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