Rosenthal, R. & Rosnow, R.L. (2008). Essentials of Behavioral Research 3rd ed. NY. McGraw-Hill

CHAPTER 12

STATISTICAL POWER AND EFFECT SIZE REVISITED

WHY ASSESS STATISTICAL POWER?

The dual purpose of this chapter is to illustrate how a power analysis is done and to delve further into the concept of an effect size indicator. In chapter 2 we first mentioned that the power of a significance test (defined as $1 - \beta$) is the probability that the test will reject a null hypothesis (H_0) that is false and *should,* therefore, be rejected. We also introduced the conceptual relationship stating that significance test = size of effect × size of study (discussed again in the previous chapter, Equation 11.10). The implication is that the larger the observed magnitude of effect, or the greater the total number of observations, the larger will be the value of the significance test, and therefore the smaller its associated p value. There are a number of strategies for improving statistical power, but the one that is probably most familiar to behavioral researchers is to increase the sample size. In this chapter we provide simple tables for estimating the number of sampling units needed to detect a particular magnitude of effect at a given p level and a specified level of statistical power. What if the effect size is so small that it may be unreasonable to think that an adequate number of subjects will be available? One possibility is to replicate the study and, assuming the results are similar, to estimate the overall (combined) p of the original study and the replication after pooling the results meta-analytically (illustrated at the end of this chapter).

However, you may be wo
is needed, as we also stated t
researchers speak of "the result
their tests of statistical significa
about the magnitude of the rese
reliability (e.g., a confidence int
true that "the initial emphasis c
now given way to a more soph
also true that "there are numero
(Steiger, 2004, pp. 178–179). F
power analysis be included in a g
will not be wasted in a study tha
chance of statistically detecting
regarding the use and interpreta
et al., 1999), have not been fully
(Fidler, Thomason, Cumming, Fin
consequence is that researchers s
eses when they mistakenly inte
detect an existing effect as indic

Some years ago, Jacob Cc
for the Behavioral Sciences (19
relative seriousness of Type I to
the risk of false H_0 acceptance), r
tend to handicap themselves by
drawing on Cohen's conception
How much statistical power is neec
but the number of units or obser
of the effect, the preferred p leve
"Isn't the purpose of doing resea
realistically anticipate its size?"
literature search and, assuming si
study, to base our estimate of t
reported in the literature. A secon
to help us make a plausible estin
option is simply to assume that "r
of investigation, which often see
(cf. Brewer, 1972; Chase & Cha
Solomon, 1982; Sedlmeier & G
effects (and also "small" and "lar

Nonetheless, we cannot str
confidence intervals rather than t
null" any p that is not greater t
.05. It may not be an exaggeratio
has acquired an ontological mys
doctoral degree, and a tenure-tr
despair, and the adviser's sudde

However, you may be wondering why a detailed discussion of statistical power is needed, as we also stated that the primary coin of the realm when behavioral researchers speak of "the results of a study" should not be whether the p values of their tests of statistical significance are .05 or less. Instead, it should be information about the magnitude of the research finding (i.e., the effect size) and its accuracy or reliability (e.g., a confidence interval around the effect size). Although it is certainly true that "the initial emphasis on power analysis spearheaded by Cohen (1962) has now given way to a more sophisticated emphasis on precision of estimation," it is also true that "there are numerous obstacles to change in behavioral studies practice" (Steiger, 2004, pp. 178–179). For example, funding agencies frequently insist that a power analysis be included in a grant application to ensure that the resources requested will not be wasted in a study that implies the use of significance testing but has little chance of statistically detecting an effect that exists. Furthermore, statistical reforms regarding the use and interpretation of effect sizes and interval estimates (Wilkinson et al., 1999) have not been fully absorbed into the mainstream of behavioral research (Fidler, Thomason, Cumming, Finch, & Leeman, 2004; Thompson, 1999). An unfortunate consequence is that researchers sometimes give up prematurely on promising hypotheses when they mistakenly interpret an underpowered significance test's failure to detect an existing effect as indicating "no effect."

Some years ago, Jacob Cohen, in his seminal book *Statistical Power Analysis for the Behavioral Sciences* (1969, 1988), noted that without a conception of the relative seriousness of Type I to Type II errors (i.e., the risk of false H_0 rejection to the risk of false H_0 acceptance), researchers who do null hypothesis significance testing tend to handicap themselves by working with ridiculously low power. We begin by drawing on Cohen's conception to operationalize the effect of the neglect of power. How much statistical power is needed? We describe Cohen's benchmark recommendation, but the number of units or observations needed depends on the anticipated magnitude of the effect, the preferred p level, and the available resources. You may be thinking; "Isn't the purpose of doing research to *find out* the size of the effect? So how can we realistically anticipate its size?" There are, in fact, several options. One is to do a literature search and, assuming similar circumstances will be operating in the planned study, to base our estimate of the anticipated effect size on the average effect size reported in the literature. A second option is to rely on preliminary data in a pilot study to help us make a plausible estimate of the size of the effect in the full study. A third option is simply to assume that "medium-sized" effects are probably typical in the area of investigation, which often seems to be the case in behavioral and social research (cf. Brewer, 1972; Chase & Chase, 1976; Cohen, 1962, 1973; Haase, Waechter, & Solomon, 1982; Sedlmeier & Gigerenzer, 1989). Cohen's definitions of "medium" effects (and also "small" and "large" effects) are discussed in this chapter.

Nonetheless, we cannot stress too strongly our preference for effect sizes and confidence intervals rather than the counterproductive practice of interpreting as "anti-null" any p that is not greater than .05 and as "pro-null" any p that is greater than .05. It may not be an exaggeration to say that, for many Ph.D. students, the .05 alpha has acquired an ontological mystique. A dissertation p less than .05 means joy, a doctoral degree, and a tenure-track position, but a p greater than .05 means ruin, despair, and the adviser's suddenly thinking of a new control group that should be

run. Gigerenzer (1987) and associates (Gigerenzer & Murray, 1987; Gigerenzer et al., 1989), in discussions that examined the development of statistical inference, noted that the idea of dichotomous significance testing was initially developed out of agricultural experimentalists' need to answer questions such as "Is the manure effective?" It may be harder to object to the necessity of an accept-reject approach when the question is phrased in exactly that way, but the composition of behavioral research data, certainly, is substantively different, as is the phraseology of the questions that behavioral researchers attempt to address. R. A. Fisher at one point (largely as a reaction against the criticisms of J. Neyman and E. S. Pearson) objected to the idea of a fixed, dichotomous decision-level approach and instead recommended a cumulative, more provisional conception of statistical data analysis in science (Gigerenzer, 1987, p. 24).

Fisher's objection is not to imply that confidence intervals around effect sizes are unaffected by power. Table 12.1 shows 99%, 95%, and 90% confidence intervals for

TABLE 12.1
99%, 95%, and 90% confidence intervals (*CI*) for $r = .1$, .3, and .5 in samples of 10 to 800

N	$r = .1$	$r = .3$	$r = .5$
10	99%*CI*: −.70 to +.79	99%*CI*: −.58 to +.86	99%*CI*: −.40 to +.91
	95%*CI*: −.57 to +.69	95%*CI*: −.41 to +.78	95%*CI*: −.19 to +.86
	90%*CI*: −.48 to +.62	90%*CI*: −.30 to +.73	90%*CI*: −.07 to +.82
20	99%*CI*: −.48 to +.62	99%*CI*: −.31 to +.73	99%*CI*: −.08 to +.83
	95%*CI*: −.36 to +.52	95%*CI*: −.16 to +.66	95%*CI*: +.07 to +.77
	90%*CI*: −.29 to +.46	90%*CI*: −.09 to +.61	90%*CI*: +.15 to +.74
30	99%*CI*: −.38 to +.53	99%*CI*: −.18 to +.67	99%*CI*: +.05 to +.78
	95%*CI*: −.27 to +.44	95%*CI*: −.07 to +.60	95%*CI*: +.17 to +.73
	90%*CI*: −.21 to +.39	90%*CI*: −.01 to +.56	90%*CI*: +.23 to +.70
40	99%*CI*: −.31 to +.48	99%*CI*: −.11 to +.62	99%*CI*: +.13 to +.75
	95%*CI*: −.22 to +.40	95%*CI*: −.01 to +.56	95%*CI*: +.22 to +.70
	90%*CI*: −.17 to +.35	90%*CI*: +.04 to +.52	90%*CI*: +.27 to +.67
50	99%*CI*: −.27 to +.44	99%*CI*: −.07 to +.59	99%*CI*: +.17 to +.73
	95%*CI*: −.18 to +.37	95%*CI*: +.02 to +.53	95%*CI*: +.26 to +.68
	90%*CI*: −.14 to +.33	90%*CI*: +.07 to +.50	90%*CI*: +.30 to +.66
100	99%*CI*: −.16 to +.35	99%*CI*: +.05 to +.52	99%*CI*: +.28 to +.67
	95%*CI*: −.10 to +.29	95%*CI*: +.11 to +.47	95%*CI*: +.34 to +.63
	90%*CI*: −.07 to +.26	90%*CI*: +.14 to +.44	90%*CI*: +.36 to +.61
200	99%*CI*: −.08 to +.28	99%*CI*: +.13 to +.46	99%*CI*: +.35 to +.62
	95%*CI*: −.04 to +.24	95%*CI*: +.17 to +.42	95%*CI*: +.39 to +.60
	90%*CI*: −.02 to +.21	90%*CI*: +.19 to +.40	90%*CI*: +.41 to +.58
400	99%*CI*: −.03 to +.23	99%*CI*: +.18 to +.41	99%*CI*: +.40 to +.59
	95%*CI*: .00 to +.20	95%*CI*: +.21 to +.39	95%*CI*: +.42 to +.57
	90%*CI*: +.02 to +.18	90%*CI*: +.22 to +.37	90%*CI*: +.44 to +.56
800	99%*CI*: +.01 to +.19	99%*CI*: +.22 to +.38	99%*CI*: +.43 to +.56
	95%*CI*: +.03 to +.17	95%*CI*: +.24 to +.36	95%*CI*: +.45 to +.55
	90%*CI*: +.04 to +.16	90%*CI*: +.25 to +.35	90%*CI*: +.46 to +.54

three levels of effect size *r*s (.1, .
samples ranging from $N = 10$ to
observed effect size r of .30, we
effect size r is between .11 and
quadrupling the size of the sampl
smaller, from .21 to .39. Within e
interval is wider as the confiden
wanted to be 100% confident, we
−1.0 (the lower limit of r) and 1

THE NEGLECT OF STATI

Although leading textbooks on ps
mentioned statistical power (e.
McNemar, 1962; Siegel, 1956; W
tions of power analysis did not m
researchers until quite recently. I
ters, culminating in his seminal
strating that null hypothesis sign
with a remarkably high risk of c
he reported the median power for
at $\alpha = .05$ in articles published
during a single year (1960) was
the odds were better than 50:50
false. In an article entitled "Do
Power of Studies?" which was
Gigerenzer (1989) reported that
slightly worse.

Cohen (1969) proposed a co
of the neglect of power in any g
α. For example, suppose a rese
significance with power = .40, in
becomes .60/.05 = 12. In other w
rejecting the null hypothesis sh
mistakenly accepting it. Table 12.
1,000 under the conditions of assu
and two levels of statistical signif
weight attached to the avoidance
increases the smaller the effect s
level of significance. Some criti
typically have ample power to det
assumption that error terms are u
samples and highly standardized
$r = .1$, a small sample (say, $N =$
significance testing will be a prec

three levels of effect size *r*s (.1, .3, and .5), assuming zero effect null hypotheses, with samples ranging from $N = 10$ to $N = 800$. For example, with a total N of 100 and an observed effect size r of .30, we can state with 95% confidence that the true population effect size $\dot{r}$ is between .11 and .47. Given the same observed effect size r of .3, but quadrupling the size of the sample to an N of 400, we see that the 95% *CI* is noticeably smaller, from .21 to .39. Within each batch of confidence intervals, it is evident that the interval is wider as the confidence level increases from 90% to 95% to 99%. If we wanted to be 100% confident, we could say the true population effect size r is between −1.0 (the lower limit of r) and 1.0 (the upper limit).

THE NEGLECT OF STATISTICAL POWER

Although leading textbooks on psychological statistics in the 1950s and 1960s routinely mentioned statistical power (e.g., Edwards, 1964; Guilford, 1956; Hays, 1963; McNemar, 1962; Siegel, 1956; Walker & Lev, 1953; Winer, 1962), the design implications of power analysis did not make their way into the consciousness of psychological researchers until quite recently. In the 1960s, in a series of articles and invited chapters, culminating in his seminal book published in 1969, Cohen pioneered in demonstrating that null hypothesis significance testing in behavioral research is conducted with a remarkably high risk of committing Type II errors. In an early meta-analysis, he reported the median power for detecting what he characterized as "medium" effects at $\alpha = .05$ in articles published in the *Journal of Abnormal and Social Psychology* during a single year (1960) was no better than flipping a coin (Cohen, 1962). Indeed, the odds were better than 50:50 that the null hypothesis would *not* be rejected when false. In an article entitled "Do Studies of Statistical Power Have an Effect on the Power of Studies?" which was published nearly three decades later, Sedlmeier and Gigerenzer (1989) reported that the median power of studies in the same journal was slightly worse.

Cohen (1969) proposed a convenient way to operationalize the relative seriousness of the neglect of power in any given situation by simply examining the ratio of β to α. For example, suppose a researcher has set α at .05 and is conducting a test of significance with power = .40, in which case β is 1 − .40 (or .60), and the β/α ratio becomes $.60/.05 = 12$. In other words, the researcher ostensibly believes that mistakenly rejecting the null hypothesis should be regarded as 12 times more serious than mistakenly accepting it. Table 12.2 shows the ratio of β/α for sample sizes from 10 to 1,000 under the conditions of assumed effect size noted in Table 12.1 (*r*s of .1, .3, and .5) and two levels of statistical significance ($p = .05$ and $p = .10$). The generally greater weight attached to the avoidance of Type I errors relative to Type II errors clearly increases the smaller the effect size r, the smaller the N, and the more stringent the level of significance. Some critics have argued that psychologists working in labs typically have ample power to detect even "small" effects (to be defined shortly), on the assumption that error terms are usually small in lab studies (because of homogeneous samples and highly standardized procedures). However, Table 12.2 shows that at $r = .1$, a small sample (say, $N = 20$, or 10 per group), and a binary decisional $p = .05$, significance testing will be a precarious exercise ($\beta/\alpha = 19$).

TABLE 12.2
Ratios of Type II to Type I error rates (β/α) for various sample sizes, effect sizes, and significance levels (two-tailed)

	Effect sizes (r) and significance levels (.05 and .10)					
	r = .10		r = .30		r = .50	
N	.05	.10	.05	.10	.05	.10
10	19	9	17	8	13[a]	5[b]
20	19	9	15	6	7[a]	2[b]
30	18	8	13	5	3	1
40	18	8	10	4	2	*
50	18	8	9	3	*	*
60	18	8	7	2	*	*
70	17	8	6	2	*	*
80	17	8	4	1	*	*
90	17	8	4	1	*	*
100	17	7	3	*	*	*
120	16	7	2	*	*	*
140	16	7	1	*	*	*
160	15	6	*	*	*	*
180	15	6	*	*	*	*
200	14	6	*	*	*	*
300	12	5	*	*	*	*
400	10	4	*	*	*	*
500	8	3	*	*	*	*
600	6	2	*	*	*	*
700	5	2	*	*	*	*
800	4	1	*	*	*	
900	3	*	*	*	*	
1000	2	*	*	*	*	

*Values less than 1.
[a] For r = .70 these ratios drop to 6 and <1, respectively.
[b] For r = .70 these ratios drop to 2 and <1, respectively.

Given that alpha is typically set at .05, Cohen (1965) recommended .80 as a baseline of the statistical power usually needed in behavioral research. With power set at .80, it follows that β = .2, and the β/α ratio (.2/.05 = 4) implies that Type I error is regarded as 4 times more serious than Type II error. Setting the power higher than .80 obviously reduces the β/α ratio. With power at .90 and α at .05, Type I error would be regarded as 2 times more serious than Type II error (.1/.05 = 2). If the

effect size is small and recruiting
number of sampling units needed m
researcher who does null hypothesis
constraints imposed by the neglect

In the next section we describe
Cohen's (1988) text on power analys
our suggestion in chapter 1 that, whe
the effect size measures to a particula
family is the most generally useful. I
we might use Cohen's d (Equation
(Equation 2.6). In practice, effect siz
on more than two groups, for exampl
pattern of three or more means (Ros
is not as natural to use a two-group-l
use a member of the r family of effec
note ways of converting certain effect
we will show how to conceptualize
directly estimated. Even when all we
to compute an interpretable effect siz

THE $r_{equivalent}$ STATISTIC

Suppose all we have are the total sar
else. A quite serviceable approach t
$r_{equivalent}$ procedure (Rosenthal & Ru
the estimated r is equivalent to a si
treatment indicator and an exactly r
experiment with $N/2$ units in each g
to compute $r_{equivalent}$ is to identify t
(usually with $df = N - 2$) and simp
is a variation on Equation 11.11):

$$r$$

where r in this case refers to r_{equiva}
such as Table B.2 or B.3 in Appendi
in meta-analytic work, or in other r
nor significance test statistics (such
only p values and sample sizes ar
generally accepted for the data-a
nonparametric statistics, such as th
estimate can be computed directly fr
sizes or severe nonnormality, the est

Say that all we have availabl
Whitney U test, which is commonly
who want to avoid the t test's ass

effect size is small and recruiting subjects is expensive, the cost in terms of the number of sampling units needed may be prohibitive. Nonetheless, it behooves every researcher who does null hypothesis significance testing to be mindful of the potential constraints imposed by the neglect of statistical power.

In the next section we describe the principal effect size indices operationalized in Cohen's (1988) text on power analysis. Before we do so, however, we want to reiterate our suggestion in chapter 1 that, when it is necessary to make a decision to convert all the effect size measures to a particular index (e.g., in meta-analytic work), the correlation family is the most generally useful. If our research calls for a comparison of two groups, we might use Cohen's d (Equation 2.4), or Hedges's g (Equation 2.5), or Glass's Δ (Equation 2.6). In practice, effect sizes are often needed for comparisons that are based on more than two groups, for example, in computing linear trends or any other predicted pattern of three or more means (Rosenthal, Rosnow, & Rubin, 2000). In such cases it is not as natural to use a two-group-based effect size indicator, but it is quite natural to use a member of the r family of effect size indicators (discussed in chapter 15). We will note ways of converting certain effect sizes to r and r-type indices (viz., Fisher's z_r), and we will show how to conceptualize the study design so that an effect size r can be directly estimated. Even when all we have are minimal raw ingredients, we still are able to compute an interpretable effect size r, as illustrated next.

THE $r_{\text{equivalent}}$ STATISTIC

Suppose all we have are the total sample size (N) and an accurate p value, but nothing else. A quite serviceable approach to estimating an interpretable effect size r is the $r_{\text{equivalent}}$ procedure (Rosenthal & Rubin, 2003). It takes its name from the fact that the estimated r is equivalent to a sample point-biserial correlation (r_{pb}) between the treatment indicator and an exactly normally distributed outcome in a two-treatment experiment with $N/2$ units in each group and the obtained p value. All that is needed to compute $r_{\text{equivalent}}$ is to identify the value of t that corresponds to the accurate p (usually with $df = N - 2$) and simply to substitute in the following equation (which is a variation on Equation 11.11):

$$r = \sqrt{\frac{t^2}{t^2 + df}}, \tag{12.1}$$

where r in this case refers to $r_{\text{equivalent}}$, and t can be obtained from a standard table such as Table B.2 or B.3 in Appendix B. This procedure is especially useful when (a) in meta-analytic work, or in other reanalyses of others' studies, neither effect sizes nor significance test statistics (such as an obtained t or F value) are provided, but only p values and sample sizes are given; (b) no effect size index has yet been generally accepted for the data-analytic procedure used (as is true of certain nonparametric statistics, such as the Mann-Whitney U test); or (c) an effect size estimate can be computed directly from the reported data, but because of small sample sizes or severe nonnormality, the estimates may be seriously misleading.

Say that all we have available from a study report is the result of a Mann-Whitney U test, which is commonly used in small sample studies by experimenters who want to avoid the t test's assumptions. Expert judges were used to rank the

performance of nine children on a reading-ability measure. Four of the children were randomly assigned to a condition (the treatment) in which they were taught by a new method, and five children were taught by an old (control) method. And let us further assume that all four treated children were ranked higher than any of the five control children. This outcome would yield an exact probability of .008, one-tailed Mann-Whitney U (Siegel, 1956, p. 271). Given $p = .008$ and $N = 9$, $t(df = 7) = 3.16$, and from Equation 12.1 ($df = N - 2$) we find

$$r_{\text{equivalent}} = \sqrt{\frac{(3.16)^2}{(3.16)^2 + 7}} = .77.$$

When reporting this result, we would also want to report an interval estimate, which is typically the 95% confidence interval. In this case we use the procedure described in the previous chapter in connection with Equation 11.3; that is, 95% $CI = z_r \pm 1.96/\sqrt{N-3}$. For $r_{\text{equivalent}} = .77$, we see in Table B.7 that Fisher $z_r = 1.02$, and so the 95% CI extends from a z_r of .22 to a z_r of 1.82. Using Table B.8 to transform this 95% CI into units of r gives us an interval from $r = .22$ to .95.

The $r_{\text{equivalent}}$ method can be used to estimate an effect size r in any two-group comparison of means of a normally distributed outcome. As we show in chapter 15, a limitation of Equation 12.1 is that with more than two treatment conditions, this formula is an estimate of what we call r_{contrast}. The distinction is that $r_{\text{effect size}}$ is the unpartialed correlation between group (or condition) membership and individual scores on the dependent measure, and r_{contrast} is the partial correlation (with noncontrast variation removed) and therefore tends to overstate what might be viewed as the more natural effect size correlation (i.e., $r_{\text{effect size}}$). In any two-group comparison, however, r_{contrast} is equivalent to $r_{\text{effect size}}$ because there is no noncontrast variation to be partialed out.

COHEN'S MULTIPURPOSE POWER TABLES

The expression *power analysis* is another umbrella term, as it may refer to estimating the number of sampling units needed to detect a particular magnitude of effect at a stipulated α, or to estimating the statistical power of a study already conducted. The way these procedures work involves estimating one of four parameters from our knowledge of the other three. For example, given a p level, an effect size value, and the number of subjects participating in a study, we can estimate the power of an already completed study, which we call the **effective power.** Suppose the study design was based on equal sample sizes, but something unexpected happened and we ended up with more no-shows in one group. In the next chapter we will show how to estimate the loss of power in an unequal-n study relative to an equal-n design. In some situations, however, a study may be specifically designed to allocate the subjects to various conditions unequally, the objective being to "optimize" statistical power by emphasizing some conditions over others. We have more to say about this approach in chapter 15, but by far the most common reason for a power analysis is to estimate the number of sampling units that will be needed in an equal-n study (given an α, the desired level of power, and an anticipated magnitude of effect).

Table 12.3 is a composite based on seven of Cohen's (1988) effect size indicators for use in a power analysis. In the first column, the effect size d is based on the t test

TABLE 12.3
Multipurpose power tables with s

	t	r	$r_1 - r_2$
Effect size index	d	r	q
Effect size:			
Small	.20	.10	.10
Medium	.50	.30	.30
Large	.80	.50	.50

A. Sample size (rounded) required to detec

Power	t	r	$r_1 - r_2$
.25	14	20	40
.50	32	42	88
.60	40	53	112
.70	50	67	140
.75	57	75	157
.80	64	85	177
.85	73	97	203
.90	85	113	236
.95	105	139	292
.99	148	195	411
Definition of n (see note below)	a	b	c

B. Sample size (rounded) required to detec

Power	t	r	$r_1 - r_2$
.25	31	40	83
.50	55	72	150
.60	66	87	181
.70	79	103	217
.75	86	113	238
.80	95	125	263
.85	106	139	293
.90	120	158	334
.95	144	189	399
.99	194	254	537

ure. Four of the children were
ich they were taught by a new
ol) method. And let us further
er than any of the five control
lity of .008, one-tailed Mann-
l $N = 9$, $t(df = 7) = 3.16$, and

= .77.

: to report an interval estimate,
his case we use the procedure
h Equation 11.3; that is, 95%
Table B.7 that Fisher $z_r = 1.02$,
z_r of 1.82. Using Table B.8 to
erval from $r = .22$ to .95.
n effect size r in any two-group
ne. As we show in chapter 15, a
reatment conditions, this formula
is that $r_{\text{effect size}}$ is the unpartialed
p and individual scores on the
tion (with noncontrast variation
t be viewed as the more natural
p comparison, however, r_{contrast} is
variation to be partialed out.

LES

rm, as it may refer to estimating
rticular magnitude of effect at a
a study already conducted. The
e of four parameters from our
p level, an effect size value, and
e can estimate the power of an
power. Suppose the study design
xpected happened and we ended
chapter we will show how to
elative to an equal-n design. In
designed to allocate the subjects
o "optimize" statistical power by
more to say about this approach
r a power analysis is to estimate
n an equal-n study (given an α,
itude of effect).
nen's (1988) effect size indicators
effect size d is based on the t test

TABLE 12.3
Multipurpose power tables with sample size estimates

	Statistic						
	t	r	$r_1 - r_2$	$P - .50$	$P_1 - P_2$	χ^2	F
Effect size index	d	r	q	g	h	w	f
Effect size:							
Small	.20	.10	.10	.05	.20	.10	.10
Medium	.50	.30	.30	.15	.50	.30	.25
Large	.80	.50	.50	.25	.80	.50	.40

A. Sample size (rounded) required to detect "medium" effect at $p = .05$ two-tailed

Power	t	r	$r_1 - r_2$	$P - .50$	$P_1 - P_2$	$\chi^2(df = 1)$	F ($df = 1$ in numerator)
.25	14	20	40	20	13	18	14
.50	32	42	88	44	31	43	32
.60	40	53	112	54	39	54	40
.70	50	67	140	67	49	69	50
.75	57	75	157	75	56	77	57
.80	64	85	177	85	63	87	64
.85	73	97	203	97	72	100	73
.90	85	113	236	113	84	117	85
.95	105	139	292	138	104	144	105
.99	148	195	411	194	147	204	148
Definition of n (see note below)	a	b	c	d	c	d	a

B. Sample size (rounded) required to detect "medium" effect at $p = .01$ two-tailed

Power	t	r	$r_1 - r_2$	$P - .50$	$P_1 - P_2$	$\chi^2(df = 1)$	F ($df = 1$ in numerator)
.25	31	40	83	44	29	40	31
.50	55	72	150	74	53	74	55
.60	66	87	181	88	64	89	66
.70	79	103	217	105	77	107	79
.75	86	113	238	115	85	117	86
.80	95	125	263	127	93	130	95
.85	106	139	293	141	104	145	106
.90	120	158	334	160	119	165	120
.95	144	189	399	191	143	198	144
.99	194	254	537	255	192	267	194

(*continued*)

TABLE 12.3 (*continued*)

C. Sample size (rounded) required to detect "small" effect at $p = .05$ two-tailed

Power	t	r	$r_1 - r_2$	$P - .50$	$P_1 - P_2$	$\chi^2(df = 1)$	F ($df = 1$ in numerator)
.25	84	168	333	166	83	165	84
.50	193	386	771	384	192	384	193
.60	246	491	983	489	245	490	246
.70	310	617	1,237	616	309	617	310
.75	348	692	1,391	692	347	694	348
.80	393	784	1,573	783	392	785	393
.85	450	896	1,799	895	449	898	450
.90	526	1,048	2,104	1,047	525	1,051	526
.95	651	1,295	2,602	1,294	650	1,300	651
.99	920	1,829	3,677	1,827	919	1,837	920

Note: The definitions of n indicated at the bottom of Section A are symbolized as follows: a = each group or condition; b = n of score pairs; c = n of each sample; and d = total N.

(The sample sizes are based on tables in *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), by J. Cohen, 1988, Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, and on Foster's (2003) METASTATS software.)

for independent means (e.g., Equation 11.9), where the null hypothesis is $M_1 - M_2 = 0$. The next effect index is r, the point-biserial correlation discussed in the previous chapter. Cohen's recommended test of the null hypothesis that $r = 0$ is again based on the t distribution (e.g., Equation 2.2). The effect size q (in the third column) refers to the difference between two independent Fisher z_r transformed correlations. The null hypothesis is that $z_{r1} - z_{r2} = 0$, and the significance test might be Z (as we illustrate shortly), from which we get a p value from which we can get the value t that will let us use Equation 12.1 to compute the $r_{equivalent}$ statistic. Next in the table is the effect size index g, which is different from Hedges's g and is instead the difference between an observed proportion (symbolized in Cohen's text by a capital P) and .50. The null hypothesis is $P - .50 = 0$, and Cohen's test of significance is the sign test, which is a nonparametric procedure that uses plus and minus signs rather than quantitative measures as its data (Siegel, 1956). Next is the effect size index h, which measures the difference between independent arcsine-transformed proportions; the null hypothesis is that this difference is zero. The effect size index w is that for the chi-square test (Equation 11.13, discussed in more detail in chapter 19), although the sample sizes in Table 12.3 refer only to 1-df χ^2 tests. Finally, the effect size f is for F tests on means in the analysis of variance, but the sample sizes in Table 12.3 refer only to F tests with numerator $df = 1$ based on a comparison of the means of two samples.

Cohen's quantitative definitions of "small," "medium," and "large" effect sizes are shown in the top panel of Table 12.3, and the values in the body of the table are the rounded sample size estimates. In Section A, the statistical power and sample size equivalences are given for an alpha level of .05 two-tailed and the assumption of "medium" effects. Section B shows equivalences for "medium" effects at $\alpha = .01$ two-tailed, and Section C shows the equivalences for "small" effects at $\alpha = .05$ two-tailed. One other important fe
same for all seven statistics. Samp
tion"; those for r are for "n of sco
of each sample"; and those for P
Whenever Table 12.3 involv
$P_1 - P_2$), the sample sizes are ass
we can use the harmonic mean s
harmonic mean sample size for tw

In an equal-n design, the harmonic
sample size, but in an unequal-n
smaller than the arithmetic mean
sample sizes of 12 and 18. The ar
15, and the harmonic mean sampl

$$n_h =$$

We also want to note that Co
"large" are for use solely with his
should be interpreted "relative no
science or even more particularly
employed in any given investiga
Cohen's benchmark labels as if the
that "the *meaning* of any given E
the context in which it is embed
that even small effects (by Cohen's
practical implications (see again T

Table 12.3 reveals a numbe
Sections A, B, and C. First, as the
subtables, the sample size needed
the more stringent or conservativ
that more units are needed with
Third, as a comparison of Sectio
the more sampling units needed
statistics call for different sample
example, it will take fewer units f

THE t TEST FOR COMPAR TWO MEANS

We have mentioned Cohen's d (Ec
Δ (Equation 2.6) as effect size i
another option in some situations
the dependent variable the daily

two-tailed. One other important feature is that the definition of sample size is not the same for all seven statistics. Sample sizes for t and F are for "each group or condition"; those for r are for "n of score pairs"; those for $r_1 - r_2$ and $P_1 - P_2$ are for "n of each sample"; and those for $P - .50$ and χ^2 are for "total N."

Whenever Table 12.3 involves a comparison of two samples (e.g., t, F, $r_1 - r_2$, $P_1 - P_2$), the sample sizes are assumed to be equal. If the sample sizes are unequal, we can use the harmonic mean sample size (n_h) to provide an approximate n. The harmonic mean sample size for two samples of n_1 and n_2 size is obtained by

$$n_h = \frac{2n_1n_2}{n_1 + n_2} \tag{12.2}$$

In an equal-n design, the harmonic mean sample size is equal to the arithmetic mean sample size, but in an unequal-n design, the harmonic mean sample size is always smaller than the arithmetic mean sample size. Suppose we have two groups with sample sizes of 12 and 18. The arithmetic mean sample size is $n = (12 + 18)/2 = 15$, and the harmonic mean sample size is

$$n_h = \frac{2(12 \times 18)}{12 + 18} = 14.4$$

We also want to note that Cohen's benchmark labels of "small," "medium," and "large" are for use solely with his power tables. Cohen (1988) advised that effect sizes should be interpreted "relative not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation" (p. 25). Nonetheless, many researchers cite Cohen's benchmark labels as if they were context-free, though he specifically cautioned that "the *meaning* of any given ES [effect size] is, in the final analysis, a function of the context in which it is embedded" (p. 535). In the previous chapter we showed that even small effects (by Cohen's definition) can sometimes be loaded with profound practical implications (see again Table 11.8).

Table 12.3 reveals a number of fundamental relationships in the subtables in Sections A, B, and C. First, as the statistical power increases from .25 to .99 in these subtables, the sample size needed for each significance test also increases. Second, the more stringent or conservative the p value, the more sampling units needed, so that more units are needed with $p = .01$ two-tailed than with $p = .05$ two-tailed. Third, as a comparison of Sections A and C will reveal, the smaller the effect size, the more sampling units needed at the same significance level. Fourth, different statistics call for different sample sizes to detect the same benchmark levels, so, for example, it will take fewer units for r than for $r_1 - r_2$ (we explain why shortly).

THE t TEST FOR COMPARING TWO MEANS

We have mentioned Cohen's d (Equation 2.4), Hedges's g (Equation 2.5), and Glass's Δ (Equation 2.6) as effect size indices for the difference between two means, but another option in some situations is the raw difference itself. Suppose we choose as the dependent variable the daily number of cigarettes smoked by experimental and

control subjects. The raw difference between M_1 and M_2 (where M is the mean number of cigarettes in a group) is intrinsically meaningful. As another example, suppose we want to compare a method of vocational rehabilitation to a control, and we have a record of the days that workers were reported absent in each condition. If we found that workers in the control condition averaged five more absences per month than did workers in the rehabilitation condition, this difference would be fraught with practical meaning. The point is that raw mean differences can, in some instances, be informative and useful as effect size indicators, whether we are analyzing differences in a specific study (Rosnow & Rosenthal, 2003) or making cross-study comparisons in the context of a meta-analysis (Bond, Wiitala, & Richard, 2003).

Nonetheless, the most popular measure of effect size for comparing two means is Cohen's *d*, where the effect size is expressed in standard deviation units. Assuming two populations with equal variability and equal sample sizes, Cohen recommended dividing the difference between the sample means ($M_1 - M_2$) by the standard deviation of either group (i.e., σ_1 or σ_2) to obtain *d*. With σ_1 and σ_2 unequal, he recommended dividing $M_1 - M_2$ by the square root of the mean of the two variances for the denominator, that is,

$$d = \frac{M_1 - M_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}. \tag{12.3}$$

As an all-purpose expression of Cohen's *d* in the case of two independent means, we recommend Equation 2.4:

$$d = \frac{M_1 - M_2}{\sigma_{\text{pooled}}},$$

where the difference between two independent means is divided by the common within-group σ. For the sample sizes in Table 12.3, we see that when the power level is no better than a coin flip (.50), we need respective samples of 32, 55, and 193 in *each group* for the three combinations of the stipulated effect size *d* and the α in Sections A, B, and C. Before we examine an extended power table, it is of interest to review how Cohen chose the benchmark values of small, medium, and large *d*s.

For $d = .2$, Cohen (1988) reasoned that "in new areas of research, the effect sizes are likely to be small (when they are not zero!)" because "the phenomena under study are typically not under good experimental or measurement control or both" (p. 25). Assuming the populations being compared are normal and have equal variability, then if *d* is zero, the two distributions will perfectly overlap. With $d = .2$, the amount of nonoverlap will be 14.7%, because the two means are separated by one-fifth of a standard deviation difference, a "small" difference. An example was the larger size of the difference in mean IQ of non-twins as opposed to twins. Another example was the magnitude of the difference between the mean height of 16-year-old girls and 15-year-old girls (about one-half inch, where $\sigma = 2.1$, for a rounded *d* of .2). For $d = .5$, Cohen (1988, p. 26) thought that a difference just "visible to the naked eye" was a good way of thinking about "medium" effects, and he theorized that *d*s of about .5 (one-half a standard deviation difference) should be visible, because there is 33% nonoverlap of the normal population curves. An

TABLE 12.4
Sample size per group (n) n differences between two ind statistical significance

	$d = .2$		
	.10	.05	.01
Power	.05	.025	.005
.50	136	193	333
.60	181	246	402
.70	236	310	482
.80	310	393	586
.85	360	450	654
.90	429	526	746
.95	542	651	892
.99	789	920	1203

Note: The sample size values in this table

example was the higher mean semiskilled workers (about 8 Cohen (1988) thought that 47 a good indicator of a "large" An example was the mean I freshmen, which he said was (1988, p. 27).

Table 12.4 focuses in o of equal size ($n_1 = n_2$). Value the sample size needed in each with power of .80, sets α at .(means will be a half standard researcher will need 64 subj direction of the difference, the .80, the researcher would nee

We will provide convers chapter, but assuming equal-si *r* as follows:

Thus, a "medium" Cohen's *d* the correlation between the treatment, usually coded as 1, dependent or outcome variabl

TABLE 12.4
Sample size per group (n) needed to detect "small," "medium," and "large" differences between two independent means at various levels of power and statistical significance

	$d = .2$			$d = .5$			$d = .8$			
	.10	.05	.01	.10	.05	.01	.10	.05	.01	two-tailed p
Power	.05	.025	.005	.05	.025	.005	.05	.025	.005	one-tailed p
.50	136	193	333	22	32	55	9	13	22	
.60	181	246	402	30	40	66	12	16	27	
.70	236	310	482	38	50	79	15	20	32	
.80	310	393	586	50	64	95	20	26	38	
.85	360	450	654	58	73	106	23	29	43	
.90	429	526	746	69	85	120	27	34	48	
.95	542	651	892	87	105	144	35	42	57	
.99	789	920	1203	127	148	194	50	58	77	

Note: The sample size values in this table are based on Foster's (2003) METASTATS program.

example was the higher mean IQ of managers and professionals versus clerical and semiskilled workers (about 8 points, where $\sigma = 15$, rounded to $d = .5$). For $d = .8$, Cohen (1988) thought that 47.4% nonoverlap of normal population curves might be a good indicator of a "large" effect (four-fifths of a standard deviation difference). An example was the mean IQ difference of typical Ph.D.s versus typical college freshmen, which he said was a "grossly perceptible and therefore large" difference (1988, p. 27).

Table 12.4 focuses in on Cohen's three benchmark levels of d for two groups of equal size ($n_1 = n_2$). Values indicated in the body of the table refer once again to the sample size needed in each group. Suppose the researcher is interested in working with power of .80, sets α at .05 two-tailed, and predicts that the difference between means will be a half standard deviation (i.e., $d = .5$). Table 12.4 indicates that the researcher will need 64 subjects in each group. Had the researcher predicted the direction of the difference, then with α set at .05 one-tailed and statistical power of .80, the researcher would need 50 subjects in each group.

We will provide conversion formulas for Hedges's g and Glass's Δ in the next chapter, but assuming equal-sized samples ($n_1 = n_2$), we can transform Cohen's d into r as follows:

$$r = \sqrt{\frac{d^2}{d^2 + 4}} \tag{12.4}$$

Thus, a "medium" Cohen's d of .50 is equivalent to $r = .24$, which is interpreted as the correlation between the independent variable of group or condition (e.g., a treatment, usually coded as 1, vs. a control, usually coded as 0) and the score on the dependent or outcome variable for sample sizes found, or assumed on theoretical

grounds, to be equal. Notice that, although $d = .5$ is the benchmark level of a "medium" d, the r of .24 is slightly smaller than the benchmark level of a "medium" r (i.e., .3). We will have more to say about this discrepancy shortly.

What if the sample sizes were inherently unequal? In that case we could use the following formula to convert Cohen's d into r:

$$r = \sqrt{\frac{d^2}{d^2 + \left(\frac{1}{PQ}\right)}}, \tag{12.5}$$

where P denotes the proportion of the overall total sample (N) or population represented by the sample in one group (n_1), and $Q = 1 - P$. Thus, $P = (n_1)/N$, and $Q = (n_2)/N$. As an illustration, say that we compare scores from patients with a rare disorder to scores from patients with common disorders, or to scores obtained from people in general. If the rare disorder occurs in 5% of people, then $P = .05$ and $Q = .95$. (Of course, when $P = Q$, then Equation 12.5 is equivalent to Equation 12.4.) The r obtained from Equation 12.5 is interpreted as the correlation between the independent variable of group or condition (e.g., a treatment, usually coded as 1, vs. control, usually coded as 0) and the score on the dependent or outcome variable for sample sizes that (a) are observed to be unequal and (b) are assumed on theoretical grounds to be inherently unequal.

THE SIGNIFICANCE OF A PRODUCT-MOMENT r

Cohen's effect size associated with the relationship between two variables is the product-moment r (Equation 11.1). Table 12.3 indicates the benchmark levels of small, medium, and large rs as .1, .3, and .5, respectively. As noted before, there is not consistently an exact correspondence between Cohen's benchmark levels for r and d. This fact is illustrated in more detail in Table 12.5, where we see that "small" rs and ds do not run afoul of Cohen's labeling convention. However, an r of .3 (a "medium" r) corresponds to a Cohen's d of .63, and an r of .5 (a "large" r) corresponds to a Cohen's d of 1.15 (a "jumbo" effect). As the third column of Table 12.5 shows, the relationship between Cohen's d and r is not a perfectly straight line. It is another reason not to mindlessly use the labeling convention when reporting effect sizes, but to specify the particular index and its precise value.

Consulting the sample sizes in Table 12.3, we see that to achieve the "flipping-the-coin" power level of .50 requires sample sizes of 42, 72, and 386 score pairs, respectively, for the three combinations of expected effect size and α shown in Sections A, B, and C. Comparing the effect sizes listed for the t statistic (i.e., effect size index d) and r reveal the sample sizes required for r to be uniformly higher. However, the entries under t are the ns for each of the two groups, whereas the entries for r are the total sample size. In fact, for most power levels, and for most effect sizes and α levels, the total sample size required by r is smaller than that required by the t statistic. Part of the reason for this difference is that the standard independent t test comparing two means cannot take advantage of both between-group and within-group linearity of regression. In chapter 15 we will describe a t test that *does* take advantage of this information and

TABLE 12.5
Relation between r and Cohen's of equal sample size

r	Cohen's d	d/r
.01	.02	2.0
.02	.04	2.0
.03	.06	2.0
.04	.08	2.0
.05	.10	2.0
.10	.20	2.0
.15	.30	2.0
.20	.41	2.1
.25	.52	2.1
.30	.63	2.1

therefore is more powerful in mult
about the pattern of the group mea
Table 12.6, which shows the score
experimental condition, where the

TABLE 12.6
Comparisons between two sets

A. No specific prediction other than M_e

Subject	Condition
1	Control
2	Control
3	Experimental
4	Experimental

B. Linear prediction that Subject 1 <

Subject	Condition
1	Control
2	Control
3	Experimental
4	Experimental

TABLE 12.5
Relation between r and Cohen's d, assuming two conditions of equal sample size

r	Cohen's d	d/r	r	Cohen's d	d/r
.01	.02	2.0	.35	.75	2.1
.02	.04	2.0	.40	.87	2.2
.03	.06	2.0	.45	1.01	2.2
.04	.08	2.0	.50	1.15	2.3
.05	.10	2.0	.55	1.32	2.4
.10	.20	2.0	.60	1.50	2.5
.15	.30	2.0	.70	1.96	2.8
.20	.41	2.1	.80	2.67	3.3
.25	.52	2.1	.90	4.13	4.6
.30	.63	2.1	1.00	∞	∞

therefore is more powerful in multiple comparisons when there is a specific prediction about the pattern of the group means. To anticipate a little, the problem is explained by Table 12.6, which shows the scores of two subjects each in a control condition and an experimental condition, where the means are 3.0 and 7.0, respectively.

TABLE 12.6
Comparisons between two sets of scores

A. No specific prediction other than $M_{control} \neq M_{experimental}$

Subject	Condition	Within-condition specific prediction	Score	M
1	Control	None	2	3.0
2	Control	None	4	
3	Experimental	None	6	7.0
4	Experimental	None	8	

B. Linear prediction that Subject 1 < Subject 2 < Subject 3 < Subject 4

Subject	Condition	Within-condition specific prediction	Score
1	Control	−3	2
2	Control	−1	4
3	Experimental	+1	6
4	Experimental	+3	8

Suppose the only prediction is that the means will be different, in which case $t = 2.83$, $df = 2$, and p is approximately .11 two-tailed. In Equation 12.1, the effect size r associated with this t test is $r = .895$. Clearly, it would make no difference to the t test whether the two scores in each condition were in the order that is shown in Part A of Table 12.6, or if the order had been reversed within conditions (Subject 2 and then Subject 1 in the control condition, and Subject 4 and then Subject 3 in the experimental condition). The means remain unchanged, and those are what the standard t test is focused upon. However, suppose we had predicted (based on some theory) that Subject 1 would have the lowest score, Subject 2 would have a higher score, Subject 3 would have an even higher score, and Subject 4 would have the highest score. We can express this linear prediction by weights of $-3, -1, +1, +3$. Correlating these four weights with the four individual scores gives $r = 1.00$, which is what it should be, as the prediction of a perfect linear relationship is confirmed by the pattern of the scores. The t test of this r is infinitely large, because there is no "within-group" variability (the "groups" are the subjects, and there is a single subject in each "group"). But if even one of the pairs of numbers of the experimental or control condition is interchanged, the r drops from 1.00 to .80 ($t = 1.89$, $df = 2$, p approximately .20 two-tailed). Therefore, where there really is more nearly perfect linear regression between the predicted and obtained results, r is likely to be more powerful than the standard t test comparing two independent means. The reason is that the t test has lost some information in the independent (or predictor) variable by dichotomizing the predictor values ($-3, -1, +1, +3$) into just two levels.

Before leaving this discussion of r, see Table 12.7, which is an extended table for use when alpha is set at .05 two-tailed (or .025 one-tailed). Suppose we had estimated that it was 95% likely that an effect size in the population would be

TABLE 12.7
Sample sizes (rounded) to detect r by t test at $p = .05$ two-tailed or .025 one-tailed

	Effect size correlation (r)													
Power	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70
.25	664	168	76	44	29	21	16	13	10	9	8	7	6	5
.50	1,538	386	172	97	63	44	33	25	20	16	14	11	10	8
.60	1,960	491	218	123	79	55	41	31	25	20	16	14	12	10
.70	2,469	617	274	154	99	68	50	38	30	24	20	16	14	12
.80	3,138	784	348	195	124	86	63	48	37	30	24	20	17	14
.85	3,589	896	397	222	142	98	71	54	42	34	27	22	19	16
.90	4,200	1,048	464	260	165	114	83	62	49	39	31	26	21	18
.95	5,193	1,295	573	320	203	140	101	76	59	47	38	31	25	21
.99	7,341	1,829	808	451	286	196	142	106	82	65	52	42	34	28

Note: Based on Arno Ouwehand's Power Calculator 2, available via UCLA Department of Statistics (http://calculators.stat.ucla.edu).

TABLE 12.8
Planning the sample size

Power	$r = .1$	$r = .2$
.80	784	195
.85	896	222
.90	1048	260

between $r = .1$ and .3. Using the
mary like that in Table 12.8 to
.80 and $p = .05$ two-tailed, the
86, depending on whether we
respectively. If subjects were rea
more comfortable setting the po
need 896, 222, or 98 total subject
set at .90, we will need a total N
or .3, respectively.

DIFFERENCES BETWEEN CORRELATION COEFFIC

Cohen operationalized the differe
by the effect size index q, where

which is the difference between t
The Fisher z_r (Equation 11.2) mak
detectable, because equal differe
instance, the difference between
statistically than is the difference
are also more accurate when Fishe
Borodkin, 1989). As Table 12.3
need, respectively, 177, 263, and 1
effect size and alpha indicated in

Why should it be so difficu
r and another r when it is so mu
of one r and zero? The answer li
around a second observed r and t
course, has no confidence interva
be overcome." Consider an r of .
r is 2.06, and $p < .05$ two-tailed
r is between .01 and .54. There is
pare this r with another study wit
The confidence interval of the

TABLE 12.8
Planning the sample size

Power	$r = .1$	$r = .2$	$r = .3$
.80	784	195	86
.85	896	222	98
.90	1048	260	114

between $r = .1$ and .3. Using the information in Table 12.7, we might create a summary like that in Table 12.8 to help us plan our study. To work with a power of .80 and $p = .05$ two-tailed, the total number of subjects we need is 784, 195, or 86, depending on whether we want to place our bet on the r of .1, .2, or .3, respectively. If subjects were readily available and not costly to run, we might be more comfortable setting the power higher than .80. With the power set at .85, we need 896, 222, or 98 total subjects, given an r of .1, .2, or .3, respectively. With power set at .90, we will need a total N of 1,048, 260, or 114 subjects, given an r of .1, .2, or .3, respectively.

DIFFERENCES BETWEEN CORRELATION COEFFICIENTS

Cohen operationalized the difference between two independent correlation coefficients by the effect size index q, where

$$\text{Cohen's } q = z_{r_1} - z_{r_2}, \tag{12.6}$$

which is the difference between the Fisher z_r transformations associated with each r. The Fisher z_r (Equation 11.2) makes equal differences between Fisher z_r values equally detectable, because equal differences between rs are not equally detectable. For instance, the difference between .90 and .70 in units of r is much more detectable statistically than is the difference between .40 and .20. Tests of significance among rs are also more accurate when Fisher's z_r transformation is used (Alexander, Scozzaro, & Borodkin, 1989). As Table 12.3 shows, to achieve a power level of .80, we would need, respectively, 177, 263, and 1573 units for *each* r for the combinations of expected effect size and alpha indicated in Sections A, B, and C.

Why should it be so difficult to detect the difference between the value of one r and another r when it is so much easier to detect the difference between the value of one r and zero? The answer lies in the difference between the confidence interval around a second observed r and that around a theoretical value of zero. The latter, of course, has no confidence interval, but the former has a real confidence interval "to be overcome." Consider an r of .30 based on an N of 45. The t associated with this r is 2.06, and $p < .05$ two-tailed. The 95% confidence interval around the obtained r is between .01 and .54. There is no overlap with zero. Suppose we wanted to compare this r with another study with an r of zero based on the same sample size of 45. The confidence interval of the latter r ranges from $-.29$ to $+.29$ and overlaps

NOTE: These Pages are Reference for Power Analysis Chapter

to rule out the prospect that a theory still waiting to be created might better account for all the existing results. Nonetheless, (g) as no experiment is entirely free of all alternative explanations, those known and those waiting to be discovered, both findings consistent and findings inconsistent with a theory's predictions can have probative value (Brinberg, Lynch, & Sawyer, 1992). Of course, for all these criteria to be applicable, the theory must also be precisely articulated so there will be no confusion or disagreement about what is asserted or predicted (H. A. Simon, 1979; Simon & Groen, 1973).

TYPE I AND TYPE II DECISION ERRORS

We turn now to null hypothesis significance testing (NHST), the dichotomous decision-making process in which a hypothesis to be nullified (called the **null hypothesis,** symbolized as $\mathbf{H_0}$) is contrasted with a specific working hypothesis (called the **alternative hypothesis,** $\mathbf{H_1}$). In most cases in behavioral research, the H_0 implies that no relationship between two variables is present in the population from which the sample data were drawn, or that there is no difference in the responses of treated and untreated subjects to an experimental manipulation, whereas H_1 does imply a relationship or real difference. Table 2.3 is a traditional way of representing four possible outcomes of NHST. The mistake of rejecting H_0 when it is true and should not have been rejected is called **Type I error,** and the mistake of not rejecting H_0 when it is false and should have been rejected is called **Type II error.** The ***p* value** (or **significance level**) indicates the probability of Type I error and is denoted as **alpha (α)** when *p* has been stipulated in advance (as a threshold or cutoff point). The probability of Type II error is symbolized as **beta (β). Confidence,** defined as $1 - \alpha$, is the probability of not making a Type I error. The **power of a test,** or $1 - \beta$, indicates the probability of not making a Type II error (i.e., the sensitivity of the significance test in providing an adequate opportunity to reject H_0 when it warrants rejection). As NHST is now construed, it is a hybrid

TABLE 2.3
Four outcomes involving the decision to reject or not to reject the null hypothesis (H_0)

	Actual state of affairs	
Scientist's decision	**Null hypothesis is true**	**Null hypothesis is false**
Reject null hypothesis	Type I error refers to a decision to reject H_0 when it is true and should not be rejected. Alpha (α) is the probability of Type I error.	No error. Statistical power ($1 - \beta$) refers to the probability of not making a Type II error.
Do not reject null hypothesis	No error. The confidence level ($1 - \alpha$) refers to the probability of not making a Type I error.	Type II error refers to a failure to reject H_0 when it is false and should be rejected. Beta (β) is the probability of Type II error.

endeavor that evolved out of the work (and arguments) of several different statisticians (for historical accounts, see Gigerenzer, Swijtink, Porter, Daston, Beatty, & Krüger, 1989; Stigler, 1986).

It has been well documented that behavioral researchers (as well as many other scientists who use NHST) have gotten into the habit of worrying more about Type I errors than about Type II errors. Some philosophers have suggested that this greater concern over Type I errors reflects the "healthy skepticism" of the scientific method (Axinn, 1966; Kaplan, 1964), the idea being that Type I error is an error of "gullibility," and Type II error is an error of "blindness." An analogy of Wainer's (1972) helps illustrate what is implied by this use of the terms *gullibility* and *blindness*. Suppose you were walking along the street and a shady character approached and said he had a quarter to sell you "for only five dollars." You might say to yourself, "He must think I'm stupid to ask me to hand over five dollars for an ordinary quarter." As if reading your mind, he says, "Don't think it's an ordinary quarter, pal, but one with special properties that make it worth five dollars. This quarter doesn't just come up heads and tails equally often; it is a biased coin. If you're as shrewd as you look, you're going to win fame and fortune by simply betting on which outcome is the more likely."

In this illustration we will think of the alternative hypothesis (H_1) as predicting that the probability of heads is not equal to the probability of tails in the long run. And since the null (H_0) hypothesis and the alternative hypothesis are mutually exclusive, we think of H_0 as predicting that the probability of heads *is* equal to the probability of tails in the long run. Thus, if H_0 is true, H_1 cannot be true. In Table 2.4 we have recast this situation into the framework of Table 2.3. The implication of Table 2.4 is that "Type I error" is analogous to being "taken in" by a false claim that the coin is biased when it is merely an ordinary coin (i.e., an error of gullibility), whereas "Type II error" is analogous to failing to see that the coin is biased as claimed (i.e., an error of blindness). We could subject H_1 to an empirical test by flipping the coin a large number of times and recording each time whether it landed heads or tails. We could state a particular probability (p value) as our alpha rejection criterion and be as stringent as we like in setting such a rejection criterion. However, we may eventually pay for this decision by failing to reject what we perhaps should reject.

TABLE 2.4
Example illustrating definitions of type I and type II errors

	Actual state of affairs	
Your decision	**The coin is fair**	**The coin is not fair**
The coin is not fair (i.e., it can win you fame and fortune, since it will not come up heads and tails equally).	Error of "gullibility"	No error
The coin is fair (i.e., it cannot win you fame and fortune, since it is just an ordinary coin).	No error	Error of "blindness"

STATISTICAL SIGN
EFFECT SIZE

Because a complete acc
report not just the p val
relationship between th
refer to again in the sec

Signific

In other words, the large
sampling units, or the la
(e.g., t, F, χ^2) and, there
is true unless the size of t
N) will not produce a re
effect sizes of exactly ze
of this general relationsh
the third can be determi
positive conclusion (i.e.,
be, we can readily deter
level of statistical powe

In fact, any parti
definitions of the effect
For example, if we wer
we could write

where $\chi^2_{(1)}$ is a chi-squ
counts), ϕ^2 is the squa
ship in the row catego
(scored 1 or 0), and
(e.g., found in the cell
in this book.)

Were we intereste
would have a choice of

and

where, in Equation 2.2
(scored 1 or 0) and c
represented as Cohen's
the pooled standard de
(usually $N - 2$). (We

STATISTICAL SIGNIFICANCE AND THE EFFECT SIZE

Because a complete account of "the results of a study" requires that the researcher report not just the p value but also the effect size, it is important to understand the relationship between these two quantities. The general relationship, which we will refer to again in the second half of this book, is given by

$$\text{Significance test} = \text{Size of effect} \times \text{Size of study}.$$

In other words, the larger the study in terms of the total number (N) of observations or sampling units, or the larger the effect size, the larger the value of the significance test (e.g., t, F, χ^2) and, therefore, the smaller (and usually more coveted) the p value. This is true unless the size of the effect is truly zero, in which case a larger study (i.e., a larger N) will not produce a result that is any more significant than a smaller study (although effect sizes of exactly zero are rarely seen in behavioral research). A further implication of this general relationship is that if we are able to specify any two of these three factors, the third can be determined. Thus, if we know the level of risk of drawing a spuriously positive conclusion (i.e., the p value) and can estimate what the size of the effect will be, we can readily determine how large a total sample we will need to achieve a desired level of statistical power. (We show how in chapter 12.)

In fact, any particular test of significance can be obtained by one or more definitions of the effect size multiplied by one or more definitions of the study size. For example, if we were interested in chi-square (discussed in detail in chapter 19), we could write

$$\chi^2_{(1)} = \phi^2 \times N, \tag{2.1}$$

where $\chi^2_{(1)}$ is a chi-square on 1 degree of freedom (e.g., from a 2 × 2 table of counts), ϕ^2 is the squared Pearson product-moment correlation between membership in the row category (scored 1 or 0) and membership in the column category (scored 1 or 0), and N (the study size) is the total number of sampling units (e.g., found in the cells of the 2 × 2 table). (We will see Equation 2.1 again later in this book.)

Were we interested in t as a test of significance (discussed in chapter 12), we would have a choice of many equations (Rosenthal, 1991a, 1994b), of which two are

$$t = \frac{r}{\sqrt{1 - r^2}} \times \sqrt{df} \tag{2.2}$$

and

$$t = \frac{M_1 - M_2}{\sigma_{\text{pooled}}} \times \frac{\sqrt{df}}{2}, \tag{2.3}$$

where, in Equation 2.2, r is the point-biserial Pearson r between group membership (scored 1 or 0) and obtained score. In Equation 2.3, the effect size indicator is represented as Cohen's d (i.e., the difference between means, M_1 and M_2, divided by the pooled standard deviation, σ). In both equations, df is the degrees of freedom (usually $N - 2$). (We will see Equations 2.2 and 2.3 later in this book as well.)

When the relationship between statistical significance and the effect size is understood, it is less likely that researchers who employ NHST will do significance testing with low power. In the 1960s and later, Jacob Cohen hammered this point home in articles and a handy reference text on the analysis of statistical power (e.g., Cohen, 1962, 1965, 1988). To illustrate, suppose Smith conducts an experiment (with $N = 80$) to show the effects of leadership style on productivity and finds that Style A is better than Style B. Jones, however, is skeptical (because he invented Style B) and repeats Smith's study with $N = 20$. Although Jones's results are clearly in the same direction as Smith's, Jones nevertheless reports a "failure to replicate" because his t was only 1.06 ($df = 18$, $p > .30$), whereas Smith's t was 2.21 ($df = 78$, $p < .05$). Although it is certainly true that Jones has not replicated Smith's t test result or p value, the magnitude of the effect obtained by Jones (as measured by the Pearson correlation statistic) is $r = .24$, which is exactly the size of the effect in Smith's study! In other words, Jones has found exactly the same relationship that Smith found even though the obtained t and p values of the two studies are not very close. Because Jones's total sample size (N) was so much smaller than Smith's total sample size, Jones's *power* to reject at $p = .05$ is substantially less than Smith's power. In this case the power (i.e., $1 - \beta$) of Jones's t test is .18, whereas the power of Smith's t test is .57.

TWO FAMILIES OF EFFECT SIZES

Two of the most important families of effect sizes in behavioral and social science are the *correlation family* and the *difference family,* and we will discuss in more detail examples of each of these classes later in this book. There is also a third family, which we call the *ratio family,* and within these three families there are subtypes as well (Rosnow & Rosenthal, 2003). Three primary members of the difference family are Cohen's d (i.e., the effect size component of Equation 2.3), Hedges's g, and Glass's Δ. All three of these effect size indices employ the same numerator (the difference between the means of the two groups that are being compared), but each uses a slightly different denominator:

$$\text{Cohen's } d = \frac{M_1 - M_2}{\sigma_{\text{pooled}}} \tag{2.4}$$

$$\text{Hedges's } g = \frac{M_1 - M_2}{S_{\text{pooled}}} \tag{2.5}$$

$$\text{Glass's } \Delta = \frac{M_1 - M_2}{S_{\text{control}}}, \tag{2.6}$$

with terms in Cohen's d as indicated previously (Equation 2.3). In Equation 2.5, S is the square root of the pooled unbiased estimate of the population variance. In Equation 2.6, the S_{control} is like the S in the denominator of Hedges's g, but it is computed only for the control group. Computing S only from the control group is a useful procedure when we know or suspect that the treatment may affect not only the mean but also the variance of the scores in the treatment condition.

The second impo
Equation 2.1 we noted
which is a special case
dichotomous. In Equation
(r_{pb}), which is the Pearso
and a continuous variab
transformation of r) and
(called the *coefficient of*
η^2 (eta squared). Becaus
treatment helping or hurt
use as effect size indices
essential. There are seve
than squared indices, and
a little, another reason is
value of small effect size

To illustrate, at a sp
to end, prematurely, a ra
in reducing heart attacks
Group, 1988). The reaso
dantly clear that aspirin
thus it would have been
The subjects in that study
had been given an ordin
der of whom (11,034) ha
are shown in Table 2.5.
who did or did not suffe
heart attack group. And
so dramatic as to call fo
compute the phi coefficie
size $r = .034$, and thus

TABLE 2.5
Aspirin's effect on hear

A. Myocardial infarctions in a	
Condition	**No he**
Aspirin	1
Placebo	1

B. Fatal and nonfatal myocard	
Condition	**Lived**
Aspirin	99
Placebo	171

The second important family of effect sizes is the *correlation family*. In Equation 2.1 we noted a popular incarnation of this family, the phi coefficient (ϕ), which is a special case of the Pearson product-moment r when both variables are dichotomous. In Equation 2.2 we noted another special case, the point-biserial correlation (r_{pb}), which is the Pearson product-moment correlation between a dichotomous variable and a continuous variable. Also included in the correlation family is z_r (the Fisher transformation of r) and various squared indices of r and r-like quantities such as r^2 (called the *coefficient of determination*), Ω^2 (omega squared), ε^2 (epsilon squared), and η^2 (eta squared). Because squared correlational indices lose their directionality (Is the treatment helping or hurting, is the correlation positive or negative?), they are of little use as effect size indices in scientific work in which information on directionality is essential. There are several other reasons that we prefer the product-moment r rather than squared indices, and we explain those reasons in chapters 11 and 12. To anticipate a little, another reason is that squared indices can be misleading in terms of the practical value of small effect sizes.

To illustrate, at a specially called meeting held in December 1987, it was decided to end, prematurely, a randomized double-blind experiment on the effects of aspirin in reducing heart attacks (Steering Committee of the Physicians' Health Study Research Group, 1988). The reason for this unusual termination was that it had become abundantly clear that aspirin prevented heart attacks (and deaths from heart attacks), and thus it would have been unethical to continue to give the control subjects a placebo. The subjects in that study were 22,071 male physicians, roughly half of whom (11,037) had been given an ordinary aspirin tablet (325 mg) every other day, and the remainder of whom (11,034) had been given a placebo. A portion of the results of the study are shown in Table 2.5. Part A shows the number of participants in each condition who did or did not suffer a heart attack, and Part B shows the survival rates in the heart attack group. And what was the magnitude of the experimental effects that were so dramatic as to call for the termination of that research? To find the answer, we compute the phi coefficient on the raw data in this table. In Part A, we find the effect size $r = .034$, and thus the corresponding $r^2 = .00$ or, to four decimal places, .0012.

TABLE 2.5
Aspirin's effect on heart attack

A. Myocardial infarctions in aspirin and placebo conditions

Condition	No heart attack	Heart attack
Aspirin	10,933	104
Placebo	10,845	189

B. Fatal and nonfatal myocardial infarctions

Condition	Lived	Died
Aspirin	99	5
Placebo	171	18

In Part B, the effect size $r = .084$, with a corresponding $r^2 = .0071$. Using a simple method described in more detail in chapter 11 (the binomial effect size display, or BESD), we would find that these r values imply a 3.4% greater success rate for aspirin than for placebo in preventing a heart attack and an 8.4% greater success rate for preventing death when a heart attack has occurred. The point is that, had we considered only the squared rs, we might have concluded there were *no benefits* of taking aspirin, a costly mistake to make in terms of human lives saved.

It is sometimes necessary to decide how to convert effect size indices to one particular index (e.g., in meta-analytic work, discussed in chapter 21). In that situation, there are several reasons to view the family of correlational indices as a more generally useful group of effect size measures. Suppose the data came to us as rs. We would not ordinarily want to convert rs to ds, gs, or Δs, because the concept of a mean difference index makes little sense in describing a linear relationship over a great many values of the independent variable of interest. On the other hand, if we were working with effect sizes reported as ds, gs, or Δs, the r index (as we show in chapter 11) makes perfectly good sense in its point-biserial form (two levels of the independent variable of interest). If the data were structured in a 2×2 table of counts, the phi form of the effect size index would be suitable. However, suppose the design involved more than two conditions. For example, suppose a hypothesis called for five levels of arousal, and the scientist predicted better performance on the outcome measure at the middle levels of arousal than at the more extreme levels, and the very best performance in the midmost level of arousal. The magnitude of an effect associated with a curvilinear trend is quite naturally indexed by r (discussed in chapter 15), but not so naturally by d, g, or Δ.

INTERVAL ESTIMATES AROUND EFFECT SIZES

Earlier we also mentioned the importance of reporting interval estimates along with effect size estimates. For example, the confidence interval of the effect size is the margin of error that surrounds the obtained value of the effect size index. For example, a 95% confidence interval around the obtained effect size r of .16 might range from a lower limit r of .10 to an upper limit r of .22. Our interpretation of this 95% confidence interval would be that there is a 95% chance the population value of the r that our obtained effect size r of .16 was trying to estimate falls between the lower and upper limits of .10 and .22. Of course, researchers need not restrict themselves to only 95% confidence intervals if they prefer working with more (or less) stringent levels of confidence. Decreasing the confidence level from 95% to 90% will shrink the interval, and vice versa. Increasing the size of the study (i.e., working with a larger total sample size) will also shrink the confidence interval.

Another type of interval estimate (described in chapter 11) is called the **null-counternull interval** (Rosenthal & Rubin, 1994). This interval estimate is based on the actual p value rather than on a previously specified alpha. The "null" anchoring one end of the interval is the effect size that is associated with the null hypothesis (and is typically zero); *counternull* refers to the non-null magnitude of the effect size that is larger than the obtained effect size and is supported by the same amount of evidence as the null val
value) can alert a resear
In this way it provides
reject the null hypothes

SUMMING UP

We will have more to sa
the results of a study alv
estimate, the researchers
behavioral and social res
than Type I error (e.g., B
Waechter, & Solomon, 1
II error can be reduced
effect size. If the estima
researcher would do we
(i.e., that "nothing happe
point to a very small effe
does not reach the resea
concluding that no nontr

Table 2.6 summari
tion of the results of sign
1991a). Suppose a nonsi
may have led to failure
probably be continued v
"nothing happened." Ha
much smaller total sampl

TABLE 2.6
Population effect sizes
ing as determinants of

Population effect size
Zero
Small
Large

[a]Low power may lead to failure to
is quite small, the costs of this err
[b]Although this is not an inferentia
very large, we may mistake a resu
of practical importance.
[c]Low power may lead to failure
true effect the costs may be very

evidence as the null value of the effect size. This interval (null value to counternull value) can alert a researcher to whether a conclusion of "no effect" might be in error. In this way it provides some protection against mistaken interpretations of failure to reject the null hypothesis (Rosenthal & Rubin, 1994; Rosnow & Rosenthal, 1996a).

SUMMING UP

We will have more to say about all these topics later on. The vital point here is that, if the results of a study always include both an estimate of the effect size and an interval estimate, the researchers better protect themselves against Type I and Type II errors. In behavioral and social research, there is little doubt that Type II error is far more likely than Type I error (e.g., Brewer, 1972; Chase & Chase, 1976; Cohen, 1962, 1988; Haase, Waechter, & Solomon, 1982; Sedlmeier & Gigerenzer, 1989). The frequency of Type II error can be reduced drastically by our attention to the magnitude of the estimated effect size. If the estimate is large and the researcher finds a nonsignificant result, the researcher would do well to avoid concluding that variables X and Y are not related (i.e., that "nothing happened"). Only if the pooled results of a good many replications point to a very small effect (on the average), and to a combined test of significance that does not reach the researcher's preferred alpha level, would a researcher be justified in concluding that no nontrivial relationship exists between X and Y.

Table 2.6 summarizes decision errors and possible consequences as a joint function of the results of significance testing and the population effect size (Rosenthal, 1983, 1991a). Suppose a nonsignificant effect. What should it tell the researcher? Low power may have led to failure to detect the true effect, and this line of investigation should probably be continued with a larger sample size before the researcher concludes that "nothing happened." Had the medical researchers in the aspirin study worked with a much smaller total sample, they would not have gotten statistical significance: It would

TABLE 2.6
Population effect sizes and results of significance testing as determinants of inferential errors

	Results of significance testing	
Population effect size	Not significant	Significant
Zero	No error	Type I error
Small	Type II error[a]	No error[b]
Large	Type II error[c]	No error

[a]Low power may lead to failure to detect the true effect; however, if the true effect is quite small, the costs of this error may not be very great.

[b]Although this is not an inferential error, if the effect size is very small and N is very large, we may mistake a result that is merely very significant for one that is of practical importance.

[c]Low power may lead to failure to detect the true effect, and with a substantial true effect the costs may be very great.

have been like trying to read small print in a very dim light and finding it harder to make out the information. On the other hand, suppose a significant but small effect. What should it tell the researcher? The answer depends on what the researcher considers the practical importance of the small estimated population effect. In the aspirin study, even a "quite small" effect was considered important, because the criterion was "who lives and who dies." The lesson is that a test of significance without an effect size estimate fails to tell the whole story. Fortunately, as we shall see later, just from the basic information that many journals require scientists to report, effect sizes (and interval estimates) can usually be directly derived even from the barest of raw ingredients (e.g., Rosenthal & Rubin, 2003; Rosnow & Rosenthal, 1996a).

Finally, we also want to mention a new statistic proposed by Peter Killeen (2005) that increases the utility of p values. This statistic, called p_{rep}, gives the probability that a same size replication (e.g., of a treatment vs. control group study) will obtain an effect in the same direction as did the original study.

Killeen's equation for estimating p_{rep} is:

$$p_{rep} = \frac{1}{1 + \left(\frac{p}{1-p}\right)^{2/3}} \tag{2.7}$$

where p is the obtained significance level. Table 2.7 shows for 15 p values the corresponding estimates of p_{rep}. It should be noted that p_{rep} is not an effect size index, nor is it intended to be. But it does give us far more useful information to learn that there is an 88%, 96%, or 99% chance of obtaining the same direction of result on replication (assuming the context and the experimental circumstances are relatively unchanged) than that our p values are .05, .01, or .001. In the end, of course, significance tests and their associated p values alone are not nearly as informative as estimates of effect sizes along with their corresponding interval estimates (e.g., 95% confidence intervals), but p_{rep} is a useful advance as well.

TABLE 2.7
Probabilities of replicating the direction of treatment effects (p_{rep}) from obtained p values

p value	p_{rep}	p value	p_{rep}
.50	.500	.01	.955
.40	.567	.005	.971
.30	.638	.001	.990
.20	.716	.0005	.994
.15	.761	.0001	.998
.10	.812	.00005	.999
.05	.877	.00001	.9995
.025	.920		

PUZZLES AND PROB

As mentioned in chapter 2, and took a position at the gave an invited talk at Caml the club was a renowned pr whose views dominated Bı was another eminent Caml seminal work had been ar Wittgenstein came to regar as antediluvian). Usually at nary remarks, and Wittger Wittgenstein harbored a de had taken on the role of a virtually unknown in Britai often, someone would pok up a little more heat.

What ensued that da entitled *Wittgenstein's Poke* sial incident that occurre Wittgenstein, who had grab