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

Cambridge University Press
978-0-521-14246-5 - The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results
Paul D. Ellis
Frontmatter
More information

This book is dedicated to Anthony (Tony) Pecotich

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