**CHAPTER 5 - The Importance of Effect Magnitude [Roger E. Kirk](http://search.credoreference.com.proxy1.ncu.edu/entry.do?id=5062797)**

**from** [**Blackwell Handbook of Research Methods in Experimental Psychology**](http://search.credoreference.com.proxy1.ncu.edu/content/title/bkhrmep)

This chapter examines the role of measures of effect magnitude in psychological research. Measures of effect magnitude fall into one of three categories as shown in Table 5.1. The categories are (a) measures of effect size (typically, standardized mean differences), (b) measures of strength of association, and (c) other measures. The measures are used for three purposes: integrating the results of empirical research studies in meta-analyses, supplementing the information provided by null hypothesis significance tests, and determining whether research results are practically significant. Practical significance is concerned with the usefulness of results. Statistical significance, the focus of null hypothesis significance tests, is concerned with whether results are due to chance or sampling variability. Null hypothesis significance testing was developed between 1915 and 1933 by three men: Ronald A. Fisher (1890-1962), Jerzy Neyman (1894-1981), and Egon S. Pearson (1895-1980). Fisher, who was employed as a statistician at a small agricultural research station 25 miles north of London, was primarily responsible for the new paradigm and for advocating 0.05 as the standard significance level . Neyman was primarily responsible for introducing confidence intervals in the 1930s -an alternative approach to statistical inference . For over 70 years, null hypothesis significance testing has been the cornerstone of research in psychology. [Cohen (1990)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.17.-#s.6.17.-) observed that “The fact that Fisher’s ideas quickly became the basis for statistical inference in the behavioral sciences is not surprising -they were very attractive. They offered a deterministic scheme, mechanical and objective, independent of content, and led to clear-cut yes-no decisions” (p. 1307). In spite of these apparent advantages, null hypothesis significance testing has been surrounded by controversy. The acrimonious exchanges between Fisher and Neyman that began in 1935 set the pattern for the debate that has continued to this day ([Box, 1978](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.7.-#s.6.7.-), p. 263). One of the earliest serious challenges to the logic and usefulness of null hypothesis significance testing appeared in a 1938 article by Joseph Berkson. Since then there has been a crescendo of challenges ([Bakan, 1966](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.2.-#s.6.2.-); [Carver, 1978](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.10.-#s.6.10.-), [1993](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.11.-#s.6.11.-); [Cohen, 1990](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.17.-#s.6.17.-), [1994](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.19.-#s.6.19.-); [Falk & Greenbaum, 1995](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.29.-#s.6.29.-); [Hunter, 1997](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.48.-#s.6.48.-); [Meehl, 1967](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.63.-#s.6.63.-); [Rozeboom, 1960](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.79.-#s.6.79.-); [Schmidt, 1996](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.81.-#s.6.81.-); and [Shaver, 1993](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.86.-#s.6.86.-)).

In this chapter I examine some of the criticism of null hypothesis significance testing and describe ways to supplement the paradigm. The fifth edition of the *Publication Manual of the American Psychological Association* (2001) explicitly recognizes that null hypothesis significance tests and p values tell only part of the story.

Neither of the two types of probability value [significance level and p value] directly reflects the magnitude of an effect or the strength of a relationship. For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section. . . . The general principle to be followed, however, is to provide the reader not only with information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship. ([APA, 2001](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.1.-#s.6.1.-), pp. 25-6)

Three Criticisms of Null Hypothesis Significance Testing

What are the major criticisms of classical null hypothesis significance testing? Three criticisms are frequently mentioned.

Answering the wrong question

[Cohen (1994)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.19.-#s.6.19.-) and others have criticized null hypothesis significance testing on the grounds that it doesn’t tell researchers what they want to know ([Berger & Berry, 1988](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.4.-#s.6.4.-); [Carver, 1978](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.10.-#s.6.10.-); [Dawes, Mirels, Gold, & Donahue, 1993](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.25.-#s.6.25.-); [Falk, 1998](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.28.-#s.6.28.-)). To put it another way, null hypothesis significance testing and scientific inference address different questions. In scientific inference, what we want to know is the probability that the null hypothesis (H0) is true given that we have obtained a set of data (D); that is, p(H0|D). What null hypothesis significance testing tells us is the probability of obtaining these data or more extreme data if the null hypothesis is true, p(D|H0). Unfortunately for researchers, obtaining data for which p(D|H0) is low does not imply that p(H0|D) also is low. [Falk (1998)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.28.-#s.6.28.-) pointed out that p(D|H0) and p(H0|D) can be equal but only under rare mathematical conditions. Researchers incorrectly reason that if the p value associated with a test statistic is suitably small, say less than 0.05, the null hypothesis is probably false. This form of deductive reasoning has been referred to by [Falk and Greenbaum (1995)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.29.-#s.6.29.-) as the “illusion of probabilistic proof by contradiction.” The logic underlying this form of reasoning has been examined by [Falk and Greenbaum (1995)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.29.-#s.6.29.-) and [Nickerson (2000)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.64.-#s.6.64.-). Associated with this form of reasoning are the incorrect, widespread beliefs that (a) the p value is the probability that the null hypothesis is true, and (b) the complement of the p value is the probability that the alternative hypothesis is true. [Nickerson (2000)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.64.-#s.6.64.-) summarized other common misconceptions regarding null hypothesis significance testing including the beliefs that (a) a small p value is indicative of a large treatment effect, (b) the complement of the p value is the probability that a significant result will be found in a replication, (c) statistical significance is indicative of practical significance, (d) failure to reject the null hypothesis is equivalent to demonstrating that it is true, and (e) a small value of p(D|H0) implies that p(D|H1) must be large, where H1 denotes the alternative hypothesis.

All null hypotheses are false

A second criticism of null hypothesis significance testing is that it is a trivial exercise. As [John Tukey (1991)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.92.-#s.6.92.-) wrote, “the effects of A and B are always different - in some decimal place - for any A and B. Thus asking ‘Are the effects different?’ is foolish” (p. 100). More recently, Jones and Tukey reiterated this view.

For large, finite, treatment populations, a total census is at least conceivable, and we cannot imagine an outcome for which μA - μB = 0 when the dependent variable (or any other variable) is measured to an indefinitely large number of decimal places. . . . The population mean difference may be trivially small but will always be positive or negative. ([Jones & Tukey, 2000](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.49.-#s.6.49.-), p. 412)

The view that null hypotheses are never true except those we construct for Monte Carlo tests of statistical procedures is shared by many researchers ([Bakan, 1966](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.2.-#s.6.2.-); [Berkson, 1938](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.5.-#s.6.5.-); [Cohen, 1990](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.17.-#s.6.17.-); [Harris, 1994](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.40.-#s.6.40.-), p. 21; [Thompson, 1998](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.90.-#s.6.90.-)). Hence, because type I errors cannot occur, statistically significant results are assured if large enough samples are used. [Thompson (1998)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.90.-#s.6.90.-) captured the essence of this view when he wrote, “Statistical testing becomes a tautological search for enough participants to achieve statistical significance. If we fail to reject, it is only because we’ve been too lazy to drag in enough participants” (p. 799). It is ironic that a ritualistic adherence to null hypothesis significance testing has led researchers to focus on controlling the type I error that cannot occur because all null hypotheses are false while allowing the type II error that can occur to exceed acceptable levels, often as high as 0.50 to 0.80 ([Cohen, 1962](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.14.-#s.6.14.-), [1969](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.15.-#s.6.15.-)).

[Rindskopf (1997)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.70.-#s.6.70.-) has argued that most researchers are not really interested in the possibility that the population effect is precisely zero, but rather in whether the effect is close enough to zero to be of no interest. This has led some researchers to suggest that instead of testing a point null hypothesis - for example, the effect is zero - researchers should test a range null hypothesis that involves designating a range of values that is considered effectively null ([Serlin, 1993](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.83.-#s.6.83.-); [Serlin & Lapsley, 1985](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.84.-#s.6.84.-), [1993](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.85.-#s.6.85.-); [Yelton & Sechrest, 1986](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.98.-#s.6.98.-)). According to [Cortina and Dunlap (1997)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.21.-#s.6.21.-), the adoption of this strategy would rarely lead to a different outcome than current practice.

Other modifications to null hypothesis significance tests have been suggested. The traditional two-tailed test does not permit a conclusion about the direction of an effect, although most researchers do draw such a conclusion. [Bohrer (1979)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.6.-#s.6.6.-), [Harris (1994](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.40.-#s.6.40.-), [1997)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.41.-#s.6.41.-), and [Kaiser (1960)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.51.-#s.6.51.-) have suggested a modification that permits conclusions about directionality. The modification, called a three-outcome test, consists of replacing the alternative hypothesis, say H1: μ1 ≠ μ2, with two alternatives: H> : μ1 > μ2 and H< : μ1 < μ2. The modified test permits a researcher to conclude that μ1 > μ2 or μ1 < μ2 or, if the null hypothesis is not rejected, that the direction of the difference between μ1 and μ2 is indeterminate. Although the three-outcome test was recommended as early as 1960, [Hunter (1997)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.48.-#s.6.48.-) reported that it has not found much acceptance. Recently, [Jones and Tukey (2000)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.49.-#s.6.49.-) proposed a similar three-alternative conclusion procedure that eliminates the null hypothesis. It is too early to determine if their proposal will find favor among researchers.

Traditional one-tailed tests do not permit researchers to conclude that an effect is statistically significant if the direction of the effect is opposite to that predicted even though the test statistic would fall in the rejection region if the prediction were reversed. [Braver (1975)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.8.-#s.6.8.-) and [Nosanchuk (1978)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.65.-#s.6.65.-) described a modification of the alternative hypothesis that circumvents the problem. They pointed out that one- and two-tailed tests are the limiting cases of split-tail tests. In a split-tail test, c × (significance level) is allocated to the predicted tail of the sampling distribution of the test statistic where c is a proportion and (1 - c) × (significance level) is allocated to the opposite tail. For example, if the alternative hypothesis states that μ1 > μ2, c = 0.8, and α = 0.5, the upper boundary of the nonrejection region is the 96th percentile of, say, the t distribution and the lower boundary the first percentile. The significance level is (1 - 0.8)(0.05) + (0.8)(0.05) = 0.01 + 0.04 = 0.05. In this example, evidence against the research hypothesis must be four times as strong under the null hypothesis before the researcher concludes that μ1 < μ2 than it has to be to conclude that μ1 > μ2. A traditional two-tailed test can be thought of as a special case of a split-tail test in which c = 0.5. If c = 0.5, the size of the two critical regions is the same. A one-tailed test occurs when c = 1, resulting in infinite bias [c /(1 - c) arbitrarily large] against concluding that the research hypothesis has placed the rejection region in the wrong tail. The choice of c presumably reflects the researcher’s prior subjective probabilities regarding the direction of the population difference. Values of c > 0.5, but < 1 are reasonable. The choice of c = 1 does not leave room for the possibility that the researcher’s prediction could be wrong and is avoided in this approach. This modification, like the three-outcomes test has found little acceptance among researchers.

Making a dichotomous decision from a continuum of uncertainty

A third criticism of null hypothesis significance testing is that by adopting a fixed level of significance, a researcher turns a continuum of uncertainty into a dichotomous reject-do not reject decision ([Frick, 1996](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.31.-#s.6.31.-); [Grant, 1962](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.35.-#s.6.35.-); [Rossi, 1997](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.78.-#s.6.78.-); [Wickens, 1998](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.94.-#s.6.94.-)). A p value only slightly larger than the level of significance is treated the same as a much larger p value. Some researchers attempt to blur the reject-do not reject dichotomy with phrases such as “the results approached significance” or “the results were marginally significant.” However, studies of the way psychologists interpret p values find a “cliff effect” at 0.05 in which reported confidence in research findings drops perceptibly when p becomes larger than 0.05 ([Beauchamp & May, 1964](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.3.-#s.6.3.-); [Rosenthal & Gaito, 1963](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.71.-#s.6.71.-), [1964](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.72.-#s.6.72.-)). The adoption of 0.05 as the dividing point between significance and nonsignificance is quite arbitrary. The comment by [Rosnow and Rosenthal (1989](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.76.-#s.6.76.-), p. 1277) is pertinent, “surely, God loves the .06 nearly as much as the .05.”

Beyond Null Hypothesis Significance Tests

These criticisms and others have led some researchers to call for a ban on significance testing ([Carver, 1978](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.10.-#s.6.10.-); [Hunter, 1997](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.48.-#s.6.48.-); [Schmidt, 1996](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.81.-#s.6.81.-)). Nickerson advocated a less radical position:

[null hypothesis significance testing] is arguably the most widely used method of analysis of data collected in psychology experiments and has been so for a long time. If it is misunderstood by many of its users in as many ways as its critics claim, this is an embarrassment for the field. A minimal goal for experimental psychology should be to attempt to achieve a better understanding among researchers of the approach, of its strengths and limitations, of the various objections that have been raised against it, and of the assumptions that are necessary to justify specific conclusions that can be drawn from its results. ([Nickerson, 2000](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.64.-#s.6.64.-), pp. 289-90)

It is clear that null hypothesis significance testing is open to criticisms, but then the alternatives are not perfect either.

Researchers want to answer three basic questions from their research: (a) Is an observed effect real or should it be attributed to chance? (b) If the effect is real, how large is it? and (c) Is the effect large enough to be useful? The first question concerning whether chance is a viable explanation for an observed effect is usually addressed with a null hypothesis significance test. A significance test tells the researcher the probability of obtaining the effect or a more extreme effect if the null hypothesis is true. The test doesn’t tell the researcher how large the effect is. This question is usually addressed with a descriptive statistic, confidence interval, and measure of effect magnitude. The third question concerning whether an effect is useful or practically significant is more difficult to answer. The answer requires a judgment that is influenced by a variety of considerations including the researcher’s value system, societal concerns, assessment of costs and benefits, and so on. One point is evident, statistical significance and practical significance address different questions. In the following sections, I describe the advantages of confidence intervals relative to null hypothesis significance tests, the importance of reporting measures of effect magnitude, and the accumulation of knowledge through meta-analysis.

Advantages of Confidence Intervals

Confidence intervals and null hypothesis significance tests are two complementary approaches to classical statistical inference. As Tukey pointed out, rejection of a null hypothesis is not very informative. We know in advance that the hypothesis is false. [Tukey (1991)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.92.-#s.6.92.-) stated that rejection of a two-sided null hypothesis simply means that a researcher is able to specify the direction of a difference. Failure to reject means that the researcher is unable to specify the direction. Is this any way to advance psychological knowledge and theory building? I think not. How far would physics have progressed if their researchers had focused on discovering ordinal relationships?

A descriptive statistic and confidence interval provide an estimate of the population parameter and a range of values - the error variation - qualifying that estimate. A 100(1 - α)% confidence interval for, say, μ1 - μ2 contains all of the values for which the null hypothesis, μ1 - μ2 = 0, would not be rejected at α level of significance. Values outside the confidence interval would be rejected. An important advantage of a confidence interval is that it requires the same assumptions and information as a null hypothesis significance test, but the interval provides much more information. Instead of simply knowing the direction of a difference as in a significance test, a confidence interval also provides a range of values within which the population parameter is likely to lie. Furthermore, a descriptive statistic and confidence interval use the same unit of measurement as the data. This facilitates the interpretation of results and makes trivial effects harder to ignore. Confidence intervals and measures of effect magnitude are especially useful in assessing the practical significance of results. However, in spite of these advantages, confidence intervals rarely appear in psychology journals. What we see in the journals is a reject-nonreject decision strategy that doesn’t tell researchers what they want to know and a preoccupation with p values that are several steps removed from examining the data. Perhaps the recommendation in the fifth edition of the *Publication Manual of the American Psychological Association* will result in greater use of confidence intervals.

The reporting of confidence intervals (for estimates of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results. Because confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy. The use of confidence intervals is therefore strongly recommended. ([APA, 2001](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.1.-#s.6.1.-), p. 22)

Measures of Effect Magnitude   
Effect size

In 1969 Cohen introduced the first effect size measure that was explicitly labeled as such. His measure, denoted by δ, expresses the size of a population contrast of means, say ψ = μE - μC, in units of the population standard deviation,

where μE and μC denote the population means of the experimental and control groups and σ denotes the common population standard deviation. The size of the contrast is influenced by the scale of measurement of the means. Cohen divided the contrast by σ to rescale the contrast in units of the amount of error variability in the data.

What made Cohen’s contribution unique is that he provided guidelines for interpreting the magnitude of δ:

 δ = 0.2 is a small effect  
 δ = 0.5 is a medium effect  
 δ = 0.8 is a large effect.

According to [Cohen (1992)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.18.-#s.6.18.-), a medium effect of 0.5 is visible to the naked eye of a careful observer. Several surveys have found that 0.5 approximates the average size of observed effects in various fields ([Cooper & Findley, 1982](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.20.-#s.6.20.-); [Haase, Waechter, & Solomon, 1982](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.39.-#s.6.39.-); [Sedlmeier & Gigerenzer, 1989](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.82.-#s.6.82.-)). A small effect of 0.2 is noticeably smaller than medium but not so small as to be trivial. A large effect of 0.8 is the same distance above medium as small is below it. These operational definitions turned Cohen’s measure of effect size into a much more useful statistic. For the first time, researchers had general guidelines for interpreting the size of treatment effects. The guidelines are particularly useful for researchers working in uncharted territory, for example, assessing the performance of animals in a new apparatus. An effect size is a valuable supplement to the information provided by a p value. A p value of 0.0001 loses its luster if the effect turns out to be trivial. Effect sizes also are useful for comparing and integrating the results of different studies. This application is described later in the section on Cumulating Knowledge through Meta-analysis.

The parameters of Cohen’s δ are rarely known. The sample means of the experimental and control groups are used to estimate μE and μC. An estimator of σ can be obtained in a number of ways. Under the assumption that σE and σC are equal, the sample variances of the experimental and control groups are pooled as follows

An estimator of δ is

where ȲE and ȲC denote, respectively, the sample mean of the experimental and control groups and δ̂Pooled denotes the pooled estimator of δ. Gene [Glass (1976)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.33.-#s.6.33.-) in his pioneering work on meta-analysis recommended using the sample standard deviation of the control group, δ̂C, to estimate δ. He reasoned that if there were several experimental groups and a control group, pairwise pooling of the sample standard deviations could result in different values of δ̂Pooled for each experimental-control contrast. Hence, the same size difference between experimental and control means would result in different effect sizes when the standard deviations of the experimental groups differed. Glass’s estimator of δ

where ȲEJ and ȲC denote, respectively, the sample means of the jth experimental group and control group. [Larry Hedges (1981)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.43.-#s.6.43.-) used a different approach to estimate σ. He observed that population variances are often homogeneous, in which case the most precise estimate of the population variance is obtained by pooling the j = 1, . . ., p sample variances. His pooled estimator

is identical to the square root of the within-groups mean square in a completely randomized analysis of variance. Hedges’ estimator of δ is

According to [Hedges (1981)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.43.-#s.6.43.-), all three estimators of δ - d, g′, and g - are biased. He recommended correcting g for bias as follows,

where J (N - 2) is the bias correction factor described in [Hedges and Olkin (1985](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.44.-#s.6.44.-), p. 80). The correction factor is approximately

[Hedges (1981)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.43.-#s.6.43.-) showed that gc is the unique, uniformly minimum variance-unbiased estimator of δ, and also described an approximate confidence interval for δ:

where zα/2 denotes the two-tailed critical value that cuts off the upper α/2 region of the standard normal distribution and

[Cumming and Finch (2001)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.24.-#s.6.24.-) describe procedures for obtaining exact confidence intervals using noncentral sampling distributions. The procedures require the use of special statistical software.

Cohen’s δ has a number of features that contribute to its popularity: (a) it is easy to understand and has a consistent interpretation across different research studies, (b) the sampling distributions of estimators of δ are well understood, and (c) estimators of δ can be readily computed from t statistics and F statistics with one degree of freedom that are reported in published articles. The latter feature is particularly attractive to researchers who do meta-analyses.

The correct conceptualization of the denominator of δ and its computation can be problematic when the treatment is a classification or organismic variable ([Grissom & Kim, 2001](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.37.-#s.6.37.-); [Olejnik & Algina, 2000](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.67.-#s.6.67.-)). For experiments with a manipulated treatment and random assignment of participants to j = 1, . . ., p levels of the treatment, the computation of an effect size such as gc is relatively straightforward. The denominator of gc is the square root of the within-groups mean square. This mean square provides an estimate of σ that reflects the variability of observations for the full range of the manipulated treatment. If, however, the treatment is an organismic variable such as gender, boys and girls, the square root of the within-groups mean square does not reflect the variability for the full range of the treatment because it is a pooled measure of the variation of boys alone and the variation of girls alone. If there is a gender effect, the within-groups mean square reflects the variation for a partial range of the gender variable. The variation for the full range of the gender variable is given by the total mean square and will be larger than the within-groups mean square. Effect sizes should be comparable across different kinds of treatments and experimental designs. Use of the square root of the total mean square to estimate σ in the gender experiment gives an effect size that is comparable to those for treatments that are manipulated. The problem of estimating σ is more complicated for multitreatment designs and designs with repeated measures and covariates. [Olejnik and Algina (2000)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.67.-#s.6.67.-) provide guidelines for computing effect sizes for such designs.

There are other problems. The three estimators of δ assume normality and a common standard deviation. The value of the estimators is greatly affected by heavy-tailed distributions and heterogeneous standard deviations ([Wilcox, 1996](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.95.-#s.6.95.-), p. 157). Considerable research has focused on ways to deal with these problems ([Olejnik & Algina, 2000](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.67.-#s.6.67.-); [Kendall, Marss-Garcia, Nath, & Sheldrick, 1999](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.54.-#s.6.54.-); [Kraemer, 1983](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.57.-#s.6.57.-); [Lax, 1985](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.58.-#s.6.58.-); [Wilcox, 1996](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.95.-#s.6.95.-), [1997](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.96.-#s.6.96.-)). Some solutions attempt to improve the estimation of δ; other solutions call for radically different ways of conceptualizing effect magnitude. In the next subsection, I describe measures that are based on the proportion of variance in the dependent variable that is explained by the variance in the independent variable.

Strength of association

Another way to supplement null hypothesis significance tests is to provide a measure of the strength of the association between the independent and dependent variables. [Carroll and Nordholm (1975)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.9.-#s.6.9.-) and [Särndal (1974)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.80.-#s.6.80.-) describe a variety of measures of strength of association. Two popular measures are omega squared, ω2, for a fixed-effects treatment and the intraclass correlation, ρI, for a random-effects treatment. A fixed-effects treatment is one in which all treatment levels about which inferences are to be drawn are included in the experiment. A random-effects treatment is one in which the p treatment levels in the experiment are a random sample from a much larger population of P levels. For a completely randomized analysis of variance design, both omega squared and the intraclass correlation are defined as

where and denote, respectively, the treatment and error variance. According to [Hays (1963)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.42.-#s.6.42.-), who introduced omega squared, ω2 and ρI indicate the proportion of the population variance in the dependent variable that is accounted for by specifying the treatment-level classification, and thus are identical in general meaning. The parameters and for a completely randomized design are generally unknown, but they can be estimated from sample data. Estimators of ω2 and ρI are

where SS and MS denote, respectively, sum of squares and mean squares, dfTreat denotes the degrees of freedom for SSTreat, and n is the number of observations in each treatment level. Both omega squared and the intraclass correlation are biased estimators because they are computed as the ratio of unbiased estimators. In general, the ratio of unbiased estimators is itself not an unbiased estimator. [Carroll and Nordholm (1975)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.9.-#s.6.9.-) showed that the degree of bias in ω̂2 is slight.

Earlier I noted that the usefulness of Cohen’s δ was enhanced when he suggested guidelines for its interpretation. Based on [Cohen’s (1988](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.16.-#s.6.16.-), pp. 284 -8) classic work, the following guidelines are suggested for interpreting omega squared:

 ω2 = 0.010 is a small association  
 ω2 = 0.059 is a medium association  
 ω2 = 0.138 or larger is a large association.

[Sedlmeier and Gigerenzer (1989)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.82.-#s.6.82.-) and [Cooper and Findley (1982)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.20.-#s.6.20.-) reported that the typical strength of association in the journals that they examined was around 0.06 - a medium association.

Omega squared and the intraclass correlation, like the measures of effect size, are not without their detractors. One criticism voiced by [O’Grady (1982)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.66.-#s.6.66.-) is that ω̂2 and ρ̂I may underestimate the true proportion of explained variance. If, as is generally the case, the dependent variable is not perfectly reliable, measurement error will reduce the proportion of variance that can be explained. It is well known that the absolute value of the product-moment correlation coefficient, rXY, cannot exceed (rXX′)½ (rYY′)½, where rXX′ and rYY′ are the reliabilities of X and Y ([Gulliksen, 1950](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.38.-#s.6.38.-), pp. 22-3). [O’Grady (1982)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.66.-#s.6.66.-) also criticized measures of strength of association on the grounds that their value is affected by the choice and number of treatment levels. In general, the greater the diversity and number of treatment levels, the larger is the strength of association. [Levin (1967)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.60.-#s.6.60.-) observed that omega squared is not very informative when an experiment contains more than two treatment levels. A large value of ω̂2 simply indicates that the dependent variable for at least one treatment level is substantially different from the other levels. As is true for all omnibus measures, ω̂2 and ρ̂I do not pinpoint which treatment level(s) is responsible for a large value.

One way to address the last criticism is to compute omega squared and the intraclass correlation for two-mean contrasts as is typically done with Hedges’ gc. This solution is in keeping with the preference of many quantitative psychologists to ask focused one-degree-of-freedom questions of their data ([Judd, McClelland, & Culhane, 1995](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.50.-#s.6.50.-); [Rosnow, Rosenthal, & Rubin, 2000](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.77.-#s.6.77.-)) and the recommendation of the *Publication Manual of the American Psychological Association*: “As a general rule, multiple degree-of-freedom effect indicators tend to be less useful than effect indicators that decompose multiple degree-of-freedom tests into meaningful one degree-of-freedom effects - particularly when these are the results that inform the discussion” ([APA, 2001](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.1.-#s.6.1.-), p. 26). The formulas for omega squared and the intraclass correlation can be modified to give the proportion of variance in the dependent variable that is accounted for by the i th contrast. The formulas are

where and the cjs are coefficients that define the contrast ([Kirk, 1995](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.55.-#s.6.55.-), p. 179). These two measures answer focused one-degree-of-freedom questions as opposed to omnibus questions about one’s data.

A measure of strength of association that is popular with meta-analysts is the familiar product-moment correlation coefficient, r. The square of r called the coefficient of determination indicates the sample proportion of variance in the dependent variable that is accounted for by the independent variable. The product-moment correlation and its close relatives can be used with a variety of variables:

|  |  |
| --- | --- |
| product-moment correlation | X and Y are continuous and linearly related |
| phi correlation, φ | X and Y are dichotomous |
| point-biserial correlation, rpb | X is dichotomous, Y is continuous |
| Spearman rank correlation, rs | X and Y are in rank form. |

The point-biserial correlation coefficient is particularly useful for answering focused questions. The independent variable is coded 0 and 1 to indicate the treatment level to which each observation belongs. [Wilcox (1996)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.95.-#s.6.95.-) provides an excellent critique of this and other potentially useful measures of correlation.

Two categories of effect magnitude have been described thus far: measures of effect size and strength of association. Researchers differ in their preferences for the measures. Fortunately, it is a simple matter to convert from one measure to another. Conversion formulas are shown in Table 5.2. The omega squared and intraclass correlation formulas in the table are for a contrast. Also shown in Table 5.2 are formulas for converting the t statistic into each of the measures of effect magnitude.

Other measures of effect magnitude

Quantitative psychologists continue to search for ways to supplement the null hypothesis significance test and obtain a better understanding of their data. Most attention has focused on measures of effect size and strength of association. But as Table 5.1 shows, there are many other ways to measure effect magnitude. Some of the statistics in the “Other measures” column of Table 5.1 are radically different from anything described thus far. One such measure for the two-group case is the probability of superiority, denoted by PS . PS is the probability that a randomly sampled member of a population given one treatment level will have a score, Y1, that is superior to the score, Y2, of a randomly sampled member of another population given the other treatment level. The computation of PS is straightforward: PS = U/n1n2, where U is the Mann-Whitney statistic and n1 and n2 are the two sample sizes. The value of U indicates the number of times that the n1 participants who are given treatment level 1 have scores that outrank those of the n2 participants who are given treatment level 2, assuming no ties or equal allocation of ties. Dividing U by n1n2, the number of possible comparisons of the two treatment levels, yields an unbiased estimator of the population Pr(Y1 > Y2). According to [Grissom (1994)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.36.-#s.6.36.-), PS does not assume equal variances and is robust to nonnormality.

Another example of a different way of assessing effect magnitude is the odds ratio. It is applicable to the two-group case when the dependent variable has only two outcomes, say, success and failure. The term odds is frequently used by those who place bets on the outcomes of sporting events. The odds that an event will occur are given by the ratio of the probability that the event will occur to the probability that the event will not occur. If an event can occur with probability p, the odds in favor of the event are p/(1 - p) to 1. For example, suppose an event occurs with probability ¾, the odds in favor of the event are (¾)/(1 - ¾) = (¾)/(¼) = 3 to 1.

The computation of the odds ratio is illustrated using the data in Table 5.3 where participants in experimental and control groups are classified as either a success or a failure. For participants in the experimental group, the odds of success are

For participants in the control group, the odds of success are

|  | **Success** | **Failure** | **Total** |
| --- | --- | --- | --- |
| Experimental group | n11 = 40 | n12 = 8 | n11 + n12 = 50 |
| Control group | n21 = 28 | n22 = 22 | n21 + n22 = 50 |
| Total | n11 + n21 = 70 | n12 + n22 = 30 |  |

The ratio of the two odds is the odds ratio, ω̂,

In this example, the odds of success for participants in the experimental group are 4.1 times greater than the odds of success for participants in the control group. When there is no association, the two rows (or two columns) are proportional to each other and ω̂ = 1. The more the groups differ, the more ω̂ departs from 1. A value of ω̂ less than 1 indicates reduced odds of success among the experimental participants; a value greater than 1 indicates increased odds of success among the experimental participants. The lower bound for ω̂ is 0 and occurs when n11 = 0; the upper bound is arbitrarily large, in effect infinite, and occurs when n21 = 0. The probability distribution of the odds ratio is positively skewed. In contrast, the probability distribution of the natural log of ω̂, ln ω̂, is more symmetrical. Hence, when calculating a confidence interval for ω̂, it is customary to work with ln ω̂. A 100(1 - α)% confidence interval for ln ω̂ is given by

where zα/2 denotes the two-tailed critical value that cuts off the upper α/2 region of the standard normal distribution and σ̂ln ω̂ denotes the standard error of ln ω̂ and is given by

Once the end points of the confidence interval are found, the values are exponentiated to find the confidence interval for ω. The computation will be illustrated for the data in Table 5.3 where ω̂ = 4.125 to three places. A 100(1 - 0.05)% confidence interval for ln ω is

1.4171 - 1.96(.04796) < ln ω < 1.4171 + 1.96(.04796) 0.4771 < ln ω < 2.3571.

The confidence interval for ω is

 e0.4771 < ω < e2.3571  
 1.6 < ω < 10.5

We can be 95 percent confident that the odds of success for participants in the experimental group are between 1.6 and 10.5 times greater than the odds of success for participants in the control group. Notice that the interval does not include 1. The odds ratio is widely used in the medical sciences, but less often in psychology. Space limitations preclude an examination of other potentially useful measures of effect magnitude. The reader is referred to the excellent overview by [Grissom and Kim (2001)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.37.-#s.6.37.-).

Before leaving this topic, I want to emphasize that important or useful results do not necessarily require large effect magnitudes. [Prentice and Miller (1992)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.69.-#s.6.69.-) and [Spencer (1995)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.87.-#s.6.87.-) provide examples of how small effect magnitudes can be both theoretically and practically significant. One needs to calibrate the magnitude of an effect by the benefit possibly accrued from that effect. In the next section, I illustrate the role of effect magnitude in the accumulation of knowledge.

Cumulating Knowledge through Meta-analysis

For far too long, rejecting null hypotheses and obtaining small p values have been the primary goals of many researchers. This overreliance on null hypothesis significance tests has, in effect, placed blinders on researchers. Consider a researcher who believes that a medication will improve the intelligence test performance of Alzheimer patients. He randomly assigns 20 patients to experimental and control groups and administers the medication to the experimental group and a placebo to the control group. In due time he administers an intelligence test to the patients and performs a t test, t (18) = 2.076, p = 0.052. To his dismay, the p value is larger than 0.05, which means that the null hypothesis cannot be rejected. What’s wrong with this typical scenario? The researcher focused on the null hypothesis and p value without asking whether the data supported the scientific hypothesis. Unfortunately, a result that is not statistically significant is interpreted as providing no support for the scientific hypothesis, even though the data are consistent with the hypothesis. Suppose that the mean for the experimental group is 13 IQ points above that for the control group. This information should make any rational researcher think that the data provides some support for the scientific hypothesis. In fact, the best guess that can be made is that the population mean difference is 13 IQ points. A 95 percent confidence interval for the population mean IQ difference indicates that it is likely to be between -0.2 and 26.2. The nonsignificant t test doesn’t mean that there is no difference between the groups in IQ; all it means is that the researcher cannot rule out chance or sampling variability as an explanation for the 13-point difference.

Suppose that instead of focusing on statistical significance, the researcher focused on what the data said about the scientific hypothesis. He computed an estimate of Cohen’s δ using formulas [1] - [3] given earlier and obtained gc = 0.89. If the 13-point difference is not attributable to chance, it is a large effect. Anyone who has worked with intelligence tests would probably agree that 13 IQ points is a large effect. A 95 percent confidence interval for the population effect size using formula [4] is from -0.03 to 1.8. The data provide considerable support for the researcher’s scientific hypothesis although he cannot rule out chance sampling variability as a possible explanation for the difference. Will the results replicate, are they real? There is only one way to find out - do a replication. Does the medication appear to have promise with Alzheimer patients? I think so. Notice the difference in our reasoning process when we shift attention from the t test and p value to deciding whether the data support our scientific hypothesis and are practically significant.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 5.4 Meta-analysis statistics for the Alzheimer experiment | | | | | | |
|  | **nE** | **nC** | **gc** | **σ̂2(gc)** | **1/σ̂2(gc)** | **gc/σ̂2(gc)** |
| Experiment 1 | 10 | 10 | 0.8893 | 0.2198 | 4.5502 | 4.0465 |
| Experiment 2 | 11 | 11 | 0.8245 | 0.1973 | 5.0692 | 4.1795 |
|  |  |  |  |  | 9.6194 | 8.2260 |

The researcher, encouraged by the large effect size, decided to repeat the experiment. This time 22 Alzheimer patients were available. In the second experiment, the mean of the experimental group was 12 IQ points above the control group. A 95 percent confidence interval for the population mean difference is -0.5 to 24.5. Because the interval includes zero, the null hypothesis for the second experiment cannot be rejected at the 0.05 level of significance. The effect size is gc = 0.82; a 95 percent confidence interval for the effect size is -0.05 to 1.7. To obtain an overall summary of the results of the two experiments, the researcher performed a meta-analysis. The terms needed for the analysis are shown in Table 5.4. In this example, the experiments share a common effect size but have different sample sizes. Following [Hedges and Olkin (1985](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.44.-#s.6.44.-), pp. 112 -13), a weighted mean of the effect sizes is

with estimated variance = 1/9.6194 = 0.1040. A 95 percent confidence interval for the common effect size is

The researcher can be 95 percent confident that the common effect size is greater than 0.2 (a small effect) and less than 1.5 (a large effect). Because the interval does not include 0, the null hypothesis can be rejected at the 0.05 level of significance. A graphical summary of the two experiments and the meta-analysis is shown in Figure 5.1. The graph drives home the point that the medication is effective.

Figure 5.1 Ninety-five percent confidence intervals for the effect sizes (gc).

Although meta-analysis is typically used as a secondary data analysis strategy, this example shows that it also is a useful primary analysis strategy. It allows researchers to accumulate results over a series of studies to obtain a better evaluation of the scientific hypothesis. It is well known that a single study rarely provides a definitive test of a scientific hypothesis. The outcomes of a series of null hypothesis significance tests also can be accumulated. But the analysis techniques - “vote-counting” of reject-nonreject decisions and synthesis of p values - are much less effective than meta-analysis. Meta-analysis has the added advantage of accumulating results in a manner that focuses on the effects of interest rather than p values.

The Alzheimer example illustrates how measures of effect size, confidence intervals, meta-analysis, and graphs can supplement null hypothesis significance testing. Bayesian analysis is yet another way of supplementing null hypothesis significance testing. For an insightful comparison of this and other approaches, the reader is referred to [Howard, Maxwell, and Fleming (2000)](http://search.credoreference.com.proxy1.ncu.edu/content/entry/bkhrmep/chapter_5_the_importance_of_effect_magnitude/0?secid=.6.46.-#s.6.46.-).

It is time for researchers to avail themselves of the full arsenal of quantitative and qualitative statistical tools that are available. It is evident that the current practice of focusing exclusively on a dichotomous reject-nonreject decision strategy of null hypothesis testing can actually impede scientific progress. I suspect that the continuing appeal of null hypothesis significance testing is that it is considered to be an objective scientific procedure for advancing knowledge. In fact, focusing on p values and rejecting null hypotheses actually distracts us from our real goals: deciding whether data support our scientific hypothesis and are practically significant. The focus of research should be on our scientific hypotheses, what data tell us about the magnitude of effects, the practical significance of effects, and the steady accumulation of knowledge.

References

American Psychological Association (2001). *Publication Manual of the American Psychological Association*, (5th edn.). Washington, DC: American Psychological Association. Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423-37. Beauchamp, K. L., & May, R. B. (1964). Replication report: Interpretation of levels of significance by psychological researchers. *Psychological Reports*, 14, 272. Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist*, 76, 159-65. Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, 33, 526-42. Bohrer, R. (1979). Multiple three-decision rules for parametric signs. *Journal of the American Statistical Association*, 74, 432-7. Box, J. F. (1978). *R. A. Fisher: The life of a scientist*. New York: Wiley. Braver, S. L. (1975). On splitting the tails unequally: A new perspective on one- versus two-tailed tests. *Educational and Psychological Measurement*, 35, 283-301. Carroll, R. M., & Nordholm, L. A. (1975). Sampling characteristics of Kelley’s e2 and Hays’ C2. *Educational and Psychological Measurement*, 35, 541-54. Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, 48, 378-99. Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, 61, 287-92. Chambers, R. C. (1982). Correlation coefficients from 2 × 2 tables and from biserial data. *British Journal of Mathematical and Statistical Psychology*, 35, 216-27. Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114, 494-509. Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-53. Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, (2nd edn.). Hillsdale, NJ: Lawrence Erlbaum. Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45, 1304-12. Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 115-59. Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, 49, 997-1003. Cooper, H., & Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin*, 8, 168-73. Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, 2, 161-72. Cowles, M. (1989). *Statistics in psychology: An historical perspective*. Hillsdale, NJ: Lawrence Erlbaum. Cramer, H. (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press. Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 532-74. Dawes, R. M., Mirels, H. L., Gold, E., & Donahue, E. (1993). Equating inverse probabilities in implicit personality judgments. *Psychological Science*, 4, 396-400. Doksum, K. A. (1977). Some graphical methods in statistics. A review and some extensions. *Statistica Neerlandica*, 31 53-68. Dunlap, W. P. (1994). Generalizing the common language effect size indicator to bivariate normal correlations. *Psychological Bulletin*, 116 509-11. Falk, R. (1998). Replication - A step in the right direction. *Theory and Psychology*, 8 313-21. Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory and Psyychology*, 5 75-98. Fisher, R. A. (1921). On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron*, 1 1-32. Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, 1 379-90. Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin*, 70 245-51. Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3-8. Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classification. *Journal of the American Statistical Association*, 49 732-64. Grant, D. A. (1962). Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, 69 54-61. Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *Journal of Applied Psychology*, 79 314-16. Grissom, R. J., & Kim, J. J. (2001). Review of assumptions and problems in the appropriate conceptualization of effect size. *Psychological Methods*, 6 135-46. Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley. Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology*, 29 58-65. Harris, R. J. (1994). *ANOVA: An analysis of variance primer*. Itasca, IL: F. E. Peacock. Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (eds.), *What if there were no significance tests?* (pp. 145-74). Hillsdale, NJ: Erlbaum. Hays, W. L. (1963). *Statistics for psychologists*. New York: Holt Rinehart and Winston. Hedges, L. V. (1981). Distributional theory for Glass’s estimator of effect size and related estimators. *Journal of Educational Statistics*, 6 107-28. Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press. Herzberg, P. A. (1969). The parameters of cross-validation. *Psychometrika Monograph Supplement*, 16, 1-67. Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*, 5 315-32. Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley. Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science*, 8 3-7. Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, 5 411-14. Judd, C. M., McClelland, G. H., & Culhane, S. E. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Reviews of Psychology*, 46 433-65. Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, 67 160-7. Kelley, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences*, 21 554-9. Kendall, M. G. (1963). *Rank correlation methods*, (3rd edn.). London: Griffin. Kendall, P. C., Marss-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, 67 285-99. Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences*, (3rd edn.). Monterey, CA: Brooks/Cole. Kirk, R. E. (2001). Promoting good statistical practices: Some suggestions. *Educational and Psychological Measurement*, 61 213-18. Kraemer, H. C. (1983). Theory of estimation and testing of effect sizes: Use in meta-analysis. *Journal of Educational Statistics*, 8 93-101. Lax, D. A. (1985). Robust estimators of scale: Finite sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, 80 736-41. Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two. *Journal of the American Statistical Association*, 88 1242-8. Levin, J. R. (1967). Misinterpreting the significance of “explained variation.”. *American Psychologist*, 22 675-6. Lord, F. M. (1950). *Efficiency of prediction when a regression equation from one sample is used in a new sample*. Research Bulletin 50-110. Princeton, NJ: Educational Testing Service. McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111 361-5. Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34 103-15. Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5 241-301. Nosanchuk, T. A. (1978). Serendipity tails: A note on two-tailed hypothesis tests with asymmetric regions of rejection. *Acta Sociologica*, 21 249-53. O’Grady, K. E. (1982). Measures of explained variation: Cautions and limitations. *Psychological Bulletin*, 92 766-77. Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, 25 241-86. Preece, P. F. W. (1983). A measure of experimental effect size based on success rates. *Educational and Psychological Measurement*, 43 763-6. Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, 112, 160-4. Rindskopf, D. M. (1997). Testing “small,” not null, hypotheses: Classical and Bayesian approaches. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (eds.), *What if there were no significance tests?* (pp. 319-32). Hillsdale, NJ: Erlbaum. Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, 55 33-8. Rosenthal, R., & Gaito, J. (1964). Further evidence for the cliff effect in interpretation of levels of significance. *Psychological Reports*, 15, 570. Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74 166-9. Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin*, 106 332-7. Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Science*, 5 329-34. Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44 1276-84. Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, 11 446-53. Rossi, J. S. (1997). A case study in the failure of psychology as cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (eds.), *What if there were no significance tests?* (pp. 175-97). Hillsdale, NJ: Erlbaum. Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57 416-28. Särndal, C. E. (1974). A comparative study of association measures. *Psychometrika*, 39 165-87. Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods*, 1 115-29. Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309-16. Serlin, R. C. (1993). Confidence intervals and the scientific method: A case for Holm on the range. *Journal of Experimental Education*, 61 350-60. Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologists*, 40 73-83. Serlin, R. C., & Lapsley, D. K. (1993). Rational appraisal of psychological research and the good-enough principle. In G. Keren & C. Lewis (eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199-228). Hillsdale, NJ: Erlbaum. Shaver, J. P. (1993). What statistical significance testing is, and what it is not. *Journal of Experimental Education*, 61 293-316. Spencer, B. (1995). Correlations, sample size, and practical significance: A comparison of selected psychological and medical investigations. *Journal of Psychology*, 129 469-75. Tang, P. C. (1938). The power function of the analysis of variance tests with tables and illustrations of their use. *Statistics Research Memorandum*, 2 126-49. Tatsuoka, M. M. (1973). *An examination of the statistical properties of a multivariate measure of strength of association*, (U.S. Office of Education, Bureau of Research, Contract No. OEG-5 -72-0027). Urbana-Champaign, IL: Author. Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist*, 53 799-800. Thompson, B. (2002). “Statistical,” “practical,” and “clinical”: How many kinds of significance do counselors need to consider? *Journal of Counseling and Development*, 80, 64-71. Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6 100-16. Wherry, R. J. (1931). A new formula for predicting the shrinkage of the coefficient of multiple correlation. *Annals of Mathematical Statistics*, 2 440-51. Wickens, C. D. (1998). Commonsense statistics. *Ergonomics in Design*, 6 (4), 18-22. Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press. Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press. Wilcox, R. R., & Muska, J. (1999). Measuring effect size: A non-parametric analogue of w2. *British Journal of Mathematical and Statistical Psychology*, 52, 93-110. Yelton, W. H., & Sechrest, L. (1986). Use and misuse of no-difference findings in eliminating threats to validity. *Evaluation Review*, 10 836-52.