass through the earth's gravity field in 2029 and then again in 2036. Some have speculated that a collision with a geosynchronous satellite during the 2029 passing may alter the asteroid's orbit just enough to put it on a collision path with the earth on its return seven years later. If this were to happen, the asteroid will plow into the Atlantic Ocean on Easter Sunday 2036, sending out city-destroying tsunamis and creating a planet-choking cloud of dust. A small collision with a satellite could thus have cataclysmic consequences for life on earth. Although NASA does not endorse these speculations, it has quantified the odds of a collision as being less than 1 in 45,000.

Propranolol and heart attack survival
In 1981 the US National Heart, Lung, and Blood Institute discontinued a study when it became apparent that propranolol, a beta-blocker used to treat hypertension, was effective for increasing the survival rates of heart attack victims. This study was based on 2,108 patients and the difference between the treatment and control groups was statistically significant ($\chi^2 = 4.2$, $p < .05$). Although the effect size was small ($r = .04$), the result could be interpreted as a 4% decrease in heart attacks for people at risk. In a large country such as the US, this could mean as many as 6,500 lives saved each year.

Tiny margins and Olympic medals
Small effects can lead to particularly dramatic outcomes in the sporting arena. At the Beijing Olympics of 2008, American swimmer Dara Torres joked that her second

placing in the 50m freestyle event was a consequence of having filed her fingernails the previous night. She had missed out on a gold medal by 1/100th of a second. A similarly small difference between first and second place in the men's 100m butterfly event was enough to propel Michael Phelps into the history books, earning him his seventh of eight gold medals. In elite sports most interventions (e.g., new swimsuits) yield only tiny effects. But when the competition is close even small differences in performance can lead to dramatic outcomes.

The end of the Premarin party?
Wyeth Pharmaceutical made a fortune selling Premarin to menopausal women. Made from horse's urine, Premarin contains estrogen and is commonly prescribed as a hormone replacement therapy (HRT), useful for alleviating osteoporosis and relieving the symptoms of menopause. HRT has been popular among menopausal women ever since a book entitled *Feminine Forever* was published in 1966. The book's author, Dr. Robert Wilson, advocated HRT as a means for delaying the effects of aging. Thanks to the book – which was paid for by Wyeth – Premarin became America's fifth-leading prescription drug by 1975. However, a large-scale clinical trial involving the drug was wound up prematurely in 2002 when it became apparent that taking estrogen did more harm than good. The researchers found that taking estrogen in combination with progestin led to tiny increases in breast cancer and heart disease. Specifically, the study found that in a group of 10,000 women taking the drug combination for one year, eight more will develop breast cancer, eight more will have strokes, and seven more will have heart attacks in comparison with women not taking the therapy. (The drug combination was also found to lead to six fewer instances of colorectal cancer and five fewer hip fractures.) These risks are numerically miniscule, but potentially deterring. According to one doctor quoted in the *New York Times*, "this is such compelling evidence that women and their physicians ought to be finding ways to get off estrogen."
Sources: FDA (2008), Kolata (1981, 2002), NEO (2008).

Third, small effects can be important if they accumulate into larger effects. During the 2008 US presidential election campaign, Barack Obama suggested that the proper inflation of car tires was a viable strategy for improving America's energy efficiency. Although keeping tires properly inflated improves gas mileage by only 3%, the logic was that if everyone did it the savings would be equivalent to tens of thousands of barrels of imported oil.

Preventive medicine as a specialty discipline exists because small effects can lead to big outcomes when large numbers of people are involved.[6] The beta-blocker propranolol is a classic example. Although the effectiveness of this drug in raising the survival rates of heart attack victims is close to nought, making it available in a large market such as the US means it has the potential to save thousands of lives. Rosenthal (1990: 776) once asked a group of eminent physicians to name a medical breakthrough of "very great practical importance." The learned doctors offered the drug cyclosporine, an immunosuppressant medication that raises the probability that a body will not reject

an organ transplant. Like propranolol, the benefits of this drug in improving patient survival are small ($r = .15$ or $r^2 = .02$) but accumulative. Other life-saving drugs with small effects that accumulate include aspirin, streptokinase, cisplatin, and vinblastine (Rosnow and Rosenthal 2003).

The accumulation of small effects into big outcomes is sometimes seen in sport where the difference between victory and defeat may be nothing more than a trimmed fingernail. In baseball Abelson (1985) found that batting skills explained only one-third of 1% of the percentage of variance in batting performance (defined as getting a hit). Although the effect of batting skill on individual batting performance is "pitifully small," "trivial," and "virtually meaningless," skilled batters nevertheless influence larger outcomes because they bat more than once per game and they bat in teams. As Abelson explained, team success is influenced by batting skill because "the effects of skill cumulate, both within individuals and for the team as a whole" (Abelson 1985: 133).[7]

Fourth, small effects can be important if they lead to technological breakthroughs or new ways of understanding the world. Many important discoveries in science (e.g., Fleming's discovery of penicillin) were the result of events that on other occasions would have passed as insignificant (e.g., moldy Petri dishes). Small, unlikely events were behind the discovery of quinine, insulin, x-rays, the Rosetta Stone, the Dead Sea Scrolls, Velcro, and corn flakes (Roberts 1989). Small effects need not be serendipitous to be significant. Many are the result of meticulous preparation and hard thinking. By removing the handle of the Broad Street water pump the Victorian physician John Snow famously established the link between sewerage-infected water and a localized outbreak of cholera. This small intervention not only saved lives but spawned a whole new branch of medical science: epidemiology.

The contribution to knowledge

Estimates of effects cannot be interpreted independently of their context. In epistemological terms context is described by the current stock of knowledge. Thus, another way to interpret a research result is to assess its contribution to knowledge. Does the observed effect differ from what others have found and if so, by how much? If sample-based studies are estimating a common population effect, and if the size of the effect remains constant, different studies using similar measures and methods should produce converging estimates. Subsequent results of this kind will make an additive contribution of diminishing returns: the more we learn, the more sure we become about what we already know.[8] But if large differences in effect size estimates are observed, and the quality of the research is not in doubt, this may stimulate new and interesting research questions. Are the different results attributable to the operation of contextual moderators? Are studies in fact observing two or more populations, each with a unique effect size?[9] The implication is that the value of any individual study's estimate will be affected by its fit with previous observations. Are we getting a more refined view of the same old thing, or are we getting a glimpse of something new and interesting?

Every doctoral candidate has had the perplexing experience of reading a study known to be a classic and finding it to be peppered with odd methodological choices, dubious analyses, even downright errors. The confused student may seek an explanation from their supervisor: "How can this paper be considered a classic when it is full of mistakes? How could work of such middling quality be published in a top-tier journal?" The supervisor will patiently explain that the paper was groundbreaking in its day, that the analysis, which now appears dated and sub-par, revealed something never seen before. The supervisor will then list all the subsequent and better-done studies that followed in the wake of this pioneering paper.[10] This leads to the next conclusion regarding the interpretation of effects: effects mean different things at different points in time. Studies which hint at new knowledge or which unveil new research possibilities will be more influential and valuable than studies which merely confirm what we already know.

In their list of recommendations to the APA, Wilkinson and the Taskforce on Statistical Inference (1999: 599) argued that the interpretation of effect sizes in the context of previously reported effects is "essential to good research." In the consolidated standards of report trials (CONSORT) used to govern randomized controlled trials in medicine, the twenty-second and final recommendation to researchers is to interpret the results in the context of current evidence (Moher et al. 2001). Many journal editors such as Bakeman (2005: 6) would agree: "In the discussion section, when authors compare their results to others, effect sizes should be mentioned. Are comparable effect sizes found in comparable studies, and if not, why not?" Fitting an independent observation to a larger set of results is the essence of meta-analytical thinking. In the explanatory notes supporting the CONSORT statement, Altman et al. (2001: 685) recommend combining the current result with a meta-analysis or systematic review of other effect size estimates. "Incorporating a systematic review into the discussion section of a trial report lets the reader interpret the results of the trial as it relates to the totality of the evidence." Authors who do this well can make a contribution to knowledge that goes beyond the individual estimate obtained in the study. Different methods for pooling effect size estimates are discussed in Chapter 5.

To assess the contribution to knowledge, authors need to do more than merely compare the results of different studies. They should also entertain alternative plausible explanations (APEs) for the cumulated findings. The researcher should ask, what are the competing interpretations for this result? In classic null hypothesis statistical testing there is only one rival hypothesis – the null hypothesis of chance. But most of the time the null is an easily demolished straw-man, making the contest unfairly biased in favor of the solitary alternative hypothesis. There might yet be other plausible explanations for the observed result. Experimental research seeks to account for these APEs through the randomized assignment of treatments to participants. Randomization is intended to control for an infinite number of rival hypotheses "without specifying what any of them are" (Campbell, writing in Yin 1984: 8). But in nonexperimental settings the explicit identification and evaluation of rival hypotheses is often essential to conclusion drawing (Yin 2000).

The use of plausible rival hypotheses in the interpretation of research results in the social sciences can be traced back to Donald Campbell (Campbell and Stanley 1963: 36; Campbell 1994; Webb et al. 1981: 46).[11] Campbell's big idea was that theories can never be confirmed by data but their degree of confirmation can be gauged by the number of remaining plausible hypotheses. We can never prove that our interpretation is infallible, but we can explicitly identify and rule out some of the alternatives. How? By judging the fit between each competing hypothesis and the data. Alternative explanations may come from the literature, critical colleagues, or stakeholders. An example of a study which does this well is Allison's (1971) analysis of the 1962 Cuban missile crisis. In his book Allison examined the actions of the United States and the Soviet Union through three explanatory lenses: a rational actor model, an organizational process model, and a governmental politics model. In separate chapters the predictions of each theory were compared against the others in terms of their ability to explain the facts of the crisis. Although Allison concluded that the models were complementary, he identified specific aspects of the crisis which were better explained by one model or another. In doing so he challenged the implicit idea that the rational actor model, then popular among political scientists, could provide a stand-alone account of the crisis.[12]

Cohen's controversial criteria

The previous discussion reveals that the importance of an effect is influenced by when it occurs, where it occurs, and for whom it occurs. But in some cases these may not be easy assessments to make. A far simpler way to interpret an effect is to refer to conventions governing effect size. The best known of these are the thresholds proposed by Jacob Cohen. In his authoritative *Statistical Power Analysis for the Behavioral Sciences*, Cohen (1988) outlined a number of criteria for gauging small, medium, and large effect sizes estimated using different statistical procedures. Table 2.1 summarizes Cohen's criteria for several types of effect size.[13] To take the first row as an example, three cut-offs are listed for interpreting effect sizes reported in the form of Cohen's *d*. Referring back to our earlier example of rugby versus soccer fans, a 2-point difference on an IQ test with a standard deviation of 15 equates to a *d* of .13. According to Cohen, this difference is too low to even register as a small effect (i.e., it is below the recommended cut-off of .20). This suggests that Cohen would side with the soccer players in concluding that a 2-point difference in IQ is trivial or essentially meaningless.[14]

Cohen's cut-offs provide a good basis for interpreting effect size and for resolving disputes about the importance of one's results. Professor Brown might believe his correlation coefficient $r = .09$ is superior to Professor Black's result of $r = .07$, but both results would be labeled trivial by Cohen as both are below the cut-off for small effects reported in the correlational form. In the Alzheimer's example mentioned in Chapter 1, the group receiving medication scored on average 13 points higher on an IQ test than the control group. Given that the standard deviation of IQ scores in the population is about 15 points, this difference is equivalent to a *d* of .87 (or 13/15). As this exceeds the recommended cut-off of .80, the observed difference indicates a

Table 2.1 *Cohen's effect size benchmarks*

Test	Relevant effect size	Effect size classes		
		Small	Medium	Large
Comparison of independent means	d, Δ, Hedges' g	.20	.50	.80
Comparison of two correlations	q	.10	.30	.50
Difference between proportions	Cohen's g	.05	.15	.25
Correlation	r	.10	.30	.50
	r^2	.01	.09	.25
Crosstabulation	w, φ, V, C	.10	.30	.50
ANOVA	f	.10	.25	.40
	η^2	.01	.06	.14
Multiple regression	R^2	.02	.13	.26
	f^2	.02	.15	.35

Notes: The rationale for most of these benchmarks can be found in Cohen (1988) at the following pages: Cohen's d (p. 40), q (p. 115), Cohen's g (pp. 147–149), r and r^2 (pp. 79–80), Cohen's w (pp. 224–227), f and η^2 (pp. 285–287), R^2 and f^2 (pp. 413–414).

large effect adding weight to the idea that additional drug trials are warranted. Had the effect been small, any request for further funding would be much less convincing.

Cohen's effect size classes have two selling points. First, they are easy to grasp. You just compare your numbers with his thresholds to get a ready-made interpretation of your result. Second, although they are arbitrary, they are sufficiently grounded in logic for Cohen to hope that his cut-offs "will be found to be reasonable by reasonable people" (1988: 13). In deciding the boundaries for the three size classes, Cohen began by defining a medium effect as one that is "visible to the naked eye of the careful observer" (Cohen 1992: 156). To use his example, a medium effect is equivalent to the difference in height between fourteen- and eighteen-year-old girls, which is about one inch. He then defined a small effect as one that is less than a medium effect, but greater than a trivial effect. Small effects are equivalent to the height difference between fifteen- and sixteen-year-old girls, which is about half an inch. Finally, a large effect was defined as one that was as far above a medium effect as a small one was below it. In this case, a large effect is equivalent to the height difference between thirteen- and eighteen-year-old girls, which is just over an inch and a half.[15]

Despite these advantages the interpretation of results using Cohen's criteria remains a controversial practice. Noted scholars such as Gene Glass, one of the developers of meta-analysis, have vigorously argued against classifying effects into "t-shirt sizes" of small, medium, and large:

> There is no wisdom whatsoever in attempting to associate regions of the effect size metric with descriptive adjectives such as "small," "moderate," "large," and the like. Dissociated from a context of decision and comparative value, there is little inherent value to an effect size of 3.5 or .2. Depending on what benefits can be achieved at what cost, an effect size of 2.0 might be "poor" and one of .1 might be "good." (Glass et al. 1981: 104)

Reliance on arbitrary benchmarks such as Cohen's hinders the researcher from thinking about what the results really mean. Thompson (2008: 258) takes the view that Cohen's cut-offs are "not generally useful" and notes the risk that scholars may interpret these numbers with the same mindless rigidity that has been applied to the $p = .05$ level in statistical significance testing. Shaver (1993: 303) agrees: "Substituting sanctified effect size conventions for the sanctified .05 level of statistical significance is not progress." Cohen himself was not unaware of the "many dangers" associated with benchmarking effect sizes, noting that the conventions were devised "with much diffidence, qualifications, and invitations not to employ them if possible" (1988: 12, 532).

Of the three interpretation routes suggested here, Cohen's criteria are rightly listed last. In an ideal world scholars would normally interpret the practical significance of their research results by grounding them in a meaningful context or by assessing their contribution to knowledge. When this is problematic, Cohen's benchmarks may serve as a last resort. The fact that they are used at all – given that they have no raison d'être beyond Cohen's own judgment – speaks volumes about the inherent difficulties researchers have in drawing conclusions about the real-world significance of their results.

Summary

In many disciplines there is an ongoing push towards relevance and engagement with stakeholders beyond the research community. Academy presidents and journal editors alike are calling for research that is "scientifically valid and practical" (Cummings 2007: 355) and which culminates in the reporting of effect sizes that are "simultaneously helpful to academics, educators, and practitioners" (Rynes 2007: 1048). These are exciting times for researchers who believe their work can and should be used to make the world a better place.

If our research is to mean something it is essential that we confront the challenge of interpretation. Historically researchers have drawn conclusions from their studies by looking at the results of statistical tests. But the importance of a result is unrelated to its statistical improbability. Indeed, statistical significance, which partly reflects sample size, may say nothing at all about the practical significance of a result. With this in mind the editors of many journals have begun pushing for the reporting of effect sizes. Knowing the size of an effect is a necessary but insufficient condition for interpretation.

To extract meaning from their results social scientists need to look beyond *p* values and effect sizes and make informed judgments about what they see. No one is better placed to do this than the researcher who collected and analyzed the data (Kirk 2001). The fact that most published effect sizes go uninterpreted shows that many researchers are either unable or reluctant to take this final step. Most of us are far more comfortable with the pseudo-objectivity of null hypothesis significance testing than we are with making subjective yet informed judgments about the meaning of our results. But led by Cohen and others like him we have already begun to steer a new course. The highly

cited researcher of tomorrow may well be the one who seizes these opportunities to explore new avenues of significance and meaning.

Notes

1 Television shows purporting to test national IQs have been broadcast in Europe, North America, Asia, and the Middle East. In a recent BBC version of this show some interesting group differences emerged: men scored three IQ points higher than women, participants aged 70 or above scored 11 points higher than 20-somethings, right-handed people did marginally better than left-handed people, and participants from Scotland did better than participants from anywhere else in the United Kingdom (BBC 2007).

2 Something which is often heard but is inaccurate is the claim that a statistically significant result reveals a real effect. This will be true most of the time but not all of the time for reasons explained in Chapter 3. A statistically significant result means the evidence is sufficient, in terms of some adopted standard (e.g., $p < .05$), for rejecting the null hypothesis. But the only way we can say for sure that a result is real is to replicate. Not only is reproducibility the litmus test of whether a result is real, but "replicated results automatically make statistical significance unnecessary" (Carver 1978: 393).

3 A distinction should be made between small real-world *effects* and small sample-based effect size *estimates*. In new or poorly understood areas of research, estimates of effect size tend to be undermined by measurement attenuation and by the inability of researchers to properly decipher causal complexity. Small observed effects may thus reflect measurement error.

4 In the case of the Asian financial crisis the prior conditions were defined, in part, by trade imbalances between Southeast Asian nations and both China and Japan that led to massive trade deficits and rising interest rates. In the three years preceding the crisis both the Chinese yuan and the Japanese yen had fallen in value relative to the US dollar. As the currencies of some Southeast Asian nations were pegged to the US dollar, regional exporters found it increasingly difficult to compete in the Japanese market. Not only were their exports becoming relatively more expensive (thanks to the declining yen), but they were being outsold by Chinese rivals (thanks to the declining yuan). Worsening trade deficits had to be paid for by borrowing money, putting pressure on local currencies. To preserve their currency pegs in the face of a strong US dollar, Southeast Asian governments had to raise interest rates, making it harder for local businesses to finance investment. But the underlying economics were unsustainable: the US was booming while Southeast Asia was hemorrhaging capital. A lack of confidence led to capital flight, compelling governments to raise interest rates even further, even those with free-floating currencies. Speculators smelled blood and began short-selling Asian currencies. Thailand was the first to fold. After it pulled the peg on July 2, 1997 its currency dropped 60% relative to the US dollar. Then the Philippine peso fell 30%, the Malaysian ringgit lost 40%, and Indonesia's rupiah lost 80% of its pre-crash value. Within a year the economies of Thailand, Malaysia, and the Philippines had all contracted by about 40% while Indonesia's economy had shrunk by 80%. It would be years before these economies began to recover.

5 In a simulated trial Efran (1974) found that good-looking defendants were less likely to be judged guilty and received less punishment than unattractive defendants. So when you have your day in court, wear something nice.

6 Because it deals in important outcomes (e.g., lives saved), medicine provides many examples of important small effects. Drugs that have only tiny effects are fast-tracked through the certification process because of their potential to radically change the quality of life for a few.

7 Coined by Cohen (1988: 535), the term "Abelson's Paradox" describes how trivial effects can accumulate into meaningful effects over time.

8 The diminishing returns of replicated results were quantified by Schmidt (1992: 1180) when he noted that "the first study conducted on a research question contains 100% of the available research information, the second study contains roughly 50%, and so on. Thus, the early studies in any area have a certain status. But the 50th study contains only about 2% of the available information, the 100th, about 1%."

9 Methods for determining whether one or more populations are being observed are described in Appendix 2.

10 A good example of this trend is the body of research surveying the statistical power of published studies. (This research is reviewed in Chapter 4.) The first such survey was Cohen's (1962) assessment of research published in the *Journal of Abnormal and Social Psychology*. Tversky and Kahneman (1971: 107) called Cohen's survey an "ingenious study" and the dozens of authors who have since been inspired by it would probably agree. But in many respects Cohen's pioneering work has been put in the shade by its successors. Cohen reviewed only a year's worth of research published in a single journal, making it difficult to comment on trends. Subsequent authors have generally reviewed many years' worth of research published in multiple, related journals within a discipline. For example, Brock (2003) surveyed eight volumes of four business journals, while Lindsay (1993) surveyed eighteen volumes of three management accounting journals.

11 In the natural sciences the use of alternative hypotheses goes back to Platt (1964), Popper (1959), Chamberlin (1897), and Francis Bacon in the early seventeenth century.

12 For more on the use of APEs in the interpretation of results, see Dixon (2003), Perrin (2000), Yin (2000), and Campbell's foreword to Yin's (1984) book.

13 Supplementing Cohen's (1988) small, medium, and large effect sizes, Rosenthal (1996) adds a classification of very large, defined as being equivalent to or greater than $d = 1.30$ or $r = .70$. Rosenthal also offers qualitative size categories for odds ratios and differences in percentages. Different odds ratios he classifies as follows: small (∼1.5), medium (∼2.5), large (∼4.0), and very large (∼10 or greater). Percentage difference is simple to use but tricky to interpret: "the difference between 2% and 12% (10 points) represents a difference of 0.88 standard deviations while that between 40% and 50% (also 10 points) represents 0.25 standard deviations" (1996: 51). Accordingly, Rosenthal proposes size conventions that apply only in the 15–85% range, as follows: small (∼7 points), medium (∼18 points), large (∼30 points), and very large (∼45 points or more). To interpret differences between percentages outside this 15–85% range, Rosenthal recommends using the odds ratio.

14 It is worth reiterating that a 2-point gap in this example is not meaningless because it is just 2 points but because the variability in the distribution of scores is much larger than 2 points. If the standard deviation of an IQ test was 1.5 points, instead of 15 points, then a 2-point difference in IQ would be very large indeed.

15 If you think these are odd examples on which to build a convention, consider the cut-offs proposed by Karl Pearson (1905). In his view, a *high* correlation ($r \geq .75$) was equivalent to the correlation between a man's left and right thigh bones; a *considerable* correlation ($.50 < r < .75$) was equivalent to the association between the height of fathers and their sons; a *moderate* correlation ($.25 < r < .75$) was equivalent to the association between the eye color of fathers and their daughters; and a *low* correlation ($r \leq .25$) was equivalent to the association between a woman's height and her pulling strength!

Part II

The analysis of statistical power

3 Power analysis and the detection of effects

When I stumbled on power analysis . . . it was as if I had died and gone to heaven. ∼ Jacob Cohen (1990: 1308)

The foolish astronomer

An astronomer is interested in building a telescope to study a distant galaxy. A critical factor in the design of the telescope is its magnification power. Seen through a telescope with insufficient power, the galaxy will appear as an indecipherable blur. But rather than figure out how much power he needs to make his observations, the astronomer foolishly decides to build a telescope on the basis of available funds. Maybe he does not know how much magnification power he needs, but he knows exactly how much money is in his equipment budget. So he orders the biggest telescope he can afford and hopes for the best.

In social science research the foolish astronomer is the one who sets sample sizes on the basis of resource availability. He is the one who, when asked "how big should your sample be?", answers "as big as I can afford." Resource constraints are a fact of research life. But if our goal is to conserve limited resources, it is essential that we begin our studies by asking questions about their power to detect the phenomena we are seeking. How big a sample size do I need to test my hypotheses? Assuming the phenomenon I'm searching for is real, what are my chances of finding it given my research design? How can I increase my chances? My sample size is only 50 (or 30 or 200); do I have enough power to run a statistical test? Power analysis provides answers to these sorts of questions.

The improbable null

In the Alzheimer's study introduced in Chapter 1, the researcher was interested in testing the hypothesis that a certain treatment would lead to improved mental health. Against this hypothesis stands another, often unstated, hypothesis: that the treatment will have no effect. In any study the "no effect" hypothesis is called the *null* hypothesis (H_0), while the hypothesis that "there is an effect" is called the *alternative* hypothesis

(H_1). Expressed in terms of effect size, the classic null hypothesis is always that the effect size equals zero, while the alternative hypothesis is that the effect size is nonzero.[1]

In undergraduate statistics classes students are taught how to run tests assessing the truthfulness of the null hypothesis. Using probability theory, statistical tests can be done to determine how likely a result would be if there was no underlying effect. The outcome of any test is a conditional probability or *p* value which is the probability of getting a result at least this large if there was no underlying effect. If the *p* value is low (e.g., $<.05$), the result is said to be statistically significant, permitting us to reject the null hypothesis of no effect. In the Alzheimer's study a statistical test would have been used to calculate the probability that the observed result was attributable to variations within the sample.[2] Such a test would answer the question: what are the chances that the 13-point gain in IQ is attributable to random fluctuations in the data? In this case the *p* value (.14) was not low enough to achieve statistical significance, so the null could not be rejected as false.[3] Perversely, this does not mean the null could be accepted as true. In practice null hypotheses are virtually never true and even if they were, statistical testing could not permit you to accept them as such.[4] About the only thing a statistical test can do with confidence is tell you when a null is probably false, which we usually already know. This limitation is one of many that have given rise to a "long and honorable tradition of blistering attacks on the role of significance testing" (Harris 1991: 377). A brief summary of these criticisms is provided in Box 3.1. Yet despite its many limitations significance testing persists because it provides a basis for checking that our results obtained from samples are not due to random fluctuations in the data.[5]

Given the two competing hypotheses – the null and the alternative – it is not hard to see that there are two possible errors researchers can make when drawing conclusions. They might wrongly conclude that there is an effect when there isn't (known as a Type I error), or they might conclude that there is no effect when there is (a Type II error). Type I errors, also known as false positives, occur when you see things that are not there. Type II errors, or false negatives, occur when you don't see things that are there (see Figure 3.1).

The need for error insurance

Type I errors – seeing things that are not there – are easier to make than you might think. The human brain is hardwired to recognize patterns and draw conclusions even when faced with total randomness. Conspiracy theorists, talk-show hosts, astrologers, data-miners and over-zealous graduate students can easily make these types of errors. Even distinguished professors have been known to draw spurious conclusions from time to time. Box 3.2 provides some examples of famous false positives.

Unfortunately, Type I errors happen to the best of us and this is why Sir Ronald Fisher decided long ago that we needed standards for deciding when a result is sufficiently improbable as to warrant the label "statistically significant" (Fisher 1925). For any test, the probability of making a Type I error is denoted by the Greek letter *alpha* (α).

Box 3.1 The problem with null hypothesis significance testing

Undergraduates taking statistics classes are routinely taught to test the null hypothesis of no effect. That is, they learn the rules which determine the conditions under which the null hypothesis can be rejected. But there are numerous shortcomings with this classical testing approach.

First, treatments will always have some tiny effect and as these effects will be detected in studies with sufficient power, the null hypothesis doesn't stand a chance. As long as a statistical test is powerful enough, it will be impossible not to reject the null. It makes little sense to test the null unless there are a priori grounds for believing the null hypothesis is true – which it almost never is.

Second, *p* values are usually (and wrongly) interpreted in such a way that hypotheses are rejected or accepted solely on the basis of the $p < .05$ cut-off. If the test result is statistically significant an effect is inferred. If the result is not significant then this is taken as evidence of no effect. But as Rosnow and Rosenthal (1989: 1277) argue, this practice of "dichotomous significance testing" is a pseudo-objective convention without an ontological basis. Alpha levels fall on a continuum and "surely, God loves the .06 nearly as much as the .05."

Third, the *p* value is a confounded index that reflects both the size of the effect and the size of the sample. Hence any information included in the *p* value is ambiguous (Lang et al. 1998). A statistically significant *p* value could reflect either a large effect or a large sample or both. Consequently *p* values cannot be used to interpret either the size or the probability of observed effects.

Fourth, even when it is achieved, statistical significance is no guarantee that a result is real. Some proportion of false positives arising from sampling variation is inevitable. The best test of whether a result is real is whether it can be replicated at different times and in different settings.

For more on the limitations associated with classical null hypothesis testing, see Abelson (1997), Bakan (1966), Carver (1978), Cortina and Dunlap (1997), Falk and Greenbaum (1995), Gigerenzer (2004), Harlow et al. (1997), Hunter (1997), Johnson (1999), Kline (2004, Chapter 3), Meehl (1967, 1978), Shaver (1993), and Ziliak and McCloskey (2008).

Alpha can range from 0 to 1, where 0 means there is no chance of making a Type I error and 1 means it is unavoidable. Following Fisher, the critical level of alpha for determining whether a result can be judged statistically significant is conventionally set at .05.[6] Where this standard is adopted the likelihood of making a Type I error – or concluding there is an effect when there is none – cannot exceed 5%. This means that out of a group of twenty scholars all searching for an effect that actually does not exist, only one is likely to make a fool of himself by seeing something that is not there.[7]

If good statistical practice is followed and alpha levels are set sufficiently low, the probability of making a Type I error is kept well below the cringe threshold. The temptation might then be to set alpha levels as stringently as possible. Lowering critical

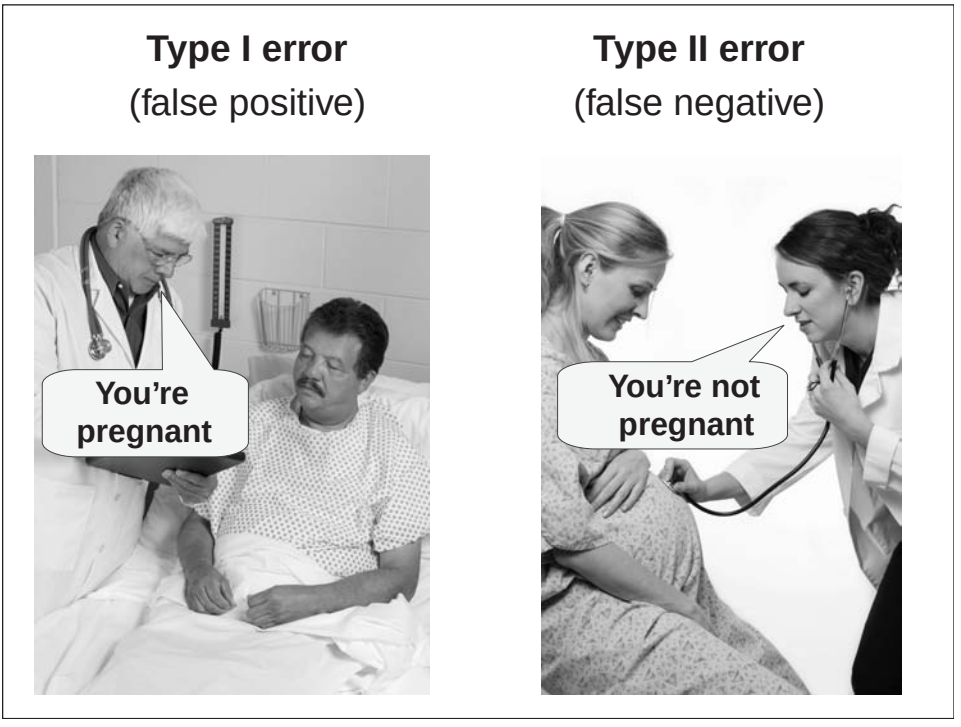


Figure 3.1 Type I and Type II errors

alpha levels to .01 or even .001 means the risk of making a Type I error falls to 1% and .1% respectively. But when we tighten alpha levels we simultaneously raise our chances of making a Type II error. Type II errors, or not seeing things that are there, are very common, as we will see in the next chapter. For any test, the probability of making a Type II error is denoted by the Greek letter *beta* (β).

Few researchers seem to realize that alpha and beta levels are related, that as one goes up, the other must go down. This ignorance is manifested in the unquestioning allegiance to the $p = .05$ level of significance and in the pride some researchers seem to take in studding their results with asterisks. In other words, all the attention is given to minimizing alpha. But while alpha safeguards us against making Type I errors, it does nothing to protect us from making Type II errors. A well thought-out research design is one that assesses the relative risk of making each type of error, then strikes an appropriate balance between them. We will return to this point below.

It is also important to note that both alpha and beta are conditional probabilities: alpha is the conditional probability of making an error when the null hypothesis is true while beta is the conditional probability of making an error when the null hypothesis is false. Because the null hypothesis cannot be both true and false, in any given test only one type of error is possible. A study cannot be afflicted by a little bit of alpha error and a little bit of beta error. If the null hypothesis is false it will be impossible

Box 3.2 Famous false positives

Astrological injuries

In a large-scale investigation of Canadian hospital records, evidence was found linking birth dates with medical afflictions (Austin et al. 2006). For example, people born under the astrological star sign Leo were found to be 15% more likely to be admitted to hospital for gastric bleeding, while Sagittarians were 38% more likely to go to hospital for broken arms. However, the authors of this study recognized these false positives for what they were. In fact, they had deliberately sought them out to show that testing multiple hypotheses increases the likelihood of detecting implausible associations.

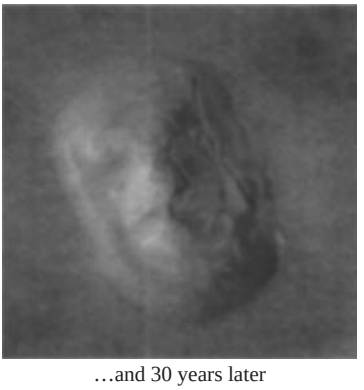
The Super Bowl stock market predictor

Historical evidence shows a correlation between the performance of the US Dow Jones Index and the outcome of the Super Bowl. This link has caused some to jump to the inductive conclusion that the two events are causally related: when a team from the old American Football League (now the American Football Conference) wins the Super Bowl, stock prices fall; when a team from the old National Football League (now the National Football Conference) wins, prices rise. All sorts of creative explanations have been offered to account for this relationship, but the link is most likely spurious. In this case the false positive is not the correlation but the conclusion that football performance affects stock market performance.

The Cydonian Face

A photograph taken by the Viking spacecraft in 1976 revealed a face-like shape on the surface of the planet Mars. Some took this image to be evidence of a vanished civilization. Others maintained that the image was an optical illusion or a geological fluke. Those in the first group thought those in the second were making a Type II error ("how can you not see it?") while those in the second thought those in the first were making a Type I error ("it's probably just a trick of light"). Subsequent imagery obtained in July 2006 through the Mars Express Probe of the European Space Agency supported the non-believers (ESA 2006). The new high-resolution evidence confirmed skeptics' conclusion that the Martian face is nothing more than a figment of human imagination.

The 1976 image…



…and 30 years later

to make a Type I error and if the null is true it will be impossible to make a Type II error. The problem is we often do not know whether the null is true or false so we do not know which type of error we are more likely to make. Most of the time we need an insurance policy that covers both error types. But sometimes there is prior evidence that an effect really exists. On such occasions an exclusive emphasis on alpha that leads to the neglect of beta is the height of folly. If an effect actually exists, the probability of making a Type I error is zero. When effects are there to be found, the only error that can be made is a Type II error, and the only way that can occur is if our study lacks statistical power.

Statistical power

Statistical power describes the probability that a test will correctly identify a genuine effect. Technically, the power of a test is defined as the probability that it will reject a false null hypothesis. Thus, power is inversely related to beta or the probability of making a Type II error. In short, power $= 1 - \beta$.

Every statistical test has a unique level of power. Other things being equal, a test based on a large sample has more statistical power (or is less likely to fall prey to Type II error) than a test involving a small sample. But how large should a sample be? If the sample is too small, the study will be underpowered, increasing the risk of overlooking meaningful effects. Consider the aspirin study discussed in Chapter 1. In that study the benefits of aspirin were found to be both small ($r^2 = .001$) and important. But the odds are that this tiny effect would have been missed if the sample had had fewer than the minimum 3,323 participants needed to detect an effect of this size.[8] But in another setting, 3,323 observations might generate far more power than necessary to detect an effect.

Both under- and overpowered studies are inefficient. Underpowered studies waste resources as they lack the power needed to reject the null hypothesis.[9] As nonsignificant results are sometimes wrongly interpreted as evidence of no effect, low-powered studies can also misdirect further research on a topic. Underpowered studies may even be unethical if they involve subjecting individuals to inferior treatment conditions. Where studies lack the power to resolve questions of treatment effectiveness, the risk of exposure to inferior treatments may not be justifiable (Halpern et al. 2002). But overpowered studies can also be wasteful and misleading. For example, any study with more than 1,000 observations will be more than capable of detecting essentially trivial effects (defined as $r < .10$ or $d < .20$). This possibility may arise when hypotheses are tested using large databases with thousands of data points. Being highly powered, such studies are apt to yield statistically significant results that are essentially meaningless (see Box 3.3). Of course researchers who are in the habit of interpreting effect sizes directly will not fall into the trap of imputing importance on the basis of p values. The wastefulness of overpowered studies lies not in the amount of data collected (the more the better!) but in the possibly unnecessary expenditure of resources. A study becomes wasteful when the costs of collecting data needed to accurately estimate effects exceed the benefits of doing so.

Box 3.3 Overpowered statistical tests

Researchers sometimes compare groups to see whether there are meaningful differences between them and, if so, to assess the statistical significance of these differences. The statistical significance of any observed difference will be affected by the power of the statistical test. As statistical power increases, the cut-offs for statistical significance fall. Taken to an extreme this can lead to the bizarre situation where two essentially identical groups are found to be statistically different. Field and Wright (2006) provide the following SPSS-generated results showing how this situation might arise:

t	df	Sig. (2-tailed)	Mean difference
−2.296	999998	.022	.00

The number in the last column tells us that the difference between two groups on a particular outcome is zero, yet this "difference" is statistically significant at the $p < .05$ level. How is it possible that two identical groups can be statistically different? In this case, the actual difference between the two groups was not zero but −.0046, which SPSS rounded up to .00. Most would agree that −.0046 is not a meaningful difference; the groups are essentially the same. Yet this microscopic difference was judged to be statistically significant because the test was based on a massive sample of a million data-points. This demonstrates one of the dangers of running overpowered tests. A researcher who is more sensitive to the *p* value than the effect size might wrongly conclude that the statistically significant result indicates a meaningful difference.

What, then, is an appropriate level of statistical power? This is not an easy question to answer as it involves a trade-off between risk and return. A couple of thought experiments will illustrate the dangers and costs of setting power too low or too high. If power is set to .50, this means a study has a 50–50 chance of rejecting a null hypothesis that happens to be false. If research success is defined as finding *something*, a study with power = .50 has, at best, a coin-flip's chance of being successful.[10] Should such a study be done? Would you commit to a multi-year research project if your chances of success were the same as tossing a coin? Most researchers would not find these odds agreeable.

If power is set at a higher level, say .90, then the chances of detecting effects are greatly improved. To be exact, the chances of making a Type II error are reduced to 10%. But statistical power is costly. To detect a small effect of $r = .12$ using a nondirectional test with alpha levels set at $p < .05$ and beta set at .10 would require a sample of $N = 725$. The question that must be asked is: does the nature of the effect warrant the expense required to uncover it?

There is nothing cast in stone regarding the appropriate level of power, but Cohen (1988) reasoned that the power levels should be set at .80. This means studies should be

designed in such a way that they have an 80% probability of detecting a real effect (or a 20% probability of making a Type II error).[11] Why 80%? Because Cohen believed that this would strike a reasonable balance between alpha and beta risk. Cohen explained that most scientists would view Type I errors to be more serious than Type II errors and therefore deserving of more stringent safeguards. "The notion that failure to find is less serious than finding something that is not there accords with the conventional scientific view" (1988: 56). Consequently Cohen proposed that Type I errors should be treated four times more seriously than Type II errors. If alpha significance levels are set at .05, then beta levels should be set at .20. If we can tolerate a 5% chance of a Type I error, then we should be able to tolerate a 20% chance of a Type II error.

Cohen's recommendation was timely and convincing. Researchers now had a standard for setting power (.80) that complemented their long and dearly held attachment to Fisher's alpha-significance criterion (.05). Together the two numbers became known as the five-eighty convention. This new convention was appealing as it conveyed a sense of objectivity while enabling the researcher to side-step the tricky challenge of balancing alpha and beta risk (Di Stefano 2003). But Cohen would have been appalled at the conventionalization of his recommendation. In his mind power levels of .80 were no more special than p values of .05 (Cohen 1994). The numbers were merely guidelines intended to make researchers think about the need to balance two competing types of risk. It was always Cohen's hope that his recommendation would be ignored by thoughtful researchers who had considered the relative risk of each error type and struck a balance appropriate to their studies.

Cohen's four-to-one weighting of beta-to-alpha risk serves as a good default that will be reasonable in many settings. But it is not difficult to conceive of research scenarios where the four-to-one ratio will represent a gross misallocation of risk. Consider a hypothetical study comparing the effectiveness of a drug with a placebo on some outcome for two groups of twenty patients. Given that the drug either has an effect or it doesn't, and given that the results of the study will either lead us to conclude that we see these effects or we don't, there are four possible conclusions that can be reached from this study (see Figure 3.2). If there is no effect (i.e., the treatment is ineffective) we will either come to the correct conclusion or we will incorrectly reject a true null hypothesis, making a Type I error of commission. Conversely, if there is a genuine effect (i.e., the treatment does work) we will either draw this conclusion or we will incorrectly accept a false null hypothesis, making a Type II error of omission. Suppose that our study is a replication study and that previous research reveals that the drug has a genuine effect equivalent to half a standard deviation ($d = .50$). What is the probability that we will come to the wrong conclusion given the design of our study? A reader mindful of alpha levels might conclude that the risk of making an error is 5%, but in fact the probability of a Type I error is zero. There is no chance that we can falsely conclude there is an effect when in fact there is an effect. In this study the only error that can be made is a Type II error. As it happens the probability of a Type II error in this study is a hefty 66%. (The maths will be explained below.) If

What conclusion is reached by the researcher?	What is true in the real world?	
	There is no effect (null = true)	There is an effect (null = false)
No effect (ES = 0)	Correct conclusion ($p = 1 - \alpha$)	Type II error (p = b)
There is an effect (ES ≠ 0)	Type I error (p = a)	Correct conclusion (p = 1– b)

Figure 3.2 Four outcomes of a statistical test

we were to proceed with this study without addressing issues of statistical power, we would be setting ourselves up to fail.[12]

In this example a four-to-one emphasis on alpha risk is not appropriate because we have prior reasons for believing that there is almost no chance of committing a Type I error. Past research tells us there is an effect to be found. But an analysis of this study's statistical power shows that there is a massive risk of making a Type II error. Given these costs (the high risk of a Type II error) and the benefits (a zero risk of a Type I error), it would be irrational to set alpha at the conventional level of .05. Doing so would be an expensive and needless drain on statistical power. A more rational approach would be to balance the error rates or even swing them in favor of protecting us against making the only type of error that can be made.[13] Some other examples of when it would be inappropriate to follow the five-eighty convention are provided in Box 3.4.

Few authors explicitly assess the relative risk of Type I and II errors, but any decision about alpha implies a judgment about beta. Sometimes the choice of a particularly stringent alpha level (e.g., $\alpha = .01$) is interpreted as being scientifically rigorous, while the adoption of a less rigorous standard (e.g., $\alpha = .10$) is considered soft. But this is misguided. As we will see in the next chapter, blind adherence to conventional levels of alpha has meant that beta levels in published research often rise to unacceptable levels. Surveys of statistical power reveal that many studies are done with less than a 50–50 chance of finding sought-after effects. When this practice is combined with publication biases favoring statistically significant results, the paradoxical outcome is an increase in the Type I error rates of published research, the very thing researchers hoped to avoid.

In many research areas the accumulation of knowledge leads to a better understanding of an effect and a reduction in the likelihood of Type I errors. The chance that the null is true diminishes with understanding. The implication is that as research in a field advances, researchers should pay increasing attention to Type II errors and statistical power (Schmidt 1996).

Box 3.4 Assessing the beta-to-alpha trade-off

The desired ratio of beta-to-alpha risk should be informed by the type of risk being considered. Medical testing done for screening purposes provides a fertile area for assessing this trade-off. Many medical tests are designed in such a way that virtually no false negatives (Type II errors) will be produced. This inevitably raises the risk of obtaining a false positive (Type I errors). Designers of these tests are implicitly saying that it is better to tell a healthy patient "we may have found something – let's test further" than to tell a diseased patient "all is well."

But in another setting the occurrence of a single Type II error may be extremely costly. Mazen et al. (1987a: 370) illustrate this with reference to the space shuttle *Challenger* explosion. Prior to launching the doomed shuttle NASA officials faced a choice between two assumptions, each with a unique risk and cost. The first assumption was that the shuttle was unsafe to fly because the performance of the O-ring in the booster was different from previous missions. The second assumption was that the performance of the O-ring was no different and therefore the shuttle was safe to fly. Had the mission been aborted and the O-ring was found to be functional, a Type I error would have been committed. The cost of this error would have been the cost of postponing a shuttle launch and carrying out unnecessary maintenance. As it happened, the shuttle was launched with a defective O-ring and a Type II error was made, leading to the loss of seven astronauts and the immediate suspension of the shuttle program. In this case the cost of the Type II error far exceeded the cost of incurring a Type I error.

The analysis of statistical power

Power analysis answers questions like "how much statistical power does my study have?" and "how big a sample size do I need?". Power analysis has four main parameters: the effect size, the sample size, the alpha significance criterion, and the power of the statistical test.

1. The effect size describes the degree to which the phenomenon is present in the population and therefore "the degree to which the null hypothesis is false" (Cohen 1988: 10).
2. The sample size or number of observations (N) determines the amount of sampling error inherent in a result.[14]
3. The alpha significance criterion (α) defines the risk of committing a Type I error or the probability of incorrectly rejecting a null hypothesis. Normally alpha is set at $\alpha = .05$ or lower and statistical tests are assumed to be nondirectional (two-tailed).[15]
4. Statistical power refers to the chosen or implied Type II error rate (β) of the test. If an acceptable level of β is .20, then desired power = .80 (or $1 - \beta$).

The four power parameters are related, meaning the value of any parameter can be determined from the other three. For example, the power of any statistical test can be expressed as a function of the alpha, the sample size, and the effect size. If the effect being sought is small, the sample is small, and the alpha is low, the resulting power of the test will be low. This is because small effects are easy to miss, smaller samples generate noisier datasets on account of sampling error, and low alphas (e.g., .01) make it harder for researchers to draw conclusions about effects they may or may not be seeing. Conversely, power will be higher for tests involving larger effects, bigger samples, and more relaxed alphas (e.g., .10).

Prospective power analyses

Power analyses are normally run before a study is conducted. A prospective or a priori power analysis can be used to estimate any one of the four power parameters but is most often used to estimate required sample sizes. In other words, sample size is cast as a dependent variable contingent upon the other three parameters.

The value of prospective power analysis can be illustrated with reference to the hypothetical Alzheimer's study described in Chapter 1. In that study the researcher conducted a test which returned an interesting result but which failed to rule out the possibility of Type II error. This most likely occurred because her total sample size ($N = 12$) was too small to detect an effect of this size. But how big should the test groups have been? Suppose she decides to repeat the study taking her first test as a pretest. Based on this pretest she might speculate that the effect of taking the medication is worth 13 IQ points, this being the result she has already obtained. If she sets power at .80 and alpha at .05 for a two-tailed test, an a priori power analysis will reveal that she will need to compare two groups of at least twenty patients each to detect an effect of this size. However, if she decides that a one-tail test is sufficient (she has reason to believe the drug only has a positive effect), she will need only sixteen patients in each group.[16] Of course these numbers are the bare minimum. If the effect is actually smaller than she anticipates, or if her measurement is unreliable, she will need a bigger sample to mitigate the loss in power.

Prospective power analyses can also be run to determine the likelihood of making a Type II error in a planned study. In other words, power is cast as an outcome contingent upon effect size, sample size, and alpha. Had our Alzheimer's researcher done this type of analysis, she might not have proceeded with her original study for she would have learned that the power to detect was only .31. In other words, the risk of making a Type II error was 69%. Prospective analyses can also be used to identify the minimum detectable effect size associated with a particular research design. In the underpowered Alzheimer's case the smallest effect size that could have been labeled statistically significant was a difference of 1.80 standard deviations. In other words, the reliance on small groups meant the researcher would not have been able to rule out sampling error as a source of bias unless the difference between the groups was at least 25 IQ points. Finally, a prospective analysis can be run to determine the alpha level that would

be required to achieve statistical significance, given the other three parameters. If she had done this she may have been shocked to learn that her results would not achieve statistical significance unless the critical alpha level was set at a high .44. However, this is the result that would be achieved with a two-tailed test and with power levels set at .80. Knowing that she had access to only twelve patients, the researcher might have felt that a little more latitude was warranted. If she settled for a one-tailed test and was prepared to accept a 30% beta risk (i.e., power = .70), then the cut-off for determining statistical significance falls to $\alpha = .15$. As it happens, her results fell on the right side of this threshold ($p = .14$). But whether she could convince a reviewer that she had adequately ruled out Type I error by adopting an unconventionally relaxed alpha level is another story!

Prospective power analyses are particularly useful when planning replication studies. By analyzing the effect and sample sizes of past research on a particular topic a researcher can make informed decisions about studies that aim to replicate or build upon earlier work. Suppose a researcher wishes to investigate a relationship between a particular X and Y. A review of the literature reveals two other studies that have examined this relationship. These studies reported correlations of .20 and .24, but in both cases the results were found to be statistically nonsignificant. The researcher suspects that the nonsignificance of these results was a consequence of insufficient statistical power. She notes that the two studies had sample sizes of seventy-eight and sixty-three respectively. Before deciding to retest this hypothesis she consults some power tables to find the sample size that would give her an 80% chance of detecting an effect size that she estimates is exactly midway between these two sample-based estimates (i.e., $r = .22$) with two-tailed alpha levels (α_2) set at .05. She learns that she will need a minimum sample size of 159. This number is greater than the combined samples of the two previous studies, reinforcing her impression that both were underpowered. The researcher has now positioned herself to make two valuable contributions to the literature. First, if she proceeds to conduct an adequately powered study she has a good chance of finding a statistically significant relationship where others have found none. Second, if she finds an effect size close to her expectations, she will have good grounds for reinterpreting the inconclusive results of the earlier studies as Type II errors arising from insufficient power. As a result of her study she may be able to revitalize interest in a relationship that others may have mistakenly dismissed as a dead-end.

The perils of post hoc power analyses

Power analyses can be helpful during the design stages of a study. In addition, power analyses are sometimes run retrospectively after the data have been analyzed and typically when the results turn out to be statistically nonsignificant. However, as we will see, analyzing the power of a study power based on data obtained in that study is usually a waste of time.

When a study returns a nonsignificant result, there is a "powerful" temptation to find out whether the study possessed sufficient statistical power. The researcher wonders:

"did my study have enough power to find what I was looking for?" A variation on this is: "my sample size was evidently too small – how much bigger should it have been?" Sometimes these sorts of questions are put to authors by journal editors. According to Hoenig and Heisey (2001), nineteen journals advocate the analysis of post-experiment power. The rationale is that a nonsignificant result returned from an underpowered test might constitute a Type II error rather than a negative result. Even though our significance test will not let us reject the null hypothesis of no effect, an effect might none the less exist.

Nonsignificant results are a researcher's bane and running a power analysis prior to a study is no guarantee that results will turn out as expected. Prospective analyses hinge on anticipating the correct effect size, but if effects are smaller than expected, then resulting power may be inadequate. Reassessing power based on the observed rather than the estimated effect size is sometimes done to determine actual power as opposed to planned power. If it can be shown that power was low, the researcher might conclude: "the results are not significant but that was because the test was not sufficiently powerful." This is called the "fair chance" rationale for post hoc analysis; if power levels were too low, then null hypotheses were not given a fair chance of rejection (Mone et al. 1996). The implication is that the conclusion of no result should not be entertained and that further, more powerful, research should follow. However, if power levels are found to be adequate, then the researcher can rule out low power as a rival explanation and definitively conclude that the result was negative. In the case of the Alzheimer's study, a retrospective analysis based on the observed effect size reveals that actual power was a low .31. The researcher – should she be unacquainted with the perils of post hoc analysis – might conclude that her nonsignificant result was a consequence of insufficient power. This would be like saying "even though my results don't say so, I believe an effect really does exist." Indeed, she may have good grounds for believing this (e.g., other studies or her expert intuition), but it is incorrect to draw this conclusion from a power analysis.

The post hoc analysis of nonsignificant results is sometimes painted as controversial (e.g., Nakagawa and Foster 2004), but it really isn't. It is just wrong. There are two small technical reasons and one gigantic reason why the post hoc analysis of observed power is an exercise in futility. The two technical concerns relate to the use of observed effect sizes and reported *p* values.

Retrospective analyses based on observed effect sizes make the dubious assumption that study-specific estimates are identical to the population effect size. The analyst may look at the correlation matrix to find the appropriate *r* or convert observed differences between groups to a Cohen's *d* and then calculate the power of the test (see, for example, Katzer and Sodt (1973) and Osborne (2008b)). But observed effect sizes are likely to be poor estimates of population effect sizes, particularly if they are based on samples that are small and biased by sampling error.[17] Can our Alzheimer's researcher be confident that the observed difference between the groups is unaffected by random variations within her small sample? If we cannot rely on the accuracy of our effect size estimates, then there is little to be gained in using them to calculate observed power.

Some have argued that post hoc power analyses are warranted for statistically nonsignificant results, that is, when *p* values are relatively high (Erturk 2005; Onwuegbuzie and Leech 2004). But calculating observed power on the basis of reported *p* values is pointless as there is a one-to-one correspondence between power and the *p* value of any statistical test (Hoenig and Heisey 2001). As *p* goes up, power goes down, and vice versa. A nonsignificant result will almost always be associated with low statistical power (Goodman and Berlin 1994).[18]

In addition to these minor difficulties, there is a much bigger reason why post hoc analyses of nonsignificant results should not be done. Consider the researcher who is confronted by a nonsignificant result. Mindful of the possibility of making a Type II error the researcher asks: does this lack of a result indicate the absence of an effect or is there a chance that I missed something? This is a fair question, but it is unanswerable with power analysis. Recall that statistical power is the probability that a test will correctly reject a false null hypothesis. Statistical power has relevance only when the null is false. The problem is that the nonsignificant result does not tell us whether the null is true or false. To calculate power after the fact is to make an assumption (that the null is false) that is not supported by the data. A retrospective analysis tells us nothing about the truthfulness of the null so we cannot proceed to calculate power. To do so would be like trying to solve an equation with two unknowns (Zumbo and Hubley 1998).[19]

Even aware of these difficulties, a researcher might still desire to calculate post hoc power by imposing a number of qualifiers. The logic might run as follows: (a) let's assume the effect is real (because other research says so) and that (b) it is the size I observed in my study, and (c) let's use alpha instead of actual *p* values to determine power: what would power be given the size of my sample? There is nothing inherently wrong with this because it is basically a prospective power analysis done after the fact. (Whether it generates good numbers or not will depend on how close the observed effect size is to the population effect size.) In fact, this is exactly what statistics packages such as SPSS do when they calculate power based on a test result. SPSS does not actually know that the study has been done so what looks like a retrospective analysis is actually prospective in nature. SPSS calculates power as if the observed effect size was identical to the hypothesized population effect size (Zumbo and Hubley 1998).

In a similar vein a researcher with a nonsignificant result may wish to know "how big a sample size should I have had?" or "what was the minimum effect size my study could have detected?". Again, when the qualifiers above are imposed, this is akin to analyzing power prospectively. It is the same as asking: "if I were to use the parameters of this study again, what effect size might I be able to detect next time?" The results cannot be used to interpret nonsignificant results, but they can be used to assess the sensitivity of future studies. For example, if the Alzheimer's researcher asked "what would be the smallest difference between the two groups that would be detectable in my study?" she is essentially asking "what is the smallest difference that could be observed in a follow-up study that had the same parameters as my first study?"

It should be clear by now that the post hoc analysis of a study's observed power is "nonsensical" (Zumbo and Hubley 1998: 387), "inappropriate" (Levine et al. 2001), and generally "not helpful" (Thomas 1997: 278). However, post hoc analyses can be useful when they are based on population effect sizes, such as might be obtained from pooling the estimates of many studies. In addition, post hoc analyses are sometimes based on a range of hypothetical effect sizes. This type of analysis is usually done to gauge the prevalence of Type II errors across a set of studies or an entire field of research. Some examples of these sorts of retrospective power surveys are described in the next chapter.

Using power analysis to select sample size

We began this chapter with the question: how big a sample size do I need to test my hypotheses? In the absence of a power analysis, this question is usually answered by falling back on to what Cohen (1962: 145) called "non-rational bases" for making sample size decisions. These include following past practice, making decisions based on data availability, relying on unaided intuition or experience, and negotiating with influential others such as PhD supervisors. Also popular are statistical rules of thumb (van Belle 2002). For example, in multivariate analysis, desired sample sizes are sometimes expressed as some multiple of the number of predictors in a regression equation (see, for example, Harris (1985: 64) and Nunnally (1978: 180)). The great drawback of these methods is that none of them can guarantee that studies will have sufficient power to mitigate beta risk.

Setting sample sizes

A prospective power analysis provides arguably the best answer to the sample size question. Hoping to detect an effect of size $r = .40$ using a two-tailed test, a researcher can look up a table to learn that he will need a sample size of at least $N = 46$ given conventional alpha and power levels. To detect a smaller effect of $r = .20$ under the same circumstances, he would need a sample of at least $N = 193$. These are definitive answers that are likely to be a lot closer to the mark than estimates obtained using rules of thumb. The only tricky part in this equation is estimating the size of the effect that one hopes to find.[20] If the expected effect size is overestimated, required sample sizes will be underestimated and the study will be inadequately powered. The researcher has several options for predicting effect sizes. The best of these is to refer to a meta-analysis of research examining the effect of interest. A meta-analysis will normally generate a pooled estimate of effect size that accounts for the sampling and measurement error attached to individual estimates (see Chapter 5). When a meta-analytically derived estimate is not available, the next best option may be for the researcher to pool the effect size estimates of whatever research is available. If no prior research has been done, the researcher may opt to run a pretest or make an estimate based on theory. Another alternative is to construct a dummy table to explore the trade-offs between

Table 3.1 *Minimum sample sizes for different effect sizes and power levels*

	Power				Power		
ES = *d*	.70	.80	.90	ES = *r*	.70	.80	.90
.10	2,471	3,142	4,205	.05	2,467	3,137	4,198
.20	620	787	1,053	.10	616	782	1,046
.30	277	351	469	.15	273	346	462
.40	157	199	265	.20	153	193	258
.50	101	128	171	.25	97	123	164
.60	71	90	119	.30	67	84	112
.70	53	67	88	.35	49	61	81
.80	41	52	68	.40	37	46	61
.90	33	41	54	.45	29	36	47
1.00	27	34	45	.50	23	29	37

Note: The sample sizes reported for *d* are combined (i.e., $n_1 + n_2$). The minimum number in each group being compared is thus half the figure shown in the table rounded up to the nearest whole number. $\alpha_2 = .05$.

different effect sizes and the sample sizes that would be required to identify them under varying levels of power. Whichever approach is used, effect size estimates should be conservative in nature and sample size predictions should err on the high side.

With some idea of the anticipated effect size, the researcher can crunch the numbers to determine the minimum sample size. Power calculations are rarely done by hand. Instead, researchers normally refer to tables of critical values in much the same way that tables of critical *t*, *F*, and other statistics are sometimes used to assess statistical significance.[21] Table 3.1 is a trimmed-down version of the power tables found in Appendix 1 at the back of the book. This table shows minimum sample sizes for both the *d* and *r* family effect sizes involving two-tailed tests with alpha levels set at .05. The columns in the table refer to three different power levels (.70, .80, and .90) and the rows refer to different effect sizes ($d = .10 – 1.00$; $r = .05 – .50$). To determine a required sample size, you find the cell that intersects the desired level of power and the anticipated effect size. For example, if you expect the difference between two groups to be equivalent to an effect size of $d = .50$, and you wish to have at least an 80% chance of detecting this difference, you will need at least 128 participants in your sample. As this effect size relates to differences between groups, the implication is that you will need a minimum of 64 participants in each group. If you wish to further reduce the possibility of overlooking real effects by increasing power to .90, you will need a minimum of 171 participants (or 86 in each group).

Power tables are not difficult to use but they can be coarse and cumbersome. A superior way to run a power analysis is to use an online power calculator or a computer program such as G*Power (Faul et al. 2007). At the time of writing the latest version of this freeware program was G*Power 3, which runs on both Windows XP/Vista/7 and Mac OS X 10.6 operating systems. This user-friendly program can be used to

run all types of power analysis for a variety of distributions. Using the interface you select the outcome of interest (e.g., minimum sample size), indicate the test type, input the parameters (e.g., the desired power and alpha levels), then click "calculate" to get an answer. This program was used to calculate the minimum sample sizes in Table 3.1.[22]

Minimum detectable effects

It should be clear by now that the number of observations in a study has a profound impact on our ability to detect effects. In many cases it is fair to say that the success or failure of a project – in terms of arriving at a statistically significant result – hinges on its sample size. For instance, if we are seeking to detect an effect of size $d = .50$ and we were to run many studies with sixty participants each, we would achieve statistical significance less than half of the time. If we wanted to reduce the risk of missing this effect to 20%, we would need group sizes to be more than twice as large.[23] Before committing to a study it is helpful for researchers to have some idea of the sensitivity of their research design. The question to ask is: what is the smallest effect my proposed study can detect? Table 3.2 shows the minimum detectable effect size for conventional levels of alpha and power. (The minimum effect sizes were calculated using G*Power 3.) If one had access to a sample of 100, the smallest effect that could be detected using a nondirectional test is $r = .28$ (or $d = .57$). Double the sample size and the sensitivity of the test increases such that the smallest detectable effect drops to $r = .20$ (or $d = .40$).

These tables should be taken as a starting point only. In practice, a number of additional factors affecting the power of a study will need to be considered. Chief of these is the issue of whether the minimum detectable effect is worth investigating. Before committing to any study the researcher should ask whether the anticipated effect size is intrinsically meaningful. This is an issue which power analysis cannot address. Statistical power is test-specific so another issue concerns the types of analysis that will be run. For instance, if subgroup analysis is to be performed, then the appropriate sample size to estimate will be the size of the smallest subgroup.[24] If a multivariate analysis such as multiple regression is to be performed, then the researcher will need to take into account factors such as the number of predictors to be used in the model. The researcher will need to assess the power required to detect the omnibus effect (i.e., R^2) along with the power required to detect targeted effects associated with specific predictors (i.e., a particular regression coefficient or part correlation) (Green 1991; Kelley and Maxwell 2008; Maxwell et al. 2008). As statistical power is test-specific, three separate power requirements are relevant for multiple regression: the power required to detect at least one effect, the power required to detect a particular effect, and the power required to detect all effects (Maxwell 2004).[25] Table 3.3 illustrates these different requirements by showing the minimum sample sizes and power values for a regression equation with five predictors when each has a medium correlation ($r = .3$) with the outcome variable. If the aim is to find at least one statistically significant effect

Table 3.2 *Smallest detectable effects for given sample sizes*

	r		*d*	
Sample size	One-tailed	Two-tailed	One-tailed	Two-tailed
10	.705	.761	1.725	2.024
20	.526	.579	1.156	1.325
30	.437	.485	.931	1.060
40	.382	.426	.801	.909
50	.344	.384	.713	.809
60	.315	.352	.650	.736
70	.292	.327	.600	.679
80	.274	.307	.561	.634
90	.259	.290	.528	.597
100	.246	.276	.501	.566
110	.235	.263	.477	.539
120	.225	.252	.457	.516
130	.216	.243	.438	.495
140	.208	.234	.422	.477
150	.201	.226	.408	.460
160	.195	.219	.395	.446
170	.189	.213	.383	.432
180	.184	.207	.372	.420
190	.179	.202	.362	.409
200	.175	.197	.353	.398
210	.171	.192	.344	.388
220	.167	.187	.336	.379
230	.163	.183	.329	.371
240	.160	.180	.322	.363
250	.157	.176	.315	.356
260	.154	.173	.309	.349
270	.151	.169	.303	.342
280	.148	.166	.298	.336
290	.145	.164	.293	.330
300	.143	.161	.288	.325

Note: Power = .80, alpha = .05.

then a sample of 100 should suffice (power will be .84). However, if the researcher has their heart set on detecting one particular effect, a sample of at least $N = 400$ will be needed to achieve satisfactory statistical power (.78).

Power and precision

So far the question of sample size has been framed as an issue of statistical power, as in "how much power do I need to detect an effect of a certain size?" A related question is: "how precise should my estimate be?" In Chapter 1 we saw how the precision of

Table 3.3 *Power levels in a multiple regression analysis with five predictors*

	Sample size		
Power to detect. . .	100	200	400
At least one effect	.84	.99	>.99
Any single specified effect	.26	.48	.78
All effects	<.01	.01	.22

Note: Every predictor has a medium correlation ($r = .3$) with the outcome variable. $\alpha = .05$.
Source: Adapted from Maxwell (2004, Table 3).

an effect size estimate can be quantified as the width of its corresponding confidence interval. The precision of an estimate has implications for interpreting the result of a study. Maxwell et al. (2008) offer the hypothetical example of a study reporting an effect of size $d = .50$ with a corresponding CI_{95} ranging from .15 to .85. How should this result be interpreted? A medium-sized effect was observed, but the estimate was so imprecise that the true effect could plausibly be smaller than small or larger than large. The lesson here is to avoid these sorts of interpretation nightmares by making sure studies are designed with precision targets in mind.

As both precision and statistical power are related to sample size, each can be mathematically related to the other. For instance, where an effect can be expressed as the observed difference between two means, Goodman and Berlin (1994) provide the following rule of thumb approximation: predicted CI_{95} = observed difference $\pm 0.7\Delta_{80}$, where Δ_{80} denotes an effect size (Δ) being sought in a test where power = .80. For example, if we ran a test with power of .80 to detect an effect of size $d = .50$, our result would have a predicted average precision of $\pm 0.7 \times .50 = \pm .35$. If our observed difference between two groups was equivalent to $d = .50$, the resulting CI would have an expected margin of error of .35, giving the results above (.15 to .85). The implication, which is fully explained by Maxwell et al. (2008), is that a sample size which is big enough to generate sufficient power may not provide a particularly accurate estimate of the population effect.

When sample sizes are set on the basis of desired power the aim is to ensure the study has a good chance of rejecting the null hypothesis of no effect. But there is no guarantee that the study will be large enough to generate accurate parameter estimates. This is because effect sizes affect power estimates but have no direct bearing on issues of accuracy and precision. A prospective power analysis done with the expectation of detecting a medium to large effect will suggest sample sizes that may be insufficient in terms of generating precise estimates. In other settings (e.g., when effects are tiny), the reverse may be true. Studies may generate precise estimates but fail to rule out the possibility of a Type II error as indicated by a narrow confidence interval that does not exclude the null value. In view of these possibilities, the researcher needs to decide in

advance whether the aim of the study is (a) to reject a null hypothesis, (b) to estimate an effect accurately, or (c) both. In most cases the researcher will desire both power and precision and this leads to the question: how precise is precise? Or, how narrow should intervals be? Smithson (2003: 87) argues that if a study is seeking a medium-sized effect then, as a bare minimum, the desired confidence interval should at least exclude the possibility of values suggesting small and large effects.[26]

Accounting for measurement error

One of the biggest drains on statistical power is measurement error. Unreliable measures introduce unrelated fluctuations or noise into the data, making it harder to detect the signal of the underlying effect. Any drop in the signal-to-noise ratio introduced by measurement error must be matched by a proportional increase in statistical power if the effect is to be accurately estimated. If X and Y are measured poorly, then the observed correlation will be less than the true correlation on account of measurement error.

To correct for measurement error we need to know something about the reliability of our measurement procedures. Reliability can be estimated using test-retest procedures, calculating split-half correlations, and gauging the internal consistency of a multi-item scale (Nunnally and Bernstein 1994, Chapter 7). The latter method is probably the most familiar to those of us educated in the era of cheap computing. Internal consistency captures the degree to which items in a scale are correlated with one another. Low scores (below .6 or .7) indicate that the scale is too short or that items have little in common and therefore may not be measuring the same thing. Knowing the internal consistency of our measures of X and Y, we can adjust our estimates of the effect size to compensate for measurement error. This is done by dividing the observed correlation by the square root of the two reliabilities multiplied together. If r_{xy} (observed) = .14 and the measurement reliability for both X and Y = .7, then r_{xy} (true) = .20.[27]

Measurement error has a profound effect on our need for statistical power. To detect a true effect of $r = .20$ with perfect measurement and conventional levels of alpha and power would require a sample of at least $N = 193$, but to capture this effect with unreliable measures that depress the observed correlation to $r = .14$ raises our minimum sample size requirement to 398. Small effects are hard enough to detect at the best of times. But add a little measurement error and the task becomes far more challenging. Table 3.4 shows the reduction from the true to the observed effect size for different levels of measurement error and the implications for sample size. For example, to detect a true effect of size $r = .30$ with perfect measures requires a sample of 84 or more. But to detect this effect when internal consistency scores are .6 would require a sample of at least 239.

Summary

Power analysis is relevant for any researcher who relies on tests of statistical significance to draw inferences about real-world effects. Conducting a power analysis

Table 3.4 *The effect of measurement error on statistical power*

	Small effect r_{xy}(true) = .10		Medium effect r_{xy}(true) = .30		Large effect r_{xy}(true) = .50	
$\sqrt{(r_{xx},r_{yy})}$	r_{xy}(observed)	Min *N*	r_{xy}(observed)	Min *N*	r_{xy}(observed)	Min *N*
1.0	0.10	782	0.30	84	0.50	29
0.9	0.09	966	0.27	105	0.45	36
0.8	0.08	1,224	0.24	133	0.40	46
0.7	0.07	1,599	0.21	175	0.35	61
0.6	0.06	2,177	0.18	239	0.30	84

Note: Power = .80 and α_2 = .05. r_{xy}(observed) = r_{xy}(true) × $\sqrt{(r_{xx},r_{yy})}$ where r_{xx} and r_{yy} denote the reliability coefficients for X and Y respectively. If r_{xy}(true) = .30 and $\sqrt{(r_{xx},r_{yy})}$ = .70, then r_{xy}(observed) = .21 and the minimum sample size required to detect = 175, not 84. Table adapted from Dennis et al. (1997, Table 8).

during the design stages of a project can protect scholars from engaging in studies that are fatally underpowered or wastefully overpowered. Running a power analysis is not difficult. Anyone who can perform a statistical test can conduct a power analysis. Neither is power analysis time-consuming. Usually no more than a few minutes are needed to check that a project stands a fair chance of achieving what it is supposed to achieve.[28] Yet surveys of research practice reveal that power analyses are almost never done (Bezeau and Graves 2001; Kosciulek and Szymanski 1993).[29] The consequence of this neglect is a body of research beset by Type I and Type II errors. Numerous power surveys done over the past few decades reveal that none of the social science disciplines has escaped this plague of missed or misinterpreted results (e.g., Brock 2003; Cohen 1962; Lindsay 1993; Mazen et al. 1987b; Rossi 1990; Sedlmeier and Gigerenzer 1989). The evidence for, and the implications of, this sorry state of affairs are discussed in the next chapter.

Notes

1 A literal interpretation of a null hypothesis of no effect may not be desirable as there is always some effect of at least minuscule size. Given sufficient statistical power even trivial effects will be detectable and this makes the literal null easy to reject. The null is almost *always* false (Bakan 1966). (Hunter (1997) reviewed the 302 meta-analyses in Lipsey and Wilson (1993) and found just three (1% of the total) that reported a mean effect size of 0, and those three were reviewing the same set of studies.) Consequently, some scholars interpret the null as being the hypothesis of no nontrivial effect, distinguishing it from the nil hypothesis that the effect size is precisely zero (e.g., Cashen and Geiger 2004). For the sake of convenience, we will ignore these distinctions here and adopt the classic interpretation of the null as indicating no effect. For more on the non-nil null hypothesis, see Cohen (1994). For an introduction to tests of the "good enough" hypothesis, see Murphy (2002: 127).

2 Sample-specific variation is referred to as sampling error and is inversely proportional to the square root of the sample size. Every sample has unique quirks that introduce noise into any

data obtained from that sample. In bigger samples these quirks tend to cancel each other out and average values are more likely to reflect actual values in the larger population. But in the Alzheimer's study the groups were very small, with just six patients in each. Consequently there is every possibility that some of the difference observed between the two groups can be attributed to the luck of the draw – perhaps certain types of people ended up together in the same group. The alpha significance criterion exists to protect against this threat and in this case the nonsignificant *p* value indicates that blind luck may have indeed affected the results.

3 A common misperception is that $p = .05$ means there is a 5% probability of obtaining the observed result by chance. The correct interpretation is that there is a 5% probability of getting a result this large (or larger) *if* the effect size equals zero. The *p* value is the answer to the question: if the null hypothesis were true, how likely is this result? A low *p* says "highly unlikely," making the null improbable and therefore rejectable.

4 Statistical significance tests can only be used to inform judgments regarding whether the null hypothesis is false or not false. This arrangement is similar to the judicial process that determines whether a defendant is guilty or not guilty. Defendants are presumed innocent; therefore, they cannot be found innocent. Similarly, a null hypothesis is presumed to be true unless the findings state otherwise (Nickerson 2000). This logic may work well in the courtroom but Meehl (1978: 817) argued that the adoption of the practice of corroborating theories by merely refuting the null hypothesis was "one of the worst things that ever happened in the history of psychology."

5 This is not to say that statistical significance testing is worth keeping, for there are better means for gauging the importance, certainty, replicability, and generality of a result. Importance can be gauged by interpreting effect sizes; certainty can be gauged by estimating precision via confidence intervals; replicability can be gauged by doing replication studies; and generality can be gauged by running meta-analyses (Armstrong 2007). Schmidt and Hunter (1997) spent three years challenging researchers to provide reasons justifying the use of statistical significance testing. They ended up with a list of eighty-seven reasons, of which seventy-nine were easily dismissed. The remaining eight reasons, after considered evaluation by Schmidt and Hunter, were also found to be meritless. Schmidt and Hunter concluded that "statistical significance testing retards the growth of scientific knowledge; it never makes a positive contribution" (1997: 37). A similar conclusion was reached by Armstrong (2007) after he made a similar appeal to colleagues for evidence in support of significance testing: "Even if properly done and properly interpreted, significance tests do not aid scientific progress." According to McCloskey (2002), the "plague" of statistical significance testing also explains the lack of progress in empirical economics. "It's all nonsense, which future generations of economists are going to have to do all over again. Most of what appears in the best journals of economics is unscientific rubbish" (McCloskey 2002: 55).

6 It will be apparent from reading Box 3.1 that this is a convention which has attracted a fair amount of criticism. Rightly or wrongly, much of this criticism has been directed at Fisher. But Gigerenzer (1998) argues that Fisher's preference for the 5% level of significance was not as strong as many think it was. Apparently Fisher's choice simply reflected his lack of tables of critical values for other levels of significance.

7 At least that is the theory. In practice the one unlucky scientist who mistakes sampling variation for a genuine effect will probably be the only one to get their work published. After all, they found something to report whereas the other nineteen found nothing. Editors prefer publishing statistically significant results, so it is the "unlucky" scientist who becomes famous and moves on to other projects before replication research discredits his original finding.

8 3,323 is the minimum sample size for detecting a correlation of $r = .034$ using a nondirectional two-tailed test with power and alpha set to .50 and .05 respectively. If desired power is set to the more conventional .80 level, the minimum sample size for this small correlation is 6,787. As it happened, more than 22,000 doctors participated in the aspirin study, ensuring that it had more than enough power to detect (Rosenthal and Rubin 1982).

9 If the researcher's aim is to test a null hypothesis, conducting an underpowered study will almost certainly be a waste of a time in the sense that the outcome will likely be nonsignificant and therefore inconclusive. But it would be going too far to claim that small studies are worthless and that large studies are always the ideal. Any study that provides an effect size estimate has intrinsic value to a meta-analyst, as we will see in Chapter 5. (However, sloppy statistical practices combined with publication biases favoring statistically significant results can give rise to a disproportionate number of false positives which can taint a meta-analysis, as we will see in Chapter 6.) This value is independent of the statistical significance of the result. Plus, small studies sometimes have disproportionately big effects on practice. This can happen when they generate timely results while larger-scales studies are still being run, or when they are based on inherently small but meaningful samples as in the case of a rare disease.

10 We say "at best" because the success of the study is defined as the probability of finding an effect which actually exists. If there is no effect to be found, then no amount of power will save the study from "failure."

11 There is a little bit of Goldilocks' logic to setting power at .80. It is higher than .50, which is definitely too low (or too risky in terms of making a Type II error), and it is lower than .90, which is probably too high (or likely to be too expensive). As .80 is neither too low nor too high, it seems just about right. But some dismiss this pragmatic approach and take the view that in the absence of mitigating factors, Type I and Type II errors should be viewed as equally serious. If alpha is set to .05, then beta should = .05 and power should be .95 (Cashen and Geiger 2004; Lipsey 1998; Di Stefano 2003).

12 Failure here means "failure to find an effect that is there to be found." But there is a broader sense that studies which fail to find are failed studies. There is ample evidence to show that studies which are published in top journals are more likely to be those which have found *something* rather than those which have found *nothing* (Atkinson et al. 1982; Coursol and Wagner 1986; Hubbard and Armstrong 1992). One of the unfortunate implications arising from this publication bias is that good quality projects reporting nonsignificant results are likely to be filed away and not submitted for consideration. This "file drawer" problem combined with a publication bias leads to the publication of effect sizes that are on average higher than they should be (more on this in Chapter 6). The *Journal of Articles in Support of the Null Hypothesis* (www.jasnh.com) represents a concerted attempt to remedy this imbalance. By offering an outlet that is biased towards the publication of statistically nonsignificant results, the editors hope to compensate for the inflation in effect sizes arising from the publication bias and the file drawer problem.

13 A rational approach to balancing error rates implies a decision should be made according to the relative costs and benefits associated with each error type. For example, if a Type II error is judged to be three times more serious than a Type I error then beta should be set at .05 and alpha should be set at .15 (Lipsey 1998: 47). But this almost never happens in practice. Indeed it is rare to find alpha being allowed to go higher than .10 (Aguinis et al. in press). However, this may say more about institutional inertia and bad habits than sound statistical thinking. As an aside, the challenges of balancing alpha and beta risk can be illustrated in the classroom with reference to judicial process (Feinberg 1971; Friedman 1972). In this context a Type I error is analogous to convicting the innocent while a Type II error is analogous to acquitting the guilty.

14 To be correct, power is related to the sensitivity of the test. There are many factors which affect test sensitivity (e.g., the type of test being run, the reliability or precision of the measures, the use of controls, etc.) but the size of the sample is usually the most important (Mazen et al. 1987a).

15 Why α and not p when most hypotheses live or die on the result of p values? Because α is the probability specified in advance of collecting data while p is the calculated probability of the observed result for the given sample size. Technically α is the conditional prior probability of a Type I error (rejecting H_0 when it is actually true) whereas p is the exact level of significance of a statistical test. If p is less than α then H_0 is rejected and the result is considered statistically

significant (see Kline 2004: 38–41). For a history of the .05 level of statistical significance, see Cowles and Davis (1982). For a summary of how researchers routinely confuse α with p, see Hubbard and Armstrong (2006).

The default assumption of nondirectional or two-tailed tests originates in the medical field where directional or one-tailed tests are generally frowned upon. Two-tailed tests acknowledge that the effects of experimental treatments may be positive or negative. But one-tailed tests are more powerful and may be preferable whenever we have good reasons for expecting effects to run in a particular direction (e.g., we expect the treatment to be always beneficial or we expect the strategy can only boost performance).

16 Where do these numbers come from? The hypothetical Alzheimer's study discussed in Chapter 1 was originally a thought experiment included in Kirk's (1996) paper on practical significance. In that paper Kirk provided information on the size of the sample ($N = 12$), the effect size (13 IQ points), and the statistical significance of the results ($t = 1.61$, $p = .14$). The other numbers can be deduced from these starting points given a few assumptions. For instance, if we assume that the patients who were treated had their IQ return to the mean population level (100), the mean IQ of the untreated group must be 87. Given these means, the only standard deviations that can generate the t and p statistics reported by Kirk are 14 (for each group). This can be worked out using an online t test calculator such as the one provided by Uitenbroek (2008) and the result is fairly close to the standard deviation of 15 normally associated with an IQ test. Plugging these means, SDs, and Ns into the calculator generates $t = 1.608$ and a double-sided $p = .1388$. Using an online effect size calculator such as Becker's (2000) we can then transform this difference between the means into a Cohen's d of 0.929. A power program such as G*Power 3 (Faul et al. 2007) can then be used to compute the minimum sample sizes. In G*Power 3 this is labeled a priori analysis. Running an a priori analysis reveals that the minimum sample size required to detect an effect of size $d = .929$ given conventional alpha and power levels is forty (or twenty in each group) if a two-tailed test is used or thirty-two (sixteen per group) if a one-tailed test is used. For what it's worth, G*Power 3 can also be used to run a post hoc analysis to compute observed power, which in the original study was just .307. A sensitivity analysis can be used to calculate the minimum detectable effect size given the other parameters: 1.796. Finally, a criterion analysis can be used to calculate alpha as a function of desired power, the effect size, and the sample size in each group. In this case the critical level of alpha is .438 when power is .80 and a nondirectional test is adopted. If a directional test is used and desired power is lowered to .70, the critical level of alpha falls to .149.

17 Even supporters of retrospective power analyses acknowledge that "the observed effect size used to compute the post hoc power estimate might be very different from the true (population) effect size, culminating in a misleading evaluation of power" (Onwuegbuzie and Leech 2004: 210). This begs the question that if these sorts of analyses are misleading, why do them?

18 Post hoc analyses are sometimes promoted as a means of quantifying the uncertainty of a nonsignificant result (e.g., Armstrong and Henson 2004). A far better way to gauge this uncertainty is to calculate a confidence interval. A confidence interval will answer the question: "given the sample size and observed effect, which plausible effects are compatible with the data and which are not?" (Goodman and Berlin 1994: 202). A confidence interval for a nonsignificant result will span the null value of zero. But it will also indicate the likelihood that the real effect size is zero. A narrow interval centered on a point close to zero will be more consistent with the null hypothesis of no effect than a broad interval that extends far from zero (Colegrave and Ruxton 2003).

19 If this is confusing, refer back to Figure 3.2 which describes the four outcomes or conclusions that can be reached in any study. Power relates to outcomes described in the right-hand column of the table. That is, statistical power is relevant only when the null is false and there is an effect to be found. But nonsignificant results are relevant to the two outcomes described in the top row of the table. We found nothing and this means that either there was nothing to be found or there was something but we missed it. Under the circumstances we might prefer to make the

Type II error – better to have the hope of something to show for all our work than nothing – hence the temptation to calculate power retrospectively. But we cannot calculate power without first assuming the existence of an effect. The inescapable fact is that a nonsignificant result is an inconclusive result.

20 Murphy and Myors (2004: 17) note that a priori power analysis is premised on a dilemma: power analysis cannot be done without knowing the effect size in advance, but if we already know the size of the effect, why do we need to conduct the study?

21 Tables of critical values can be found in Cohen (1988, see pp. 54–55 (*d*), 101–102 (*r*), 253–257 (*w*), 381–389 (*f*)) and Kraemer and Thiemann (1987: 105–112), Friedman (1968: Table 1), and Machin et al. (1997, see pp. 61–66 (Δ), 172–173 (*r*)). Murphy and Myors (2004) provide tables for the non-central F distribution, as do Overall and Dalal (1965) and Bausell and Li (2002). Tables for *q* (the effect size index for the difference between two correlations) and *h* (the index for the difference between two proportions) can be found in Rossi (1985). Instead of tables, Miles and Shevlin (2001: 122–125) present some graphs showing different sample sizes needed for different effect sizes and varying numbers of predictors.

22 Freeware power calculators can be found online by using appropriate search terms. The following is a sample: G*Power 3 can run all four types of power analysis and can be downloaded free from www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/; Daniel Soper of Arizona State University has easy-to-use calculators for all sorts of statistical calculations, including power analyses relevant for multiple regression (www.danielsoper.com/statcalc); Russ Lenth of the University of Iowa has a number of intuitive Java applets for running power analysis (www.cs.uiowa.edu/~rlenth/Power); DSS Research has calculators useful for determining sample size and statistical power (www.dssresearch.com/toolkit/default.asp). A number of sample size calculators are offered by Creative Research Systems (www.surveysystem.com/sscalc.htm), MaCorr Research Solutions (www.macorr.com/ss_calculator.htm), and the Australian-based National Statistical Service (www.nss.gov.au/nss/home.NSF/pages/Sample+Size+Calculator). Chris Houle has an online calculator relevant for differences in proportions (www.geocities.com/chrishoule). The calculation of statistical power for multiple regression equations featuring categorical moderator variables requires some special considerations, as explained by Aguinis et al. (2005). An online calculator for this sort of analysis can be found at Herman Aguinis's site at the University of Colorado at Denver (http://mypage.in.edu/~haguinis/mmrindex.html).

23 In this example, with thirty participants per group (or sixty per study) and alpha levels set at .05 with nondirectional tests, statistical power equals .48. In this arrangement we would need a minimum of sixty-four subjects per group to achieve a conventional power level of .80. If we viewed Type I and Type II errors as being equally serious we might desire a power level of .95. In this case we would need 105 participants per group.

24 Dividing samples into subgroups can slash statistical power, making statistical significance tests meaningless. Consider the survey researcher who wishes to estimate nonresponse bias by comparing early and late respondents in the belief that late respondents resemble nonrespondents (Armstrong and Overton 1977). The researcher divides respondents into early and late thirds or quartiles and then compares the groups on a number of demographic variables. Although some differences are observed between the two groups, none turns out to be statistically significant. The researcher heaves a sigh of relief and concludes that the results are unaffected by nonresponse bias. But in these comparisons the chances of finding any statistical difference are likely to be remote because of low statistical power (Wright and Armstrong 2008). In fact there probably *is* some meaningful difference between early and late respondents and the failure to detect this signals a Type II error.

Subgroup analyses are also likely to lead to trouble when lots of them are done on the same dataset. Here the "curse of multiplicity" comes into play. Run a large number of subgroup analyses and something is bound to be found – even if it is nothing more than random sampling variation.

Multiple analyses of the same data raise the risk of obtaining false positives. Consequently, in their explanatory notes to the CONSORT statement, Altman et al. (2001: 683) recommend that authors "resist the temptation to perform many subgroup analyses." The multiplicity curse is discussed further in Chapter 4.

25 Maxwell (2004, Table 4) provides a thought-provoking example of three multiple regression studies, each examining the effects of the same five predictors on a common outcome. Each of the studies was based on a sample of 100 subjects. Each reported at least one statistically significant coefficient but generally there was little agreement in their results. Viewed in isolation each study would lead to a different conclusion regarding the predictors. What is particularly interesting about these three hypothetical studies is that their data came from a single database where all of the variables had a medium correlation ($r = .30$) with each other. In other words, the statistically significant coefficients generated by the multiple regression analysis were purely the result of sampling variation. The power of each study was such that the probability of finding at least one statistically significant result was .84, but the power relevant for the detection of individual effects was just .26. Maxwell's (2004) point was this: a regression study may have sufficient power to obtain statistical significance somewhere without having sufficient power to obtain significance for specific effects. The predictor that happens to be significant will vary randomly from sample to sample. This makes it difficult to interpret the results of single studies and leads to inconsistent results across studies. Although there is a trend in some journals to include increasing numbers of independent variables (e.g., control variables), Schwab and Starbuck (2009, forthcoming) advocate the analysis of simple models. They reason that large numbers of predictors (>3) make it difficult for researchers to make sense of their findings.

26 For more on the relationship between precision and sample size see Smithson (2003, Chapter 7) and Maxwell et al. (2008). Formulas for calculating sample sizes based on desired confidence levels can be found in Daly (2000), Malhotra (1996, Chapter 12), and most research methods textbooks.

27 Here is the equation: $r_{xy}(\text{true}) = r_{xy}(\text{observed})/\sqrt{(r_{xx},r_{yy})}$ where r_{xx} and r_{yy} denote the reliability coefficients for X and Y respectively. If $r_{xy}(\text{observed}) = .14$ and r_{xx} and r_{yy} both $= .70$, then $r_{xy}(\text{true}) = .14/\sqrt{(.7 \times .7)} = .20$. See also Schmidt and Hunter (1996, 1999b).

28 Not everyone would agree that that is an hour well spent. Shaver (1993) is a well-known critic of both null hypothesis statistical testing and power analysis. His dismissal of the latter as nothing but an "empty exercise" stems from his disregard of the former. He sees little value in manipulating a research design merely to ensure that a result will be statistically significant. Instead, "the concern should be whether an anticipated effect size is obtained and whether it is obtained repeatedly" (1993: 303).

29 One reason for the neglect of power analysis is that it is not given adequate coverage in undergraduate or graduate-level statistics and methods classes. In a survey of methods instructors cited by Onwuegbuzie and Leech (2004), statistical power was found to rank thirty-fourth out of thirty-nine topics. Teachers and students who prefer a plain English introduction to the subject of statistical power will benefit from reading the short papers by Murphy (2002) and Lipsey (1998). (The latter is a trimmed down version of Lipsey's (1990) authoritative text.) Discipline-specific introductions to statistical power can be found in the following areas: management accounting (Lindsay 1993), physical therapy (Derr and Goldsmith 2003; Norton and Strube 2001), sports management (Parks et al. 1999), management information systems (Baroudi and Orlikowski 1989), marketing (Sawyer and Ball 1981), international business (Brock 2003), health services research (Dennis et al. 1997), medical research (Livingston and Cassidy 2005; Zodpey 2004), and headache research (Houle et al. 2005).

4 The painful lessons of power research

Low-powered studies that report insignificant findings hinder the advancement of knowledge because they misrepresent the 'true' nature of the phenomenon of interest, and might thereby lead researchers to overlook an increasing number of effects. ~ Jürgen Brock (2003: 96–97)

The low power of published research

How highly would you rate a scholarly journal where the majority of articles had more than a 50% chance of making a Type II error, where one out of every eight papers mistook randomness for a genuine effect, and where replication studies falsifying these Type I errors were routinely turned away by editors uninterested in reporting nonsignificant findings? You would probably think this was a low-grade journal indeed. Yet the characteristics just described could be applied to top-tier journals in virtually every social science discipline. This is the implicit verdict of studies that assess the statistical power of published research.

Power analyses can be done both for individual studies, as described in the previous chapter, and for sets of studies linked by a common theme or published in a particular journal. Scholars typically analyze the power of published research to gauge the "power of the field" and assess the likelihood of Type II errors. They avoid the usual perils of post hoc power analyses by using alpha instead of reported *p* values and by calculating power for a range of hypothetical effect sizes instead of observed effect sizes. In making these decisions analysts are essentially asking: "what was the power of a study to detect effects of different size given the study's sample size, the types of tests being run, and conventional levels of alpha?" By averaging the results across many studies analysts can draw conclusions about their power to reject null hypotheses for predefined effect sizes. Box 4.1 describes the procedures for surveying the statistical power of published research.

Surveying the power of the field

As with many of the methodological innovations described in this book, surveys of statistical power originated with Jacob Cohen. Cohen's (1962) original idea was to calculate the average statistical power of all the research published in the 1960 volume

Box 4.1 How to survey the statistical power of published research

Most retrospective analyses of statistical power in published research follow the method developed by Cohen (1962). This procedure can be described in terms of four activities. First, identify the set of studies to be surveyed. This set might be limited to publications within a certain journal or journals over a certain number of years. For example, Sawyer and Ball (1981) assessed all the research published in the *Journal of Marketing Research* in 1979. Assessments of statistical power can be done for any study reporting sample sizes and effect size estimates obtained from statistical tests. When the journal article is adopted as the unit of analysis, the aim is to calculate an average power figure for each article.

Second, for each study record the sample size and the type of statistical tests performed. Unless specified otherwise by the individual study authors, assume statistical tests to be nondirectional (two-tailed). If a variety of statistical tests is reported, record only those results which bear on the hypotheses being tested or which relate to relationships between the core constructs. Peripheral tests (e.g., factor analyses, manipulation checks, tests of statistical assumptions, etc.) can be ignored.

Third, given the above information, and assuming alpha levels of .05, calculate the minimum statistical power of each study relevant for the detection of three hypothetical effects corresponding to Cohen's (1988) thresholds for small, medium, and large effect sizes. For example, if a study reports the difference between two independent groups of twenty participants each, the mean power to detect would be .09, .34, and .69 for *d* effects of size .2, .5, and .8 respectively.

Fourth, average the results across all the studies in the database to arrive at the mean power figures for detecting effects of three different sizes. As mean results are usually inflated by a small number of high-powered studies, it is also a good idea to calculate median power levels.

of the *Journal of Abnormal and Social Psychology*. He did this to assess the prevalence of Type II errors in this body of research. Like all great ideas, this was one whose value was immediately apparent to others. Cohen's study was followed by power surveys done in the following areas:

- accounting information systems (McSwain 2004)
- behavioral accounting (Borkowski et al. 2001)
- communication (Chase and Baran 1976; Chase and Tucker 1975; Katzer and Sodt 1973; Kroll and Chase 1975)
- counseling research (Kosciulek and Szymanski 1993)
- education (Brewer 1972; Brewer and Owen 1973; Christensen and Christensen 1977; Daly and Hexamer 1983)
- educational psychology (Osborne 2008b)

- health psychology (Maddock and Rossi 2001)
- international business (Brock 2003)
- management (Cashen and Geiger 2004; Mazen et al. 1987a/b; Mone et al. 1996)
- management information systems (Baroudi and Orlikowski 1989)
- marketing (Sawyer and Ball 1981)
- neuropsychology (Bezeau and Graves 2001)
- psychology (Chase and Chase 1976; Clark-Carter 1997; Cohen 1962; Rossi 1990; Sedlmeier and Gigerenzer 1989)
- social work (Orme and Combs-Orme 1986).

The general conclusion returned by these analyses is that published research is underpowered, meaning average statistical power levels are below the recommended level of .80. In many cases statistical power is woefully low (see Table 4.1). In the business disciplines, the average statistical power for detecting small effects has been found to vary between .16 for accounting research (Lindsay 1993) and .41 for marketing research (Sawyer and Ball 1981), with results for management in between and toward the low end (Mazen et al. 1987b; Mone et al. 1996). The implication is that published business research ran the risk of overlooking small effects 59–84% of the time. The results are similarly poor in other disciplines. For education research average power levels relevant to the detection of small effects were found to be in the .14–.22 range (Brewer 1972; Daly and Hexamer 1983), in the .16–.34 range for communication research (Chase and Baran 1976; Kroll and Chase 1975), and .31 for social work research (Orme and Combs-Orme 1986). In other words, studies in these disciplines risked missing small effects 69–86% of the time.

In the field of psychology the results are more dispersed, with average power levels relevant for small effects ranging from .17 (Rossi 1990) to a table-topping .50 (Bezeau and Graves 2001). In absolute terms this last number is not particularly high, but it stands out for being more than double the mean power value recorded for small effects in the table.[1]

Significantly, many of these results were obtained from research published in prestigious journals.[2] For example, in separate surveys of the *Academy of Management Journal*, average statistical power was found to be in the .20–.31 range for small effects (Mazen et al. 1987a; Mone et al. 1996). Instead of having an 80% chance of detecting small effects, contributors to the *AMJ* were prepared to take up to an 80% risk of missing them. Low power figures have also been reported for the *Strategic Management Journal* (.18 and .28 for Mone et al. (1996) and Brock (2003) respectively), *Administrative Science Quarterly* (.32 for Mone et al. 1996), the *Journal of Applied Psychology* (.24 and .35 for Rossi (1990) and Mone et al. (1996) respectively), *Behavioral Research in Accounting* (.20 for Borkowski et al. 2001), the *Journal of Management Information Systems* (.21 for McSwain 2004), *Research Quarterly* (.18 for Christensen and Christensen 1977), and the *British Journal of Psychology* (.20 for Clark-Carter 1997). The best journal-specific figure comes from the *Journal of Marketing Research*, but even here the typical study achieved only half of the recommended minimum power level

Table 4.1 *The statistical power of research in the social sciences*

Study	Journal(s) surveyed	Year(s) of survey	# Articles	Mean no. of tests per study	Mean power		
					Small	Medium	Large
Cohen (1962)	*Journal of Abnormal and Social Psychology*	1960	70	69.0	.18	.48	.83
Brewer (1972)	*American Educational Research Journal*	1969–1971	47	7.9	.14	.58	.79
Brewer & Owen (1973)	*Journal of Educational Measurement*	1969–1971	13	20.5	.21	.72	.96
Katzer & Sodt (1973)	*Journal of Communication*	1971–1972	31	53.9	.23	.56	.79
Chase & Tucker (1975)	9 communication journals	1973	46	28.2	.18	.52	.79
Kroll & Chase (1975)	2 communication journals	1973–1974	62	16.7	.16	.44	.73
Chase & Baran (1976)	2 communication journals	1974	48	14.6	.34	.76	.91
Chase & Chase (1976)	*Journal of Applied Psychology*	1974	121	27.9	.25	.67	.86
Christensen & Christensen (1977)	*Research Quarterly*	1975	43	–	.18	.39	.62
Sawyer & Ball (1981)	*Journal of Marketing Research*	1979	23	–	.41	.89	.98
Daly & Hexamer (1983)	*Research in the Teaching of English*	1978–1980	57	21.6	.22	.63	.86
Orme & Combs-Orme (1986)	*Social Work Research and Abstracts*	1977–1984	79	39.4	.31	.76	.92
Mazen et al. (1987b)	*Academy of Management Journal, Strategic Management Journal*	1982–1984	44	83.3	.23	.59	.83
Baroudi & Orlikowski (1989)	4 MIS journals	1980–1985	57	2.6	.19	.60	.83
Sedlmeier & Gigerenzer (1989)	*Journal of Abnormal Psychology*	1984	54	–	.21	.50	.84
Rossi (1990)	3 psychology journals	1982	221	27.9	.17	.57	.83
Lindsay (1993)	3 management accounting journals	1970–1987	43	43.5	.16	.59	.83
Kosciulek & Szymanski (1993)	4 rehabilitation counseling journals	1990–1991	32	–	.15	.63	.90
Mone et al. (1996)	7 leading management journals	1992–1994	210	126.1	.27	.74	.92
Clark-Carter (1997)	*British Journal of Psychology*	1993–1994	54	23.0	.20	.60	.82
Borkowski et al. (2001)	3 behavioral accounting journals	1993–1997	96	18.6	.23	.71	.93
Maddock & Rossi (2001)	3 healthy psychology journals	1997	187	44.2	.36	.77	.92
Bezeau & Graves (2001)	2 neuropsychology journals	1998–1999	66	29.5	.50	.77	.96
Brock (2003)	2 international business & 2 management journals	1990–1997	374	3.0	.29	.77	.93
McSwain (2004)	2 MIS journals	1996–2000	45	3.9	.22	.74	.92

(Sawyer and Ball 1981). Getting published in a top-tier journal is certainly no indicator that a study has adequately addressed the risk of making a Type II error.

Although low, the average power levels reported in nearly all of these surveys were inflated by a handful of high-powered studies. Given a skewed distribution of power scores, a better indicator of a typical study's power is the middle or median score, rather than the mean. Median scores lower than the mean were found in every survey where both scores were provided, save one. (Chase and Tucker (1975) reported a mean equal to the median.) For example, in Mazen et al.'s (1987b) survey, the mean power score relevant for small effects was .23, but the median score was much lower at .13. This indicates that the typical study had an 87% rather than a 77% risk of missing small effects. Even in Bezeau and Graves' (2001) study of power-sensitive neuropsychologists, the median power (.45) was lower than the mean (.50). In none of the disciplines surveyed did the typical study have even a coin-flip's chance of detecting small effects. For medium-sized effects, the power to detect improved but not sufficiently. Only studies published in the *Journal of Marketing Research* had, on average, a reasonable chance of detecting medium-sized effects, where reasonable is defined as power $\geq$.80.[3]

Unless sought-after effects were large, the majority of studies in the social sciences would not have found what they were looking for. Yet curiously, most studies published in top tier journals *have* found something, otherwise they would not have been published.[4] This begs the question: if average statistical power is so low, and the odds of missing effects are so high, how is it possible to fill a journal issue with studies that have detected effects? There are only two plausible explanations for this state of affairs. Either these studies are detecting whopper-sized effects, or they are detecting effects that simply aren't there. The whopper-effect explanation is easily dismissed. Meta-analyses of social science research routinely reveal effects that are sometimes medium sized but are more often small (e.g., Churchill et al. 1985; Haase et al. 1982; Lipsey and Wilson 1993). Wang and Yang (2008) found the mean effect size obtained from more than 1,000 estimates in the field of international marketing was small ($r = .19$). In their meta-analysis of research investigating the link between organizational slack and performance, Daniel et al. (2004) calculated mean effects that ranged from trivial ($r = .05$) to small ($r = .17$) in size. Mazen et al. (1987a) reviewed twelve published meta-analyses encompassing hundreds of management studies and concluded that the overall mean or meta-effect size was small ($d = .39$). Aguinis et al. (2005) examined thirty years worth of research examining the moderating effects of categorical variables in psychology research and found that the average effect size was $f^2 = .002$, well below Cohen's (1988: 412) cut-off (.02) for a small effect of this type. If large effects are the exception rather than the norm in underpowered social science research, the odds are good that many authors are mistaking sampling variability for genuine effects. This can happen because of what Wilkinson and the Taskforce on Statistical Inference (1999: 599) called "the curse of the social sciences," namely, the multiplicity problem.

The curse of multiplicity

The multiplicity problem arises when studies report the results of multiple statistical tests raising the probability that at least some of the results will be found to be statistically significant even if there is no underlying effect. The likelihood of this type of error occurring is referred to as the familywise or experimentwise error rate which can be distinguished from the test-specific alpha level introduced in the previous chapter. The familywise error rate becomes relevant whenever two or more tests are run on the same set of data (Keppel 1982). Consider a study that tests fourteen null hypotheses, all of which happened to be true (meaning there are no underlying effects). If each test is assessed according to a conventional alpha level of .05, the odds are better than even that one of the hypotheses will be found to be statistically significant purely as a result of chance.[5] Even when power is extremely low, if you run enough tests you will eventually get a statistically significant result.

An example of the multiplicity problem may have been unintentionally provided by Peterson et al. (2003) in their study of the link between chief executive officer (CEO) traits and the dynamics of their top management teams. In this study the sample consisted of just seventeen CEOs. This small sample is not unusual given that busy CEOs are difficult to survey. But the limited data that Peterson et al. managed to obtain were used for a large number of statistical tests. Altogether forty-eight correlations were examined, of which seventeen were found to be statistically significant at the $p < .05$ level. However, in a re-analysis of these data Hollenbeck et al. (2006) showed that all but one of the original results were highly unstable, meaning they could not be replicated. Hollenbeck et al. concluded that the combination of a small sample size with a large number of tests led to unstable effect size estimates, dubious inferences, and a weak foundation for further research on the topic.

In the power surveys summarized in Table 4.1 the average statistical power to detect small or medium-sized effects was 0.64. This meant that there was a fair chance (36%) that the average study would fail to detect even medium-sized effects. But with the average study running thirty-five statistical tests, the probability was very high (83%) that at least one result would turn out to be statistically significant even if the null hypothesis were true in every single case.[6] In practice, the chances of a universally true null are negligible. In many cases there was at least a small effect to detect. Even though studies lacked the power to detect small effects, the large number of tests being run meant that on average at least eight results would turn out to be statistically significant every time.[7]

The large number of statistical tests reported in social science research suggests that some authors may be fishing around in their data in the hope of finding a publishable result. This can lead to the brilliantly named sin of HARKing or hypothesizing after the results are known (Kerr 1998). HARKing is what happens when the researcher plays with the numbers, finds a statistically significant result, then positions the paper as if that particular result was the original object of the study. Accidental findings have occasionally led to scientific breakthroughs and there is nothing intrinsically wrong with hypothesizing after the results are in. But such hypotheses must be clearly labeled

as post hoc. Post hoc hypotheses are moderately radioactive when they lack a theoretical foundation larger than the study itself. They emerged from the sample rather than from pre-existing theory. The researcher may be delighted at finding something unexpected and the temptation to spin a story explaining the new result may be overwhelming. But circumspection and squinty-eyed skepticism are called for as unexpected results may be nothing more than random sampling variation. As always, the litmus test of any result is replication.[8]

The unintended consequences of adjusting alpha

The standard cure for the multiplicity problem is to adjust alpha levels to account for the number of tests being run on the data. One way to do this is to apply the Bonferroni correction of α/N where α represents the critical test level that would have been applied if only one hypothesis was being tested and N represents the number of tests being run on the same set of data.[9] In the study of CEO traits mentioned earlier, a small sample size meant that alphas were set at the relatively relaxed level of .10. Adjusting this alpha level for the number of hypotheses ($N = 15$) or the number of actual tests ($N = 48$) would have meant the critical alpha level for inferring statistical significance would have been in the .007 (or .10/15) to .002 (or .10/48) range. But if Peterson et al. had adjusted alpha to compensate for the familywise error rate, none of their results would have been judged statistically significant. They certainly would not have held to their original conclusion that the data "provide broad support" for their hypothesis linking CEO affects management team dynamics (2003: 802).

In the psychology literature a growing awareness of the multiplicity problem has led to an increased used of alpha-adjustment procedures. This has had the unfortunate and unintended consequence of further reducing the average statistical power of psychology research (Sedlmeier and Gigerenzer 1989). (Recall from the previous chapter than when critical alpha levels are tightened, the power of a test is reduced. Adjusting alpha to compensate for familywise error will make it harder to assign statistical significance to both chance fluctuations and genuine effects.) This is an alarming trend. Instead of dealing with the very credible threat of Type II errors, researchers have been imposing increasingly stringent controls to deal with the relatively unlikely threat of Type I errors (Schmidt 1992). In view of these trade-offs, adjusting alpha may be a bit like spending $1,000 to buy insurance for a $500 watch.[10]

Statistical power and errors of gullibility

In addition to being a sad commentary on the level of statistical power in published research, the numbers in Table 4.1 provide some insight into the risk preferences of researchers. These preferences can be gauged by considering the implied beta-to-alpha ratios relevant for medium-sized effects. (As truly large effects are rare in the social sciences, most researchers probably initiate projects expecting to find medium-sized effects (Sedlmeier and Gigerenzer 1989).) As we saw in Chapter 3, this ratio reflects

our tolerance for Type II to Type I errors. Following Cohen's (1988) guarded recommendation, this ratio is conventionally set at 4:1, meaning the risk of being duped is considered four times as serious as the risk of missing effects. However, one consequence of low statistical power is an increase in this ratio. In the case of the studies summarized in the table, the average ratio is 7:1.[11] The implication is that researchers publishing in social science journals implicitly treat the risk of wrongly concluding there is an effect to be seven times more serious than wrongly concluding there isn't.

At first glance this may not seem to be such a bad thing; better to err on the side of caution (risking a Type II error) and be wrong than to claim to see effects that don't exist (risking a Type I error). But while low statistical power boosts the probability of making Type II errors for individual studies, it paradoxically has the effect of raising the Type I error rates of published research. A thought experiment will illustrate how this can happen.

Consider two journals that publish only studies reporting statistically significant effects. (This scenario is not far removed from reality as studies have shown that editors and reviewers exhibit a preference for publishing statistically significant results (Atkinson et al. 1982; Coursol and Wagner 1986).) Owing to the vagaries of statistical inference-making it is inevitable that some proportion of these results will reflect nothing more than sampling variation. In other words, most statistically significant results will be genuine but a few will reflect Type I errors. Imagine that the editor of Journal A publishes only articles that satisfy the five-eighty convention introduced in Chapter 3. The five-eighty convention refers to the desired balance between alpha and power. By following the five-eighty rule, the editor of Journal A aims to publish those studies that have at most a 5% chance of incorrectly detecting an effect when there was no effect to detect and at least an 80% chance of detecting an effect when there was one. Consequently the ratio of false to legitimate conclusions published in Journal A will be about .05 to .80, or 1:16. For every sixteen studies that correctly reject a false null, Journal A will publish one that wrongly claims to have found an effect. In other words, one false positive will be published for every sixteen true positives. In contrast, the editor of Journal B, while shunning research that fails to meet the $p < .05$ threshold, has no expectations regarding desired levels of statistical power. A retrospective survey reveals that the average statistical power of studies published in Journal B is .40. This figure is low but hardly unusual when compared with the journals listed in Table 4.1. The implication is that the ratio of false to legitimate conclusions in Journal B is .05 to .40, or one false positive for every eight true positives. As a consequence of low statistical power Journal B will publish twice as many false positives as Journal A.[12]

Of course, no one sets out to publish a bad study. But dubious results are the inevitable by-product of combining low statistical power with high numbers of tests in a business where the prevailing incentive structure ("publish or perish") encourages HARKing. A publication bias in favor of significant results places enormous pressures on researchers to find something, anything. Not only does this lead to the reporting of Type I errors,

as we have seen, but it inflates the size of legitimate published effects (Ioannidis 2008). This was demonstrated in a simulation of 100,000 experiments run by Brand et al. (2008). The difference between published and true effects can be substantial. For instance, if published effects are medium sized ($d = .58$), then true effects may be much smaller ($d = .20$). This has serious implications for researchers planning replication studies. If prospective power analyses are based on inflated estimates of effect size, then resulting sample sizes will be insufficient for detecting actual effects.

How to boost statistical power

Although there are risks associated with having too much statistical power, the pressing need for most social science research is to have much more of it. Fortunately there are several ways researchers can boost the statistical power of their studies. Often the best way is to search for bigger effects. If a researcher is interested in measuring the benefits of advertising, stronger effects are more likely to be observed for marketing-related outcomes such as brand recall than more generic outcomes such as sales revenues. This is because brand recall has a clearly identifiable connection with advertising expenditures while sales levels are affected by a variety of internal and external factors. Bigger effects can sometimes be obtained by increasing the scale of the treatment or intervention. An educational researcher interested in measuring the effect of a remedial class might observe a stronger result by running two classes instead of one.

In situations where researchers do not have any control over the effects they are seeking, the next best way to boost power may be to increase the size of the sample. In some cases doubling the number of observations can lead to a greater than doubling of the power of a study.[13] But sample sizes should not be increased without a careful analysis of the trade-off between additional sampling costs, which are additive, and corresponding gains in power, which may be incremental and diminishing. Fortunately when sample size increases are not possible or desirable there are other ways to increase power.

Statistical power is related to the sensitivity of procedures used to measure effects. Like dirt on an astronomer's telescope, unreliable measures are observationally noisy, making it harder to detect an effect's true signal. For this reason observed effects will appear to be smaller than true effects and will require greater power to detect. By giving careful thought to their measurement procedures researchers can reduce the discrepancy between observed and true effects, reducing their need for additional power. There are many well-known methods for reducing measurement error. These include better controls of extraneous sources of variation, more reliable measures of constructs, and repeated measurement designs (Rossi 1990; Sutcliffe 1980).[14]

Statistical power is also affected by the type of test being run. Parametric tests are more powerful than non-parametric tests; directional (one-tailed) tests are more powerful than nondirectional (two-tailed) tests; and tests with metric data are more powerful than tests involving nominal or ordinal data. To boost statistical power researchers should choose the most powerful test permitted by their data and theory.[15]

Another way to increase power is to relax the alpha significance criterion. In many studies alpha levels are set without any consideration for their impact on statistical power. This can happen when authors confuse low alpha levels (.01, .001) with scientific rigor. Instead of focusing on alpha risk while ignoring beta risk, a better approach is to explicitly assess the relative seriousness of Type I and Type II errors (Aguinis et al. in press; Cascio and Zedeck 1983). Swinging the balance in favor of mitigating beta risk might be justifiable in settings where a long history of research shows the null hypothesis to be false and therefore the risk of a Type I error is virtually non-existent. It might also be justifiable when the other power parameters are relatively fixed and there is a reasonable fear of reporting a false negative (e.g., in the Alzheimer's study where there was limited access to sufficient numbers of patients). Relaxing alpha is sometimes done when there is an expectation that an important effect will be small. For example, it is not uncommon to see significance levels of $p = .10$ in studies analyzing moderator effects which tend to be small and difficult to detect.[16]

Summary

Time and again power surveys have revealed the low statistical power of published research. If published research is biased in favor of studies reporting statistically significant findings, then the power of unpublished research is likely to be lower still. That studies are designed in such a way that effects will be missed most of the time is a serious shortcoming indeed. Underpowered studies incur an opportunity cost in terms of the misallocation of limited resources. By reporting nonsignificant findings that are the result of Type II errors, underpowered studies may also misdirect leads for future research. Potentially interesting lines of inquiry may be wrongly dismissed as dead ends. When low power is combined with an editorial preference for statistically significant findings, the result is the publication of effect sizes that are sometimes false or inflated. This in turn leads to adverse spillover effects for meta-analysts and those engaged in replication research.

In view of these dangers it is no wonder that Cohen (1992), writing thirty years after his pioneering power survey, remained mystified that researchers routinely ignored statistical power when designing studies. It seems that bad habits are hard to change, as evidenced by the low number of studies that even mention power (Fidler et al. 2004; Kosciulek and Szymanski 1993; Osborne 2008b). If change does come it is likely to be initiated by editors wary of publishing false positives, funding agencies concerned about the misallocation of resources, and researchers keen to avoid committing to studies that lack a reasonable chance of success.

The analysis of statistical power can lead to informed judgments about sample size, minimum detectable effect sizes, and the trade-off between alpha and beta risk. The key to a good power analysis is to have a fair idea of the size of the effect being sought. This information is ideally found by pooling the results of several studies. Different methods for doing this are described in the next chapter.

Notes

1 Is there something unique about Bezeau and Graves's (2001) survey of neuropsychology research that accounts for this relatively high number? One plausible explanation is that neuropsychology research attracts a disproportionate level of funding from external agencies that require prospective power analyses. Having been compelled to do these sorts of analyses, neuropsychology researchers would be sensitized to the dangers of pursuing underpowered studies and therefore less likely to do so. This is consistent with Maddock and Rossi's (2001) finding that externally funded studies tend to have more statistical power than unfunded studies.

2 Journal-specific power surveys are available for the following journals: the *Academy of Management Journal* (Brock 2003; Cashen and Geiger 2004; Mazen et al. 1987a; Mone et al. 1996), the *Accounting Review* (Lindsay 1993), *Administrative Science Quarterly* (Cashen and Geiger 2004; Mone et al. 1996), the *American Educational Research Journal* (Brewer 1972), *Behavioral Research in Accounting* (Borkowski et al. 2001), the *British Journal of Psychology* (Clark-Carter 1997), *Communications of the ACM* (Baroudi and Orlikowski 1989), *Decision Sciences* (Baroudi and Orlikowski 1989), the *Journal of Abnormal Psychology* (Rossi 1990), the *Journal of Abnormal and Social Psychology* (Cohen 1962; Sedlmeier and Gigerenzer 1989), the *Journal of Accounting Research* (Lindsay 1993), the *Journal of Applied Psychology* (Chase and Chase 1976; Mone et al. 1996), the *Journal of Clinical and Experimental Neuropsychology* (Bezeau and Graves 2001), the *Journal of Communication* (Chase and Tucker 1975; Katzer and Sodt 1973), the *Journal of Consulting and Clinical Psychology* (Rossi 1990), the *Journal of Educational Measurement* (Brewer and Owen 1973), the *Journal of Educational Psychology* (Osborne 2008), the *Journal of Information Systems* (McSwain 2004), the *Journal of International Business Studies* (Brock 2003), the *Journal of the International Neuropsychology Society* (Bezeau and Graves 2001), the *Journal of Management* (Cashen and Geiger 2004; Mazen et al. 1987a; Mone et al. 1996), the *Journal of Management Accounting* (Borkowski et al. 2001), the *Journal of Management Information Systems* (McSwain 2004), the *Journal of Management Studies* (Cashen and Geiger 2004), the *Journal of Marketing Research* (Sawyer and Ball 1981), the *Journal of Personality and Social Psychology* (Rossi 1990), *Journalism Quarterly* (Chase and Baran 1976), *Management Sciences* (Baroudi and Orlikowski 1989), *MIS Quarterly* (Baroudi and Orlikowski 1989), *Neuropsychology* (Bezeau and Graves 2001), *Research Quarterly* (Christensen and Christensen 1977; Jones and Brewer 1972), *Research in the Teaching of English* (Daly and Hexamer 1983), and the *Strategic Management Journal* (Brock 2003; Cashen and Geiger 2004; Mazen et al. 1987b; Mone et al. 1996).

3 In their assessment of the statistical power of studies reporting nonsignificant results only, Hubbard and Armstrong (1992) calculated similarly decent power levels relevant for the detection of medium-sized effects for the *Journal of Marketing Research* (.92), the *Journal of Marketing* (.86) and the *Journal of Consumer Research* (.87). These results suggest that marketing scholars are the standard bearers among social science researchers when it comes to designing studies with sufficient power for the detection of medium-sized effects.

4 Many prestigious journals will consider only submissions that advance knowledge in some original way. Whether or not this is stated explicitly in the submission guidelines, this is universally taken to mean "if you found nothing, send your paper somewhere else." The controversial implication for researchers is that the likelihood of getting studied is inversely proportion to the *p* values arising from their statistical tests.

5 If N independent tests are examined for statistical significance, and all of the individual null hypotheses are true, then the probability that at least one of them will be found to be statistically significant is equal to $1 - (1 - \alpha)^N$. If the critical alpha level for a single test is set at .05, this means the probability of erroneously attributing significance to a result when the null is true is .05. But if two or three tests are run, the probability of at least one significant result rises to .10

and .14 respectively. For a study reporting fourteen tests, the probability that at least one result will be found to be statistically significant is $1 - (1 - .05)^{14} = .51$.

6 In some of the studies surveyed an extremely high number of tests made the chance of returning at least one statistically significant a dead certainty. The maximum number of tests for a single study was found to be 224 in Orme and Combs-Orme's (1986) survey, 256 in Rossi's (1990) survey, 334 in Maddock and Rossi's (2001) survey, and 861 for a study included in Cohen's (1962) survey.

7 Average power for detecting small effects (.24) times the average number of tests per study (35) equals 8.4 statistically significant results.

8 A surefire way to get a publication is to buy a large database, run lots of tests, then fish like mad. Run enough tests and you will surely find something. If you can then develop some plausible hypotheses to account for these accidental results you just might be able to pass off a Type I error as something real. But be warned, Kerr (1998) provides editors and reviewers with a number of diagnostic symptoms that might indicate the practice of HARKing. These include the just-too-good-to-be-true theory, the too-convenient qualifier (e.g., "we expect this effect to occur only for ___ because of ___"), and the glaring methodological gaffe (e.g., the non-optimal measurement of a key construct may suggest opportunistic hypothesizing). Other tell-tale signs of HARKing are provided by Wilkinson and the Taskforce on Statistical Inference (1999: 600): "Fishing expeditions are often recognizable by the promiscuity of their explanations. They mix ideas from scattered sources, rely heavily on common sense, and cite fragments rather than trends." As with all sample-based results, the definitive test for HARKing is replication. If a result cannot be reproduced in a separate sample, it was probably nothing more than sampling error.

9 This is admittedly a simplistic application of the Bonferroni correction. For more sophisticated variations of this remedy, see Keppel (1982: 147–149). Other alpha-adjustment procedures include the Scheffé, Dunnet, Fisher, and Tukey methods which are described in Keppel (1982, Chapter 8), Keller (2005, Chapter 15), and McClave and Sincich (2009, Chapter 10). A Bonferroni calculator can be found at www.quantitativeskills.com/sisa/calculations/bonfer.htm.

10 Rothman (1990) argues against adjusting alpha for two reasons. First, alpha adjustment provides insurance against the fictitious universal null. In other words, it assumes the null is true in every case, which is unrealistic. Second, the practice of adjusting alpha rests on the flawed idea that the truthfulness of the null hypothesis can be calculated as an objective probability. A better means for assessing the tenability of the null is to refer to both the evidence and the plausibility of alternative explanations.

11 Where does this ratio come from? The average power score for medium effects is .64, indicating that the mean beta rate is .36. Consequently, the beta-to-alpha ratio is $.36/.05 = 7.2{:}1$. Given that alpha and beta are inversely and directly related, the implication is that researchers' tolerance for Type II errors is 7.2 times as great as their tolerance for Type I errors. In other words, alpha is implicitly judged to be 7.2 times as serious as beta. However, researchers publishing in the *Journal of Marketing Research* seem to be the exception in this regard. With average power levels relevant for the detection of medium-sized effects found to be a healthy .89 by Sawyer and Ball (1981), the implied beta-to-alpha ratio is (1–.89)/.05 or 2.2:1. Similar numbers obtained for the *Journal of Marketing* and the *Journal of Consumer Research* (Hubbard and Armstrong 1992) suggest marketing researchers in general implicitly judge alpha to be only twice as serious as beta.

12 These ratios assume that the proportion of false to not-false nulls is equal and that effects are there to be found only half of the time. But in established areas of research, where the balance of evidence indicates that there is an effect to detect, the number of published false positives, and therefore the ratio of false to legitimate conclusions, will be lower.

That editorial policies favoring the publication of significant findings can lead to an increased prevalence of Type I errors has been known since at least the time of Sterling (1959). But an

interesting twist on this idea comes from Thompson (1999a), a former journal editor. Thompson notes that prevailing publication policies combined with a low statistical power favors the publication of Type I errors then hinders the publication of replication studies revealing the previously published Type I error.

13 Doubling the size of the sample will more than double the power of the test with the following parameters: $\alpha_2 = .01$, ES $(r) = .30$, $N = 50$. The power of this test is .33; after doubling the sample size power rises to .68, representing a gain of 106%. But doubling the size of the sample will lead to a relatively smaller increase in power for a test with these parameters: $\alpha_2 = .05$, ES $(r) = .10$, $N = 50$. The power of this test is .11; after doubling the sample size power rises to .17, representing a gain of just 54%.

14 Measurement error can be introduced at almost any point in a study – during the selection of the sample, the design and administration of a survey, data editing, and entry. To illustrate the relationship between measurement reliability and power, Boruch and Gomez (1977: 412) contrast a test conducted within a well-controlled laboratory setting with a retest done out in the field. In the lab measurement was perfectly reliable and the power of the test was high at .92. But when the same test was run in the field by indifferent staff, reliability dipped to .80, and power fell to .30.

15 Note that when running multiple regression the researcher will need to consider at least two levels of statistical power – the power required to detect the omnibus effect (i.e., R^2) and the power required to detect an individual targeted effect (i.e., a particular regression coefficient) (see Green 1991; Kelley and Maxwell 2008; Maxwell et al. 2008: 547). Structural equation modeling also presents some additional concerns. For notes on estimating power when running LISREL, EQS, and AMOS, see Dennis et al. (1997: 397–399) and Miles (2003).

16 How far can we go with relaxing alpha? Theoretically, we might be able to make a case for setting alpha at any level that leads to a good balance between the two sources of error. But in practice it is hard to conceive of anything higher than .10 getting past a typical journal reviewer. Although respectable methodologists such as Lipsey (1998: 47) can conceive of scenarios where one might accept alpha = .15, institutional regard for the sacred .05 level remains high. (In their survey of five years worth of research published in leading business journals, Aguinis et al. (in press) found that 87% of studies used conventional levels of alpha, defined as $\alpha = .10$ or less. The modal level of alpha, used in 80% of studies, was .05.) Radical deviations from this standard – no matter how well argued – are unlikely to be successful. In such cases, other methods for boosting power will be needed. For more general treatments of this issue, see Bausell and Li (2002: Chapter 2), Baroudi and Orlikowski (1989: 98ff), Sawyer and Ball (1981: 284ff), Boruch and Gomez (1977), and Allison et al. (1997).

Part III

Meta-analysis

5 Drawing conclusions using meta-analysis

Many discoveries and advances in cumulative knowledge are being made not by those who do primary research studies, but by those who use meta-analysis to discover the latent meaning of existing research literatures. ~ Frank L. Schmidt (1992: 1179)

The problem of discordant results

A researcher is interested in the effect of X on Y so she collects all the available literature on the topic. She organizes all the relevant research into three piles according to their results. On one side she puts those studies reporting results that were statistically significant and positive. On the other side she puts those studies reporting results that were statistically significant and negative. In the middle she puts those studies that reported results that were statistically nonsignificant. She is unable to draw any conclusions from these disparate results and decides that this is a topic in need of a first-rate study to settle the issue. She conducts her own study and observes that X has a significant negative effect on Y. She writes that her result is consistent with other studies that observed the same effect. However, she is not sure what to make of those studies which found something completely different so she makes some vague comments about "the need for further research before firm conclusions can be drawn." In the back of her mind she is a little disappointed that she was unable to settle the matter, but she has little time to reflect on this as she is already planning her next study.

The moral of this tale is that single studies seldom resolve inconsistencies in social science research. When there are no large-scale randomized controlled trials, scientific knowledge advances through the accumulation of results obtained from many small-scale studies. But as any reviewer of research knows, extant results are sometimes discordant, making it difficult to draw conclusions or find baselines against which future results can be compared. A marginally better situation arises when there is some consensus regarding the direction of effects, as when results are consistently found to be "significantly positive" or "significantly negative." But without knowing the magnitude of these effects the scientist interested in doing a replication study cannot

Table 5.1 *Discordant conclusions drawn in market orientation research*

	MO effect on performance	
Study	Direction	Magnitude
Narver & Slater (1990: 34)	+	"strongly related"
Slater & Narver (2000: 71)	+	"significant predictor"
Pelham (2000: 55)	+	"strong relationship"
Megicks & Warnaby (2008: 111)	+	"highly significant"
Jaworski and Kohli (1993: 64)	+	"mixed support"
Chan & Ellis (1998: 133)	+	"weak association"
Atuahene-Gima (1996: 99)	+	"minimal"
Ellis (2007: 381)	+	"rather weak"
Greenley (1995: 7)	NS	"no main effect"
Harris (2001: 28)	NS	"no main effect"

NB: + denotes positive and statistically significant, NS denotes nonsignificant, MO denotes market orientation.

tell whether extant benchmarks are small, medium, or large. These research scenarios can be summarized as two questions:

1. How do I draw definitive conclusions from studies reporting disparate results?
2. How do I identify non-zero benchmarks from past research?

Answers to these questions may be sought using either qualitative or quantitative approaches or some mixture of the two. The qualitative approach, also known as the narrative review, is useful for documenting the unfolding story of a particular research theme. The aim is to summarize and synthesize the conclusions of others into a compelling narrative about the effect of interest. In short, the narrative reviewer interprets the words of others using words of their own. In contrast, the quantitative approach, better known as meta-analysis, completely ignores the conclusions that others have drawn and looks instead at the effects that have been observed. The aim is to combine these independent observations into an average effect size and draw an overall conclusion regarding the direction and magnitude of real-world effects. In short, the meta-analyst looks at the numbers of others to come up with a number of their own.

Reviewing past research – two approaches

Scholars review past research in order to circumscribe the boundaries of existing knowledge and to identify potential avenues for further inquiry. By reviewing the literature scholars also hope to insure themselves against the prospect of repeating mistakes that others have made. One purpose of a literature review is to draw conclusions about the nature of real-world phenomena and to use these conclusions as a basis for further work. But how do we draw conclusions from results that appear to be discordant? Consider the set of conclusions summarized in Table 5.1. These verbatim conclusions

Table 5.2 *Seven fictitious studies examining PhD students' average IQ*

Study	Mean	SD	*n*	*p*	CI_{95} for mean	*d*
1	100.7	14.0	46	0.736	96.54–104.86	0.047
2	104.2	14.5	39	0.078	99.50–108.90	0.280
3	102.1	15.2	158	0.084	99.71–104.49	0.140
4	103.9	14.5	55	0.051	99.98–107.82	0.260
5	103.9	14.5	56	0.049	100.02–107.78	0.260
6	102.8	14.7	110	0.048	100.02–105.58	0.187
7	93.2	10.1	38	0.002	89.88–96.52	−0.453

were all taken from studies examining the effect of market orientation on organizational performance. As can be seen the results of these studies led to a variety of conclusions, with some authors reporting a strong relationship or effect while others reported none. This inconsistency makes it difficult, if not impossible, to draw a general conclusion regarding the effect of market orientation on performance. Even when similar conclusions were drawn there is nothing to indicate that they were based on a common standard. What constitutes a "strong" relationship? How weak is "weak"? Was the same definition used by all authors?

1. The narrative review – warts and all

Even when reviewers have access to the raw study data, the narrative approach places severe restrictions on the types of conclusions that can be drawn. Consider a hypothetical set of studies examining whether PhD students are, on average, smarter than everybody else. The summary results of seven fictitious studies are reported in Table 5.2.[1] The table shows the mean IQ scores and standard deviations of seven samples of PhD students which can be compared with the population mean and standard deviation of 100 and 15 points respectively. A mean score greater than 100 in the table suggests that PhD students are smarter than average and vice versa. What conclusions can we draw from these numbers?

There are at least four ways to interpret these results. We might: (i) summarize the conclusions of the published literature only, (ii) do a vote-count of all the available results, (iii) graph the confidence intervals to gauge the precision of each estimate, and (iv) calculate an average effect size. As we will see, the conclusions we draw are greatly affected by the methods we choose to review the literature.

First, if we limit our review of the literature to studies that have been published, it is likely that we will miss most of the relevant research on the topic. As none of the first four studies achieved statistical significance there is a good chance that they were filed away rather than submitted for peer review (the so-called file drawer problem) or, if they had been submitted, that they did not survive the review process (owing to a publication bias against nonsignificant results). Thus, any conclusion we form from reading the published literature is likely to be based on an incomplete representation of relevant research, that is, studies 5–7 only. What conclusion would we draw from these

three studies? The first two reported positive differences that just achieved statistical significance while the third reported a negative difference that was very statistically significant. It is erroneous but not uncommon for authors to infer meaning from tests of statistical significance, so it is possible that the authors of study 7 concluded that there was a large, negative difference while the other two studies' authors concluded there was only a small, positive difference. From this we might draw the conservative conclusion that the results are mixed, and therefore we cannot say whether there is any difference between PhD students and others. But because reviewers do not like to sound indecisive, and because big, confidently proclaimed results are more impressive than small, timid ones, the chances are we will swing in favor of the "strong" negative conclusion. In other words, we will be inclined to conclude that PhD students are dumber than the rest of society. Three cheers for Joe Six-pack!

Second, if we were able to obtain a complete summary of all the research on the topic, that is, all seven studies, we could try to reach a conclusion using the vote-counting method. Under the traditional vote-counting procedure discordant findings are decided on the basis of the modal result (Light and Smith 1971). As the majority of studies in the table report nonsignificant results, this would be interpreted as a win for the nonsignificant conclusion. We would be inclined to conclude that there is no difference between PhD students and Joe Six-pack. Yet we might have some misgivings about this conclusion. Given the relatively small samples involved we might suspect that some of the nonsignificant results reflect Type II errors. We note that the results for studies 4 and 5 are identical, yet one result was judged to be statistically significant while the other was not. The difference was that the sample for study 5 had one more person in it. Suspecting an epidemic of underpowered research, we decide to revise the critical level of alpha to .10. At a stroke, three more positive results become statistically significant. Suddenly the positive group is in the clear majority, leading us to conclude that PhD students are indeed smarter than everyone else. Given that this conclusion is based on a clear majority of all the available studies (five out of seven), this seems be a step forward. But we can go no further. The *p* values of the study tell us nothing about the size of the difference or the precision of the estimates. We have no way of telling whether PhD students are a tiny bit smarter or are relative Einsteins.

Third, abandoning the narrative review we could pursue a more quantitative approach by graphing confidence intervals for each of the seven means. This would enable us to gauge the precision of each study's estimate. Confidence intervals for the seven mean PhD scores are shown in Figure 5.1. For each study the reported mean is placed within an interval of plausible values and the width of the interval corresponds to the precision of the estimate – narrow intervals obtained from larger samples are more precise. Looking at the seven intervals we can draw some new conclusions about the set of results. Immediately we can see that study 7 is the odd one out. Study 7's estimate of the mean is well below the estimates reported for the other studies and its confidence interval does not overlap any of the other intervals. Comparing the intervals in this way should cause us to consider reasons why this result was different from the rest.

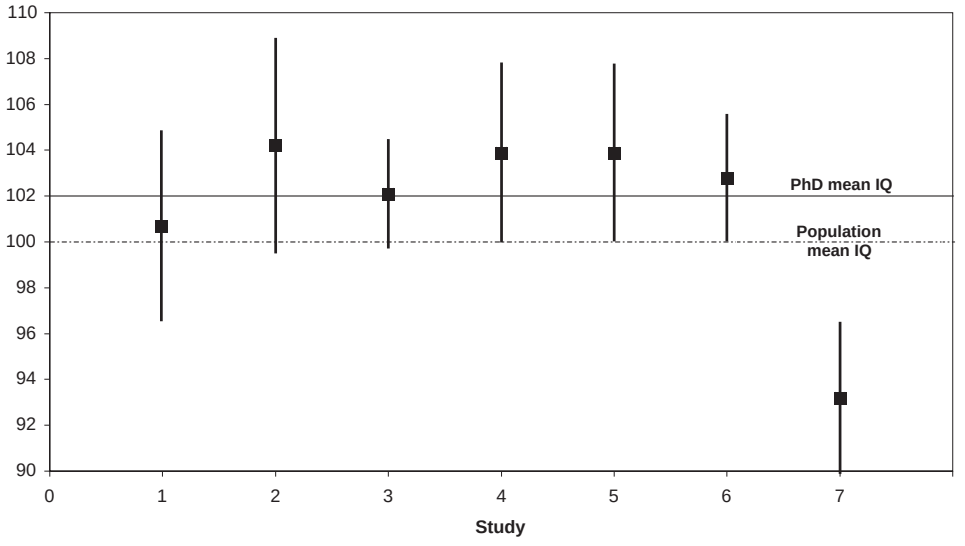


Figure 5.1 Confidence intervals from seven fictitious studies

Examining the confidence intervals leads naturally to meta-analytic thinking. We might conclude that the true mean for the population of PhD students is to be found in the range of overlapping values for the six studies that reported mean values higher than the population mean. (In making this choice we are dismissing study 7 as anomalous. The authors of study 7 might have something to say about that, but we have empirical grounds for reaching this conclusion – the non-overlapping interval.) Although the precision of the first six estimates is variable, the observed means for this group all fall between 100.7 and 104.2. Consequently we might conclude that the true mean is somewhere within this 3.5 point range. This is certainly a more definitive conclusion than what we had before, but the intervals cannot tell us much more than this. We know there is a positive difference and that it is probably greater than .7 IQ points but less than 4.2 points, but we cannot tell exactly how big it is.

Fourth, we could convert the observed differences into standardized effect sizes (d) and calculate an average effect. Seven ds, ranging from −.45 to .28, are listed in the final column of Table 5.2.[2] To interpret these results we could weight each result by the relative sample size and calculate a weighted mean effect size. (The methods for doing this are explained later in this chapter.) This would give us a weighted mean of $d = .13$ which corresponds to a mean IQ of 102.0. A line reflecting this weighted mean effect size has been included in Figure 5.1 and it runs through six of the seven confidence intervals. Summarizing the research in this way would allow us to conclude that PhD students are, on average, 2 IQ points smarter than the general population. Calculating the 95% confidence interval for this mean estimate would permit us to judge the difference as statistically significant as the interval (CI = 100.7–103.3) excludes the null value of 100. Based on this analysis we can conclude that the difference in IQ is real, statistically significant, and, according to Cohen's (1988) benchmarks, utterly

trivial. With reference to the televised IQ test mentioned in Chapter 2 it is a difference no bigger than that separating blondes from brunettes, or rugby fans from soccer fans.[3]

The purpose of this whimsical exercise is to highlight the severe limitations of narrative reviews. Using the narrative approach we found evidence to support all four possible conclusions: that there is no difference, there is a positive difference, there is a negative difference, and we cannot say whether there is any difference. This gives us considerable scope to introduce reviewer bias into our conclusion. If we held the view that PhD students are just as smart but no smarter than everybody else, we might see no reason for adjusting the alpha significance criterion. We would dismiss the results of the first four studies on the grounds that there is a credible risk of Type I errors, meaning none of them met the $p < .05$ cut-off. The evidence that remains – the three statistically significant studies – would be sufficient to confirm our prior belief that PhD students are no different from other people. But we could just as easily marshal evidence to support the alternative view that PhD students are different. If we believed that PhD students are smarter we might be tempted to relax the alpha criterion and accept as valid the three marginally significant results that support our position. If challenged, we would defend our decision on the reasonable suspicion of low statistical power. If these studies had been just a little bigger, chances are their results would have achieved statistical significance. Or if we believed PhD students are dumber, we could highlight the very statistically significant effect reported in study 7. After all, this was the biggest, least equivocal of all the results. As this example shows, it is not hard for narrative reviewers to reach different conclusions when reviewing the same body of research.

Narrative summaries are probably the most common form of literature review but their shortcomings are legion. They are rarely comprehensive, they are highly susceptible to reviewer bias, and they seldom take into account differences in the quality of studies. But the chief limitation of narrative reviews is that they often come to the wrong conclusion. This can happen as a consequence of the vote-counting method. The statistical power of the vote-counting method is inversely related to the number of apparently contradictory studies being compared. The surprising implication is that the probability of detecting an effect using this method falls as the amount of evidence increases (Hedges and Olkin 1980). Wrong conclusions also arise because narrative reviewers typically ignore differences in the precision of estimates. Large effects estimated with low precision are more likely to attract attention than small or null effects estimated with high precision, even though the latter are more likely to be true. In summary, narrative reviews generally cannot provide answers to the two questions posed at the beginning of this chapter, questions that every reviewer seeks to answer.

2. Meta-analysis as a means for generalizing results

A more effective means for assessing the generalizability of results is provided by meta-analysis. Meta-analysis, literally the statistical analysis of statistical analyses, describes a set of procedures for systematically reviewing the research examining a

particular effect, and combining the results of independent studies to estimate the size of the effect in the population. Before they are combined, study-specific effect size estimates are weighted according to their degree of precision. To reduce the variation attributable to sampling error, estimates obtained from small samples are given less weight than estimates obtained from large samples. Individual estimates may also be adjusted for measurement error. The outcome of a meta-analysis is a weighted mean effect size which reflects the population effect size more accurately than any of the individual estimates. In addition, a meta-analysis will generate information regarding the precision and statistical significance of the pooled estimate and the variation in the sample of observed effects.

Although the roots of meta-analysis extend back into the dim history of statistics, the first modern meta-analysis is generally acknowledged as being Gene Glass and Mary Lee Smith's pioneering study of psychotherapy treatments (Glass 1976; Smith and Glass 1977). Like all breakthroughs in research, this one has a good story behind it. In Glass's (2000) version of the tale, indignation was the mother of invention.

In the early 1970s Glass had been fired up by a series of "frequent and tendentious reviews" regarding the merits of psychotherapy written by the eminent psychologist Hans Eysenck. Eysenck had read all the literature on the topic and concluded that psychotherapy was pure bunk. Glass, who had personally benefited from therapy, was miffed by this and set out to "annihilate Eysenck and prove that psychotherapy really works." In Glass's own words, this "was personal."

According to Glass, Eysenck's review of the literature had been compromised by some bad decisions. For one thing, Eysenck considered only the results of published research. This led him to miss evidence reported in dissertations and unpublished project reports. Eysenck also gauged the effectiveness of psychotherapy treatments solely on the basis of statistical test results. A result was judged to indicate "no effect" if it failed to exceed the critical $p < .05$ level. No thought was given to the size of the effect or whether the study had sufficient statistical power to detect it.

Unhappy with both Eysenck's conclusions and methods, Glass and Smith decided to review the literature for themselves. Together they set out to collect all the available evidence assessing the effectiveness of psychotherapy. They ended up analyzing 833 effects obtained from 375 studies. (In contrast, Eysenck's conclusion was based on the evidence of just eleven studies.) The initial results of this meta-analysis, which Glass presented at his 1976 presidential address to the American Educational Research Association, showed that the combined effect of psychotherapy was equivalent to .68 of a standard deviation when comparing treated and untreated groups.[4] Coming at a time when many doubted the benefits of psychotherapy, this was considered a profound validation of the intervention. Just as significant was the process by which this conclusion had been reached. Although Eysenck (1978) and others took the view that combining the results of dissimilar studies was an "exercise in mega-silliness," meta-analysis was widely received as a valid means for reviewing research. Within a short time meta-analyses were being used to examine all sorts of unresolved issues, particularly in the field of psychology (see Box 5.1). Meta-analysis had arrived.

Box 5.1 Is psychological treatment effective?

Glass and Smith's pioneering meta-analysis (Glass 1976; Smith and Glass 1977) spawned hundreds of follow-up meta-analyses and it was only a matter of time before someone thought to assess the results of these meta-analyses meta-analytically. Lipsey and Wilson (1993) did this in their unprecedented study of 302 separate meta-analyses pertaining to various psychological treatments. Within this large set of reviews Lipsey and Wilson identified a smaller, better quality set of 156 meta-analyses from which they drew their conclusions. This smaller set still represented 9,400 individual studies and more than one million study participants. The mean effect size for this set of meta-analyses was 0.47 standard deviations. In plain language this means that a group of clients undergoing psychological treatment will experience a 62% success rate in comparison with a 38% success rate for untreated clients. Lipsey and Wilson (1993: 1198) then presented data showing that while psychologists rarely deal with life and death issues, the benefits of psychological treatment are none the less comparable in magnitude to the benefits obtained with medical treatment.

Meta-analysis offers several advantages over the traditional narrative review. First, meta-analysis brings a high level of discipline to the review process. Many decisions made during the review process are subjective, but the meta-analyst, unlike the narrative reviewer, is obliged to make these decisions explicit. Reading a narrative review one usually cannot tell whether it provides a full or partial survey of the literature. Were awkward findings conveniently ignored? How did the reviewer accommodate outliers or extreme results? But a meta-analysis is like an audit of research. Each step in the process needs to be recorded, justified, and rendered suitable for scrutiny by others. Second, with its emphasis on cumulating data as opposed to conclusions, meta-analysis compels reviewers to become intimately acquainted with the evidence (Rosnow and Rosenthal 1989). Lifting conclusions from abstracts is not enough; reviewers need to evaluate each study's methods and data. Third, and most significantly, meta-analyses can provide definitive answers to questions regarding the nature of an effect even in the presence of conflicting findings.

Consider again the diverse conclusions that have been drawn regarding the effects of market orientation summarized in Table 5.1. We noted that some authors have concluded that market orientation is "strongly related" to organizational performance while others reported that there is only a "minimal" or "rather weak" effect. Still others concluded that there was no effect at all. These inconsistent findings make it virtually impossible to draw a conclusion or estimate the size of the effect using a narrative review. However, four separate meta-analyses of this literature have independently revealed that market orientation does indeed have a positive effect on performance and that the magnitude of that effect is in the $r = .26$–$.35$ range, with 95% confidence intervals ranging from .25–.37.[5] These results tell us that market orientation has a statistically significantly positive effect on organizational performance that is robust

across diverse cultural and industrial settings. (Significance can be inferred from the fact that none of the reported confidence intervals includes zero.) This effect is non-trivial and may even be considered fairly substantial in comparison with other performance drivers studied in the business disciplines.

In principle, meta-analysis offers a more objective, disciplined, and transparent approach to assimilating extant findings than the traditional narrative review. However, in practice meta-analysis can be undermined by all sorts of bias leading to the calculation of precise but erroneous conclusions.

Meta-analysis in six (relatively) easy steps

The purpose of a meta-analysis is to collect individual effect size estimates from different studies and combine them into a mean effect size. The primary output is a single number. To help us interpret this number we would normally compute three other numbers relating to the statistical significance and the precision of the result, and the variability in the sample of observations. To someone who lacks numerical skills, the prospect of crunching these four numbers may seem daunting. But the statistical analyses associated with meta-analysis are not difficult. If you can add, subtract, multiply, and divide, you can combine effect sizes using a variety of approaches. Textbooks on the subject make it look harder than it is.[6]

The meta-analytic process can be broken down into six steps:

1. Collect the studies.
2. Code the studies.
3. Calculate a mean effect size.
4. Compute the statistical significance of the mean.
5. Examine the variability in the distribution of effect size estimates.
6. Interpret the results.

Step 1: Collect the studies

Having selected an effect to study, the reviewer begins by conducting a census of all relevant research on the topic. Relevant research is defined as any study that examines the effect of interest using comparable procedures and which reports effects in statistically equivalent forms. Ideally relevant research would include both published and unpublished research written in any language.

Identifying published research usually involves scanning bibliographic databases such as ABI/Inform, EconLit, Psychological Abstracts, Sociological Abstracts, the Educational Resources Information Center (ERIC) database, MEDLINE, and any other database the reviewer can think of. Access to these sorts of databases has become considerably easier over the years thanks to the emergence of web-based service providers such as EBSCO, ProQuest, Ovid, and JSTOR. Now a reviewer can scan multiple databases in a single afternoon without even leaving their office.

Electronic databases make it relatively easy to identify published research, but a good meta-analysis will also include relevant unpublished research such as dissertations, conference papers, technical reports written for government agencies, rejected manuscripts, unsubmitted manuscripts, and uncompleted manuscripts. Unpublished dissertations can be located using databases such as Dissertation Abstracts International and the Comprehensive Dissertation Index. Conference papers can be found by scanning conference programs which are increasingly available online. The reviewer can post requests for working papers and other unpublished manuscripts on academy websites, discussion groups, or list servers. Informal requests for unpublished papers can also be made to scholars known to be actively researching in the area. Other strategies for identifying the "fugitive literature" of unpublished studies are outlined by Rosenthal (1994).

The search process, which should be fully documented, could lead to hundreds of papers being identified and retrieved. Inevitably, many of these papers will be unsuitable for inclusion in a meta-analysis as they will not report the collection of original, quantitative data. The reviewer will need to weed out all those papers that do not report data (e.g., conceptual papers, research reviews, and research proposals) as well as those studies that are based on the analysis of qualitative data (e.g., ethnographies, naturalistic inquiries, and case studies). Getting rid of these types of papers is straightforward, but the next step involves a judgment call. Of the studies that remain, how does the reviewer decide which to include in the meta-analysis?

The ideal meta-analytic opportunity is a well-defined set of studies examining a common effect using identical measures and analytical procedures. But in practice it is virtually impossible to find even two studies sharing identical measures and procedures.[7] The temptation will be to throw all the evidence into the mix to see what comes out. But mixing studies indiscriminately gives rise to the concern that meta-analysis seeks to compare apples with oranges.

There are several tactics for dealing with the apples and oranges problem. One tactic is to articulate clear criteria for deciding which studies can be included in the meta-analysis. At a minimum, eligibility criteria should cover measurement procedures and research designs. For example, the reviewer might include only those studies that collected experimental data based on the random assignment of subjects. Or the reviewer might analyze only those studies that collected survey data and that measured key constructs using a similar set of scales. Other eligibility criteria might relate to the characteristics of respondents and the date of publication (e.g., studies published after a certain date). Other criteria that are more contentious include publication type (e.g., peer-reviewed research only) and publication language (e.g., English language studies only). Criteria of this type can introduce bias into a meta-analysis, as we will see in the next chapter.

Step 2: Code the studies

From the initial set of papers, the reviewer will identify a smaller set of empirical studies that has used comparable procedures and that reports effects in statistically equivalent

forms. There could be anywhere from a few to several hundred studies in this group. If there are only a few studies the possibility of abandoning the meta-analysis should be considered as there may be insufficient statistical power to detect effects even after the studies have been combined. Guidelines for deciding whether to proceed when only a handful of studies has been found are discussed in Chapter 6. However, if there are a large number of studies in the database the reviewer might want to consider coding only a portion of them. The issue is one of diminishing returns. While four studies are definitely better than two studies, 400 studies are only marginally better than 200 studies. Will the benefits of including an additional 200 studies offset the cost in time of coding them? Cortina (2002) makes the case that if one has found many studies, one may be able to exclude some of them as long as (a) one retains a sufficient number to test all the relationships of interest and (b) one can show that coded and uncoded studies do not differ on variables that might affect the calculation of the mean effect size. For instance, it would be misguided in the extreme to code only the published studies (because they were easy to find) and ignore unpublished studies.

If the reviewer decides to proceed with the meta-analysis, the next step is to prepare the studies by assigning numerical codes to study-specific characteristics. Coding renders raw study data manageable and enables the reviewer to turn a large pile of papers into a single database. At a minimum, the reviewer will need to code the results of each study (e.g., the effect types and sizes) along with those study characteristics that affect the accuracy of the results (e.g., the sample size and the reliability of key measures). Locating this information for a large set of studies may require hundreds of hours of careful reading. Hunt (1997) compares this work to panning for gold – tiresome work punctuated by the burst of exhilaration at finding the occasional nugget. In this case the nuggets are quantifiable effects that can be included in the meta-analysis.

It is likely that effects will be reported in a variety of forms. Some will be reported as effects in the *r* family (e.g., Pearson correlation coefficients, R^2s, beta coefficients, Cramér's Vs, omega-squares) while others will be reported as effects in the *d* family (e.g., odds ratios, relative risks, Glass's deltas). Before these effects can be combined they will need to be transformed into a common metric. The reviewer may choose to convert all the *r* effects to *d* effects or vice versa. The easiest approach is to adopt the metric most frequently reported in the research being reviewed. If most of the effects are reported as correlation coefficients or their derivatives, and only a few of the effects are reported as group differences, it makes sense to transform the latter into *r*s. However, if the modal effect is expressed in terms of group differences, then it makes sense to transform all the *r*s into *d*s. Any *d* can be transformed into *r* or vice versa using the equations found on page 16. If roughly an equal mix of *r* and *d* effects is reported, the best approach is to convert everything into *r*s. Effect sizes expressed in the *r* form have several advantages over *d* and converting an *r* into a *d* usually involves some loss of information (Cohen 1983). Measures of association also have the advantage of being bounded from zero to one whereas Cohen's *d* has no upper limit.

Where effect sizes are not reported directly, the reviewer may have to do some calculations. For example, both *r* and *d* can also be computed from certain test statistics as well as *p* values.[8] In some cases information regarding the size of the effect may be missing from a research paper. This can happen when a result is reported as nonsignificant or NS with no further information provided. This can also happen when papers report only omnibus effects for a set of predictors (e.g., R^2) and provide no data on individual effects (e.g., bivariate or part correlations). Faced with incomplete data the reviewer may need to contact the study authors directly to solicit information on the size of the effect observed. In the case of *r* effects this may entail little more than asking for a correlation matrix.

Effect sizes combined in the meta-analysis need to be based on independent observations. This means the reviewer will need to be aware of multiple papers that report the results of the same study. Only one set of results should be included in the analysis. A related issue is when a single paper reports multiple effects drawn from the same set of data. Some studies report dozens, even hundreds of effects. Recording all these effects separately can lead to the problem of "inflated *N*s" with adverse consequences for the generalizability of the meta-analysis (Bangert-Drowns 1986). This problem can be avoided by calculating an average effect size for each study. However, if there are potentially interesting differences in the ways in which effects are reported within studies, these differences can be coded and examined. For instance, if the outcome of interest is performance and studies tend to report two distinct measures of performance (e.g., objective and subjective performance), the reviewer might want to record the individual performance effects along with an average effect (e.g., overall performance) for each study. This would give the reviewer the option of calculating a main effect for overall performance across all studies and then running a moderator analysis to see whether that effect is affected by the type of performance being measured. This could be done by comparing the mean effect size obtained when performance was measured objectively with the mean result found when performance was measured subjectively. Differences between these two means would reveal the operation of a measurement moderator.

Apart from converting effect sizes into a common metric, the reviewer may also want to adjust study-specific estimates for measurement error. Measurement error attenuates effect sizes by adding random noise into the estimates. We can compensate for this if studies provide information regarding the reliability of measures. Effect size estimates are adjusted by dividing by the square root of the reliabilities, as shown in Chapter 3. If some studies neglect to provide information on the reliability of measures, then a mean reliability value obtained from all the other studies can be used.

The reviewer may also wish to code various study-specific features such as the data-collection methods, the sampling techniques, the measurement procedures, the research setting, the year of publication, the mode and language of publication. Coding this information makes it possible to analyze the impact of potential moderators. For example, to assess the possibility of publication bias the reviewer may compare the mean effect sizes reported in published versus unpublished studies. If the mean of the

published group is substantially higher than the mean of the unpublished group, this could be interpreted as evidence of a publication bias favoring statistically significant results. Coding measurement procedures and research settings would also enable the reviewer to assess whether effect size estimates had been affected by the choice of instrument or the location of the study. Judicious coding thus offers the reviewer another remedy for the apples and oranges problem.[9]

Whatever can be coded can be analyzed later as a potential moderator. But coding is hard, mind-numbing work. It starts out being fun but often ends with the reviewer abandoning the project out of frustration or fatigue. Many of those who make it through the coding process never wish to repeat the experience. In the first modern meta-analysis, Glass, Smith, and four research assistants scanned 375 studies for 100 items of information, some of which had 10–20 different coding options. Smith later said of the exercise, "it was incredibly tedious and I would never do it again" (Hunt 1997: 40).[10]

The coding of a set of studies presents at least three challenges to the reviewer. The first challenge is deciding what not to code. The problem is that initially everything looks promising and the reviewer will want to code it all. But as each new code increases the coding burden, the reviewer will need to quickly decide which codes are most likely to bear fruit in the eventual meta-analysis. This is a tough decision because often there is no way of knowing in advance which codes will prove to be useful. Erring on the side of caution the reviewer will be inclined to include more codes than necessary. The upshot is that the reviewer will spend unnecessary days and weeks engaged in coding knowing full well that much of the work will be for naught.

The second challenge is to devise a set of clear, unambiguous coding definitions that are interpreted in the same way by independent coders. The best way to test this is to measure the proportion of interrater agreement. This can be done by getting two or more reviewers to code the same subset of studies (at least twenty) and then comparing their coding assignments. Interrater agreement is defined as the number of agreements divided by the sum of agreements plus disagreements. High scores close to one indicate that coding definitions are sufficiently clear. Often several rounds of coding and definition revising are needed before acceptable levels of interrater agreement are reached.[11]

Third, and hopefully with assistance from others, the reviewer has to actually code all the studies. During the process of coding studies, it is likely that the reviewer will identify additional variables or more efficient ways to code. These discoveries are a mixed blessing for they improve the efficiency of the coding exercise while compelling the reviewer to recode studies that have already been done.

Step 3: Calculate a mean effect size

At the end of step 2 the reviewer will have a database of effect sizes and will be ready to begin crunching numbers. If the first two steps have been done carefully, and the reviewer has survived the sheer drudgery of coding, then the anticipation of calculating a mean effect size can be quite thrilling. Months of searching, reading, and coding will

Table 5.3 *Kryptonite and flying ability – three studies*

Study	*r*	*p*	*n*	DV reliability*
Luthor (1940)	−.48	0.02	80	0.70
Brainiac (1958)	−.58	<0.001	112	0.92
Zod et al. (1961)	.05	0.33	32	–

* Cronbach's alpha

have led up to this point where one is on the verge of being able to answer, how big is this effect?

The worst way to combine the individual effect sizes is to simply average them. A far better alternative is to calculate a weighted mean effect size after each individual estimate has been corrected for measurement error. There are different procedures for weighting effect size estimates, but the easiest method, and arguably the best, is to weight estimates by their corresponding sample size. (Other methods are described in Appendix 2.) A simple example will show how this is done.

Let's assume that we have a special interest in the effect of kryptonite on flying ability. Anecdotal reports in the *Daily Planet* newspaper suggest that exposure to kryptonite has an adverse effect on the ability to fly, but definitive conclusions are lacking. We do a census of all the available research and identify three studies reporting effects, as in Table 5.3. With only three studies this is going to be a very small meta-analysis! But the sad fact is that the effects of kryptonite on various superpowers have been examined by only a few highly motivated individuals operating well outside the scientific establishment.

Looking at the table we see that each study reports a unique effect size estimate (r) and sample size (n) and that two of the studies also report the measurement reliability (Cronbach's α) of the dependent variable. (We can ignore the p value of each study. The statistical significance of extant hypothesis tests is generally irrelevant to meta-analysis.) Using these data we can calculate a mean effect size three different ways. The easiest way is to calculate the *simple mean effect size* of the three correlations, which is $(-.48 + -.58 + .05)/3 = -.337$. Note how this mean effect size is considerably smaller than two of the three observed correlation coefficients, which hints at how the simple mean can be biased. In our example the correlation coefficient reported by Zod et al. seems to be unusual. These authors reported an effect which was trivially small and positive in direction in contrast with the much larger, negative effects of the other two studies. We have good reasons to be somewhat suspicious of the result reported by General Zod and his allies. We note that their study seems underpowered (it has a relatively small sample of observations) and that they did not take much care in either measuring effects or reporting the reliabilities of their measures.

As this third estimate may be dampening our mean result, we might be tempted to discard it as an outlier. If we did this Zod and his colleagues would no doubt accuse us

of introducing reviewer bias into the analysis. A more justifiable approach is to retain all the studies but place greater weight on the results obtained from larger samples. This is reasonable because estimates obtained from larger samples will be less biased by sampling error. To calculate a *weighted mean effect size* we multiply each effect size estimate by its corresponding sample size and divide by the total sample size, as shown in equation 5.1. In this equation n_i and r_i refer to the sample size and correlation in study i respectively:

$$\begin{aligned}\bar{r} &= \frac{\sum n_i r_i}{\sum n_i} \\ &= \frac{(80 \times -.48) + (112 \times -.58) + (32 \times 0.05)}{80 + 112 + 32} \\ &= \frac{(-38.4) + (-65.0) + (1.6)}{224} \\ &= -.454\end{aligned} \tag{5.1}$$

Note how the weighted average (−.454) is larger in absolute terms than the unweighted average (−.337) and is closer to the two effect size estimates returned by the more credible studies of Luthor and Brainiac. In other words, the weighted estimate is better.

We can further improve the quality of our weighted mean by accounting for the measurement error attenuating each estimate. We can see from Table 5.3 that the procedure used to measure the dependent variable in Luthor's study was less reliable than the procedure used in Brainiac's. We know this by looking at the Cronbach's alphas in the last column (high alphas indicate internally consistent measures). Both estimates of effect size will be suppressed because of measurement error, but Luthor's will be more so than Brainiac's. No doubt Zod et al.'s estimate is also attenuated, but as they provided no information on reliability we will have to substitute the mean Cronbach's alpha obtained from the other two studies.

We correct for measurement error by dividing each study's effect size by the square root of the reliability of the measure used in that study ($r\sqrt{\alpha}$). The corrected estimate for Luthor's study is $-.48/\sqrt{.70} = -.574$ and the corrected estimate for Brainiac's study is $-.58/\sqrt{.92} = -.605$. The mean reliability value is $(.70 + .92)/2 = .81$ so the corrected estimate for Zod et al.'s study is $.05/\sqrt{.81} = .056$.[12] With these corrected estimates we can calculate the *weighted mean corrected for measurement error* as follows:

$$\begin{aligned}&= \frac{(80 \times -.574) + (112 \times -.605) + (32 \times 0.056)}{80 + 112 + 32} \\ &= \frac{(-45.92) + (-67.76) + (1.79)}{224} \\ &= -.500\end{aligned}$$

We can see from this exercise that our meta-analytic results are affected by the calculation used. Our mean estimates ranged from −0.337 to −0.500. However, we can have the greatest confidence in the third result as it is the least biased by the sampling and measurement error of the individual studies.

Step 4: Compute the statistical significance of the mean
There are two complementary ways to calculate the statistical significance of the mean effect size: (1) convert the result to a *z* score and then determine whether the probability of obtaining a score of this size is less than .05, or (2) calculate a 95% confidence interval and see whether the interval excludes the null value of zero.[13] In either case we will need to know the standard error associated with our mean effect size. Recall from Chapter 1 that the standard error describes the spread or variability of the sampling distribution. In other words, it is a special kind of standard deviation. In the kryptonite example the sampling distribution consists of just three effect size estimates. It may be small, but it has a spread or variance. Some authors prefer the term variance to standard error but the terms are interchangeable as the standard error is the square root of the variance. The variance of the sample of correlations ($v_{.r}$) can be found by multiplying the square of the difference between each estimate and the mean by the sample size, summing the lot, and then dividing the result by the total sample size, as follows:

$$
\begin{aligned}
v_{.r} &= \frac{\sum n_i (r_i - \overline{r})^2}{\sum n_i} \qquad (5.2)\\
&= \frac{(80 \times (-.574 - -.500)^2) + (112 \times (-.605 - -.500)^2) + (32 \times (.056 - -.500)^2)}{80 + 112 + 32}\\
&= \frac{(80 \times .005) + (112 \times .011) + (32 \times .309)}{224}\\
&= \frac{.400 + 1.232 + 9.888}{224}\\
&= .051
\end{aligned}
$$

An important point which we will return to in Chapter 6 is to consider whether there are not one but two samples, namely, the sample of observations (or estimates) and a higher-level sample of population effect sizes (or parameters). Many meta-analyses are done as if there was just one actual effect size, but often there are many. Real-world effects may be bigger or smaller for different groups. Consequently reviewers will often need to account for the variance in the sample of estimates as well as the variance in the sample of effect sizes. If this second source of variance is not accounted for, confidence intervals will be too narrow and tests of statistical significance will be susceptible to Type I errors. To keep things moving along for now, we will account for the variability in both distributions in the calculation of the standard error ($SE_{\overline{r}}$). We do this by dividing the observed variance by the number of studies (k) in the meta-analysis and

then taking the square root (Schmidt and Hunter 1999a). The equation is as follows:

$$\begin{aligned} SE_{\bar{r}} &= \sqrt{\frac{v_{\cdot r}}{k}} \\ &= \sqrt{\frac{.051}{3}} \\ &= .130 \end{aligned} \tag{5.3}$$

With this standard error we can convert the mean correlation into its standard normal equivalent or standard score. A standard score, or *z* score, conveys the magnitude of an effect in terms of standard deviation units. To convert the *r* score into its corresponding *z* score we divide the absolute value of the mean effect size by its standard error, as follows:

$$\begin{aligned} z &= \frac{|\bar{r}|}{SE_{\bar{r}}} \\ &= \frac{.500}{.130} \\ &= 3.85 \end{aligned} \tag{5.4}$$

To interpret the statistical significance of this result we compare it with the critical value of *z* for our chosen standard of significance. If $\alpha_2 = .05$, the critical value of *z* (or $z_{\alpha/2}$) is 1.96. If $\alpha_2 = .01$, then $z_{\alpha/2} = 2.58$. We would reject the null hypothesis in a two-tailed test whenever our *z* score exceeds $z_{\alpha/2}$. In this case 3.85 is greater than the critical value of $z_{\alpha/2}$ permitting us to conclude that the result is statistically significant.

The second way to assess the statistical significance of a result is to calculate a 95% confidence interval. In this regard a 95% confidence interval is analogous to the alpha significance criterion of $p < .05$. A 95% confidence interval that excludes 0 puts the odds of $\bar{r} = 0$ beyond reasonable possibility and indicates that the mean effect size is statistically significant at $\alpha = .05$.

In Chapter 1 we saw that the width of an interval is the mean plus or minus the standard error multiplied by a critical *t* value. But in meta-analysis we are more likely to use a critical *z* value. Both values come from central distributions, but the *t* distribution is preferred when sample sizes are small or "less" normal. As sample sizes increase, the *t* distribution begins to resemble the *z* distribution and the critical values of both distributions converge. Most of the time in meta-analysis we will be dealing with healthy sample sizes, so it is just easier to adopt the critical *z* value of 1.96 instead of fishing around with *t* tables.

The equation for calculating the width of a confidence interval is: $\bar{r} \pm z_{(\alpha/2)} SE_{\bar{r}}$. Using the numbers obtained above, the lower and upper bounds of a 95% interval can be determined, as follows:

$$\begin{aligned} \text{CI}_{95\text{lower}} &= -.500 - (1.96 \times 0.13) = -.755 \\ \text{CI}_{95\text{upper}} &= -.500 + (1.96 \times 0.13) = -.245 \end{aligned}$$

Box 5.2 Credibility intervals versus confidence intervals

The difference between a confidence interval and a credibility interval relates to two distinct distributions: the distribution of effect size *estimates* reported in each study and the higher-level distribution of actual *effect sizes* in the population. The distribution of estimates is centered on the mean observed effect size (e.g., $\bar{r}$), while the distribution of effect sizes is centered on the mean population effect size ($\bar{\rho}$). Intervals which bound the observed mean are called confidence intervals while intervals which bound the population mean are called credibility intervals. The width of a confidence interval is determined by standard error of the observed mean ($SE_{\bar{r}}$) and reflects the amount of sampling error in the estimate of the mean. In contrast, a credibility interval is based on the standard deviation of the population effect size ($SD_{\bar{\rho}}$) and is unaffected by sampling error.

The equations for calculating standard errors and confidence intervals are based on the variance observed in the sample of effect size estimates ($v_{.r}$). In contrast, credibility intervals are based on the variance in the distribution of population effects ($v_{.\rho}$). This variance is the difference between the variance observed in the sample correlations minus the variance attributable to sampling error ($v_{.e}$), or

$$v_{.\rho} = v_{.r} - v_{.e}$$

The variance in sample correlations is the frequency-weighted average squared error, given in equation 5.2 in the text. The variance attributable to sampling error is calculated using the average uncorrected correlation ($\bar{r}$) and the average sample size ($\overline{N}$), as follows:

$$v_{.e} = (1 - \bar{r}^2)^2/(\overline{N} - 1)$$

Subtracting the sampling error variance ($v_{.e}$) from the variance in the sample correlations ($v_{.r}$) gives the population variance ($v_{.\rho}$). The square root of the population variance is the standard deviation of the population effect size ($SD_{\bar{\rho}}$).

The upper and lower bounds of both types of interval are found by multiplying $SE_{\bar{r}}$ or $SD_{\bar{\rho}}$ by the critical value of $z_{\alpha/2}$ and adding or subtracting the result to the mean effect size. The standard error of the observed mean is found by dividing $v_{.r}$ by the number of studies and taking the square root as in equation 5.3. For a 95% interval the corresponding equations are:

$$\text{Confidence interval: } = \bar{r} \pm 1.96 SE_{\bar{r}}$$
$$\text{Credibility interval: } = \bar{r} \pm 1.96 SD_{\bar{\rho}}$$

To calculate credibility intervals for *d* effects, the variation in the data attributable to sampling error variance is subtracted from the observed variance as before. However the equation for calculating sampling error variance is different, as follows:

$$v_{.e} = [(\overline{N} - 1)/(\overline{N} - 3)][(4/\overline{N})(1 + \bar{d}^2/8)]$$

where $\overline{N}$ is the total sample size divided by the number of studies and $\bar{d}$ is the average effect size.

Sources: Hunter and Schmidt (2004: 81, 88–89, 205, 288); Whitener (1990)

As this interval excludes the null value of zero, we can conclude that the result is statistically significant.

In this example we calculated a confidence interval for a mean effect size that has been corrected for measurement error. Technically, this is not an appropriate thing to do because the disattenuation of estimates, while improving the accuracy of the mean, increases the sampling error in the variance, making confidence intervals wider than they should be. The standard error calculated using the corrected estimates was .130, but a standard error calculated on uncorrected estimates would be .122. The implication is that the confidence intervals just calculated are about 7% too big. While this is not a big difference in absolute terms, in borderline cases it could mean that an otherwise good result is judged to be statistically nonsignificant. To remedy this we can create an interval that is unaffected by sampling error variance. This can be done by isolating and removing the variation in the distribution of correlations that is attributable to sampling error. What is left is the variation attributable to the natural distribution of effect sizes. Taking the square root of this natural or population variance enables us to calculate a credibility interval, as explained in Box 5.2.

Step 5: Examine the variability in the distribution of effect size estimates

A wide confidence interval indicates that the distribution of effect sizes is likely to be heterogeneous. This would normally be interpreted as meaning that effect sizes are not centered on a single population mean but are dispersed around several population means – more on this in Chapter 6. To test the hypothesis that the distribution is homogeneous (i.e., that there is only a single population mean), the reviewer can calculate a Q statistic to quantify the degree of difference between the observed and expected effect sizes. The results are interpreted using a chi-square distribution for $k - 1$ degrees of freedom, where k equals the number of effect sizes in the meta-analysis. A Q statistic that exceeds the critical chi-square value would lead to the rejection of the hypothesis that population effect sizes are homogeneous. Effect size samples that are found to be heterogeneous then become candidates for moderator analysis.

To calculate a Q statistic we multiply the difference between the observed (r_i) and expected effect sizes ($\bar{r}$) for each study by some weight and sum the results. When dealing with correlations the relevant weight is usually the sample size minus one ($n_i - 1$). A Q statistic can be calculated from the kryptonite data, as follows:

$$
\begin{aligned}
Q &= \sum (n_i - 1)(r_i - \bar{r})^2 \qquad (5.5)\\
&= ((80 - 1) \times (-.574 - -.500)^2) + ((112 - 1) \times (-.605 - -.500)^2)\\
&\quad + ((32 - 1) \times (.056 - -.500)^2)\\
&= (79 \times .005) + (111 \times .011) + (31 \times .309)\\
&= 0.395 + 1.221 + 9.579\\
&= 11.195
\end{aligned}
$$

To interpret this result we need to consult a table listing critical values of the chi-square distribution. Such a table can usually be found in the back of any statistics or research methods text and will list values by various levels of alpha and degrees of freedom. The critical value that intersects the upper tail area or alpha of .05 and two degrees of freedom is 5.991. As the *Q* statistic of 11.195 exceeds this critical value, the homogeneity hypothesis is rejected. We can conclude that the population of effect sizes is heterogeneous, meaning different effects have been observed for different groups.[14]

Under other circumstances this *Q* statistic would motivate us to search for moderators, but with so few studies in the kryptonite meta-analysis there may be little point. If we had more studies we might consider organizing them into subgroups to assess the relative impact of various measurement and contextual moderators. For example, if there were two ways of measuring flying ability, we could group studies according to their measurement choice and then calculate a weighted mean effect size for each group. Statistically significant differences between the group means would indicate that effect of kryptonite on flying ability is moderated by the procedures used to measure flying ability.

Step 6: Interpret the results

At the end of step 5, with the review essentially finished, the temptation will be to report the results and get the study published. However, one more step is needed – to interpret the results of the meta-analysis. If steps 3 to 5 revealed how big the effect is, whether it is statistically significant, and whether it is moderated by contextual or other variables, then step 6 answers the question, what does it all mean? The challenge here is for the reviewer to interpret the practical significance of the meta-analytic results.

Extracting meaning from research results is a challenge that many researchers avoid. This may be because interpretation is an inherently subjective process – what means one thing to you may mean something else to another. But no one is better positioned to interpret the data than the reviewer who has spent months reading it, coding it, and combining it.

Interpretation is increasingly a requirement for publication in top-tier journals and this is particularly true of meta-analyses. Ten years ago a meta-analysis that was technically sound was almost guaranteed to fly at a good journal. As long as you were meticulous in the collection and coding of studies and knew how to pool the results, you could be reasonably confident of getting a nice publication. Times have changed. Now editors expect meta-analyses to make a clearly identifiable contribution to theory. In other words, editors want reviewers to interpret the theoretical significance of their results. Consider the following advice which comes from a former editor at the prestigious *Academy of Management Journal*:

> AMJ will publish the meta-analyses that fulfill the promise of the method's champions: advancing theoretical knowledge. A meta-analysis that merely tallies the existing literature quantitatively but provides no new insights into the nature of the relationships so tallied will not be favored. A meta-analysis that sheds new light on how or why a relationship or set of relationships occurs should be (re)viewed favorably. (Eden 2002: 844)

Identifying the contribution to theory is just one part of the interpretation challenge. It is likely that in years to come editors will want more, that is, they will ask for a broader evaluation of the practical significance of the results. This means reviewers will need to consider questions such as those raised in Chapter 2: Are the results reported in non-arbitrary metrics that can be understood by nonspecialists? What is the context of this effect? Who is affected or who potentially could be affected by this result and why does it matter? What are the consequences of this effect and do they cumulate? What is the net contribution to knowledge? Does this result confirm or disconfirm what was already known or suspected? And, when all else fails, what would Cohen say? Would he consider this result to be small, medium, or large?

The interpretation challenge is here to stay. Researchers with an eye on the future will recognize an opportunity to explore new methods for extracting and communicating the meaning of a study's results. Confidence intervals and other graphical displays are likely to become more common, but these are only an initial step. New methods and protocols will be developed. New books will be written and new Cohens will emerge. This bodes well for the future of social science research as more thoughtful interpretation of empirical results will ultimately lead to the posing of more interesting research questions and the development of better theory.

Other types of meta-analysis

Within ten years of Glass and Smith's pioneering study, there were at least five different methods for running a meta-analysis (Bangert-Drowns 1986). Since then the number of methods has increased further, but two methods have emerged, like Coke and Pepsi, to dominate the market. These are the methods developed by Hunter and Schmidt (see Hunter and Schmidt 2000; Schmidt and Hunter 1977, 1999a) and by Hedges and his colleagues (see Hedges 1981, 1992, 2007; Hedges and Olkin 1980, 1985; Hedges and Vevea 1998). The kryptonite studies above were aggregated following the "bare bones" meta-analysis of Hunter and Schmidt (2004).[15] To illustrate the differences between the methods, the same data are combined in Appendix 2 using the procedures developed by Hedges et al.

Meta-analysis as a guide for further research

The gradual accumulation of evidence pertaining to effects is essential to scientific progress. In a research environment characterized by low statistical power, an interesting result may be sufficient to get a paper published in a top journal, but ultimately it counts for little until it has been replicated. The results of many replications can be subsequently combined using meta-analysis and this, in turn, can stimulate new ideas for research and theoretical development.

Meta-analysis and replication research

There are very few exact replications in the social sciences, but many studies contain at least a partial replication of some earlier study. These replications are essential to

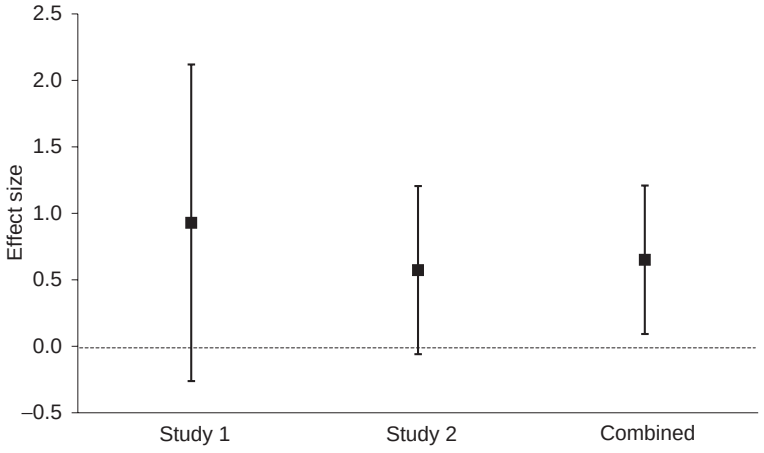


Figure 5.2 Combining the results of two nonsignificant studies

meta-analysis for without repeated studies there would be nothing for reviewers to combine. Yet the relationship goes both ways because the value of any replication is often not fully realized until someone does a meta-analysis. Some would even say that individual studies have no value at all except as data points in future meta-analyses (Schmidt 1996). The implication of this extreme view is that authors of individual studies need not waste their time drawing conclusions or running tests of significance as these will be ignored by meta-analysts. While this view is certainly controversial, most would agree that meta-analysis provides the best means for generalizing the results of replication studies.

To illustrate the symbiotic relationship between replication research and meta-analysis, recall the "failed" Alzheimer's study introduced in Chapter 1. In that study the sample size was twelve, the observed effect was equivalent to 13 IQ points, and the results were statistically nonsignificant ($t = 1.61$, $p = .14$). Imagine that the Alzheimer's researcher had followed up with a second and larger study. Based on a prospective power analysis she set the sample size of the replication study to forty, which should have been sufficient to detect an effect of similar size as that observed in the first study. However, in the second study the observed effect was smaller as the treatment led to an improvement of only 8 IQ points. As with the first study the difference between the treatment and control groups was not statistically significant ($t = 1.81$, $p = .08$). With two nonsignificant results in her pocket our researcher might be tempted to throw up her hands and abandon the project. But meta-analyzing these two results would reveal this to be a bad decision. The effect sizes (ds) for the first and second study were 0.93 and 0.57 respectively. Weighting and combining these results generates a mean effect size of 0.65 and a 95% confidence interval (0.09–1.21) that excludes the null value of zero (see Figure 5.2).[16] In contrast with the results of the two studies on which it is based, the result of the meta-analysis is statistically

significant and conclusive. The meta-analysis provides the best evidence yet regarding the effectiveness of the experimental treatment.[17]

As this thought experiment illustrates, there can be no meta-analysis without a replication study, but the value of any replication cannot be fully appreciated without a meta-analysis. In this example the meta-analysis generated a conclusion that could not have been reached in either of the individual studies. Viewed in isolation, neither of the two Alzheimer's studies provided unequivocal evidence for the experimental treatment. But viewed meta-analytically the results are a compelling endorsement. The treatment works.

Meta-analysis informs further research in four ways. First, as we have just seen, meta-analysis can be used to draw conclusions even from inconclusive studies. Meta-analysis combines fragmentary findings into a more solid evidentiary foundation for further research. If the Alzheimer's researcher was to apply for additional research funds, the statistically significant meta-analytic result would provide a stronger justification for continuing the investigation than the two nonsignificant results. Second, meta-analysis provides the best effect size estimates for prospective power analyses. Prior to running the second Alzheimer's study the researcher ran a power analysis using the effect size estimate obtained from the first study. In doing so she was setting herself up to fail because the second study was empowered only to detect similarly large effects. But given the meta-analytic revelation that the true effect is not large but medium in size, future studies are less likely to be designed with insufficient power. Third, meta-analysis provides non-zero benchmarks against which future results can be compared. This can lead to more meaningful hypothesis tests than merely testing against the null. If we already know that a treatment is effective, what need is there for further research except to identify those conditions under which the treatment is more or less effective? Fourth, and by virtue of its scale, meta-analysis provides an opportunity to test hypotheses that were not tested, or could not have been tested, in the individual studies being combined. This can lead to new discoveries and stimulate the development of theory.

Meta-analysis as a tool for theory development

In addition to aiding the design and interpretation of replication research, meta-analysis can promote theory development. Meta-analysis does this by providing a clear understanding of those effects that can be explained by theory and by generating new leads for further theoretical development. Good theory building requires a solid empirical foundation. Meta-analysis provides this by "cleaning up and making sense" of the research literature (Schmidt 1992: 1179). Loose ends are tied up, relational links are brought into sharp focus, and potentially interesting directions for further work are highlighted. These new leads often take the form of situational or contextual moderators whose operation may not have been discernable at the level of the individual study. For example, in his meta-analysis of market orientation research, Ellis (2006) observed that effects were relatively large when measured in mature, western markets

and relatively small when measured in underdeveloped economies that are culturally distant from the US. Not surprisingly, this moderating effect had gone unnoticed in all of the studies included in the meta-analysis. As the majority of studies were set in a single country, cross-country comparisons were impossible. It was only by combining the results of studies done in twenty-eight separate countries that the moderating effect became apparent. This, in turn, led to a number of original conjectures and hypotheses that were examined in subsequent studies (see Ellis 2005, 2007).

Meta-analysis should not be viewed as the culmination of a stream of research but as a periodic stock take of current knowledge. The really attractive feature of meta-analysis is not that it settles issues but that it leads to the discovery of wholly new knowledge and the posing of new questions. Even a meta-analysis that fails to estimate a statistically significant mean effect size can achieve this outcome if the analysis of moderator variables stimulates hypotheses that can be examined in the next generation of studies (Kraemer et al. 1998). The implication is that the value of any meta-analysis is more than the sum of its empirical parts. Value is also created to the extent that the meta-analysis promotes the development of new theory.

Summary

Back in the mid-1970s when they were coding studies it is possible that Glass and Smith imagined that their pioneering meta-analysis would be the final word on the benefits of psychotherapy. After all, who could argue with a meta-analysis based on the combined findings of nearly 400 studies? They may have thought they were settling an argument, but in reality they were providing reviewers with a new method for systematically assessing the generalizability of results. Within fifteen years there were more than 300 meta-analyses measuring the benefits of various psychological treatments alone (Lipsey and Wilson 1993). Initially the attraction of meta-analysis was that it offered a powerful alternative to the narrative review for drawing conclusions even from studies reporting disparate results. Meta-analysis also led to better designed replication research by providing effect size estimates that could be plugged into prospective power analyses and that could also serve as non-zero benchmarks. More recently meta-analysis has been found useful in stimulating the development of new theory and signaling promising directions for further research.

Meta-analysis has become valued as a tool for researchers looking for accurate effect size estimates. In the medical field meta-analyses, or systematic reviews as they are sometimes called, are considered among the highest levels of evidence available to practitioners (Hoppe and Bhandari 2008).[18] Meta-analyses also reduce the volume of reading that researchers must do to stay abreast of new developments.[19] As more journals are launched and more studies are done, meta-analysis will become an even more essential means for coping with information overload (Olkin 1995).

The attractions of meta-analysis are compelling and easily sold. But here is the fine print: even a carefully done meta-analysis can be ruined by a variety of biases. Doorways to bias line the review process at every stage. Keeping these doors closed

requires vigilance and improvisation on the part of the reviewer. Despite this, it is virtually inevitable that some types of bias (e.g., the exclusion of results not submitted for publication or not reported in English) will affect the calculation of mean effect sizes. Minimizing the risks and consequences of bias is the subject of the next chapter.

Notes

1 A similar thought experiment is discussed by Vacha-Haase and Thompson (2004). Theirs involves nine studies examining the happiness of psychologists.

2 The standard deviation used in determining the standardized mean difference was that of the larger population, so technically this difference should be labeled Glass's Δ rather than *d*. As elsewhere in this book, *d* is used here in a generic sense to signify that we are comparing the means of two groups.

3 If the result from study 7 is excluded from the meta-analysis, Cohen's *d* rises to .18, representing an IQ difference of 2.7 points. The interpretation does not change: this is a trivial difference.

4 Hunt (1997: 34) interpreted Glass's (1976) result in terms of the benefits accruing from twenty hours of psychotherapy. "While the median treated client (at the middle of the curve) was as mentally ill before therapy as the median control client – healthier than half of them, unhealthier than the other half – after therapy, the treated client was healthier than three-quarters of the untreated group. In the social sciences, so large an effect of any intervention . . . is almost unheard of."

5 The weighted mean effect sizes returned from the four meta-analyses are as follows: $\bar{r} = .26$ (Ellis 2006: CI $\bar{r} = .25$–$.28$), $\bar{r} = .28$ (Shoham et al. 2005: CI not provided), $\bar{r} = .32$ (Kirca et al. 2005: CI not provided), $\bar{r} = .35$ (Cano et al. 2004: CI = .33–37). The mean effect sizes are not identical because each analysis adopted slightly different criteria for defining comparable effects. For example, Ellis (2006) limited his analysis to those studies which measured market orientation using either of the scales developed by Narver and Slater (1990) or Kohli et al. (1993). In contrast, Cano et al. (2004) adopted more relaxed criteria for inclusion and also analyzed studies which relied on proprietary measures of market orientation.

6 For researchers and teachers looking for online resources relating to meta-analysis, a good place to start is the meta-analysis page on psychwiki.com (www.psychwiki.com/wiki/Meta-analysis). An amusing "Bluffer's Guide to Meta-Analysis" by Field and Wright (2006) can be downloaded from here: www.psypag.co.uk/quarterly/Q60.pdf. This guide contains several numerical examples, distinguishes between fixed- and random-effects models, and shows how to calculate both confidence and credibility intervals. See also Field (2003a). Some excellent notes and PowerPoint slides can be found on David Wilson's website: http://mason.gmu.edu/~dwilsonb/ma.html. Something that is often heard by people new to meta-analysis is: "Where can I find software to help me pool effect sizes?" Hopefully it is clear from the examples presented in this chapter and in Appendix 2 that meta-analysis can be done with nothing more than a spreadsheet program. However, for those who think they need them, a number of specialized software programs are listed on the psychwiki.com page.

7 Earlier the results of four meta-analyses investigating market orientation effects were discussed. A false sense may have been conveyed that there exists a well-defined body of research that has measured market orientation and its effects in identical ways. In truth, a variety of measurement approaches has been followed and diverse effects have been assessed in a variety of settings. Rather than ignoring these differences, many of them were examined as potential moderators in the four meta-analyses.

8 Equations for calculating *d* and *r* from various test statistics can be found in McCartney and Rosenthal (2000) and Rosenthal and DiMatteo (2001).

9 The apples and oranges problem reflects the concern that it is illogical to compare dissimilar studies. By indiscriminately lumping studies together meta-analysts confound the detection of effects. Glass et al. (1981: 218–220) offered probably the first rebuttal to this argument. They observed that primary studies also mix apples with oranges by lumping different people together in samples. If legitimate results can be pooled from samples of dissimilar people, why can't they be obtained from samples of dissimilar studies? Glass et al. further argued that it is the very differences between studies examining a common effect that make meta-analysis pregnant with moderator-testing possibilities. If studies were identical in their procedures the only differences between them would be those attributable to sampling error. Nothing other than increased precision would be gained by pooling their results. But mixing apples with oranges is a good idea if one wants to learn something about fruit (Rosenthal and DiMatteo 2001: 68). Thus, meta-analysis offers unique opportunities for knowledge discovery. The best strategy for dealing with apples and oranges is selective coding. As long as there are enough apples and oranges to make comparisons worthwhile, the reviewer can assess the degree to which any variation in effect sizes is attributable to the effects of various contextual and measurement moderators.

10 The 211 columns of codes used by Smith and Glass are reproduced in Glass et al. (1981, Appendix A). In his recounting of the first modern meta-analysis, Hunt (1997, Chapter 2) highlights the challenges Glass and Smith faced in collecting studies, their need for "Solomonic" wisdom in coding results, and the subsequent impact of their work on psychology in general. Interestingly, Hunt reports that both Glass and Smith lost interest in meta-analysis after writing a couple of books on the subject in the early 1980s. As anyone who has labored through a large-N meta-analysis knows, this is not a surprising end to the story.

11 There is a considerable debate as to what constitutes acceptable levels of intercoder reliability. The issue is affected by a number of factors such as the complexity of the coding form and the reporting standards in the research being reviewed. These issues and other coding-related matters are thoroughly covered in Orwin (1994) and Stock (1994).

12 If the authors had reported reliability data for their measurements of the independent variable, we could have accounted for this as well by dividing each effect size estimate by the square root of the product of the two reliability coefficients: $r_{xy}(\text{observed})/\sqrt{(r_{xx},r_{yy})}$.

13 A third and lesser known method for determining statistical significance is to combine the p values of the individual studies (see Becker (1994) and Rosenthal (1991, Chapter 5)).

14 Calculating a Q statistic is just one way to assess the variability in the distribution of effect sizes. Another way is to examine the standard deviations of the individual effect sizes, plot them on a graph, and look for natural groupings (Rosenthal and DiMatteo 2001).

15 However, note that the kryptonite meta-analysis departed from Hunter and Schmidt orthodoxy in three ways: (1) z tests were used to calculate statistical significance, (2) a confidence interval was calculated about a corrected mean, and (3) a Q statistic was calculated to assess the homogeneity of the distribution. Hunter and Schmidt (2004) are highly dismissive of tests of statistical significance and so have little time for z scores; they prefer credibility intervals to confidence intervals; and they provide no equations for testing the homogeneity of sample effect sizes in the second edition of their text. Z scores, confidence intervals, and Q statistics were included here to illustrate some of the analytical options available to meta-analysts and advocated by textbook writers such as Glass et al. (1981), Hedges and Olkin (1985), Lipsey and Wilson (2001), and Rosenthal (1991).

16 The descriptive statistics for the two hypothetical Alzheimer's studies are as follows: study 1 ($n = 12$, SD $= 14$, $d = .929$); study 2 ($n = 40$, SD $= 14$, $d = .571$). The equations for calculating a weighted mean effect size and confidence interval for the two Alzheimer's studies were those developed by Hedges et al. (These are discussed in full in Appendix 2.) To calculate the variance (v_i) of d for each study, the following equation was used: $v_i = 4(1 + d_i^2/8)/n_i$, where n_i denotes the sample size of each study (Hedges and Vevea 1998: 490). The weights (w_i) used in the

meta-analysis are the inverse of the variance observed in each study (see equation 1 in Appendix 2). The study-specific weights were multiplied by their respective effect sizes ($w_i d_i$) prior to pooling. The relevant numbers for the two studies are as follows: study 1 ($v_i = .369$, $w_i = 2.708$, $w_i d_i = 2.515$); study 2 ($v_i = .104$, $w_i = 9.608$, $w_i d_i = 5.490$). Using these numbers we can calculate the weighted mean effect size using equation 2 in Appendix 2: $(2.515 + 5.490)/(2.708 + 9.608) = 8.005/12.316 = 0.650$. The variance of the mean effect size (v) is the inverse of the sum of the weights (see equation 3 in Appendix 3), or $1/(2.708 + 9.608) = 0.081$. The bounds of the 95% confidence interval are found using equation 4 in Appendix 2 and can be calculated as follows: $\text{CI}_{\text{lower}} = 0.65 - (1.96 \times \sqrt{.081}) = 0.091$; $\text{CI}_{\text{upper}} = 0.65 + (1.96 \times \sqrt{.081}) = 1.208$.

17 The meta-analysis also highlights the absurdity of using null hypothesis significance testing to draw conclusions about effects. Tversky and Kahneman (1971) observed that the degree of confidence researchers place in a result is often related to its level of statistical significance. This can give rise to the paradoxical situation where researchers place more confidence in pooled data than in the same data split over two or more studies. The source of the paradox lies in the mistaken view that *p* values are an indicator of a result's credibility or replicability. But as we saw in Chapter 3, *p* values are confounded indexes that reflect both effect size and sample size. Consequently, a statistically significant result cannot be interpreted as constituting evidence of a genuine effect. The best test of whether a result obtained from a sample is real is whether it replicates.

18 When evidence-based medical practitioners portray the quality of evidence hierarchically, it is usual for meta-analyses and systematic reviews to be at the top of the list (e.g., Hoppe and Bhandari 2008, Figure 1; Urschel 2005, Table 1). But missing from these lists are large-scale randomized control trials. As will be shown in the next chapter, meta-analyses have a tendency to generate inflated mean effect size estimates, an unfortunate outcome that large-scale randomized control trials avoid.

19 Sauerland and Seiler (2005) note that a surgeon who desires to stay abreast of new knowledge would have to scan some 200 surgical journals, each publishing about 250 articles per year. This works out to 137 articles per day.

6 Minimizing bias in meta-analysis

The appearance of misleading meta-analysis is not surprising considering the existence of publication bias and the many other biases that may be introduced in the process of locating, selecting, and combining studies. ~ Matthias Egger et al. (1997: 629)

Four ways to ruin a perfectly good meta-analysis

In science, the large-scale randomized controlled trial is considered the gold standard for estimating effects. But as such trials are expensive and time consuming, new research typically begins with small-scale studies which may be subsequently aggregated using meta-analysis. Relatively late in the game a randomized controlled trial may be run to provide the most definitive evidence on the subject, but in many cases the meta-analysis, for better or worse, will provide the last word on a subject. In those relatively rare instances where a large-scale randomized trial follows a meta-analysis, an opportunity emerges to compare the results obtained by the two methods. Most of the time the results are found to be congruent (Cappelleri et al. 1996; Villar and Carroli 1995). But there have been notable exceptions. In the medical field LeLorier et al. (1997) matched twelve large randomized controlled trials with nineteen meta-analyses and found several instances where a statistically significant result obtained by one method was paired with a nonsignificant result obtained by the other. Given the way in which decisions about new treatments are made in medicine, these authors concluded that if there had been no randomized controlled trials, the meta-analyses would have led to the adoption of an ineffective treatment in 32% of cases and to the rejection of a useful treatment in 33% of cases. As these numbers show, meta-analyses sometimes generate misleading conclusions.

Although meta-analysis has an aura of objectivity about it, in practice it is riddled with judgment calls. How do we decide which studies to include in our analysis? How far do we go to deal with the file drawer problem? Do we exclude results not reported in the English language? Do we need to weight effect size estimates by the quality of the study? How do we gauge the quality of studies? How do we deal with the apples and oranges issue? Unfortunately, many meta-analyses are done mechanically, with

little attention given to these issues. This leads to reviews that are undermined by both Type I and Type II errors.

Anyone with basic numeracy skills can do the statistical pooling of effect sizes that lies at the heart of meta-analysis. But the real challenge is in identifying and dealing with multiple sources of bias. Bias can be introduced at any stage of the review and inattention to these matters can result in conclusions that are spectacularly wrong. A reviewer can introduce bias into a meta-analysis in four ways: (1) by excluding relevant research, (2) by including bad results, (3) by fitting inappropriate statistical models to the data, and (4) by running analyses with insufficient statistical power. The first three sources of bias will lead to inflated effect size estimates and an increased risk of Type I errors. The fourth will lead to imprecise estimates and an increased risk of Type II errors.

In this chapter these four broad sources of bias are discussed along with measures that can be taken to minimize their impact. It is the nature of meta-analysis that some bias will almost invariably end up affecting the final result. A good meta-analysis is therefore one where the likely sources of bias have been identified, their consequences measured, and mitigating strategies have been adopted.

1. Exclude relevant research

A meta-analysis will ideally include all the relevant research on an effect. The exclusion of some relevant research can lead to an availability bias. An availability bias arises when effect size estimates obtained from studies which are readily available to the reviewer differ from those estimates reported in studies which are less accessible. An availability bias is seldom intentional and usually arises as a result of a reporting bias, the file drawer problem, a publication bias, and the Tower of Babel bias. These issues are examined below.

Reporting bias and the file drawer problem

A reporting bias and the file drawer problem are opposite sides of the same coin. Consider a researcher who conducts a study and collects data examining four separate effects. Two of the results turn out to be statistically significant while the other two do not achieve statistical significance. A reporting bias arises when the researcher reports only the statistically significant results (Hedges 1988). The researcher's decision to file away the nonsignificant results, while understandable, creates a file drawer problem (Rosenthal 1979). The problem is that evidence which is relevant to the meta-analytic estimation of effect sizes has been kept out of the public domain. Reviews that exclude these unreported and filed away results are likely to be biased.

In their survey of members of the American Psychological Association, Coursol and Wagner (1986) found that the decision to submit a paper for publication was significantly related to the outcome achieved in study. The raw counts for their study are reproduced in Table 6.1. As can be seen from Part A of the table, 82% of the studies

Table 6.1 *Selection bias in psychology research*

Outcome	Submission decision		
A. ($\Phi = .40$)	Yes	No	Total
Positive	106	23	129
Negative or neutral	28	37	65
Total	134	60	194
	Publication decision		
B. ($\Phi = .28$)	Accepted	Not accepted	Total
Positive	85	21	106
Negative or neutral	14	14	28
Total	99	35	134
	Final outcome		
C. ($\Phi = .42$)	Published	Not published	Total
Positive	85	44	129
Negative or neutral	14	51	65
Total	99	95	194

Source: Raw data from Coursol and Wagner (1986), analysis by the author.

that found therapy had a positive effect on client health were submitted for publication in comparison with just 43% of the studies showing neutral or negative outcomes. This selective reporting behavior is substantial, equivalent to a phi coefficient of .40 (or halfway between a medium- and large-sized effect according to Cohen's (1988) benchmarks).

Meta-analysts are interested in all estimates of an effect, irrespective of their statistical significance. The exclusion of non-reported research is biasing because such research typically provides estimates that are small in size. Recall that statistical power is partly determined by effect size. When effects are small, statistical significance is harder to achieve. Consequently, studies which observe small effects are less likely to achieve statistical significance and are therefore less likely to be written up and reported. If the reviewer is unable to identify these non-reported results, the mean effect size calculated from publicly available data will be higher than it should be.

At best the file drawer problem will lead to some inflation in mean estimates. At worst, it will lead to Type I errors. This could happen when the null hypothesis of no effect happens to be true and the majority of studies which have reached this conclusion have gone unreported or have been filed away rather than published. Statistically there will be a small minority of studies that confuse sampling variability with natural variation in the population (that is, their authors report an effect where none exists),

and these are much more likely to be submitted for publication. If the reviewer is only aware of this second, aberrant group of studies, any meta-analysis is likely to generate a false positive.

Publication bias

A publication bias arises when editors and reviewers exhibit a preference for publishing statistically significant results in contrast with methodologically sound studies reporting nonsignificant results. To test whether such a bias exists, Atkinson et al. (1982) submitted bogus manuscripts to 101 consulting editors of APA journals. The submitted manuscripts were identical in every respect except that some results were statistically significant and others were nonsignificant. Editors received only one version of the manuscript and were asked to rate the manuscripts in terms of their suitability for publication. Atkinson et al. found that manuscripts reporting statistically nonsignificant findings were three times more likely to be recommended for rejection than manuscripts reporting statistically significant results. A similar conclusion was reached by Coursol and Wagner (1986) in their survey of APA members. These authors found that 80% of submitted manuscripts reporting positive outcome studies were accepted for publication in contrast with a 50% acceptance rate for neutral or negative outcome studies (see Part B of Table 6.1).

The existence of a publication bias is a logical consequence of null hypothesis significance testing. Under this model the ability to draw conclusions is essentially determined by the results of statistical tests. As we saw in Chapter 3, the shortcoming of this approach is that *p* values say as much about the size of a sample as they do about the size of an effect. This means that important results are sometimes missed because samples were too small. A nonsignificant result is an inconclusive result. A nonsignificant *p* tells us that there is either no effect or there is an effect but we missed it because of insufficient power. Given this uncertainty it is not unreasonable for editors and reviewers to exhibit a preference for statistically significant conclusions.[1] Neither should we be surprised that researchers are reluctant to write up and report the results of those tests that do not bear fruit. Not only will they find it difficult to draw a conclusion (leading to the awful temptation to do a post hoc power analysis), but the odds of getting their result published are stacked against them. Combine these two perfectly rational tendencies – selective reporting and selective publication – and you end up with a substantial availability bias. In Coursol and Wagner's (1986) assessment of research investigating the benefits of therapy, a study which found a positive effect ultimately had a 66% chance of getting published and making it into the public domain, while a study which returned a neutral or negative effect had only a 22% chance (see Part C of Table 6.1). The likelihood of publication was thus three times greater for positive results.

Given the direct relationship between effect size and statistical power, results which make it all the way to publication are likely to be bigger on average than unpublished results.[2] Fortunately the extent of this bias can be assessed as long as the reviewer has managed to find at least some unpublished studies that can be used as a basis for

comparison. For example, in their review Lipsey and Wilson (1993) found that published studies reported effect sizes that were on average 0.14 standard deviations larger than unpublished studies. Knowing the difference between published and unpublished effect sizes, reviewers can make informed judgments about the threat of publication bias and adjust their conclusions accordingly.

Tower of Babel bias

A Tower of Babel bias can arise when results published in languages other than English are excluded from the analysis (Grégoire et al. 1995). This exclusion can be biasing because there is evidence that non-English speaking authors are reluctant to submit negative or nonsignificant results to English-language journals. The thinking is that if the results are strong, they will be submitted to good international (i.e., English-language) journals, but if the results are unimpressive they will be submitted to local (i.e., non-English-language) journals. Evidence for the Tower of Babel bias was provided by Grégoire et al. (1995). These authors reviewed sixteen meta-analyses that had explicitly excluded non-English results. They then searched for non-English results that were relevant to the reviews. They found one paper (written in German and published in a Swiss journal) that, had it been included in the relevant meta-analysis, would have turned a nonsignificant result into a statistically significant conclusion. Grégoire et al. (1995) interpreted this as evidence that linguistic exclusion criteria can lead to biased analyses.

Quantifying the threat of the availability bias

It should be noted that problems with accessing relevant research on a topic affect reviewers of all stripes. But meta-analysts can be distinguished from narrative reviewers by their explicit desire to collect all the relevant research and by the corresponding need to quantify and mitigate the threat of the availability bias. There are several ways to assess this threat: compare mean estimates obtained from published and unpublished results, as discussed above; examine a funnel plot showing the distribution of effect sizes; and calculate a fail-safe *N*.

A funnel plot is a scatter plot of the effect size estimates combined in the meta-analysis. Each estimate is placed on a graph where the X axis corresponds to the effect size and the Y axis corresponds to the sample size. The logic of the funnel plot is that the precision of estimates will increase with the size of the studies. Relatively imprecise estimates obtained from small samples will be widely scattered along the bottom of the graph while estimates obtained from larger studies will be bunched together at the top of the graph. Under normal circumstances, the dispersion of results will describe a symmetrical funnel shape. However, in the presence of an availability bias, the plot will be skewed and asymmetrical.[3]

An example of how to detect an availability bias using a funnel plot is provided in Figure 6.1. This chart shows the results of seven small-scale studies (the black diamonds) and one meta-analysis (the white diamond) examining the link between the

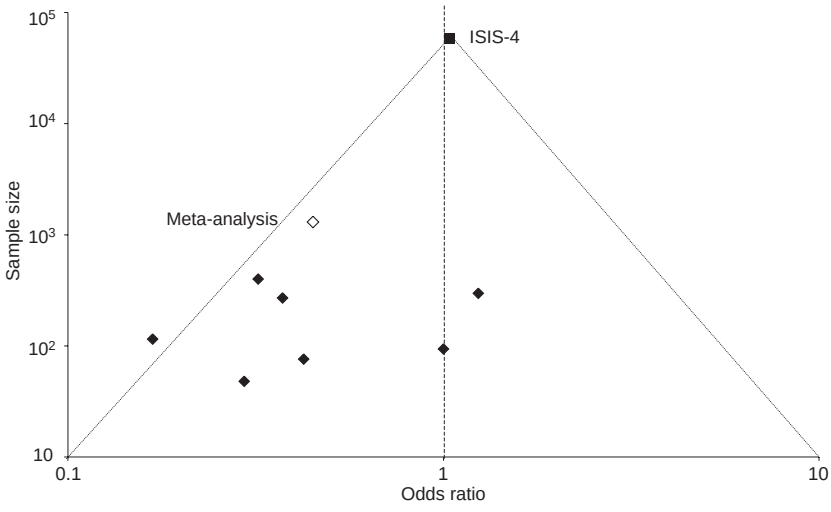


Figure 6.1 Funnel plot for research investigating magnesium effects

injection of magnesium and survival rates for heart attack victims. The effect sizes in the figure reflect the relative odds of dying in the treated versus untreated groups.[4] Ratios less than one indicate that the injection of magnesium improved the odds of surviving a heart attack. As can be seen in the figure, five of the seven small-study results seemed to indicate the beneficial effects of magnesium. Although only two of these results achieved statistical significance, Teo et al. (1991) were able to calculate a statistically significant mean effect size by pooling the results of all seven studies. The mean effect indicated that intravenous magnesium reduced the odds of death by about half, leading Teo et al. to conclude that their study had provided "strong evidence" of a "substantial benefit."

Unfortunately for these authors, they were wrong. A few years after the publication of their study, the large-scale ISIS-4 trial overturned their meta-analytic conclusion by showing that magnesium has no effect on survival rates (Yusuf and Flather 1995). (The ISIS-4 result is indicated by the black square at the top of Figure 6.1.) What went wrong with the meta-analysis? The best explanation seems to be that Teo et al.'s estimate of the mean was inflated by an availability bias. This is the conclusion we get from examining the plot of the results in Figure 6.1. The dispersion of the results ought to describe a funnel shape but it does not. There is a distinct gap in the bottom right side of the funnel indicating the absence of small studies reporting negative results (Egger and Smith 1995). Somehow, data that would have nullified the positive results on the other side of the funnel were excluded from the review. Where were these negative studies? Were they filed away? Were they victims of a publication bias? In this case, publication bias seems less a culprit than reporting bias as Teo et al. (1991) explicitly tried to include unpublished studies in their review. They even asked other investigators working in the area to help them identify unpublished trials. Yet despite their efforts the only studies they found were those which, on average, erroneously pointed towards

an effect. The best conclusion that can be drawn is that negative results from other studies were never written up.

Another way to quantify the bias arising from the incomplete representation of relevant research is to calculate the "fail-safe *N*." The fail-safe *N* is the minimum number of additional studies with conflicting evidence that would be needed to overturn the conclusion reached in the review. Conflicting evidence is usually defined as a null result. If the meta-analysis has generated a statistically significant finding, the fail-safe *N* is the number of excluded studies averaging null results that would be needed to render that finding nonsignificant (Rosenthal 1979). The fail-safe *N* is directly related to the size of the effect and the number of studies (*k*) combined to estimate it in the meta-analysis. For example, if the results of fourteen studies were combined to yield a statistically significant mean effect size of $r = .15$, $p = .018$, it would require the addition of only nine further studies averaging a null effect to render this result statistically nonsignificant. If we could accept the possibility that there are at least nine "no effect" results buried in filing cabinets or published in obscure non-English journals, then we should be skeptical of the meta-analytic conclusion. However, if the fourteen studies returned a mean effect size of $r = .30$, then the fail-safe number would be a much higher seventy-eight studies. Thus, the fail-safe *N* describes the tolerance level of a result. The larger the *N*, the more tolerant the result will be of excluded null results.[5]

The aim is to make the fail-safe *N* as high as possible and ideally higher than Rosenthal's (1979) suggested threshold level of $5k + 10$. The higher the fail-safe *N*, the more confidence we can have in the result. The fail-safe *N* rises as the number of the studies being combined increases. In our earlier example a meta-analysis of fourteen studies returned a mean correlation of .15 and had a fail-safe *N* of just nine studies, well below the recommended minimum of 80 ($14 \times 5 + 10$). If this mean correlation had been obtained by combining sixty studies, the fail-safe *N* would be 1,736 studies, well above the recommended minimum of 310. In both cases the mean effect size is statistically significant, but we would have far more confidence in the second result because the number of excluded studies required to render it nonsignificant is much higher.[6]

Four sources of availability bias have been discussed and different methods for gauging their consequences have been described. One lesson stands out: when collecting studies investigating an effect, make every effort to include the results of relevant unpublished research as well as research published in languages other than English. The threat of the availability bias is inversely proportional to the ratio of published, effect-reporting studies to unpublished, null result studies. A reviewer who collects only published studies will be unable to gauge how resistant their result is to the availability bias. But a reviewer who is able to get even just a few unpublished results will be able to assess the risk and severity of the problem while at the same time improving the tolerance of their result to further null findings.

2. Include bad results

It has been argued that a meta-analysis should include all the relevant research on an effect, but this is a controversial claim. Intelligent critics have long argued that

meta-analyses are compromised by the injudicious mixing of good and bad *studies*. But what makes one study good and another bad and where does one draw the line? As we will see, making these sorts of judgment calls can do more harm than good. A separate issue concerns the mixing of good and bad *results*. Bad results mislead reviewers and confound the estimation of mean effect sizes. A recent idea that has emerged in the research synthesis literature is the notion that potentially misleading results can be red flagged and removed, leading to better mean estimates. Although no clear standards for classifying results have yet been developed, a good starting point may be to assess the statistical power of studies being combined. Reviewers can then make informed judgments about the merits of excluding those results that are tainted by a reasonable suspicion of Type I error.

Mixing good and bad studies

From the very beginning a criticism made against meta-analysis is that it is based on the indiscriminate mixing of good and bad studies. This garbage-in, garbage-out complaint originated with Eysenck (1978: 517), who was dismayed with the apparently low standards of inclusion used in meta-analysis. "A mass of reports – good, bad, indifferent – are fed into the computer in the hope that people will cease caring about the quality of the material on which the conclusions are based." According to Eysenck, there was little to be gained by trying to distill knowledge from poorly designed studies. A similar view was espoused by Shapiro (1994: 771) in his article entitled "Meta-analysis/shmeta-analysis." Shapiro argued that the quality of a meta-analysis was contingent upon the quality of the individual studies being combined. As the highest standard of research is the experimental design, he proposed that meta-analyses based on the accumulation of nonexperimental data should be abandoned. Feinstein (1995) was also concerned with the mixing of good and bad studies which he considered statistical alchemy. He argued that insufficient attention to issues of quality control had given rise to the situation where reviewers could dredge up data to support almost any hypothesis. The solution, according to Feinstein, was to exclude biased studies and combine only "excellent individual studies" or "studies that seem unequivocally good" (1995: 77).

That poor studies can sabotage a review leads logically to the conclusion that poor studies should not be combined with good studies, or at least should not be given equal weight. But there are at least four reasons why we should hesitate to discriminate studies on the basis of quality. First, making judgments about the quality of past research introduces reviewer bias into the analysis. Quality means different things to different people. Even critics of meta-analysis are unable to agree on definitions of quality. For Shapiro (1994) quality research is synonymous with experimental research, but Feinstein (1995) would include non-experimental studies as long as they were "unequivocally good."[7]

Second, even if we could agree on quality standards, a restriction applied to certain types of studies (e.g., nonexperimental research or research that does not rely on randomized assignment) would amount to scientific censorship. This is because studies

have value only to the degree to which they contribute evidence that can be used to establish or refute an effect. As small-scale studies seldom provide definitive evidence, the full value of any study can be realized only when it is combined with others investigating the same effect. Thus, discussions about quality and selectivity inevitably lead to thornier debates about scientific value. For these reasons Greenland (1994) interpreted Shapiro's (1994) proposal to ban observational studies from meta-analysis as effectively constituting a ban on observational research.

Third, as there are virtually no fault-free studies, excluding nonexcellent research from meta-analysis would lead to the dismissal of masses of evidence on a subject. Excluded research is wasted research. But even low-quality studies may provide information that can be meaningfully combined. After all, if studies are estimating a common effect, then the evidence obtained from different studies should converge. This was Glass and Smith's (1978: 518) experience. In their pioneering review these authors noted that both good and bad studies produced "almost exactly the same results."

Fourth, some differences in study quality, such as sampling error and reliability, can be recorded and controlled for. As we saw in Chapter 5, meta-analysts can correct for measurement error and give greater weight to estimates obtained from larger samples. Thus, the question of whether and how much the results are biased by study quality is partly an empirical one that meta-analysis can readily answer.

Advocates of meta-analysis disagree with the premise that "bad" studies undermine the quality of statistical inferences drawn from a meta-analysis. They would argue that there is little to be gained by restricting reviews to only a subset of all the relevant research. The more evidence that can be analyzed the better because "many weak studies can add up to a strong conclusion" (Glass et al. 1981: 221). In his twenty-fifth year assessment of meta-analysis, Glass (2000: 10) reiterated his belief in "the idea that meta-analyses must deal with studies, good, bad, and indifferent." Ironically, a good meta-analysis is one that includes both good and bad research while a bad meta-analysis will include only good research.

Mixing good and bad results

While a case can be made for including research of all levels of quality, a separate issue concerns the mixing of good and bad results. A bad result is one which is likely to be false. As we saw in Chapter 4, false negatives are a result of low statistical power while false positives are a consequence of the multiplicity problem (or the testing of many hypotheses without adjusting the familywise error rate) and the temptation to HARK (or hypothesize after the results are known). Although it is inevitable that some proportion of the results being combined will be false, this proportion is higher than it should be on account of lax statistical practices and biased publication policies.

Statistical power is directly related to the probability that a study will detect a genuine effect. When effects are present, the chance of making a Type II error rises as power falls. Given that underpowered studies are the norm in social science research, a high proportion of false negatives is to be expected. Combine low power with

Table 6.2 *Does magnesium prevent death by heart attack?*

Study	Raw data (No. dead/no. patients)		Sample size (n)	p	Effect size (r)	Statistical power
	Magnesium	Control				
Ceremuzynski	1/25	3/23	48	0.26	0.16	0.09
Morton et al.	1/40	2/36	76	0.49	0.08	0.11
Abraham et al.	1/48	1/46	94	0.98	0.00	0.13
Schecter et al.	1/59	9/56	115	0.01	0.26	0.15
Rasmussen et al.	9/135	23/135	270	0.01	0.16	0.29
Feldstedt et al.	10/150	8/148	298	0.65	−0.03	0.32
Smith et al.	2/200	7/200	400	0.09	0.08	0.41
Crude total/mean	25/657	53/654	1,311	0.001	0.09	0.21

Note: Raw data came from Teo et al. (1991). Muncer et al. (2002) converted the raw results into the effect sizes shown here. The statistical power to detect the weighted mean effect size ($r = .086$) with $\alpha_2 = 0.05$ was calculated by the author.

editorial policies favoring statistically significant results and the surprising outcome is an increase in Type I error rates boosting the proportion of false positives among published studies.[8] This happens because underpowered studies have to detect much larger effects to achieve statistical significance. As large effects are rare in social science, there is a fair chance that many of the effects reported in low-powered studies are flukes attributable to sampling variation.

If researchers had access to all the relevant research on a topic, individually misleading conclusions would have no effect on the estimate of the mean and it would be wasteful to exclude any relevant studies from the analysis. Reviewers would simply weight and pool the individual effect sizes without regard for the statistical power of the underlying studies. But because access to past results is affected by the availability bias, power matters. Available research, being a subset of all relevant research, will consist of good results, mostly obtained from adequately powered studies, and bad results, mostly arising from underpowered studies which have chanced upon unusual samples characterized by extreme values. It is the over-representation of these bad results that can scuttle a meta-analysis like the magnesium study mentioned earlier.

In that study Teo et al. (1991) combined the results of seven clinical trials involving a total of 1,301 patients and found "strong evidence" that the injection of magnesium saved lives. Four years later, data from the ISIS-4 trial involving 58,050 patients revealed that magnesium has no effect on mortality rates (Yusuf and Flather 1995). How did Teo et al. get it wrong? The analysis of a funnel plot revealed that their conclusion was biased by the over-representation of positive results. The study-specific results combined by Teo et al. (1991) are reproduced in Table 6.2. The results clearly show that patients who received magnesium had a better chance of survival than patients in the control group. The total number of deaths in the combined control group ($N = 53$)

was twice the number of deaths in the treatment group ($N = 25$). With these data it is hard to avoid the flawed conclusion that magnesium saves lives. But the data also contain a warning that seems to have been missed. Not one of the studies in the sample had even close to enough power to detect the effect that Teo et al. believed was there. The fact that an effect was detected tells us that either the effect was large and easily found or the reported results came from unusual samples. A glance at the effect sizes listed in the table should dismiss the first possibility. The majority of results were trivial in size according to Cohen's benchmarks. No result was larger than small. That Teo et al. could combine small and trivial effects and come up with "strong evidence" that magnesium lowers the odds of death by half says a lot about the dangers of including results from underpowered studies.

Of course, this is easy to say in hindsight. The real trick is to tell in advance when a conclusion is likely to be biased by bad data. To that end, Teo et al. could have calculated the statistical power for each of the seven studies, thus determining the probability each had of correctly identifying a genuine effect. Power figures are provided in the right-hand column of Table 6.2. These figures show the power of each study to detect an effect equivalent in size to the weighted mean ($r = .086$) obtained from all seven studies. As can be seen, none of the studies achieved satisfactory levels of power. None even had the proverbial coin-flip's chance of detecting an effect of this size. The average power of the seven studies was 0.21. Assume for a moment that magnesium does reduce the mortality rate of heart attack sufferers and that the magnitude of this effect is equivalent to the weighted mean correlation of .086. To have a reasonable probability of detecting this effect, a comparison study would need to have a minimum of 528 patients in each group. None of the seven studies included in this review came close to achieving this.[9] In contrast, the large-scale ISIS-4 study which discredited the magnesium treatment had statistical power of .999 to detect an effect one-quarter of this size.

The moral of the magnesium tale is that results from over-represented and underpowered studies can bias a review. The implication is that excluding such results will lead to better inferences and stronger conclusions (e.g., Hedges and Pigott 2001; Kraemer et al. 1998; Muncer et al. 2002, 2003).

In a sense, power is related to the confidence one can have in the result of a study. Greater confidence can be placed in a result obtained from a high-powered study than a result obtained from a low-powered study. This is because high-powered studies are more likely to reach conclusions while any conclusion drawn in a low-powered study will be tainted with the suspicion of Type I error.[10] But what is less obvious is whether confidence in results accumulates. Muncer et al. (2003) make the interesting point that two underpowered studies should not be viewed with the same confidence as one adequately powered study. But what about three underpowered studies? Or ten? There is no clear answer because one cannot easily tell when the number of studies is sufficient to provide a clear picture of the true mean and ameliorate quirks associated with individual results. Again, the availability bias rears its ugly head, leading to the usual recommendations about collecting unpublished, filed-away studies.

In the previous chapter we saw how it is important to weight estimates in terms of their precision. Estimates obtained from small samples are more likely to be biased by sampling error and so are given less weight than estimates obtained from large samples. But there is also a case to be made for excluding estimates obtained from underpowered studies on the grounds that the results from such studies may be anomalous and convey more information about sampling variability than natural variability in populations. To identify underpowered studies, Muncer et al. (2003) propose an iterative analysis where a weighted mean effect size, calculated from the initial sample of studies, is used to determine the average statistical power of those studies. Although this looks a lot like a post hoc power analysis, it differs in one important respect. In Chapter 3 we saw that the retrospective analysis of statistical power for individual studies is an exercise in futility because there is no guarantee that study-specific estimates of an effect size are reliable. But combining the estimates of many studies provides a surer basis for estimating the population effect size and therefore retrospective assessments of statistical power. By running this type of power analysis the reviewer is asking, what was the power of each study to detect an effect size equivalent to the weighted mean? If the average power of studies is low, Muncer et al. recommend recalculating the weighted mean using estimates obtained from sufficiently powered studies, that is, studies with power levels in excess of .80. But this could amount to excluding most, if not all, of the available evidence. A more realistic recommendation would be to define as adequate power levels that are greater than .50 (Kraemer et al. 1998).[11]

The notion that some results should be thrown out is inconsistent with meta-analysts' belief that data from all studies are valuable. But the idea has merit when reviewers have only selective access to relevant research. Low statistical power combined with limited access leads to misleading meta-analyses, as Teo et al. discovered. Interestingly, if these authors had excluded low-powered studies from their review, they would have discarded every study in their database and abandoned their fatally flawed meta-analysis.

3. Use inappropriate statistical models

In the kryptonite meta-analysis done in Chapter 5, we calculated a *Q* statistic to quantify the variation in the sampling distribution and concluded that there was more than one population mean. This is quite a radical thought. In most places in this book we have assumed that study-specific estimates all point towards a common population effect size. But real-world effects come in different sizes. They may be bigger for one group than another. Most of the time there will not be one effect size but many. Consequently, it makes sense to talk about a sample of study-specific estimates and a higher-level sample of population effect sizes. Each sample will have its own distribution and this has ramifications for the way in which we calculate standard errors and confidence intervals.

Think of a set of studies, each providing an independent estimate of a population effect size. Following Hedges and Vevea (1998) we can distinguish between the population effect size, represented by the Greek letter theta, θ, and the study-specific estimate

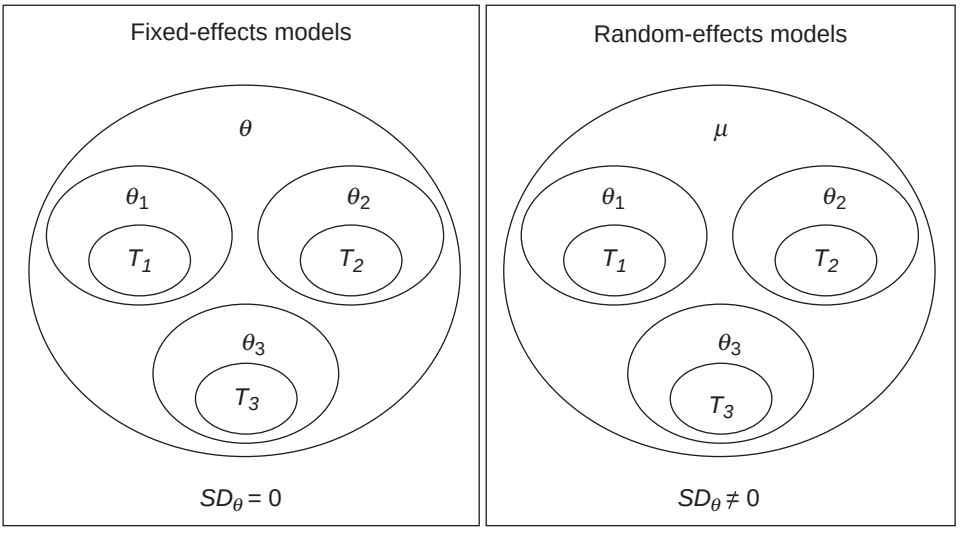


Figure 6.2 Fixed- and random-effects models compared
Notes: T_1 = estimate of effect size from study 1, θ_1 = population effect size in study 1, θ = mean of the distribution of effect sizes estimates, μ = mean of the distribution of population effect sizes. SD_θ refers to the standard deviation of the population effect sizes.

of that effect size, represented by *T*. The population effect size for study *i* is denoted θ_i and the corresponding estimate is denoted T_i. The question we are seeking to answer is whether the study-specific estimates (all the *T*s) are pointing toward a common or fixed-effect size (a single θ) or a sample of randomly distributed effect sizes (a set of dissimilar θs). If effect sizes are fixed on a single mean, then the calculation of standard errors should follow what is known as a fixed-effects procedure. However, if effect sizes are randomly distributed, then a random-effects procedure is required.

The main assumption underlying the fixed-effects model is that the value for the population effect size is the same in every study. In the fixed-effects model, $\theta_1 = \theta_2 = \ldots = \theta_k = \theta$. No such assumption is made in the random-effects model. Rather, effect sizes are presumed to be randomly distributed around some super-mean which is designated with the Greek letter mu (μ). The difference between the two models is illustrated in Figure 6.2. In the figure three independent studies have provided effect size estimates (T_1, T_2, and T_3), each of which corresponds to a population effect size (θ_1, θ_2, and θ_3). Under the fixed-effects approach shown in the left-hand side of the figure, population effect sizes are assumed to be identical. Thus the mean of θ_1, θ_2, and θ_3 is θ and the standard deviation for the sample of population effect sizes is zero. But in the random-effects approach shown on the right-hand side of the figure, the mean of θ_1, θ_2, and θ_3 can take on any value (hence μ) and the standard deviation for the sample of effect sizes is likely to be something other than zero.

In a fixed-effects analysis we use the study-specific effect size estimates to calculate the mean population effect size (θ). But in the random-effects model we need to take an additional step to calculate the mean (μ) of the effect size distribution.[12] In the fixed-effects approach we have only one distribution to think about, namely the distribution

of estimates. But in the random-effects approach we have two: the distribution of estimates and the distribution of population effect sizes. As each distribution comes with a unique dispersion, the distinguishing characteristic of the random-effects procedure is the need to account for two sources of variability – the variation in the spread of estimates (called within-study variance) plus the variation in the spread of effect sizes (called between-study variance). In the random-effects procedure these two types of variance are added together and this makes the standard errors bigger than in the case of fixed-effects methods. Bigger standard errors mean wider confidence intervals and more conservative tests of statistical significance. Consider the mean effect sizes and confidence intervals that would be obtained for the kryptonite data using the fixed- and random-effects procedures of Hedges and Vevea (1998):

Fixed-effects: $\bar{r} = -.48$ (CI $-.57$ to $-.36$)
Random-effects: $\bar{r} = -.39$ (CI $-.64$ to $-.07$)

(The calculations for these results are provided in Appendix 2.) The interval generated by the random-effects procedure is more than double the width of the fixed-effects interval. It is less precise because it is wider, but whenever population effects vary it will lead to more accurate inferences.

Fixed or random effects?

How can we tell whether population effect sizes are fixed or are randomly distributed for a set of studies? One way is to test the homogeneity of the variance in the distribution of population effect sizes (this is step 5 in the meta-analysis described in Chapter 5). If the *Q* statistic reveals that the sample of effect sizes is homogenous, then population effect sizes are likely to be homogenous and fixed-effects analyses will be sufficient. But if the sample of population effect sizes is found to be heterogeneous, random-effects methods which account for the additional variability in population effect sizes will be superior. However, one limitation of this approach is that chi-square tests normally associated with tests of homogeneity often lack the statistical power to detect between-study variation in population parameters (Hedges and Pigott 2001).

Hedges and Vevea (1998) argue that the choice between fixed- or random-effects procedures should be contingent upon the type of inference the reviewer wishes to draw. Meta-analyses based on fixed-effects models generate conditional inferences that are limited to the set of studies included in the analysis. In contrast, inferences made from random-effects models are unconditional and may be applied to a population of studies larger than the sample. Given that most reviewers will be interested in making unconditional inferences that apply to studies that were not included in the meta-analysis or that have not yet been done, then the random-effects model is unquestionably the better choice.

The consequences of using the wrong model

There is evidence to indicate that population effects vary in nature (Field 2005). Thus, random-effects procedures will be appropriate in most cases. Yet the vast majority of published meta-analyses rely on fixed-effects procedures, presumably because they

are easier to do (Hunter and Schmidt 2000). This misapplication of models to data has serious consequences for inference-making. If fixed-effects models are applied to heterogeneous data, the total variance in the data will be understated, confidence intervals will be narrower than they should be, and significance tests will be susceptible to Type I errors (Hunter and Schmidt 2000). In some cases the increase in the risk of Type I errors will substantial. In their re-examination of sixty-eight meta-analyses published in the *Psychological Bulletin*, Schmidt et al. (2009) found that fixed-effects analyses were on average 52% narrower than their actual width. Based on a Monte Carlo simulation, Field (2003b) estimated that anywhere between 43% and 80% of meta-analyses that misapply fixed-effects models to heterogeneous data will generate a statistically significant mean effect size even when no effect exists in the population. Given that most published reviews used fixed-effects procedures to estimate population effects that are normally variable, the conclusion is that a fair proportion of positive results is false.

The remedy for this problem is obvious: avoid fixed-effects procedures. Hunter and Schmidt (2004) reason that the random-effects model will always be preferable because it is the more general one. Fixed-effects procedures are but a special case of random-effects models in which the standard deviation of the population mean happens to equal zero. As this will be true only some of the time, it makes sense to master random-effects procedures which will be valid all of the time. The calculations used for both procedures are described in Appendix 2.

4. Run analyses with insufficient statistical power

Insufficient statistical power is an odd problem to associate with meta-analysis. Most of the time meta-analyses have megawatts of power, and certainly far more power than the studies on which they are based. Even so, there is no guarantee that a meta-analysis will have enough power to detect effects and the lack of it can lead to Type II errors, just as it does for individual studies. Consider the dust-mite study reported by Gøtzsche et al. (1998). This meta-analysis pooled the results of five studies examining the effect of asthma treatments in houses with dust mites. Altogether the number of people who improved as a result of treatment was found to be 41 out of 113 patients in comparison with 38 out of 117 in the control group. As the numbers were similar in each group, Gøtzsche et al. (1998) concluded that the treatment was ineffective. But in a re-analysis of these data Muncer (1999) raised the possibility that a Type II error had been made. If a small effect had been assumed ($\Phi = 0.10$), then data from an additional 552 subjects would have been needed to have an 80% chance of detecting this effect using a two-tailed test. As it happened, the mean effect size estimated in the meta-analysis was smaller than small ($\Phi = 0.04$). If this is an accurate estimate of the population effect size then the meta-analysis had only a one in eleven chance of returning a statistically significant result. In short, the meta-analysis was grossly underpowered and the possibility that the result is a false negative cannot be ruled out.

The dust-mite study illustrates the need to analyze statistical power prior to commencing a meta-analysis. Doing so helps the reviewer assess the likelihood of detecting a statistically significant effect given the number of studies being combined and the average sample size within studies (Hedges and Pigott 2001). After running a prospective power analysis the reviewer may decide that the chances of detecting an effect are too low and abandon the meta-analysis. As with power analyses done for individual studies, the challenge for the reviewer will be to calculate power without knowing the anticipated effect size. The options are to substitute the smallest effect size considered to be of substantive importance (Hedges and Pigott 2001) or to use an estimate derived from the studies themselves (Muncer et al. 2003). If the reviewer decides that there is sufficient power to proceed, the next challenge will be to determine whether the addition of new studies increases or decreases the overall power of the meta-analysis. For individual studies, the addition of more sampling units always raises statistical power. But this is not necessarily the case with meta-analysis.

Meta-analyses draw their statistical power from the studies being combined and this is why confidence intervals for pooled results are narrower than intervals for individual results. But the determination of power for a meta-analysis depends on the methods used to weight study-specific effect size estimates. Estimates can be weighted by either the sample size or the variation in the distribution of the study data. The different weighting methods affect the calculation of standard errors and confidence intervals. If estimates are weighted by sample size, as in the Hunter and Schmidt approach, then more weight is given to studies with bigger samples and smaller sampling errors. Conversely, if estimates are weighted by the inverse of the variance, as in the Hedges et al. approach, then studies with small variances will contribute more to the mean effect size than estimates based on large variances (Cohn and Becker 2003). Under this method the variance of the mean effect size is calculated as the inverse of the sum of all the weights: $v_\cdot = 1/\sum w_i$. This means that as more studies get added, the sum of the weights goes up and the variance of the mean goes down. In the fixed-effects procedure the addition of studies will always lead to a decrease in the variance ($v_\cdot$) and therefore the standard error ($\sqrt{v_\cdot}$) associated with the mean effect size. The result will be tighter confidence intervals.[13] However, this will not always be true when the random-effects procedure is used because such procedures incorporate additional sources of between-study variance. If the addition of a study leads to an increase in the total variance, standard errors will rise and confidence intervals will become larger. This leads to the paradoxical situation where the inclusion of studies with small sample sizes can reduce the overall statistical power of the meta-analysis (Hedges and Pigott 2001). Small studies do this by introducing power-sapping heterogeneity into the sample that exceeds the value of the information they provide regarding the estimate of the effect size.[14]

Summary

In science, small studies are sometimes followed by meta-analyses and eventually large-scale randomized controlled trials. Although a meta-analysis is no substitute for a large

randomized controlled trial, it is not uncommon for the former to reveal effects that are subsequently confirmed by the latter. Meta-analysis does this by filtering massive amounts of evidence and revealing those research opportunities that are worthy of larger-scale investigation. Meta-analyses thus provide an important link between small and large studies.

Yet randomized trials sometimes overturn the conclusions of meta-analyses, leading to questions about the validity of combining results from small, imperfect studies. These discrepancies highlight the need to control for at least four broad sources of bias. Three of these – the selective access to relevant research, the over-representation of underpowered studies that have chanced upon unusual samples, and the misapplication of fixed-effects models to heterogeneous population data – will conspire to inflate mean effect sizes, raising the likelihood of Type I errors. For these reasons it is not unusual for meta-analyses to generate effect size estimates which are bigger than those obtained from randomized controlled trials (Villar and Carroli 1995). Less common is when a meta-analysis generates a Type II error, as can happen when effects are sought with insufficient statistical power.

And so we come full circle.

To do a good meta-analysis one must know how to analyze statistical power. But to do a power analysis one must know something about the anticipated effect size and how to judge the quality of existing estimates. To do good research one must know how to do both.

Notes

1 In response to the publication bias some in the medical field have argued that editors have an obligation to publish the results of small, methodologically solid studies (Lilford and Stevens 2002). Whether editors of medical journals heed this call remains to be seen, but given the competition for citations and the corresponding need to publish conclusive research, it is highly unlikely that editors of social science journals will start devoting journal pages to the reporting of inconclusive studies. A better recommendation for editors would be to insist that authors provide information regarding the precision and size of all estimated effects along with evidence that statistical tests had enough power to do what was being asked of them. In short, the adverse consequences of a publication bias could be mitigated if only editors heeded the recommendations made in the APA's (2010) *Publication Manual*.

2 Young et al. (2008) compare the publication bias with the winner's curse in auction theory. In an auction the winning bid represents an extreme estimate of the true value of the item being sold. A more accurate estimate would be the mean bid of all the participants. Hence the winner's curse – the one who wins probably paid too much. Analogously, in science the mean effect size estimate of a pool of study-specific estimates will most closely reflect the true value, but extreme and spectacular results are more likely to get published. Published estimates thus tend to be exaggerated. In some cases published effects can be more than twice as large as actual effect sizes (Brand et al. 2008). The "curse" of these unrepresentative results falls on the consumers of research – other researchers, graduate students, indeed, all of society.

3 Authors who provide detailed instructions on how to use funnel plots and related graphical methods to interpret availability bias include Begg (1994), Egger et al. (1997), and Sterne et al. (2001, 2005).

4 Odds ratios for each study were calculated using the formula $e^{((O-E)/V)}$ where O is the number of deaths observed in the treatment group, E is the number of deaths that would be expected if the treatment had no effect, and V is the variance. All the data used for calculating the odds ratios come from Teo et al. (1991).

5 To calculate the fail-safe *N* we first need to transform the mean effect size into its standard normal equivalent ($\bar{z}$). For a correlation we can use the equation $\bar{z} = \bar{r}\sqrt{k}$ where $\bar{r}$ denotes the mean correlation, and *k* refers to the number of studies in the analysis. If the mean effect size is reported in the *d* metric we would use the following equation: $\bar{z} = [\bar{d}^2/(\bar{d}^2 + 4)]^{1/2}(k)^{1/2}$. Both equations are adapted from Rosenthal (1979, footnote 1). The formula for calculating the fail-safe *N* or N_{fs} for a set of *k* studies is $N_{fs} = (k/z_c^2)(k\bar{z}^2 - z_c^2)$ where z_c is the one-tailed critical value of *z* when $\alpha = .05$, or 1.645.

6 Rosenthal's (1979) fail-safe *N* and other versions of it (e.g., Gleser and Olkin 1996; Orwin 1983) is a useful heuristic for gauging the tolerance of a result to the file drawer problem, publication bias, and other types of availability bias. But it has been criticized for ignoring the size of, and variation in, observed effects, which would also have a bearing on the tolerance of results (Becker 2005). For more sophisticated approaches to assessing the availability bias see Iyengar and Greenhouse (1988) and Sterne and Egger (2005).

7 Feinstein (1995) did not provide a definition of excellent research but acknowledged that the challenge of developing quality criteria is about as difficult as that faced by a quadriplegic person trying to climb Mount Everest.

8 As we saw in Chapter 4, an editor who prefers to publish statistically significant results will, on average, publish one false positive for every sixteen true positives. But this proportion will increase to the extent to which papers are accepted for publication without regard for their statistical power. In the magnesium meta-analysis described above, two out of the seven studies reported false positives, a proportion nearly five times higher than what would have occurred if negative results had been reported and published in equal measure.

9 The fact that two studies managed to achieve statistical significance despite their small samples suggests that their samples were unusual, that random variation within these samples was mistakenly interpreted as natural variation in the underlying population.

10 Again, this is because power and Type I errors are *indirectly* related through the availability bias. Low power leads to an increased risk of Type II errors, but low power combined with the selective availability of research (e.g., arising from a publication bias favoring statistically significant results) leads to an increased risk of Type I errors, as explained in Chapter 4.

11 Strictly speaking we should assess the statistical power of tests, not studies. For instance, a study which reports both main effects and effects for subgroups will have at least two levels of power. The tests for main effects may be adequately powered but this may not be true for tests based on smaller subgroups.

12 Just to make things really confusing, both θ and μ are commonly referred to as the mean effect size, and indeed they are, even if they are not the same.

13 As long as the population effect size does not equal zero the addition of more studies will always improve the chances that the confidence interval excludes zero and boosts the power of a fixed-effects meta-analysis. Cohn and Becker (2003) provide a complex formula for calculating statistical power in these circumstances, but the main point is that an increase in power will always occur when the variance associated with the mean effect size decreases or the population effect size increases. Additional formulas for calculating the power of meta-analyses are provided by Hedges and Pigott (2001).

14 While you would generally expect the addition of studies to increase the statistical power of a random-effects meta-analysis, Cohn and Becker (2003, Table 4) provide an example where this was not the case.

Last word: thirty recommendations for researchers

The lessons of this book can be distilled into the thirty recommendations listed below. The numbers in brackets refer to the relevant chapters in this book.

Before doing the study:

1. Quantify your expectations regarding the effect size. Ask yourself, what results do I expect to see in this study? Be explicit. Develop a rationale for doing another study given extant results. If there is no past relevant research, ask: How big an effect would I need to see to make this study worthwhile? Would the rejection of the null hypothesis of no effect be sufficiently interesting? (1)
2. Identify the range of effect sizes observed in prior studies. When reviewing past research, do not be distracted by the conclusions of others that may have been mistakenly drawn from *p* values. Rather, examine the evidence and draw your own conclusions. The relevant evidence includes the size and direction of the estimated effect, the precision of the estimate, and the reliability of the measurement procedures. To minimize the threat of the availability bias make every effort to examine the evidence from unpublished, as well as published, research. (5,6)
3. Look for meta-analyses that are relevant to the effect you are interested in or consider doing one yourself. Meta-analyses are useful for providing non-zero benchmarks that may be more meaningful than testing the null hypothesis of no effect. A good meta-analysis will also reveal unexplored avenues for further research. (5)

When designing the study:

4. Conduct a prospective power analysis to determine the minimum sample sizes needed to detect the expected effect size. Carefully assess the trade-off between sample size and power. Ask yourself, do the anticipated benefits of detecting an effect of this magnitude exceed the costs required to detect it? (3, Appendix 1)
5. Quantify your expectations regarding the precision of the estimate. Ask, what is my desired margin of error and what sample size will be needed to achieve this? (3)
6. When calculating the minimum sample size, budget for the possibility of conducting subgroup or multivariate analysis. Minimum sample sizes should be based on

the size of the smallest group tested or on the number of predictors in the model. On top of this allow yourself some wiggle room to compensate for over-stated estimates (in other studies) and measurement error (in your own study). Err on the side of over-sampling. (3)

7. If conducting replication research assess the statistical power of prior studies that have failed to find statistically significant results. (But do not calculate power based on the results obtained in those studies. Instead, use the weighted mean effect size obtained from all the available research.) Do you have good reasons to suspect that prior nonsignificant results were affected by Type II error? If so, note the sample sizes and tests types used in these studies. Ask yourself: Will I be able to adopt more powerful tests? Will I have access to a bigger sample? If there is no suspicion of Type II error, rethink the need for a replication study – there may be no meaningful effect to detect. (3)

When collecting the data:

8. Run a small-scale pilot study to obtain an estimate of the effect size and to test-drive your data-collection procedures. Information on the likely effect size can be used to fine-tune decisions about the sampling frame and minimum sample size. (3)
9. Give careful thought to ways of reducing measurement error. Measurement error can be a substantial drain on statistical power. (3)
10. If your study is sample-based ensure that your sample comes from the population it is supposed to represent and not some mixture of populations. If you inadvertently try to measure two or more effects you will undermine the power of your study. (4)
11. Keep your required sample size in view. Unforeseen events which may prevent you reaching this number could undermine your ability to draw conclusions about the effects you hope to observe. (3)

When analyzing the data:

12. Choose the most powerful statistical tests permitted by the data and the theory. Parametric tests are more powerful than non-parametric tests; directional (one-tailed) tests are more powerful than nondirectional (two-tailed) tests; and tests involving metric data are more powerful than tests involving nominal or ordinal data. (4)
13. Resist the temptation to perform multiple analyses of the same data (e.g., subgroup analyses). If you run enough tests you will eventually find statistically significant results even when the null hypothesis is true. Be aware that adjusting alpha to compensate for the familywise error rate will dampen power and increase the likelihood of Type II errors. Clearly distinguish between pre-specified and post hoc hypotheses. View accidental findings with circumspection. Better still, see if they will replicate. (4)

14. Evaluate the stability of your results either by analyzing data from a second sample (replication) or by splitting the data into two parts and analyzing each part separately (cross-validation). Do not draw conclusions about the credibility or replicability of results from tests of statistical significance. (3,4)
15. Assess the relative risk of Type I and Type II errors. Understand that these risks are mutually exclusive – a study can make only one type of error. If your results turn out to be statistically significant, assess the possibility that you have still made a Type I error. Do not assume that just because $p < .05$ you have not drawn a false positive. If your results turn out to be statistically nonsignificant, consider whether there are good reasons for suspecting a Type II error (e.g., consistent effects found in past research). If so, see if a compelling case can be made for relaxing alpha significance levels. If no case can be made, evaluate the possibility of collecting additional data to increase the power of your study. Do not assume that just because $p > .05$ there is no underlying effect. Acknowledge the fact that your nonsignificant result is inconclusive. (3)

When reporting the results of the study:

16. Clearly indicate the method used for setting the sample size and provide a rationale. (3)
17. Describe the data collected. Provide the reader with enough information to both understand the data (e.g., means, standard deviations, typical and extreme cases) and independently determine whether anything appears anomalous in the dataset. (2)
18. Test the assumptions underlying your chosen statistical tests and report the results. Also report the results of tests assessing the measurement procedures used (e.g., internal consistency). (3)
19. Report the size and direction of estimated effects. Do this even if the results were found to be statistically nonsignificant and your effects are miserably small. Make your results meta-analytically friendly and report effect size estimates in standardized form (i.e., *r* or *d* equivalents). If the measure being used is meaningful in practical terms (e.g., number of lives saved by the treatment), also report the effect in its unstandardized form. Clearly indicate the type of effect size index being reported. (1,2,5)
20. Provide confidence intervals to quantify the degree of precision associated with your effect size estimates. (1)
21. Report exact *p* values for all statistical tests, including those with nonsignificant results. (3)
22. Report the power of your statistical tests. Reported power should be a priori power and not calculated from the effect sizes or *p* values observed in the study. (3)
23. If reporting the results of multivariate analyses (e.g., multiple regression), report the zero-order correlations for all variables. (Future researchers and meta-analysts may be interested in the relationship between only one pair of variables in your study.) A correlation matrix serves this purpose well but there is no need to stud

it with asterisks. A note indicating that correlations larger than X are statistically significant at the $p = .05$ level is more than sufficient. (5)

24. Clearly label as post hoc any hypotheses developed to account for accidental or unexpected findings. Entertain the possibility that unexpected findings may reflect random sampling variation rather than natural variation in the population. (4)

When interpreting the results of the study:

25. Assess the practical significance of your results. Ask yourself: What do the results mean and for whom? In what contexts might the observed effect be particularly meaningful? Who might be affected? What is the net contribution to knowledge? If the estimated effect is small, under what circumstances might it be judged to be important? Do the effects accumulate over time? Do not confuse practical with statistical significance. Always use a qualifier when discussing significance. (1,2)
26. If it aids interpretation, report your effect size estimates in language familiar to the layman. For example, if reporting a measure of association, consider using a binomial effect size display. If comparing differences between groups, consider calculating a risk ratio or a probability. (1)
27. Explicitly compare your results with prior estimates and intervals obtained in other studies. Is your effect size estimate bigger, smaller, or about the same? Are the different estimates converging on a common population effect or are there reasons to suspect that several effects are being measured? Are you seeing something new or verifying something known? Consider calculating a weighted mean effect size based on all the available estimates. If multiple intervals are reported in the literature, consider presenting them along with your own in a graph. (1,2,5)
28. When comparing results meta-analytically, ensure that the statistical model used to pool the individual estimates is appropriate for the data. If population effect sizes are variable, do not use fixed-effects methods. If you wish to draw inferences that are not limited to the results in hand, use random-effects methods. (6, Appendix 2)

Other recommendations:

29. Make your data and results publicly available. If your study is not likely to get published, put your results online as a working paper or present a conference paper. If your study does get published, make your dataset publicly accessible (e.g., by putting it on your website). (6)
30. Before submitting your finished paper, review the publication manuals of the APA (2010) or AERA (2006) as appropriate. Alternatively, review the twenty-one guidelines of Wilkinson and the Taskforce on Statistical Inference (1999) or the fifteen guidelines of Bailar and Mosteller (1988) or go through an “article review checklist” such as the one provided by Campion (1993).

Appendix 1 Minimum sample sizes

Table A1.1 *Minimum sample sizes for detecting a statistically significant difference between two group means (d)*

	Power									
d (α_1)	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.10	1,084	1,256	1,443	1,650	1,884	2,154	2,475	2,878	3,427	4,331
0.20	272	315	362	414	472	540	620	721	858	1,084
0.30	122	141	162	185	211	241	277	321	382	483
0.40	70	80	92	105	120	136	156	182	216	272
0.50	45	52	60	68	77	88	101	117	139	175
0.60	32	37	42	48	54	62	71	82	97	122
0.70	24	28	31	36	40	46	52	61	72	90
0.80	19	22	24	28	31	36	41	47	55	70
0.90	15	17	20	22	25	29	32	37	44	55
1.00	13	15	16	18	21	23	27	31	36	45
1.10	11	12	14	16	18	20	22	26	30	38
1.20	10	11	12	14	15	17	19	22	26	32
1.30	9	10	11	12	13	15	17	19	22	28
1.40	8	9	10	11	12	13	15	17	19	24
1.50	7	8	9	10	11	12	13	15	17	21
1.60	7	7	8	9	10	11	12	13	15	19
1.70	6	7	7	8	9	10	11	12	14	17
1.80	6	6	7	7	8	9	10	11	13	15
1.90	6	6	6	7	8	8	9	10	12	14

	Power									
d (α_2)	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.10	1,539	1,742	1,962	2,203	2,471	2,779	3,142	3,594	4,205	5,200
0.20	387	437	492	552	620	696	787	900	1,053	1,302
0.30	173	196	220	247	277	311	351	401	469	580
0.40	98	111	125	140	157	176	199	227	265	327
0.50	64	72	81	90	101	113	128	146	171	210
0.60	45	51	57	64	71	80	90	102	119	147
0.70	34	38	42	47	53	59	67	76	88	109
0.80	27	30	33	37	41	46	52	59	68	84
0.90	22	24	27	30	33	37	41	47	54	67
1.00	18	20	22	25	27	30	34	38	45	54
1.10	15	17	19	21	23	26	29	32	37	45
1.20	13	15	16	18	20	22	24	28	32	39
1.30	12	13	14	16	17	19	21	24	27	33
1.40	11	12	13	14	15	17	19	21	24	29
1.50	10	11	11	13	14	15	17	19	21	26
1.60	9	10	10	11	12	14	15	17	19	23
1.70	8	9	10	10	11	12	14	15	17	21
1.80	8	8	9	10	10	11	12	14	16	19
1.90	7	8	8	9	10	11	11	13	14	17

Note: $\alpha = .05$; α_1 refers to one-tailed tests; α_2 refers to two-tailed (nondirectional) tests. The sample sizes reported for d are combined (i.e., $n_1 + n_2$). The minimum number in each independent sample is thus half the figure shown in the table rounded up to the nearest whole number.

Table A1.2 *Minimum sample sizes for detecting a correlation coefficient (r)*

	Power									
$r\ (\alpha_1)$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.05	1,083	1,254	1,441	1,648	1,881	2,150	2,471	2,873	3,422	4,324
0.10	271	314	360	412	470	537	616	716	853	1,077
0.15	121	140	160	183	208	238	273	317	377	476
0.20	68	79	90	103	117	133	153	177	211	266
0.25	44	50	58	66	75	85	97	113	134	168
0.30	31	35	40	45	51	59	67	77	92	115
0.35	23	26	29	33	38	43	49	56	67	83
0.40	18	20	23	25	29	32	37	42	50	63
0.45	14	16	18	20	22	25	29	33	39	49
0.50	12	13	14	16	18	20	23	26	31	38
0.55	10	11	12	13	15	16	19	21	25	31
0.60	8	9	10	11	12	14	15	17	20	25
0.65	7	8	9	9	10	11	13	14	17	21
0.70	6	7	7	8	9	10	11	12	14	17
0.75	6	6	6	7	7	8	9	10	12	14
0.80	5	5	6	6	6	7	8	8	10	11
0.85	4	5	5	5	6	6	6	7	8	9
0.90	4	4	4	5	5	5	5	6	6	8
0.95	4	4	4	4	4	4	4	5	5	6

	Power									
$r\ (\alpha_2)$	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.05	1,536	1,740	1,959	2,199	2,467	2,774	3,137	3,588	4,198	5,192
0.10	384	435	489	549	616	692	782	894	1,046	1,293
0.15	171	193	217	243	273	306	346	396	462	571
0.20	96	108	122	136	153	171	193	221	258	319
0.25	62	69	78	87	97	109	123	140	164	202
0.30	43	48	54	60	67	75	84	96	112	138
0.35	31	35	39	44	49	54	61	70	81	100
0.40	24	27	30	33	37	41	46	53	61	75
0.45	19	21	23	26	29	32	36	41	47	58
0.50	15	17	19	21	23	26	29	32	37	46
0.55	13	14	15	17	19	21	23	26	30	37
0.60	11	12	13	14	15	17	19	21	24	30
0.65	9	10	11	12	13	14	16	18	20	24
0.70	8	9	9	10	11	12	13	15	17	20
0.75	7	7	8	9	9	10	11	12	14	16
0.80	6	6	7	7	8	8	9	10	11	13
0.85	5	6	6	6	7	7	8	8	9	11
0.90	5	5	5	5	6	6	6	7	8	9
0.95	4	4	4	5	5	5	5	5	6	7

Note: $\alpha = .05$; α_1 refers to one-tailed tests; α_2 refers to two-tailed (nondirectional) tests.

Appendix 2 Alternative methods for meta-analysis

The two mainstream methods for running a meta-analysis are the methods developed by Hunter and Schmidt (see Hunter and Schmidt 2000; Schmidt and Hunter 1977, 1999) and by Hedges and his colleagues (see Hedges 1981, 1992, 2007; Hedges and Olkin 1980, 1985; Hedges and Vevea 1998). The kryptonite meta-analysis in Chapter 5 was an example of how to apply a stripped-down version of the Hunter and Schmidt method for combining effects reported in the correlational metric (*r*). In this appendix it will be shown how to meta-analyze both *r* and *d* effects using the Hedges et al. method and how to compute mean effect sizes using both fixed- and random-effects procedures. Some comparisons between the methods by Hedges et al. and Hunter and Schmidt will be drawn in the final section.

Combining *d* effects using Hedges et al.'s method

Let's assume we are interested in the effect of gender on map-reading ability and we have identified ten studies reporting sample sizes (*n*) and effect sizes (*d*) as summarized in the first two columns of Table A2.1. The direction of the effect is irrelevant to our meta-analysis and, in the interests of maintaining marital harmony, should probably not receive too much attention anyway – particularly when driving.[1]

Our meta-analysis will generate four outcomes; (1) a mean effect size, (2) a confidence interval for the mean effect size, (3) a *z* score which can be used to assess the statistical significance of the result, and (4) a *Q* statistic to quantify the variability in the sample of effect sizes. This last result will be useful in deciding whether we should ultimately rely on fixed- or random-effects procedures. Following Hedges and Vevea (1998) an asterisk will be used to distinguish equations done for the random-effects procedures. As w_i will be used to denote the weights assigned to study *i* in the fixed-effects procedure, its counterpart in the random-effects procedure will be designated w_i^*. Similarly, if $\overline{d}$ denotes the mean effect size generated by the fixed-effects analysis, then $\overline{d}^*$ will indicate the mean effect size generated by the random-effects analysis.

The fixed-effects analysis depends on the sums of three sets of variables: the individual study weights (w_i for individual estimates but *w* when summed), the weights multiplied by their corresponding effect sizes (*wd*), and the weights multiplied by the

Table A2.1 *Gender and map-reading ability*

	Raw study data			Fixed-effects sums			Random-effects sums		
Study	n	d	v_i	w	wd	wd^2	w^2	w^*	w^*d
1	35	0.17	0.11	8.72	1.48	0.25	76.01	5.50	0.94
2	76	0.30	0.05	18.79	5.64	1.69	353.01	8.32	2.50
3	80	1.02	0.06	17.70	18.05	18.41	313.23	8.10	8.26
4	44	0.22	0.09	10.93	2.41	0.53	119.55	6.31	1.39
5	50	0.23	0.08	12.42	2.86	0.66	154.20	6.78	1.56
6	105	0.87	0.04	23.98	20.86	18.15	575.09	9.20	8.00
7	168	0.79	0.03	38.96	30.78	24.32	1,517.93	10.79	8.53
8	32	0.60	0.13	7.66	4.59	2.76	58.61	5.06	3.04
9	94	0.18	0.04	23.41	4.21	0.76	547.80	9.11	1.64
10	62	0.12	0.06	15.47	1.86	0.22	239.39	7.60	0.91
Totals				178.03	92.74	67.75	3,954.83	76.78	36.76

Note: The data are fictitious. The procedures for combining them are adapted from Hedges and Vevea (1998, Table 1). The variance (v_i) of d_i for study i was calculated as: $v_i = 4(1 + d_i^2/8)/n_i$.

square of the effect sizes (wd^2). In the Hedges et al. method the weights used in the estimation are the inverse of the variance observed in each study, as expressed in the following equation:

$$w_i = \frac{1}{v_i} \tag{1}$$

The methods for calculating the sampling variance for each study depend on whether the effect size is being measured as d or r. In the case of d effects, and if we can assume the groups being compared within each study are approximately equal in size, we can compute the variance using the equation $v_i = 4(1 + d_i^2/8)/n_i$ where d_i and n_i refer to the effect size and sample size for study i respectively (Hedges and Vevea 1998). All of the calculations can be done using a spreadsheet package such as Excel. In Table A2.1 the variance scores for each study are listed in the third column under v_i and the sums for the fixed-effects procedures are shown in the middle three columns.

To calculate a mean effect size using fixed-effects procedures, we multiply the weights and effect sizes for each study, sum them, and then divide by the sum of the weights, as follows:

$$\overline{d} = \frac{\sum_{i=1}^{k} w_i d_i}{\sum_{i=1}^{k} w_i} \tag{2}$$

$$= 92.74/178.03 = 0.52$$

To calculate the confidence interval and z score for the mean effect size, we need to estimate the sampling variance ($v.$) of the mean. This is measured as the inverse of the sum of the study weights:

$$v. = \frac{1}{\sum_{i=1}^{k} w_i} = 1/178.03 = 0.006 \tag{3}$$

The width of the confidence interval is related to the standard error of the fixed-effects mean ($SE_{\overline{d}}$). The standard error is the square root of the variance, or $\sqrt{.006} = .077$. To calculate the width of the interval we also need to know the two-tailed critical value of the standard normal distribution ($z_{\alpha/2}$) for our chosen level of alpha. For a 95% interval this value is 1.96. The upper and lower bounds are measured from the mean by adding or subtracting the standard error multiplied by this critical value, as follows:

$$\overline{d} \pm z_{\alpha/2} SE_{\overline{d}} \tag{4}$$

$$\text{CI}_{\text{lower}} = 0.52 - (1.96 \times .077) = 0.37$$

$$\text{CI}_{\text{upper}} = 0.52 + (1.966 \times .077) = 0.67$$

To assess the statistical significance of this result we would normally test the null hypothesis that the mean effect size equals 0. To do this we calculate a z score by taking the absolute difference between the mean effect size and null value and dividing by the standard error of the mean. This can be expressed in an equation as follows:

$$z = \left(\left|\overline{d} - 0\right|\right) / SE_{\overline{d}} = 0\,.52/.077 = 6.75 \tag{5}$$

We would reject the hypothesis of no effect in a two-tailed test whenever the z score exceeds the critical z value for $\alpha_2 = .05$, or 1.96. In this case $6.75 > 1.96$ so we can conclude that the result is statistically significant. This same conclusion could have been reached by noting that the 95% confidence interval excluded the null value of 0.

By gauging the heterogeneity of the distribution of effect sizes, we are essentially asking, do the individual effect size estimates reflect a common population effect size? Formally, this is a test of the hypothesis H_0: $\theta_1 = \theta_2 = \ldots = \theta_{\text{k}}$ versus the alternative hypothesis that at least one of the population effect sizes θ_i differs from the rest (Hedges and Vevea 1998). To test this hypothesis we can calculate a Q statistic which is the weighted sum of squares of the effect size estimates about the weighted mean effect size, as follows:

$$Q = \sum_{i=1}^{k} w_i \left(d_i - \overline{d}\right)^2 = wd^2 - (wd)^2/w \tag{6}$$

$$= 67.75 - (92.74)^2/178.03 = 19.44$$

To interpret this result we need to compare it against the critical value of the chi-square distribution for $k - 1$ degrees of freedom where k equals the number of estimates being

pooled. With ten studies in our sample there are $10 - 1 = 9$ degrees of freedom in our test. By consulting a table listing values in the chi-square distribution, we learn that the critical chi-square value for 9 degrees of freedom when $\alpha = .05$ is 16.92. As our Q statistic exceeds this critical value we reject the hypothesis that the population effect sizes are equal. From this we infer that the sample of effect sizes is not fixed on a common mean but is randomly distributed about some super-mean. A more appropriate procedure for calculating the mean effect size is therefore one which takes into account the variance in the sample of estimates and the additional variance in the sample of effect sizes. A random-effects analysis does this by accounting for both within-study variance (v_i) and between-study variance (τ^2). Under the fixed-effects approach, individual effect sizes are weighted by the inverse of the within-study variance, as in equation 1. But under the random-effects approach the relevant weights are the inverse of both types of variance added together:

$$w_i^* = \frac{1}{v_i + \tau^2} \tag{7}$$

To do the meta-analysis using the random-effects procedure we need three more sums. The additional sums are shown under the columns headed "Random-effects sums" in Table A2.1. These refer to the sums of three sets of variables; the square of the fixed-effects weights (w^2), the random-effects weights (w^*), and the random-effects weights multiplied by the effect sizes (w^*d).

Following the procedures laid out by Hedges and Vevea (1998), the first step in our random-effects analysis is to estimate the between-studies variance component using the following equation:

$$\tau^2 = \frac{Q - (k-1)}{c} \tag{8}$$

The Q statistic was calculated in the fixed-effects analysis as $Q = 19.44$ and $k - 1 = 9$. To calculate the constant c we use the equation:

$$\begin{aligned} c &= \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} \\ &= 178.03 - 3{,}954.83/178.03 = 155.82 \end{aligned} \tag{9}$$

From this we can estimate that the between-studies variance $\tau^2 = (19.44 - 9)/155.82 = 0.067$. If this equation had generated a negative value, then we would estimate τ^2 as 0 as variance cannot be negative. We can now calculate the individual weights to be used in our random-effects analysis using equation 7: ${w_i}^* = 1/(v_i + 0.067)$. While we are at it, we can also calculate the final column of our table to get the sum of w^*d.

To calculate a mean effect size using random-effects procedures, we would use the following equation:

$$\begin{aligned}\overline{d}^{*} &= \frac{\sum_{i=1}^{k} w_i^{*} d_i}{\sum_{i=1}^{k} w_i^{*}} \\ &= 36.76/76.78 = 0.48\end{aligned} \tag{10}$$

The variance of this mean effect size is calculated using the following equation:

$$\begin{aligned}v.^{*} &= \frac{1}{\sum_{i=1}^{k} w_i^{*}} \\ &= 1/76.78 = 0.013\end{aligned} \tag{11}$$

The standard error of the random-effects mean ($SE_{\overline{d}^{*}}$) is the square root of the variance, or $\sqrt{.013} = .114$. The upper and lower bounds of the 95% confidence interval are calculated using equation 4 above but are calculated using $\overline{d}^{*}$ instead of $\overline{d}$ and $SE_{\overline{d}^{*}}$ instead of $SE_{\overline{d}}$. This generates an interval with the following bounds:

$$\begin{aligned}\text{CI}_{\text{lower}} &= 0.48 - (1.96 \times .114) = 0.26 \\ \text{CI}_{\text{upper}} &= 0.48 + (1.96 \times .114) = 0.70\end{aligned}$$

To calculate the statistical significance of this random-effects-generated result we would use equation 5 making the same changes, as follows:

$$= 0.48/.114 = 4.21$$

As our *z* score (4.21) exceeds the critical value of $z_{\alpha/2}$ (1.96), we can conclude that the random-effects result is statistically significant.

Comparing the results side by side, we can see that the random-effects procedure produced a more conservative estimate of the mean effect size with a wider confidence interval:

Fixed-effects mean: .52 (CI_{95} .37 to .67)

Random-effects mean: .48 (CI_{95} .26 to .70)

Estimates calculated using random-effects procedures will usually be smaller than their fixed-effects counterparts because of the additional between-study variance included in the analysis. For the same reason random-effects intervals will usually be wider. Although we can put more faith in the random-effects results, the accommodation

Table A2.2 *Kryptonite and flying ability – part II*

	Raw study data			Fixed-effects sums			Random-effects sums		
Study	n	r	z	w	wz	wz^2	w^2	w^*	w^*z
Luthor (1940)	80	−.48	−.523	77	−40.27	21.06	5,929	11.35	−5.94
Brainiac (1958)	112	−.58	−.662	109	−72.21	47.84	11,881	11.86	−7.86
Zod et al. (1961)	32	.05	.050	29	1.45	0.07	841	9.12	0.46
Totals				215	−111.03	68.97	18,651	32.34	−13.34

of additional variance has adverse implications for statistical power, as discussed in Chapter 6.

Combining *r* effects using Hedges et al.'s method

In the meta-analysis just done the effect size being estimated was expressed in the *d* metric. Using the Hedges et al. approach, the procedures for combining effects of the *r* family differ in two respects. First, and prior to aggregation, raw effect sizes are transformed into standard scores using the Fisher *r*-to-*z* transformation. This transformation can be done by hand using equation 12, but a simpler method is to use an online calculator such as those provided by Lane (2008) and Lowry (2008a). Where large numbers of correlations are involved, a more efficient procedure would be to use a spreadsheet such as Excel and transform raw correlations *en masse* using the formula: =FISHER(r).

$$z_r = \frac{1}{2}\ln\frac{1+r}{1-r} \tag{12}$$

Second, the variances used in this procedure (and hence the weights and standard errors) derive from the variance of the z distribution which is $1/n$. Specifically, the variance (v_i) for each study's estimate is calculated using the more accurate equation $1/(n_i - 3)$ where n_i refers to sample size of study i. As the weights are the inverse of the variance in the Hedges et al. method (see equation 1), the optimal weights (w_i) in this approach are just $n_i - 3$.

To illustrate the Hedges et al. method for cumulating correlations, we will use the kryptonite data originally reported in Chapter 5, and reproduced here in Table A2.2. The raw study data denoting sample sizes (n) and effect sizes (r) along with the transformed correlations (z) are found in the left-hand columns. The table also shows the three sums relevant for the fixed-effects equations in the middle columns and the three additional sums required for random-effects analyses in the right-hand columns.

To calculate a mean effect size the transformed effect sizes (now denoted z_i) are weighted and combined as follows:

$$\begin{aligned}\overline{z} &= \frac{\sum_{i=1}^{k} w_i z_i}{\sum_{i=1}^{k} w_i} \\ &= -111.03/215 = -0.52\end{aligned} \tag{13}$$

To calculate the variance of this mean we use equation 3 above, which gives $v_\cdot = 1/215 = 0.005$. The standard error of the mean ($SE_{\overline{z}}$) is the square root of the variance or $\sqrt{.005} = 0.071$. With this standard error a 95% confidence interval can be calculated using equation 4 above but substituting $\overline{z}$ for $\overline{d}$:

$$\begin{aligned}\text{CI}_{\text{lower}} &= -0.52 - (1.96 \times .071) = -0.66 \\ \text{CI}_{\text{upper}} &= -0.52 + (1.96 \times .071) = -0.38\end{aligned}$$

To calculate the probability that the mean effect size differs from the null value of 0, we use equation 5 to calculate a z score making a similar substitution, $.52/.071 = 7.32$. As this z score exceeds the critical value of $z_{\alpha/2}$ (1.96 when $\alpha = .05$), we can conclude that the result is statistically significant at the $p < .05$ level.

To test the homogeneity of the distribution of the correlations, we use the following variation on equation 6:

$$\begin{aligned}Q &= \sum_{i=1}^{k} w_i (z_i - \overline{z})^2 = wz^2 - (wz)^2/w \\ &= 68.97 - (-111.03)^2/215 = 11.63\end{aligned} \tag{14}$$

Again, this result is interpreted using the chi-square distribution with $k - 1$ degrees of freedom. As there are just three studies in this meta-analysis, we interpret the result as having two degrees of freedom. Consulting a table of values in the chi-square distribution at different levels of alpha and for various degrees of freedom, we would learn that the critical upper-tail value is 5.99. As the Q statistic exceeds this critical value, the hypothesis of homogeneity is rejected. We now have good grounds for calculating a revised mean correlation using a random-effects procedure. To do this we need a variance component that accounts for both the within- (v_i) and between-studies (τ^2) variance. The equation for τ^2 is the same as equation 8 above, and the calculation for the constant c is the same as equation 9. To calculate c we rely on the weights that are derived from the z-based variance, as follows:

$$\begin{aligned}c &= 215 - 18{,}651/215 = 128.25 \\ \frac{\tau^2 = 11.63 - (3-1)}{128.25} &= 0.075\end{aligned}$$

Again, if the estimate of τ^2 turns out to be less than zero it is truncated at zero as variance cannot be negative in value.

We can now calculate the weights to be used in our random-effects analysis using the equation $w_i{}^* = 1/v_i{}^*$ where $v_i{}^* = (v_i + 0.075)$. We can also calculate the information needed for the final column of Table A2.2. After summing these two columns we can compute a random-effects mean effect size as follows:

$$\begin{aligned} \overline{z}_r^* &= \frac{\sum_{i=1}^{k} w_i^* z_i}{\sum_{i=1}^{k} w_i^*} \\ &= -13.34/32.34 = -0.41 \end{aligned} \tag{15}$$

The variance for this mean is calculated as shown in equation 11 and is $1/32.34 = 0.031$. Consequently the standard error of the random-effects mean of the transformed correlation $(SE_{\overline{z}^*}) = \sqrt{.031} = .176$. From this we can calculate confidence intervals using equation 4 with the appropriate substitutions:

$$\text{CI}_{\text{lower}} = -0.41 - (1.96 \times .176) = -0.76$$
$$\text{CI}_{\text{upper}} = -0.41 + (1.96 \times .176) = -0.07$$

Again, a *z* score (which should not be confused with our transformed correlations or z_is) can be calculated using equation 5. In this case the *z* score ($.41/.176 = 2.33$) exceeds the critical value of $z_{\alpha/2}$ (1.96), permitting us to conclude that the result is statistically significant at the $p < .05$ level.

Comparing the results side by side, we can see that the random-effects procedure has again produced a more conservative estimate of the mean effect size with a wider confidence interval:

Fixed-effects mean: –.52 (CI_{95} –.66 to –.38)

Random-effects mean: –.41 (CI_{95} –.76 to –.07)

However, before we interpret these results, we would need to transform them back to the *r* metric using the inverse of the Fisher transformation (equation 16). This can be done using an online calculator or the inverse Fisher formula in Excel: =FISHERINV(*z*).

$$\overline{r} = \frac{e^{2\overline{z}} - 1}{e^{2\overline{z}} + 1} \tag{16}$$

Expressed in the correlational metric the meta-analytic results are as follows:

Fixed-effects mean: –.48 (CI_{95} –.57 to –.36)

Random-effects mean: –.39 (CI_{95} –.64 to –.07)

Comparing Hedges et al. with Hunter and Schmidt

The three kryptonite studies have now been meta-analyzed four ways using the approaches developed by Hedges et al. (above) and Hunter and Schmidt (in Chapter 5). There are some important differences between these two approaches which may not be immediately obvious by looking at the equations. Neither does it help that each set of authors has a particular preference for notation. For example, Hedges and Vevea (1998) use d^* to denote a mean effect size calculated using random-effects procedures, but in Hunter and Schmidt (2004: 284) d^* denotes an unbiased estimator of d – which Hedges et al. would label as g!

To identify substantive differences it is helpful to list the main equations of both methods alongside their generic, non-branded alternatives. Table A2.3 shows the different ways of presenting the four most important equations used in meta-analysis. The four equations are used to calculate a weighted mean effect size along with its corresponding variance and z score, and to assess the homogeneity of the sample of effect sizes. Standardized versions of these equations are listed in the column headed "Generic". The other columns show how Hunter and Schmidt and Hedges et al. adapt these generic equations for their own purposes. To keep things manageable, only the equations for fixed-effects procedures are shown in the Hedges et al. side of the table.

At first glance this table may seem to convey a bewildering array of information. But there are really just two things which distinguish the two methods. First, in the Hedges et al. approach to combining r effects, raw correlations are transformed into zs prior to aggregation while Hunter and Schmidt use untransformed rs. Correlations are transformed to correct a small negative bias in average rs, but the transformation introduces a small positive bias into the results. Different meta-analytic circumstances, such as the number of studies being combined, affect whether the swing is more one way than the other, but the choice between using transformed or raw correlations may depend on whether one prefers a slightly underestimated or overestimated result (Strube 1988).

Second, the two approaches differ in the way they accommodate the variance in the data and this has consequences for the weights given to estimates, the computation of sampling variance, standard errors, and confidence intervals. When correlations are being combined the weights used by Hunter and Schmidt are based on the sample size of each study, or n_i, while the weights used by Hedges et al. are $n_i - 3$. But that is where the similarities end. A good Hunter and Schmidt analysis would modify the weights to account for any number of study-specific artifacts such as measurement reliability ($n_i r_{yy}$), while a random-effects version of Hedges et al. would factor in the additional between-studies variance.[2] These differences explain the dissimilar results produced by the four kryptonite meta-analyses:

Hunter and Schmidt (uncorrected): –.454 (CI_{95} –.693 to –.215)
Hunter and Schmidt (corrected): –.500 (CI_{95} –.755 to –.245)
Hedges et al. (fixed-effects): –.478 (CI_{95} –.572 to –.365)
Hedges et al. (random-effects): –.391 (CI_{95} –.639 to –.068)

Table A2.3 *Alternative equations used in meta-analysis*

		Hunter and Schmidt		Hedges et al.	
Output	Generic	*d*	*r*	*d*	*r*
Weighted mean ES	$\overline{ES} = \frac{\sum w_i ES_i}{\sum w_i}$	$\overline{d} = \frac{\sum w_i d_i}{\sum d_i}$	$\overline{r} = \frac{\sum n_i r_i}{\sum n_i}$	$\overline{d} = \frac{\sum w_i d_i}{\sum w_i}$	$\overline{z} = \frac{\sum w_i z_i}{\sum w_i}$
Variance of sample ESs	$v_\cdot = \frac{1}{\sum w_i}$	$v_{\cdot d} = \frac{\sum w_i (d_i - \overline{d})^2}{\sum w_i}$	$v_{\cdot r} = \frac{\sum n_i (r_i - \overline{r})^2}{\sum n_i}$	$v_\cdot = \frac{1}{\sum w_i}$	$v_\cdot = \frac{1}{\sum w_i}$
z score	$z = \frac{\lvert\overline{ES}\rvert}{SE_{\overline{ES}}}$	$z = \frac{\lvert\overline{d}\rvert}{SE_{\overline{d}}}$	$z = \frac{\lvert\overline{r}\rvert}{SE_{\overline{r}}}$	$z = \frac{\lvert\overline{d}\rvert}{SE_{\overline{d}}}$	$z = \frac{\lvert\overline{z}\rvert}{SE_{\overline{r}}}$
Homogeneity statistic	$Q = \sum w_i (ES_i - \overline{ES})^2$	–	$\chi^2_{k-1} = \sum \frac{(n_i - 1)(r_i - \overline{r})^2}{(1 - \overline{r}^2)^2}$	$Q = \sum w_i (d_i - \overline{d})^2$	$Q = \sum w_i (z_i - \overline{z})^2$
Usual weights	–	$w_i = n_i$ (or variations thereon, e.g., $n_i a_i^2$, $n_i r_{yy}$)		$w_i = 1/\text{vi}$ where $v_i = 4(1 + d_i^2/8)n_i$	$w_i = 1/\text{vi}$ where $v_i = 1/(n_i - 3)$

Notes: ES = effect size and $\overline{ES}$ = the mean ES observed for a sample of effect sizes which may be expressed in the form of *d*, *r*, or *z* (*z* = Fisher transformed *r*); *k* = number of independent estimates being pooled; n_i = the sample size of study *i*; $n_i a_i^2$ and $n_i r_{yy}$ denote weights based on the sample size multiplied by the square of some attenuation multiplier (a_i^2) such as the measurement reliability (r_{yy}) of the dependent variable *y*; *Q* and χ^2 are both homogeneity test statistics and are interpreted in the same way; *SE* = the standard error of the mean effect size and is the square root of the sampling variance (*v.*) in nearly every case (but note that Schmidt and Hunter (1999) advocate $\sqrt{(v./k)}$ instead); v_i = variance of estimate from study *i*; *v.* = the sampling variance of the mean effect size; w_i = weight assigned to the estimate from study *i*. Hunter and Schmidt (2004: 459ff) are dismissive of tests for the homogeneity of sample effect sizes and provide no equations in the second edition of their text. The equation included here for *r* effects comes from page 111 of their 1990 book and is sometimes described by others as being part of the Hunter and Schmidt method (e.g., Johnson et al. 1995, Table 2; Schulze 2004: 67). Tests of statistical significance are also unpopular in the Hunter and Schmidt approach (see Schmidt and Hunter 1999, footnote 1).

Sources of equations: Hedges and Vevea (1998), Hunter and Schmidt (1990, 2004), Lipsey and Wilson (2001), Schmidt and Hunter (1999), Schulze (2004).

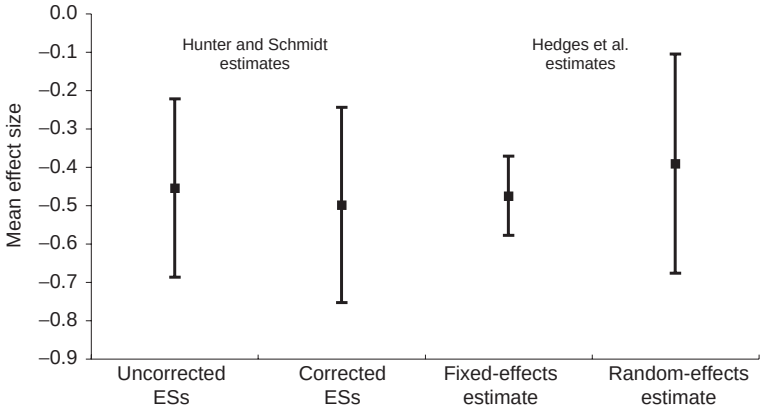


Figure A2.1 Mean effect sizes calculated four ways

The variation in these results is particularly noticeable when they are portrayed graphically, as in Figure A2.1. There is a noticeable difference in the highest and lowest means. So which result is the most accurate? And relatedly, which approach to meta-analysis generally produces the best results? This question has received attention from several scholars (e.g., Field 2005; Hall and Brannick 2002; Schulze 2004). Based on his extensive simulation Field (2005) concluded that the Hedges et al. method tends to produce the most accurate intervals, while the Hunter and Schmidt method tends to produce the most accurate mean estimates. Field noted that intervals calculated using the Hunter and Schmidt method were narrower than they should have been, meaning they would exclude the true effect a little more often than they should. This conclusion is consistent with our observations. Out of our set of four intervals, the widest was generated using the random-effects procedure of Hedges et al. This interval was 12% wider than the larger of the two intervals produced using the Hunter and Schmidt approach. But what about the narrow third interval produced using Hedges et al.'s fixed-effects analysis? Doesn't this tiny interval challenge Field's conclusion? No. As this interval is the result of misapplying fixed-effects methods to random-effects data it is much narrower than it should be and conveys a false sense of precision.

In terms of generating accurate estimates of the mean effect size, Field's findings suggest we should put our money on Hunter and Schmidt. In research settings where effects are likely to be suppressed by measurement error, Hall and Brannick (2002) concur. In this case the second estimate stands out for it is the only one that has been modified to accommodate differences in measurement reliability. Consequently it is probably the most accurate mean out of the four.

So if Hedges et al.'s method produces better intervals while Hunter and Schmidt's method produces better means, which approach to meta-analysis is better overall? The general conclusion seems to be "it depends," and certainly there is more to the debate than what has been covered here.[3] Field (2005) reasons that reviewers will need to make their own decisions based on the anticipated size of the effect, the variability in

its distribution, and the number of estimates being combined. The conclusion provided by Schulze (2004: 196) is also worth noting. At the end of his book comparing the two methods, Schulze writes: "Some approaches are better than others for various tasks but a single best set of procedures has yet to be established."

Notes

1 The data in the table are fictitious but the link between gender and navigational ability has received serious attention from scholars such as Silverman et al. (2000).
2 This additional variance is based on differences between the observed and expected values of r captured in the Q statistic and directly contributes to the value of between-studies variance (τ^2).
3 One gets the impression from reading the literature that it could be another 10–20 years before a clear winner emerges. This is because there is a general lack of awareness of the different methods and because the differences between them are so tiny that even independent reviewers can come to opposing conclusions. The random-effects procedure is clearly the superior of the two Hedges et al. methods, yet relatively few scholars use it. As Hunter and Schmidt (2000) observed, most published meta-analyses are done using the inferior fixed-effects approach. Two of the most thorough comparisons are those provided by Field (2005) and Hall and Brannick (2002). Both studies compared the methods using Monte Carlo simulations yet came to different conclusions. According to Hall and Brannick (2002: 386), the Hunter and Schmidt method produces better and "more realistic" intervals, while the wider intervals produced using Hedges et al. were more likely to "falsely contain zero." Field (2005: 463–464) drew the opposite conclusion, noting that coverage proportions for intervals generated by Hunter and Schmidt "were always too low" while those produced by Hedges et al. "were generally on target."

Bibliography

Abelson, R.P. (1985), "A variance explanation paradox: When a little is a lot," *Psychological Bulletin*, 97(1): 129–133.

Abelson, R.P. (1997), "On the surprising longevity of flogged horses," *Psychological Science*, 8(1): 12–15.

AERA (2006), "Standards for reporting on empirical social science research in AERA publications," American Educational Research Association website www.aera.net/opportunities/?id=1850, accessed 11 September 2008.

Aguinis, H., J.C. Beaty, R.J. Boik, and C.A. Pierce (2005), "Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30 year review," *Journal of Applied Psychology*, 90(1): 94–107.

Aguinis, H., S. Werner, J. Abbott, C. Angert, J.H. Park, and D. Kohlhausen (in press), "Customer-centric science: Reporting significance research results with rigor, relevance, and practical impact in mind," *Organizational Research Methods*.

Algina, J. and H.J. Keselman (2003), "Approximate confidence intervals for effect sizes," *Educational and Psychological Measurement*, 63(4): 537–553.

Algina, J., H.J. Keselman, and R.D. Penfield (2005), "An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case," *Psychological Methods*, 10(3): 317–328.

Algina, J., H.J. Keselman, and R.D. Penfield (2007), "Confidence intervals for an effect size measure in multiple linear regression," *Educational and Psychological Measurement*, 67(2): 207–218.

Allison, D.B., R.L. Allison, M.S. Faith, F. Paultre, and F.X. Pi-Sunyer (1997), "Power and money: Designing statistically powerful studies while minimizing financial costs," *Psychological Methods*, 2(1): 20–33.

Allison, G.T. (1971), *Essence of Decision: Explaining the Cuban Missile Crisis*. Boston, MA: Little, Brown.

Altman, D.G., D. Machin, T.N. Bryant, and M.J. Gardner (2000), *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. London: British Medical Journal Books.

Altman, D.G., K.F. Schulz, D. Moher, M. Egger, F. Davidoff, D. Elbourne, P.C. Gøtzsche, and T. Lang (2001), "The revised CONSORT statement for reporting randomized trials: Explanation and elaboration," *Annals of Internal Medicine*, 134(8): 663–694.

Andersen, M.B., P. McCullagh, and G.J. Wilson (2007), "But what do the numbers really tell us? Arbitrary metrics and effect size reporting in sport psychology research," *Journal of Sport and Exercise Psychology*, 29(5): 664–672.

Anesi, C. (1997), "The Titanic casualty figures," website www.anesi.com/titanic.htm, accessed 3 September 2008.

APA (1994), *Publication Manual of the American Psychological Association, 4th Edition*. Washington, DC: American Psychological Association.

APA (2001), *Publication Manual of the American Psychological Association, 5th Edition*. Washington, DC: American Psychological Association.

APA (2010), *Publication Manual of the American Psychological Association, 6th Edition*. Washington, DC: American Psychological Association.

Armstrong, J.S. (2007), "Significance tests harm progress in forecasting," *International Journal of Forecasting*, 23(2): 321–327.

Armstrong, J.S. and T.S. Overton (1977), "Estimating nonresponse bias in mail surveys," *Journal of Marketing Research*, 14(3): 396–402.

Armstrong, S.A. and R.K. Henson (2004), "Statistical and practical significance in the IJPT: A research review from 1993–2003," *International Journal of Play Therapy*, 13(2): 9–30.

Atkinson, D.R., M.J. Furlong, and B.E. Wampold (1982), "Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship?" *Journal of Counseling Psychology*, 29(2): 189–194.

Atuahene-Gima, K. (1996), "Market orientation and innovation," *Journal of Business Research*, 35(2): 93–103.

Austin, P.C., M.M. Mamdani, D.N. Juurlink, and J.E. Hux (2006), "Testing multiple statistical hypotheses resulted in spurious associations: A study of astrological signs and health," *Journal of Clinical Epidemiology*, 59(9): 964–969.

Bailar, J.C. (1995), "The practice of meta-analysis," *Journal of Clinical Epidemiology*, 48(1): 149–157.

Bailar, J.C. and F.M. Mosteller (1988), "Guidelines for statistical reporting in articles for medical journals: Amplifications and explanations," *Annals of Internal Medicine*, 108(2): 266–273.

Bakan, D. (1966), "The test of significance in psychological research," *Psychological Bulletin*, 66(6): 423–437.

Bakeman, R. (2001), "Results need nurturing: Guidelines for authors," *Infancy*, 2(1): 1–5.

Bakeman, R. (2005), "Infancy asks that authors report and discuss effect sizes," *Infancy*, 7(1): 5–6.

Bangert-Drowns, R.L. (1986), "Review of developments in meta-analytic method," *Psychological Bulletin*, 99(3): 388–399.

Baroudi, J.J. and W.J. Orlikowski (1989), "The problem of statistical power in MIS research," *MIS Quarterly*, 13(1): 87–106.

Bausell, R.B. and Y.F. Li (2002), *Power Analysis for Experimental Research: A Practical Guide for the Biological, Medical and Social Sciences*, Cambridge, UK: Cambridge University Press.

BBC (2007), "Test the nation 2007," website www.bbc.co.uk/testthenation/, accessed 5 May 2008.

Becker, B.J. (1994), "Combining significance levels," in H. Cooper and L.V. Hedges (editors), *Handbook of Research Synthesis*. New York: Russell Sage Foundation, 215–230.

Becker, B.J. (2005), "Failsafe *N* or file-drawer number," in H.R. Rothstein, A.J. Sutton, and M. Borenstein (editors), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Chichester, UK: John Wiley and Sons, 111–125.

Becker, L.A. (2000), "Effect size calculators," website http://web.uccs.edu/lbecker/Psy590/escalc 3.htm, accessed 5 May 2008.

Begg, C.B. (1994), "Publication bias," in H. Cooper and L.V. Hedges (editors), *Handbook of Research Synthesis*. New York: Russell Sage Foundation, 399–409.

Bezeau, S. and R. Graves (2001), "Statistical power and effect sizes of clinical neuropsychology research," *Journal of Clinical and Experimental Neuropsychology*, 23(3): 399–406.

Bird, K.D. (2002), "Confidence intervals for effect sizes in analysis of variance," *Educational and Psychological Measurement*, 62(2): 197–226.

Blanton, H. and J. Jaccard (2006), "Arbitrary metrics in psychology," *American Psychologist*, 61(1): 27–41.

Borkowski, S.C., M.J. Welsh, and Q. Zhang (2001), "An analysis of statistical power in behavioral accounting research," *Behavioral Research in Accounting*, 13: 63–84.

Boruch, R.F. and H. Gomez (1977), "Sensitivity, bias, and theory in impact evaluations," *Professional Psychology*, 8(4): 411–434.

Brand, A., M.T. Bradley, L.A. Best, and G. Stoica (2008), "Accuracy and effect size estimates from published psychological research," *Perceptual and Motor Skills*, 106(2): 645–649.

Breaugh, J.A. (2003), "Effect size estimation: Factors to consider and mistakes to avoid," *Journal of Management*, 29(1): 79–97.

Brewer, J.K. (1972), "On the power of statistical tests in the American Educational Research Journal," *American Educational Research Journal*, 9(3): 391–401.

Brewer, J.K. and P.W. Owen (1973), "A note on the power of statistical tests in the Journal of Educational Measurement," *Journal of Educational Measurement*, 10(1): 71–74.

Brock, J. (2003), "The 'power' of international business research," *Journal of International Business Studies*, 34(1): 90–99.

Bryant, T.N. (2000), "Computer software for calculating confidence intervals (CIA)," in D.G. Altman, D. Machin, T.N. Bryant, and M.J. Gardner (editors), *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. London: British Medical Journal Books, 208–213.

Callahan, J.L. and T.G. Reio (2006), "Making subjective judgments in quantitative studies: The importance of using effect sizes and confidence intervals," *Human Resource Development Quarterly*, 17(2): 159–173.

Campbell, D.T. (1994), "Retrospective and prospective on program impact assessment," *Evaluation Practice*, 15(3): 291–298.

Campbell, D.T. and J.C. Stanley (1963), *Experimental and Quasi-Experimental Designs for Research*, Boston, MA: Houghton-Mifflin.

Campbell, J.P. (1982), "Editorial: Some remarks from the outgoing editor," *Journal of Applied Psychology*, 67(6): 691–700.

Campion, M.A. (1993), "Article review checklist: A criterion checklist for reviewing research articles in applied psychology," *Personnel Psychology*, 46(3): 705–718.

Cano, C.R., F.A. Carrillat, and F. Jaramillo (2004), "A meta-analysis of the relationship between market orientation and business performance," *International Journal of Research in Marketing*, 21(2): 179–200.

Cappelleri, J.C., J.P. Ioannidis, C.H. Schmid, S.D. de Ferranti, M. Aubert, T.C. Chalmers, and J. Lau (1996), "Large trials vs meta-analysis of smaller trials: How do their results compare?" *Journal of the American Medical Association*, 276(16): 1332–1338.

Carver, R.P. (1978), "The case against statistical significance testing," *Harvard Educational Review*, 48(3): 378–399.

Cascio, W.F. and S. Zedeck (1983), "Open a new window in rational research planning: Adjust alpha to maximize statistical power," *Personnel Psychology*, 36(3): 517–526.

Cashen, L.H. and S.W. Geiger (2004), "Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies," *Organizational Research Methods*, 7(2): 151–167.

Chamberlin, T.C. (1897), "The method of multiple working hypotheses," *Journal of Geology*, 5(8): 837–848.

Chan, H.N. and P. Ellis (1998), "Market orientation and business performance: Some evidence from Hong Kong," *International Marketing Review*, 15(2): 119–139.

Chase, L.J. and S.J. Baran (1976), "An assessment of quantitative research in mass communication," *Journalism Quarterly*, 53(2): 308–311.

Chase, L.J. and R.B. Chase (1976), "A statistical power analysis of applied psychological research," *Journal of Applied Psychology*, 61(2): 234–237.

Chase, L.J. and R.K. Tucker (1975), "A power-analytic examination of contemporary communication research," *Speech Monographs*, 42(1): 29–41.

Christensen, J.E. and C.E. Christensen (1977), "Statistical power analysis of health, physical education, and recreation research," *Research Quarterly*, 48(1): 204–208.

Churchill, G.A., N.M. Ford, S.W. Hartley, and O.C. Walker (1985), "The determinants of salesperson performance: A meta-analysis," *Journal of Marketing Research*, 22(2): 103–118.

Clark-Carter, D. (1997), "The account taken of statistical power in research published in the British Journal of Psychology," *British Journal of Psychology*, 88(1): 71–83.

Clark-Carter, D. (2003), "Effect size: The missing piece in the jigsaw," *The Psychologist*, 16(12): 636–638.

Coe, R. (2002), "It's the effect size, stupid: What effect size is and why it is important," Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, 12–14 September, accessed from www.leeds.ac.uk/educol/documents/00002182.htm on 24 January 2008.

Cohen, J. (1962), "The statistical power of abnormal-social psychological research: A review," *Journal of Abnormal and Social Psychology*, 65(3): 145–153.

Cohen, J. (1983), "The cost of dichotomization," *Applied Psychological Measurement*, 7(3): 249–253.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1990), "Things I have learned (so far)," *American Psychologist*, 45(12): 1304–1312.

Cohen, J. (1992), "A power primer," *Psychological Bulletin*, 112(1): 155–159.

Cohen, J. (1994), "The earth is round ($p < .05$)," *American Psychologist*, 49(12): 997–1003.

Cohen, J., P. Cohen, S.G. West, and L.S. Aiken (2003), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 3rd Edition*. Mahwah, NJ: Lawrence Erlbaum.

Cohn, L.D. and B.J. Becker (2003), "How meta-analysis increases statistical power," *Psychological Methods*, 8(3): 243–253.

Colegrave, N. and G.D. Ruxton (2003), "Confidence intervals are a more useful complement to nonsignificant tests than are power calculations," *Behavioral Ecology*, 14(3): 446–450.

Cortina, J.M. (2002), "Big things have small beginnings: An assortment of 'minor' methodological understandings," *Journal of Management*, 28(3): 339–362.

Cortina, J.M. and W.P. Dunlap (1997), "Logic and purpose of significance testing," *Psychological Methods*, 2(2): 161–172.

Coursol, A. and E.E. Wagner (1986), "Effect of positive findings on submission and acceptance rates: A note on meta analysis bias," *Professional Psychology: Research and Practice*, 17(2): 136–137.

Cowles, M. and C. Davis (1982), "On the origins of the .05 level of significance," *American Psychologist*, 37(5): 553–558.

Cumming, G., F. Fidler, M. Leonard, P. Kalinowski, A. Christiansen, A. Kleinig, J. Lo, N. McMenamin, and S. Wilson (2007), "Statistical reform in psychology: Is anything changing?" *Psychological Science*, 18(3): 230–232.

Cumming, G. and S. Finch (2001), "A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions," *Educational and Psychological Measurement*, 61(4): 532–574.

Cumming, G. and S. Finch (2005), "Inference by eye: Confidence intervals and how to read pictures of data," *American Psychologist*, 60(2): 170–180.

Cummings, T.G. (2007), "2006 Presidential address: Quest for an engaged academy," *Academy of Management Review*, 32(2): 355–360.

Daly, J.A. and A. Hexamer (1983), "Statistical power research in English education," *Research in the Teaching of English*, 17(2): 157–164.

Daly, L.E. (2000), "Confidence intervals and sample sizes," in D.G. Altman, D. Machin, T.N. Bryant, and M.J. Gardner (editors), *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. London: British Medical Journal Books, 139–152.

Daniel, F., F.T. Lohrke, C.J. Fornaciari, and R.A. Turner (2004), "Slack resources and firm performance: A meta-analysis," *Journal of Business Research*, 57(6): 565–574.

Dennis, M.L., R.D. Lennox, and M.A. Foss (1997), "Practical power analysis for substance abuse health services research," in K.J. Bryant, M. Windle, and S.G. West (editors), *The Science of Prevention*, Washington, DC: American Psychological Association, 367–404.

Derr, J. and L.J. Goldsmith (2003), "How to report nonsignificant results: Planning to make the best use of statistical power calculations," *Journal of Orthopaedic and Sports Physical Therapy*, 33(6): 303–306.

Di Paula, A. (2000), "Using the binomial effect size display to explain the practical importance of correlations," *Quirk's Marketing Research Review* (Nov): website www.nrgresearchgroup.com/media/documents/BESD_000.pdf, accessed 1 April 2008.

Di Stefano, J. (2003), "How much power is enough? Against the development of an arbitrary convention for statistical power calculations," *Functional Ecology*, 17(5): 707–709.

Dixon, P. (2003), "The *p*-value fallacy and how to avoid it," *Canadian Journal of Experimental Psychology*, 57(3): 189–202.

Duarte, J., S. Siegel, and L.A. Young (2009), "Trust and credit," SSRN working paper: http://ssrn.com/abstract=1343275, accessed 15 March 2009.

Dunlap, W.P. (1994), "Generalizing the common language effect size indicator to bivariate normal correlations," *Psychological Bulletin*, 116(3): 509–511.

Eden, D. (2002), "Replication, meta-analysis, scientific progress, and *AMJ*'s publication policy," *Academy of Management Journal*, 45(5): 841–846.

Efran, M.G. (1974), "The effect of physical appearance on the judgment of guilt, interpersonal attraction, and severity of recommendation in a simulated jury task," *Journal of Research in Personality*, 8(1): 45–54.

Egger, M. and G.D. Smith (1995), "Misleading meta-analysis: Lessons from an 'effective, safe, simple' intervention that wasn't," *British Medical Journal*, 310(25 March): 751–752.

Egger, M., G.D. Smith, M. Schneider, and C. Minder (1997), "Bias in meta-analysis detected by simple graphical test," *British Medical Journal*, 315(7109): 629–634.

Eisenach, J.C. (2007), "Editor's note," *Anesthesiology*, 106(3): 415.

Ellis, P.D. (2005), "Market orientation and marketing practice in a developing economy," *European Journal of Marketing*, 39(5/6): 629–645.

Ellis, P.D. (2006), "Market orientation and performance: A meta-analysis and cross-national comparisons," *Journal of Management Studies*, 43(5): 1089–1107.

Ellis, P.D. (2007), "Distance, dependence and diversity of markets: Effects on market orientation," *Journal of International Business Studies*, 38(3): 374–386.

Ellis, P.D. (2009), "Effect size calculators," website http://myweb.polyu.edu.uk/nmspaul/calculator/calculator.html, accessed 31 December 2009.

Embretson, S.E. (2006), "The continued search for nonarbitrary metrics in psychology," *American Psychologist*, 61(1): 50–55.

Erceg-Hurn, D.M. and V.M. Mirosevich (2008), "Modern robust statistical methods: An easy way to maximize the accuracy and power of your research," *American Psychologist*, 63(7): 591–601.

Erturk, S.M. (2005), "Retrospective power analysis: When?" *Radiology*, 237(2): 743.

ESA (2006), "European Space Agency news," website www.esa.int/esaCP/SEM09F8LURE_index_0.html, accessed 25 April 2008.

Eysenck, H.F. (1978), "An exercise in mega-silliness," *American Psychologist*, 33(5): 517.

Falk, R. and C.W. Greenbaum (1995), "Significance tests die hard: The amazing persistence of a probabilistic misconception," *Theory and Psychology*, 5(1): 75–98.

Fan, X.T. (2001), "Statistical significance and effect size in education research: Two sides of a coin," *Journal of Educational Research*, 94(5): 275–282.

Fan, X.T. and B. Thompson (2001), "Confidence intervals about score reliability coefficients, please: An *EPM* guidelines editorial," *Educational and Psychological Measurement*, 61(4): 517–531.

Faul, F., E. Erdfelder, A.G. Lang, and A. Buchner (2007), "G∗Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior Research Methods*, 39(2): 175–191.

FDA (2008), "Estrogen and estrogen with progestin therapies for postmenopausal women," website www.fda.gov/CDER/Drug/infopage/estrogens_progestins/default.htm, accessed 7 May 2008.

Feinberg, W.E. (1971), "Teaching the Type I and Type II errors: The judicial process," *The American Statistician*, 25(3): 30–32.

Feinstein, A.R. (1995), "Meta-analysis: Statistical alchemy for the 21st century," *Journal of Clinical Epidemiology*, 48(1): 71–79.

Fidler, F., G. Cumming, N. Thomason, D. Pannuzzo, J. Smith, P. Fyffe, H. Edmonds, C. Harrington, and R. Schmitt (2005), "Toward improved statistical reporting in the Journal of Consulting and Clinical Psychology," *Journal of Consulting and Clinical Psychology*, 73(1): 136–143.

Fidler, F., N. Thomason, G. Cumming, S. Finch, and J. Leeman (2004), "Editors can lead researchers to confidence intervals, but can't make them think," *Psychological Science*, 15(2): 119–126.

Field, A.P. (2003a), "Can meta-analysis be trusted?" *The Psychologist*, 16(12): 642–645.

Field, A.P. (2003b), "The problems in using fixed-effects models of meta-analysis on real-world data," *Understanding Statistics*, 2(2): 105–124.

Field, A.P. (2005), "Is the meta-analysis of correlation coefficients accurate when population correlations vary?" *Psychological Methods*, 10(4): 444–467.

Field, A.P. and D.B. Wright (2006), "A bluffer's guide to effect sizes," *PsyPAG Quarterly*, 58(March): 9–23.

Finch, S., G. Cumming, and N. Thomason (2001), "Reporting of statistical inference in the *Journal of Applied Psychology:* Little evidence of reform," *Educational and Psychological Measurement*, 61(2): 181–210.

Fisher, R.A. (1925), *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fleiss, J.L. (1994), "Measures of effect size for categorical data," in H. Cooper and L.V. Hedges (editors), *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 245–260.

Fleiss, J.L., B. Levin, and M.C. Paik (2003), *Statistical Methods for Rates and Proportions, 3rd Edition*. Hoboken, NJ: Wiley-Interscience.

Friedman, H. (1968), "Magnitude of experimental effect and a table for its rapid estimation," *Psychological Bulletin*, 70(4): 245–251.

Friedman, H. (1972), "Trial by jury: Criteria for convictions by jury size and Type I and Type II errors," *The American Statistician*, 26(2): 21–23.

Gardner, M.J. and D.G. Altman (2000), "Estimating with confidence," in D.G. Altman, D. Machin, T.N. Bryant, and M.J. Gardner (editors), *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. London: British Medical Journal Books, 3–5.

Gigerenzer, G. (1998), "We need statistical thinking, not statistical rituals," *Behavioral and Brain Sciences*, 21(2): 199–200.

Gigerenzer, G. (2004), "Mindless statistics," *Journal of Socio-Economics*, 33(5): 587–606.

Glass, G. (1976), "Primary, secondary, and meta-analysis of research," *Educational Researcher*, 5(10): 3–8.

Glass, G.V. (2000), "Meta-analysis at 25," website http://glass.ed.asu.edu/gene/papers/meta25.html, accessed 7 May 2008.

Glass, G.V., B. McGaw, and M.L. Smith (1981), *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage.

Glass, G.V. and M.L. Smith (1978), "Reply to Eysenck," *American Psychologist*, 33(5): 517–518.

Gleser, L.J. and I. Olkin (1996), "Models for estimating the number of unpublished studies," *Statistics in Medicine*, 15(23): 2493–2507.

Gliner, J.A., G.A. Morgan, and R.J. Harmon (2002), "The chi-square test and accompanying effect sizes," *Journal of the American Academy of Child and Adolescent Psychiatry*, 41(12): 1510–1512.

Goodman, S.N. and J.A. Berlin (1994), "The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results," *Annals of Internal Medicine*, 121(3): 200–206.

Gøtzsche, P.C., C. Hammarquist, and M. Burr (1998), "House dust mite control measures in the management of asthma: Meta-analysis," *British Medical Journal*, 317(7166): 1105–1110.

Green, S.B. (1991), "How many subjects does it take to do a regression analysis?" *Multivariate Behavioral Research*, 26(3): 499–510.

Greenland, S. (1994), "Can meta-analysis be salvaged?" *American Journal of Epidemiology*, 140(9): 783–787.

Greenley, G.E. (1995), "Market orientation and company performance: Empirical evidence from UK companies," *British Journal of Management*, 6(1): 1–13.

Grégoire, G., F. Derderian, and J. LeLorier (1995), "Selecting the language of the publications included in a meta-analysis: Is there a Tower of Babel bias?" *Journal of Clinical Epidemiology*, 48(1): 159–163.

Grissom, R.J. (1994), "Probability of the superior outcome of one treatment over another," *Journal of Applied Psychology*, 79(2): 314–316.

Grissom, R.J. and J.J. Kim (2005), *Effect Sizes for Research: A Broad Practical Approach*. Mahwah, NJ: Lawrence Erlbaum.

Haase, R., D.M. Waechter, and G.S. Solomon (1982), "How significant is a significant difference? Average effect size of research in counseling psychology," *Journal of Counseling Psychology*, 29(1): 58–65.

Hadzi-Pavlovic, D. (2007), "Effect sizes II: Differences between proportions," *Acta Neuropsychiatrica*, 19(6): 384–385.

Hair, J.F., R.E. Anderson, R.L. Tatham, and W.C. Black (1998), *Multivariate Data Analysis, 5th Edition*. Upper Saddle River, NJ: Prentice-Hall.

Hall, S.M. and M.T. Brannick (2002), "Comparison of two random-effects methods of meta-analysis," *Journal of Applied Psychology*, 87(2): 377–389.

Halpern, S.D., J.H.T. Karlawish, and J.A. Berlin (2002), "The continuing unethical conduct of underpowered trials," *Journal of the American Medical Association*, 288(3): 358–362.

Hambrick, D.C. (1994), "1993 presidential address: What if the Academy actually mattered?" *Academy of Management Review*, 19(1): 11–16.

Harlow, L.L., S.A. Mulaik, and Steiger, J.H. (editors) (1997), *What if There Were No Significance Tests?* Mahwah, NJ: Lawrence Erlbaum.

Harris, L.C. (2001), "Market orientation and performance: Objective and subjective empirical evidence from UK companies," *The Journal of Management Studies*, 38(1): 17–43.

Harris, M.J. (1991), "Significance tests are not enough: The role of effect-size estimation in theory corroboration," *Theory and Psychology*, 1(3): 375–382.

Harris, R.J. (1985), *A Primer of Multivariate Statistics, 2nd Edition*. Orlando, FL: Academic Press.

Hedges, L.V. (1981), "Distribution theory for Glass's estimator of effect size and related estimators," *Journal of Educational Statistics*, 6(2): 106–128.

Hedges, L.V. (1988), "Comment on 'Selection models and the file drawer problem'," *Statistical Science*, 3(1): 118–120.

Hedges, L.V. (1992), "Meta-analysis," *Journal of Educational Statistics*, 17(4): 279–296.
Hedges, L.V. (2007), "Meta-analysis," in C.R. Rao and S. Sinharay (editors), *Handbook of Statistics, Volume 26*. Amsterdam: Elsevier, 919–953.
Hedges, L.V. and I. Olkin (1980), "Vote-counting methods in research synthesis," *Psychological Bulletin*, 88(2): 359–369.
Hedges, L.V. and I. Olkin (1985), *Statistical Methods for Meta-Analysis*. London: Academic Press.
Hedges, L.V. and T.D. Pigott (2001), "The power of statistical tests in meta-analysis," *Psychological Methods*, 6(3): 203–217.
Hedges, L.V. and J.L. Vevea (1998), "Fixed- and random-effects models in meta-analysis," *Psychological Methods*, 3(4): 486–504.
Hoenig, J.M. and D.M. Heisey (2001), "The abuse of power: The pervasive fallacy of power calculations for data analysis," *The American Statistician*, 55(1): 19–24.
Hollenbeck, J.R., D.S. DeRue, and M. Mannor (2006), "Statistical power and parameter stability when subjects are few and tests are many: Comment on Peterson, Smith, Martorana and Owens (2003)," *Journal of Applied Psychology*, 91(1): 1–5.
Hoppe, D.J. and M. Bhandari (2008), "Evidence-based orthopaedics: A brief history," *Indian Journal of Orthopaedics*, 42(2): 104–110.
Houle, T.T., D.B. Penzien, and C.K. Houle (2005), "Statistical power and sample size estimation for headache research: An overview and power calculation tools," *Headache: The Journal of Head and Face Pain*, 45(5): 414–418.
Hubbard, R. and J.S. Armstrong (1992), "Are null results becoming an endangered species in marketing?" *Marketing Letters*, 3(2): 127–136.
Hubbard, R. and J.S. Armstrong (2006), "Why we don't really know what 'statistical significance' means: A major educational failure," *Journal of Marketing Education*, 28(2): 114–120.
Huberty, C.J. (2002), "A history of effect size indices," *Educational and Psychological Measurement*, 62(2): 227–240.
Hunt, M. (1997), *How Science Takes Stock: The Story of Meta-Analysis*. New York: Russell Sage Foundation.
Hunter, J.E. (1997), "Needed: A ban on the significance test," *Psychological Science*, 8(1): 3–7.
Hunter, J.E. and F.L. Schmidt (1990), *Methods of Meta-Analysis*. Newbury Park, CA: Sage.
Hunter, J.E. and F.L. Schmidt (2000), "Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge," *International Journal of Selection and Assessment*, 8(4): 275–292.
Hunter, J.E. and F.L. Schmidt (2004), *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings, 2nd Edition*. Thousand Oaks, CA: Sage.
Hyde, J.S. (2001), "Reporting effect sizes: The role of editors, textbook authors, and publication manuals," *Educational and Psychological Measurement*, 61(2): 225–228.
Iacobucci, D. (2005), "From the editor," *Journal of Consumer Research*, 32(1): 1–6.
Ioannidis, J.P.A. (2005), "Why most published research findings are false," *PLoS Med*, website http://medicine.plosjournals.org/ 2(8): e124, 696–701, accessed 1 April 2007.
Ioannidis, J.P.A. (2008), "Why most discovered true associations are inflated," *Epidemiology*, 19(5): 640–648.
Iyengar, S. and J.B. Greenhouse (1988), "Selection models and the file drawer problem," *Statistical Science*, 3(1): 109–135.
Jaworski, B.J. and A.K. Kohli (1993), "Market orientation: Antecedents and consequences," *Journal of Marketing*, 57(3): 53–70.
JEP (2003), "Instructions to authors," *Journal of Educational Psychology*, 95(1): 201.
Johnson, D.H. (1999), "The insignificance of statistical significance testing," *Journal of Wildlife Management*, 63(3): 763–772.

Johnson, B.T., B. Mullen, and E. Salas (1995), "Comparisons of three meta-analytic approaches," *Journal of Applied Psychology*, 80(1): 94–106.

Jones, B.J. and J.K. Brewer (1972), "An analysis of the power of statistical tests reported in the Research Quarterly," *Research Quarterly*, 43(1): 23–30.

Katzer, J. and J. Sodt (1973), "An analysis of the use of statistical testing in communication research," *Journal of Communication*, 23(3): 251–265.

Kazdin, A. (1999), "The meanings and measurements of clinical significance," *Journal of Consulting and Clinical Psychology*, 67(3): 332–339.

Kazdin, A.E. (2006), "Arbitrary metrics: Implications for identifying evidence-based treatments," *American Psychologist*, 61(1): 42–49.

Keller, G. (2005), *Statistics for Management and Economics*. Belmont, CA: Thomson.

Kelley, K. and S.E. Maxwell (2008), "Sample size planning with applications to multiple regression: Power and accuracy for omnibus and targeted effects," in P. Alasuutari, L. Bickman, and J. Brannen (editors), *The Sage Handbook of Social Research Methods*. London: Sage, 166–192.

Kendall, P.C. (1997), "Editorial," *Journal of Consulting and Clinical Psychology*, 65(1): 3–5.

Keppel, G. (1982), *Design and Analysis: A Researcher's Handbook, 2nd Edition*. Englewood Cliffs, NJ: Prentice-Hall.

Kerr, N.L. (1998), "HARKing: Hypothesizing after the results are known," *Personality and Social Psychology Review*, 2(3): 196–217.

Keselman, H.J., J. Algina, L.M. Lix, R.R. Wilcox, and K.N. Deering (2008), "A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes," *Psychological Methods*, 13(2): 110–129.

Kieffer, K.M., R.J. Reese, and B. Thompson (2001), "Statistical techniques employed in AERJ and JCP articles from 1988 to 1997: A methodological review," *Journal of Experimental Education*, 69(3): 280–309.

Kirca, A.H., S. Jayachandran, and W.O. Bearden (2005), "Market orientation: A meta-analytic review and assessment of its antecedents and impact on performance," *Journal of Marketing*, 69(2): 24–41.

Kirk, R.E. (1996), "Practical significance: A concept whose time has come," *Educational and Psychological Measurement*, 56(5): 746–759.

Kirk, R.E. (2001), "Promoting good statistical practices: Some suggestions," *Educational and Psychological Measurement*, 61(2): 213–218.

Kirk, R.E. (2003), "The importance of effect magnitude," in S.F. Davis (editor), *Handbook of Research Methods in Experimental Psychology*. Oxford, UK: Blackwell, 83–105.

Kline, R.B. (2004), *Beyond Significance Testing: Reforming Data Analysis Methods in Behavioral Research*. Washington DC: American Psychological Association.

Kohli, A.J., B.J. Kaworski, and A. Kumar (1993), "MARKOR: A measure of market orientation," *Journal of Marketing Research*, 30(4): 467–477.

Kolata, G.B. (1981), "Drug found to help heart attack survivors," *Science*, 214(13): 774–775.

Kolata, G.B. (2002), "Hormone replacement study a shock to the medical system," *New York Times on the Web*, website www.nytimes.com/2002/07/10health/10/HORM.html, accessed 1 May 2008.

Kosciulek, J.F. and E.M. Szymanski (1993), "Statistical power analysis of rehabilitation research," *Rehabilitation Counseling Bulletin*, 36(4): 212–219.

Kraemer, H.C. and S. Thiemann (1987), *How Many Subjects? Statistical Power Analysis in Research*. Newbury Park, CA: Sage.

Kraemer, H.C., J. Yesavage, and J.O. Brooks (1998), "The advantages of excluding under-powered studies in meta-analysis: Inclusionist vs exclusionist viewpoints," *Psychological Methods*, 3(1): 23–31.

Kroll, R.M. and L.J. Chase (1975), "Communication disorders: A power analytic assessment of recent research," *Journal of Communication Disorders*, 8(3): 237–247.

La Greca, A.M. (2005), "Editorial," *Journal of Consulting and Clinical Psychology*, 73(1): 3–5.

Lane, D. (2008), "Fisher r-to-z calculator," website http://onlinestatbook.com/calculators/fisher_z.html, accessed 27 November 2008.

Lang, J.M., K.J. Rothman, and C.I. Cann (1998), "That confounded *p*-value," *Epidemiology*, 9(1): 7–8.

LeCroy, C.W. and Krysik, J. (2007), "Understanding and interpreting effect size measures," *Journal of Social Work Research*, 31(4): 243–248.

LeLorier, J., G. Grégoire, A. Benhaddad, J. Lapierre, and F. Derderian (1997), "Discrepancies between meta-analyses and subsequent large scale randomized, controlled trials," *New England Journal of Medicine*, 337(21 Aug): 536–618.

Lenth, R.V. (2001), "Some practical guidelines for effective sample size determination," *The American Statistician*, 55(3): 187–193.

Levant, R.F. (1992), "Editorial," *Journal of Family Psychology*, 6(1): 3–9.

Levine, M. and M. Ensom (2001), "Post hoc analysis: An idea whose time has passed?" *Pharmacotherapy*, 21(4): 405–409.

Light, R.J. and P.V. Smith (1971), "Accumulating evidence: Procedures for resolving contradictions among different research studies," *Harvard Educational Review*, 41(4): 429–471.

Lilford, R. and A.J. Stevens (2002), "Underpowered studies," *British Journal of Sociology*, 89(2): 129–131.

Lindsay, R.M. (1993), "Incorporating statistical power into the test of significance procedure: A methodological and empirical inquiry," *Behavioral Research in Accounting*, 5: 211–236.

Lipsey, M.W. (1990), *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, CA: Sage.

Lipsey, M.W. (1998), "Design sensitivity: Statistical power for applied experimental research," in L. Bickman and D.J. Rog (editors), *Handbook of Applied Social Research Methods*. Thousand Oaks, CA: Sage, 39–68.

Lipsey, M.W. and D.B. Wilson (1993), "The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis," *American Psychologist*, 48(12): 1181–1209.

Lipsey, M.W. and D.B. Wilson (2001), *Practical Meta-Analysis*. Thousand Oaks, CA: Sage.

Livingston, E.H. and L. Cassidy (2005), "Statistical power and estimation of the number of required subjects for a study based on the *t*-test: A surgeon's primer," *Journal of Surgical Research*, 128(2): 207–217.

Lowry, R. (2008a), "Fisher r-to-z calculator," website http://faculty.vassar.edu/lowry/tabs.html#fisher, accessed 27 November 2008.

Lowry, R. (2008b), "z-to-P calculator," website http://faculty.vassar.edu/lowry/tabs.html#z, accessed 27 November 2008.

Lustig, D. and D. Strauser (2004), "Editor's comment: Effect size and rehabilitation research," *Journal of Rehabilitation*, 70(4): 3–5.

Machin, D., M. Campbell, P. Fayers, and A. Pinol (1997), *Sample Size Tables for Clinical Studies, 2nd Edition*. Oxford, UK: Blackwell.

Maddock, J.E. and J.S. Rossi (2001), "Statistical power of articles published in three health psychology-related journals," *Health Psychology*, 20(1): 76–78.

Malhotra, N.K. (1996), *Marketing Research: An Applied Orientation, 2nd Edition*. Upper Saddle River, NJ: Prentice-Hall.

Masson, M.E.J. and G.R. Loftus (2003), "Using confidence intervals for graphically based data interpretation," *Canadian Journal of Experimental Psychology*, 57(3): 203–220.

Maxwell, S.E. (2004), "The persistence of unpowered studies in psychological research: Causes, consequences, and remedies," *Psychological Methods*, 9(2): 147–163.

Maxwell, S.E., K. Kelley, and J.R. Rausch (2008), "Sample size planning for statistical power and accuracy in parameter estimation," *Annual Review of Psychology*, 59: 537–563.

Mazen, A.M., L.A. Graf, C.E. Kellogg, and M. Hemmasi (1987a), "Statistical power in contemporary management research," *Academy of Management Journal*, 30(2): 369–380.

Mazen, A.M., M. Hemmasi, and M.F. Lewis (1987b), "Assessment of statistical power in contemporary strategy research," *Strategic Management Journal*, 8(4): 403–410.

McCartney, K. and R. Rosenthal (2000), "Effect size, practical importance and social policy for children," *Child Development*, 71(1): 173–180.

McClave, J.T. and T. Sincich (2009), *Statistics, 11th Edition*. Upper Saddle River, NJ: Prentice-Hall.

McCloskey, D. (2002), *The Secret Sins of Economics*. Chicago, IL: Prickly Paradigm Press, website www.prickly-paradigm.com/paradigm4.pdf.

McCloskey, D.N. and S.T. Ziliak (1996), "The standard error of regressions," *Journal of Economic Literature*, 34(March): 97–114.

McGrath, R.E. and G.J. Meyer (2006), "When effect sizes disagree: The case of *r* and *d*," *Psychological Methods*, 11(4): 386–401.

McGraw, K.O. and S.P. Wong (1992), "A common language effect size statistic," *Psychological Bulletin*, 111(2): 361–365.

McSwain, D.N. (2004), "Assessment of statistical power in contemporary accounting information systems research," *Journal of Accounting and Finance Research*, 12(7): 100–108.

Meehl, P.E. (1967), "Theory testing in psychology and physics: A methodological paradox," *Philosophy of Science*, 34(June): 103–115.

Meehl, P.E. (1978), "Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology," *Journal of Consulting and Clinical Psychology*, 46(4): 806–834.

Megicks, P. and G. Warnaby (2008), "Market orientation and performance in small independent retailers in the UK," *International Review of Retail, Distribution and Consumer Research*, 18(1): 105–119.

Melton, A. (1962), "Editorial," *Journal of Experimental Psychology*, 64(6): 553–557.

Mendoza, J.L. and K.L. Stafford (2001), "Confidence intervals, power calculation, and sample size estimation for the squared multiple correlation coefficient under the fixed and random regression models: A computer program and useful standard tables," *Educational and Psychological Measurement*, 61(4): 650–667.

Miles, J.M. (2003), "A framework for power analysis using a structural equation modelling procedure," *BMC Medical Research Methodology*, 3(27), website www.biomedcentral.com/1471–2288/3/27, accessed 1 April 2008.

Miles, J.M. and M. Shevlin (2001), *Applying Regression and Correlation*. London: Sage.

Moher, D., K.F. Schulz, and D.G. Altman (2001), "The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials," *Lancet*, 357(9263): 1191–1194.

Mone, M.A., G.C. Mueller, and W. Mauland (1996), "The perceptions and usage of statistical power in applied psychology and management research," *Personnel Psychology*, 49(1): 103–120.

Muncer, S.J. (1999), "Power dressing is important in meta-analysis," *British Medical Journal*, 318(27 March): 871.

Muncer, S.J., M. Craigie, and J. Holmes (2003), "Meta-analysis and power: Some suggestions for the use of power in research synthesis," *Understanding Statistics*, 2(1): 1–12.

Muncer, S.J., S. Taylor, and M. Craigie (2002), "Power dressing and meta-analysis: Incorporating power analysis into meta-analysis," *Journal of Advanced Nursing*, 38(3): 274–280.

Murphy, K.R. (1997), "Editorial," *Journal of Applied Psychology*, 82(1): 3–5.

Murphy, K.R. (2002), "Using power analysis to evaluate and improve research," in S.G. Rogelberg (editor), *Handbook of Research Methods in Industrial and Organizational Psychology*. Oxford, UK: Blackwell, 119–137.

Murphy, K.R. and B. Myors (2004), *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests, 2nd Edition*. Mahwah, NJ: Lawrence Erlbaum.

Nakagawa, S. and T.M. Foster (2004), "The case against retrospective statistical power analyses with an introduction to power analysis," *Acta Ethologica*, 7(2): 103–108.

Narver, J.C. and S.F. Slater (1990), "The effect of a market orientation on business profitability," *Journal of Marketing*, 54(4): 20–35.

Neeley, J.H. (1995), "Editorial," *Journal of Experimental Psychology: Learning, Memory and Cognition*, 21(1): 261.

NEO (2008), "NASA statement on student asteroid calculations," Near-Earth Object Program, website http://neo.jpl.nasa.gov/news/news158.html, accessed 17 April 2008.

Newcombe, R.G. (2006), "A deficiency of the odds ratio as a measure of effect size," *Statistics in Medicine*, 25(24): 4235–4240.

Nickerson, R.S. (2000), "Null hypothesis significance testing: A review of an old and continuing controversy," *Psychological Methods*, 5(2): 241–301.

Norton, B.J. and M.J. Strube (2001), "Understanding statistical power," *Journal of Orthopaedic and Sports Physical Therapy*, 31(6): 307–315.

Nunnally, J.C. (1978), *Psychometric Theory, 2nd Edition*. New York: McGraw-Hill.

Nunnally, J.C. and I.H. Bernstein (1994), *Psychometric Theory, 3rd Edition*. New York: McGraw-Hill.

Olejnik, S. and J. Algina (2000), "Measures of effect size for comparative studies: Applications, interpretations, and limitations," *Contemporary Educational Psychology*, 25(3): 241–286.

Olkin, I. (1995), "Statistical and theoretical considerations in meta-analysis," *Journal of Clinical Epidemiology*, 48(1): 133–146.

Onwuegbuzie, A.J. and N.L. Leech (2004), "Post hoc power: A concept whose time has come," *Understanding Statistics*, 3(4): 201–230.

Orme, J.G. and T.D. Combs-Orme (1986), "Statistical power and Type II errors in social work research," *Social Work Research and Abstracts*, 22(3): 3–10.

Orwin, R.G. (1983), "A fail-safe *N* for effect size in meta-analysis," *Journal of Educational Statistics*, 8(2): 157–159.

Orwin, R.G. (1994), "Evaluating coding decisions," in H. Cooper and L.V. Hedges (editors), *Handbook of Research Synthesis*. New York: Russell Sage Foundation, 139–162.

Osborne, J.W. (2008a), "Bringing balance and technical accuracy to reporting odds ratios and the results of logistic regression analyses," in J.W. Osborne (editor), *Best Practices in Quantitative Methods*. Thousand Oaks, CA: Sage, 385–389.

Osborne, J.W. (2008b), "Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices," *Educational Psychology*, 28(2): 151–160.

Overall, J.E. and S.N. Dalal (1965), "Design of experiments to maximize power relative to cost," *Psychological Bulletin*, 64(Nov): 339–350.

Pampel, F.C. (2000), *Logistic Regression: A Primer*. Thousand Oaks, CA: Sage.

Parker, R.I. and S. Hagan-Burke (2007), "Useful effect size interpretations for single case research," *Behavior Therapy*, 38(1): 95–105.

Parks, J.B., P.A. Shewokis, and C.A. Costa (1999), "Using statistical power analysis in sport management research," *Journal of Sport Management*, 13(2): 139–147.

Pearson, K. (1905), "Report on certain enteric fever inoculation statistics," *British Medical Journal*, 2(2288): 1243–1246.

Pelham, A. (2000), "Market orientation and other potential influences on performance in small and medium-sized manufacturing firms," *Journal of Small Business Management*, 38(1): 48–67.

Perrin, B. (2000), "Donald T. Campbell and the art of practical 'in-the-trenches' program evaluation," in L. Bickman (editor), *Validity and Social Experimentation: Donald Campbell's Legacy, Volume 1*. Thousand Oaks, CA: Sage, 267–282.

Peterson, R.S., D.B. Smith, P.V. Martorana, and P.D. Owens (2003), "The impact of chief executive officer personality on top management team dynamics: One mechanism by which leadership affects organizational performance," *Journal of Applied Psychology*, 88(5): 795–808.

Phillips, D.W. (2007), "The Titanic numbers game," website www.titanicsociety.com/readables/main/articles_04–20-1998_titanic_numbers_game.asp, accessed 3 September 2008.

Platt, J.R. (1964), "Strong inference," *Science*, 146(3642): 347–353.

Popper, K. (1959), *The Logic of Scientific Discovery*. New York: Harper and Row.

Prentice, D.A. and D.T. Miller (1992), "When small effects are impressive," *Psychological Bulletin*, 112(1): 160–164.

Randolph, J.J. and R.S. Edmondson (2005), "Using the binomial effect size display (BESD) to present the magnitude of effect sizes to the evaluation audience," *Practical Assessment, Research and Evaluation*, 10(14), electronic journal: http://pareonline.net/pdf/v10n14.pdf, accessed 17 April 2008.

Roberts, J.K. and R.K. Henson (2002), "Correction for bias in estimating effect sizes," *Educational and Psychological Measurement*, 62(2): 241–253.

Roberts, R.M. (1989), *Serendipity: Accidental Discoveries in Science*. New York: John Wiley and Sons.

Rodgers, J.L. and W.A. Nicewander (1988), "Thirteen ways to look at the correlation coefficient," *The American Statistician*, 42(1): 59–66.

Rosenthal, J.A. (1996), "Qualitative descriptors of strength of association and effect size," *Journal of Social Service Research*, 21(4): 37–59.

Rosenthal, M.C. (1994), "The fugitive literature," in H. Cooper and L.V. Hedges (editors), *Handbook of Research Synthesis*. New York: Russell Sage Foundation, 85–94.

Rosenthal, R. (1979), "The 'file drawer problem' and the tolerance for null results," *Psychological Bulletin*, 86(3): 638–641.

Rosenthal, R. (1990), "How are we doing in soft psychology?" *American Psychologist*, 45(6): 775–777.

Rosenthal, R. (1991), *Meta-Analytic Procedures for Social Research*. Newbury Park, CA: Sage.

Rosenthal, R. and M.R. DiMatteo (2001), "Meta-analysis: Recent developments in quantitative methods for literature reviews," *Annual Review of Psychology*, 52(1): 59–82.

Rosenthal, R. and D.R. Rubin (1982), "A simple, general purpose display of magnitude of experimental effect," *Journal of Educational Psychology*, 74(2): 166–169.

Rosenthal, R., R.L. Rosnow, and D.B. Rubin (2000), *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*. Cambridge, UK: Cambridge University Press.

Rosnow, R.L. and R. Rosenthal (1989), "Statistical procedures and the justification of knowledge in psychological science," *American Psychologist*, 44(10): 1276–1284.

Rosnow, R.L. and R. Rosenthal (2003), "Effect sizes for experimenting psychologists," *Canadian Journal of Experimental Psychology*, 57(3): 221–237.

Rossi, J.S. (1985), "Tables of effect size for z score tests of differences between proportions and between correlation coefficients," *Educational and Psychological Measurement*, 45(4): 737–745.

Rossi, J.S. (1990), "Statistical power of psychological research: What have we gained in 20 years?" *Journal of Consulting and Clinical Psychology*, 58(5): 646–656.

Rothman, K.J. (1986), "Significance testing," *Annals of Internal Medicine*, 105(3): 445–447.

Rothman, K.J. (1990), "No adjustments are needed for multiple comparisons," *Epidemiology*, 1(1): 43–46.

Rothman, K.J. (1998), "Writing for Epidemiology," *Epidemiology*, 9(3): 333–337.

Rouder, J.N. and R.D. Morey (2005), "Relational and arelational confidence intervals," *Psychological Science*, 16(1): 77–79.

Rynes, S.L. (2007), "Editor's afterword: Let's create a tipping point – what academics and practitioners can do, alone and together," *Academy of Management Journal*, 50(5): 1046–1054.

Sauerland, S. and C.M. Seiler (2005), "Role of systematic reviews and meta-analysis in evidence-based medicine," *World Journal of Surgery*, 29(5): 582–587.

Sawyer, A.G. and A.D. Ball (1981), "Statistical power and effect size in marketing research," *Journal of Marketing Research*, 18(3): 275–290.

Sawyer, A.G. and J.P. Peter (1983), "The significance of statistical significance tests in marketing research," *Journal of Marketing Research*, 20(2): 122–133.

Scarr, S. (1997), "Rules of evidence: A larger context for the statistical debate," *Psychological Science*, 8(1): 16–17.

Schmidt, F.L. (1992), "What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology," *American Psychologist*, 47(10): 1173–1181.

Schmidt, F.L. (1996), "Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers," *Psychological Methods*, 1(2): 115–129.

Schmidt, F.L. and J.E. Hunter (1977), "Development of a general solution to the problem of validity generalization," *Journal of Applied Psychology*, 62(5): 529–540.

Schmidt, F.L. and J.E. Hunter (1996), "Measurement error in psychological research: Lessons from 26 research scenarios," *Psychological Methods*, 1(2): 199–223.

Schmidt, F.L. and J.E. Hunter (1997), "Eight common but false objections to the discontinuation of significance testing in the analysis of research data," in L.L. Harlow, S.A. Mulaik, and J.H. Steiger (editors), *What if There Were No Significance Tests?*. Mahwah, NJ: Lawrence Erlbaum, 37–64.

Schmidt, F.L. and J.E. Hunter (1999a), "Comparison of three meta-analysis methods revisited: An analysis of Johnson, Mullen and Salas (1995)," *Journal of Applied Psychology*, 84(1): 144–148.

Schmidt, F.L. and J.E. Hunter (1999b), "Theory testing and measurement error," *Intelligence*, 27(3): 183–198.

Schmidt, F.L., I.S. Oh, and T.L. Hayes (2009), "Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results," *British Journal of Mathematical and Statistical Psychology*, 62(1): 97–128.

Schulze, R. (2004), *Meta-Analysis: A Comparison of Approaches*. Cambridge, MA: Hogrefe and Huber.

Schwab, A. and W.H. Starbuck (2009), "Null-hypothesis significance tests in behavioral and management research: We can do better," in D. Bergh and D. Ketchen (editors), *Research Methodology in Strategy and Management, Volume 5*. Emerald, 29–54.

Sechrest, L., P. McKnight, and K. McKnight (1996), "Calibration of measures for psychotherapy outcome studies," *American Psychologist*, 51(10): 1065–1071.

Sedlmeier, P. and G. Gigerenzer (1989), "Do studies of statistical power have an effect on the power of studies?" *Psychological Bulletin*, 105(2): 309–316.

Seth, A., K.D. Carlson, D.E. Hatfield, and H.W. Lan (2009), "So what? Beyond statistical significance to substantive significance in strategy research," in D.D. Bergh and D.J. Ketchen (editors), *Research Methodology in Strategy and Management*, Volume 5. Emerald, 3–27.

Shapiro, S. (1994), "Meta-analysis/shmeta-analysis," *American Journal of Epidemiology*, 140(9): 771–778.

Shaughnessy, J.J., E.B. Zechmeister, and J.S. Zechmeister (2009), *Research Methods in Psychology, 8th Edition*. New York: McGraw-Hill.

Shaver, J.M. (2006), "Interpreting empirical findings," *Journal of International Business Studies*, 37(4): 451–452.

Shaver, J.M. (2007), "Interpreting empirical results in strategy and management research," in D. Ketchen and D. Bergh (editors), *Research Methodology in Strategy and Management, Volume 4*. Elsevier 273–293.

Shaver, J.M. (2008), "Organizational significance," *Strategic Organization*, 6(2): 185–193.
Shaver, J.P. (1993), "What statistical significance testing is, and what it is not," *Journal of Experimental Education*, 61(4): 293–316.
Shoham, A., G.M. Rose, and F. Kropp (2005), "Market orientation and performance: A meta-analysis," *Marketing Intelligence & Planning*, 23(5): 435–454.
Sigall, H. and N. Ostrove (1975), "Beautiful but dangerous: Effects of offender attractiveness and nature of the crime on juridic judgment," *Journal of Personality and Social Psychology*, 31(3): 410–414.
Silverman, I., J. Choi, A. Mackewn, M. Fisher, J. Moro, and E. Olshansky (2000), "Evolved mechanisms underlying wayfinding: Further studies on the hunter-gatherer theory of spatial sex differences," *Evolution and Human Behavior*, 21(3): 210–213.
Simon, S. (2001), "Odds ratio versus relative risk," website www.childrensmercy.org/stats/journal/oddsratio.asp, accessed 17 April 2008.
Sink, C.A. and H.R. Stroh (2006), "Practical significance: The use of effect sizes in school counseling research," *Professional School Counseling*, 9(5): 401–411.
Slater, S.F. and J.C. Narver (2000), "The positive effect of a market orientation on business profitability: A balanced replication," *Journal of Business Research*, 48(1): 69–73.
Smith, M.L. and G.V. Glass (1977), "Meta-analysis of psychotherapy outcome studies," *American Psychologist*, 32(9): 752–760.
Smithson, M. (2001), "Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals," *Educational and Psychological Measurement*, 61(4): 605–632.
Smithson, M. (2003), *Confidence Intervals*. Thousand Oaks, CA: Sage.
Snyder, P. and S. Lawson (1993), "Evaluating results using corrected and uncorrected effect size estimates," *Journal of Experimental Education*, 61(4): 334–349.
Steering Committee of the Physicians' Health Study Research Group (1988), "Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study Research Group," *New England Journal of Medicine*, 318(4): 262–264.
Steiger, J.H. (2004), "Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis," *Psychological Methods*, 9(2): 164–182.
Sterling, T.D. (1959), "Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa," *Journal of the American Statistical Association*, 54(285): 30–34.
Sterne, J.A.C., B.J. Becker, and M. Egger (2005), "The funnel plot," in H.R. Rothstein, A.J. Sutton, and M. Borenstein (editors), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Chichester, UK: John Wiley and Sons, 75–98.
Sterne, J.A.C. and M. Egger (2005), "Regression methods to detect publication and other bias in meta-analysis," in H.R. Rothstein, A.J. Sutton, and M. Borenstein (editors), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. Chichester, UK: John Wiley and Sons, 99–110.
Sterne, J.A.C., M. Egger, and G.D. Smith (2001), "Investigating and dealing with publication and other biases," in M. Egger, G.D. Smith, and D.G. Altman (editors), *Systematic Reviews in Health Care: Meta-Analysis in Context*. London: BMJ, 189–208.
Stock, W.A. (1994), "Systematic coding for research synthesis," in H. Cooper and L.V. Hedges (editors), *Handbook of Research Synthesis*. New York: Russell Sage Foundation, 125–138.
Strube, M.J. (1988), "Averaging correlation coefficients: Influence of heterogeneity and set size," *Journal of Applied Psychology*, 73(3): 559–568.
Sudnow, D. (1967), "Dead on arrival," *Transaction*, 5(Nov): 36–43.
Sullivan, M. (2007), *Statistics: Informed Decisions Using Data*. Upper Saddle River, NJ: Prentice-Hall.

Sutcliffe, J.P. (1980), "On the relationship of reliability to statistical power," *Psychological Bulletin*, 88(2): 509–515.

Teo, K.T., S. Yusuf, R. Collins, P.H. Held, and R. Peto (1991), "Effects of intravenous magnesium in suspected acute myocardial infarction: Overview of randomized trials," *British Medical Journal*, 303(14 Dec): 1499–1503.

Thalheimer, W. and S. Cook (2002), "How to calculate effect sizes from published research articles: A simplified methodology," website http://work-learning.com/effect_sizes.htm, accessed 23 January 2008.

Thomas, L. (1997), "Retrospective power analysis," *Conservation Biology*, 11(1): 276–280.

Thompson, B. (1999a), "If statistical significance tests are broken/misused, what practices should supplement or replace them?" *Theory and Psychology*, 9(2): 165–181.

Thompson, B. (1999b), "Journal editorial policies regarding statistical significance tests: Heat is to fire as *p* is to importance," *Educational Psychology Review*, 11(2): 157–169.

Thompson, B. (1999c), "Why 'encouraging' effect size reporting is not working: The etiology of researcher resistance to changing practices," *Journal of Psychology*, 133(2): 133–140.

Thompson, B. (2002a), "'Statistical,' 'practical,' and 'clinical': How many kinds of significance do counselors need to consider?" *Journal of Counseling and Development*, 80(1): 64–71.

Thompson, B. (2002b), "What future quantitative social science research could look like: Confidence intervals for effect sizes," *Educational Researcher*, 31(3): 25–32.

Thompson, B. (2007a), "Effect sizes, confidence intervals, and confidence intervals for effect sizes," *Psychology in the Schools*, 44(5): 423–432.

Thompson, B. (2007b), "Personal website," www.coe.tamu.edu/~bthompson/, accessed 4 September 2008.

Thompson, B. (2008), "Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes," in J.W. Osborne (editor), *Best Practices in Quantitative Methods*. Thousand Oaks, CA: Sage, 246–262.

Todorov, A., A.N. Mandisodza, A. Goren, and C.C. Hall (2005), "Inferences of competence from faces predict election outcomes," *Science*, 308(10 June): 1623–1626.

Tryon, W.W. (2001), "Evaluating statistical difference, equivalence and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests," *Psychological Methods*, 6(4): 371–386.

Tversky, A. and D. Kahneman (1971), "Belief in the law of small numbers," *Psychological Bulletin*, 76(2): 105–110.

Uitenbroek, D. (2008), "T test calculator," website www.quantitativeskills.com/sisa/statistics/t-test.htm, accessed 27 November 2008.

Urschel, J.D. (2005), "How to analyze an article," *World Journal of Surgery*, 29(5): 557–560.

Vacha-Haase, T. (2001), "Statistical significance should not be considered one of life's guarantees: Effect sizes are needed," *Educational and Psychological Measurement*, 61(2): 219–224.

Vacha-Haase, T., J.E. Nilsson, D.R. Reetz, T.S. Lance, and B. Thompson (2000), "Reporting practices and APA editorial policies regarding statistical significance and effect size," *Theory and Psychology*, 10(3): 413–425.

Vacha-Haase, T. and B. Thompson (2004), "How to estimate and interpret various effect sizes," *Journal of Counseling Psychology*, 51(4): 473–481.

Van Belle, G. (2002), *Statistical Rules of Thumb*. New York: John Wiley and Sons.

Vaughn, R.D. (2007), "The importance of meaning," *American Journal of Public Health*, 97(4): 592–593.

Villar, J. and C. Carroli (1995), "Predictive ability of meta-analyses of randomized controlled trials," *Lancet*, 345(8952): 772–776.

Volker, M.A. (2006), "Reporting effect size estimates in school psychology research," *Psychology in the Schools*, 43(6): 653–672.

Wang, X. and Z. Yang (2008), "A meta-analysis of effect sizes in international marketing experiments," *International Marketing Review*, 25(3): 276–291.

Webb, E.T., D.T. Campbell, R.D. Schwartz, L. Sechrest, and J.B. Grove (1981), *Nonreactive Measures in the Social Sciences, 2nd Edition*. Boston, MA: Houghton Mifflin.

Whitener, E.M. (1990), "Confusion of confidence intervals and credibility intervals in meta-analysis," *Journal of Applied Psychology*, 75(3): 315–321.

Wilcox, R.R. (2005), *Introduction to Robust Estimation and Hypothesis Testing, 2nd Edition*. Amsterdam: Elsevier.

Wilkinson, L. and the Taskforce on Statistical Inference (1999), "Statistical methods in psychology journals: Guidelines and expectations," *American Psychologist*, 54(8): 594–604.

Wright, M. and J.S. Armstrong (2008), "Verification of citations: Fawlty towers of knowledge?" *Interfaces*, 38(2): 125–139.

Yeaton, W. and L. Sechrest (1981), "Meaningful measures of effect," *Journal of Consulting and Clinical Psychology*, 49(5): 766–767.

Yin, R.K. (1984), *Case Study Research*. Beverly Hills, CA: Sage.

Yin, R.K. (2000), "Rival explanations as an alternative to reforms as 'experiments'," in L. Bickman (editor), *Validity and Social Experimentation: Donald Campbell's Legacy, Volume 1*. Thousand Oaks, CA: Sage, 239–266.

Young, N.S., J.P. Ioannidis, and O. Al-Ubaydli (2008), "Why current publication practices may distort science," *PLoS Medicine*, website http://medicine.plosjournals.org/, 5(10): e201: 1–5.

Yusuf, S. and M. Flather (1995), "Magnesium in acute myocardial infarction: ISIS 4 provides no grounds for its routine use," *British Medical Journal*, 310(25 March): 751–752.

Ziliak, S.T. and D.N. McCloskey (2004), "Size matters: The standard error of regressions in the American Economic Review," *Journal of Socio-Economics*, 33(5): 527–546.

Ziliak, S.T. and D.N. McCloskey (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor, MI: University of Michigan Press.

Zodpey, S.P. (2004), "Sample size and power analysis in medical research," *Indian Journal of Dermatology*, 70(2): 123–128.

Zumbo, B.D. and A.M. Hubley (1998), "A note on misconceptions concerning prospective and retrospective power," *The Statistician*, 47(Part2): 385–388.

Index