# Contrasts and Effect Sizes in Behavioral Research

## A CORRELATIONAL APPROACH

**ROBERT ROSENTHAL**

*Department of Psychology, University of California, Riverside, and Harvard University*

**RALPH L. ROSNOW**

*Department of Psychology, Temple University*

**DONALD B. RUBIN**

*Department of Statistics, Harvard University*

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK http://www.cup.cam.ac.uk
40 West 20th Street, New York, NY 10011-4211, USA http://www.cup.org
10 Stamford Road, Oakleigh, Melbourne 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain



First published 2000

Printed in the United States of America

*Typeface* Stone Serif 9.5/13 pt. *System* LaTeX 2ε [TB]

*A catalog record for this book is available from the British Library.*

*Library of Congress Cataloging in Publication data*
Rosenthal, Robert, 1933 –
Contrasts and effect sizes in behavioral research : a
correlational approach / Robert Rosenthal, Ralph L. Rosnow, Donald
B. Rubin.
p. cm.
Includes bibliographical references and index.
ISBN 0-521-65258-8 (hc). – ISBN 0-521-65980-9 (pb)
1. Psychometrics. 2. Analysis of variance. 3. Psychology –
Statistical methods. 4. Social sciences – Statistical methods.
I. Rosnow, Ralph L. II. Rubin, Donald B. III. Title.
BF39.2.A52R67 1999
150′.7′27 – dc21 99-24199
CIP

ISBN 0 521 65258 8 hardback
ISBN 0 521 65980 9 paperback

# *Contents*

CHAPTER 1

# *Basic Concepts of Focused Procedures*

This chapter discusses the basic distinction between contrasts and omnibus tests of significance. Omnibus tests seldom address questions of real interest to researchers and are typically less powerful than focused procedures. Contrasts accompanied by effect size estimates address focused questions, and the effect size tells us something very different from the *p* value.

## FOCUSED VERSUS OMNIBUS QUESTIONS

Contrasts are statistical procedures for asking focused questions of data. Compared to diffuse or omnibus questions, focused questions are characterized by greater conceptual clarity, and the statistical procedure by greater statistical power when examining those focused questions. That is, if an effect truly exists, we are more likely to discover it and to believe it to be real when asking focused questions rather than omnibus ones. Contrast analyses yield both estimates of the magnitude of the effects investigated and the associated significance levels.

## AN EXAMPLE

Suppose developmental researchers interested in psychomotor skills had a total of fifty children at five age levels (11, 12, 13, 14, 15) play a new video game. The specific question of interest to the researchers was whether age is an effective predictor of proficiency in this game. The mean performance scores of ten children at each of the five age levels were 25, 30, 40, 50, and 55, respectively. These values are plotted in Figure 1.1, whereas Table 1.1 shows the overall analysis of variance (ANOVA) computed on the individual scores. Should the researchers conclude, on the basis of the omnibus $F$ with $p = .40$ noted in Table 1.1, that age was not an effective predictor variable?

If they did so, they would be ignoring what we can see very plainly in Figure 1.1. We observe that the mean performance increased in a monotonic fashion from the lowest to the highest age. In fact, the product-moment correlation for the relation between the five age levels and five performance means is .99. We

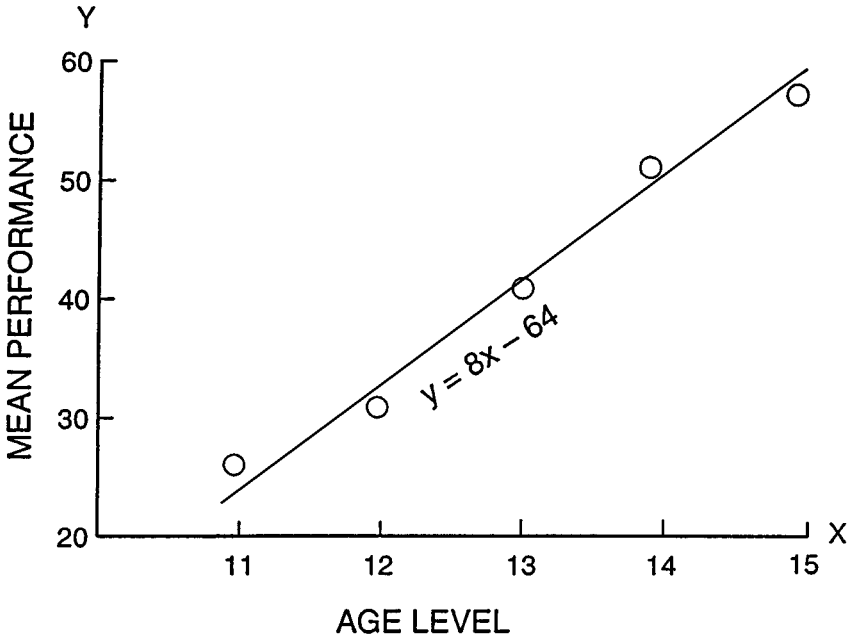


FIGURE 1.1
Graph of group performance data.

call this correlation between an indicator of the "treatment level" (e.g., age) and mean response (e.g., mean performance) an *alerting correlation* ($r_{alerting}$) because it can alert us to overall trends of interest. This alerting correlation is a poor estimate of the relation between *individual* children's ages and performances, because correlations based on aggregated date (e.g., group means) can be dramatically larger or smaller (even in the opposite direction) than correlations based on individual scores. Typically, however, in behavioral research, alerting correlations tend to be larger than correlations based on individual scores. Nonetheless, this correlation alerts us that, although the omnibus $F$ for age levels was far from significant at the conventional .05 level, we should not simply dismiss the idea that age was an effective predictor variable. Notice also in Figure 1.1 that the circles signifying the group means coincide very closely with the straight-line graph with slope 8 and

TABLE 1.1
Summary of Overall ANOVA

| **Source** | ***SS*** | ***df*** | ***MS*** | ***F*** | ***p*** |
|---|---|---|---|---|---|
| Age levels | 6,500 | 4 | 1,625 | 1.03 | .40 |
| Within error | 70,875 | 45 | 1,575 | | |

$Y$ value at the mean age, 13, equal to 40. Clearly the mean performance can be predicted with little error as a linear function of age level.

Common as omnibus significance tests are, this hypothetical case illustrates that they typically do not tell us anything we really want to know. The omnibus $F$ addressed the diffuse question of whether there were any differences among the five age levels, disregarding entirely their ordinal arrangement. The number of possible permutations of the five age levels is 120. Any of these 120 orderings (e.g., 15, 14, 13, 12, 11 or 13, 12, 14, 15, 11) would have yielded the same $F$ with the numerator degrees of freedom of 4. On the other hand, had the researchers performed a contrast analysis to address the specific question of interest, their finding would have been more illuminating. For example, had they performed a $t$ test to assess the linear pattern between age and performance (using a contrast procedure described in Chapter 3), their finding would have been that $t(45) = 2.02$, $p = .025$ one-tailed. Not surprisingly (because squaring the value of a $t$ statistic gives an $F$ statistic with 1 $df$ in the numerator), their finding, had they addressed the predicted linear trend by a focused $F$, would have been that $F(1, 45) = 4.06$, $p = .05$ (also described in Chapter 3).

Even though a linear pattern among the means was clearly evident to the naked eye, the researchers' omnibus $F$ test was insufficiently focused to reject the null hypothesis that the five means were statistically identical. To be sure, this kind of dramatic change – in which the result of the omnibus $F$ has an "insignificant" $p$ value but the result of the contrast is clearly associated with a "significant" $p$ – cannot be expected always, or perhaps even very often, to occur. But what we can usually expect is an increase both in the conceptual clarity concerning the question being asked and in statistical power when, instead of automatically employing omnibus tests of diffuse hypotheses, we formulate precise questions using planned contrasts with associated focused effect size estimates and $p$ values. Because researchers generally want to use statistical tests that will lead to more significant $p$ values when the null hypothesis is false, tests based on contrasts can be said to be more "useful" or more "successful" than omnibus tests.

As we will describe in Chapter 3, our focused procedures compared ("contrasted") the group means of 25, 30, 40, 50, 55 with fixed weights of $-2$, $-1$, 0, $+1$, $+2$ (called *lambda weights*) representing a linear increase in group means. In other words, contrasts are simply focused comparisons of actual group means with predicted lambda ($\lambda$) weights, with the predictions made on the basis of theory, hypothesis, or hunch. Such a comparison may include all the condition means or only some of the means (e.g., results at age 11 and age 15 using lambda weights of $-1, 0, 0, 0, +1$). The only formal stipulation for a contrast is that the lambda weights must sum to zero (i.e., $\Sigma\lambda = 0$). For example, suppose the hypothesis were that economic incentive improves the productivity of work groups but that a boomerang effect can result from excessive external reward. Because the predicted quadratic pattern is $\cap$-shaped (i.e., an upside-down $\cup$), we might select contrast weights of $-2$, $+1$, $+2$, $+1$, $-2$. Now that we have introduced the idea of contrast weights, we can more precisely define $r_{alerting}$ as the correlation between the means of the

various conditions or groups investigated and the contrast weights with which the conditions or groups are associated.

## EFFECT SIZES AND SIGNIFICANCE LEVELS

The basic lesson so far is that contrasts usually give us greater substantive interpretation of research results and greater power for tests of significance. Another advantage of contrasts is that effect sizes can often be easily computed from data in published reports as well as from raw data. Indeed, a maxim of data analysis is that when reporting results, we should give the sizes of the effects as well as the *p* values.

It is important to realize that the effect size tells us something very different from the *p* level. A result that is statistically significant at conventional levels is not necessarily "practically significant" as judged by the magnitude of the effect. Consequently, highly significant *p* values should not be interpreted as automatically reflecting large effects. In the case of *F* ratios, a numerator mean square (*MS*) may be large relative to a denominator *MS* because (a) the effect size is large, (b) the sample size per condition is large, or (c) both. On the other hand, even if the effect size were considered quantitatively unimpressive, it might still have important practical implications. In the next chapters, we will see why that is so, but for now, we will simply sketch some broad ideas about effect sizes and *p* levels.

Table 1.2 shows four possible outcomes of *p* levels and effect sizes as joint determinants of inferential evaluations (Rosnow & Rosenthal, 1988). The tag labels of "acceptable" and "unacceptable" simply imply that the notion that a particular value of an effect size or a significance level is large enough or sufficiently stringent to detect the presence of a "real" effect or "real" difference is not cut in stone. Unfortunately, many researchers operate as if the only proper significance-testing decision should automatically be antinull if *p* is not greater than .05 and pronull if *p* is greater than .05 (Nelson, Rosenthal, & Rosnow, 1986; Rosenthal & Gaito, 1963; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993). It may not be

TABLE 1.2
Potential Problems of Inference as a Function of Obtained Effect Sizes and Significance Levels

| | **Effect Size** | |
|---|---|---|
| **Level of significance** | **"Acceptable" (large enough)** | **"Unacceptable" (too small)** |
| "Acceptable" (low enough) | No inferential problem | Mistaking statistical significance for practical importance |
| "Unacceptable" (too high) | Failure to perceive practical importance of "nonsignificant" results | No inferential problem |

an exaggeration to say that for many Ph.D. students, for whom the .05 alpha has acquired a kind of mystique, it can mean joy, doctoral degrees, and tenure track positions at major universities if their dissertation *p*s are less than .05. On the other hand, *p*s greater than .05 can mean ruin, despair, and their advisers' suddenly thinking of new control conditions that should be run. Logically, of course, there is no justification for a sharp line between a "significant" and a "nonsignificant" difference; the distressing implications of drawing a sharp line have been discussed (e.g., Bakan, 1966; Cohen, 1990; Meehl, 1978; Rosnow & Rosenthal, 1989, 1996b; Schmidt, 1996).

Interestingly, R. A. Fisher chose the 5%, 1%, and 0.1% levels for his tables simply because he regarded them as convenient points on a continuous scale (Yates, 1990, p. xviii). Because it was cumbersome to interpolate between tabulated levels of significance (in the days before powerful hand calculators and desktop computers), writers got into the habit of indicating rough interpolations by stars and asterisks. In behavioral science, it was the 5% (in some cases, the 1%) level that became entrenched in the minds of leading journal editors, textbook authors, and researchers themselves as a "fixed critical level" of significance (Gigerenzer, 1993). Nonetheless, significance in statistics, like the significance of a value in the universe of values, varies continuously between extremes (Boring, 1950; Gigerenzer & Murray, 1987).

Furthermore, as many, including Robert Abelson (1962), have wisely cautioned, significance tests should always be used "for guidance rather than for sanctification" (p. 9). We recommend the practice, endorsed by many statisticians, of reporting precise *p* values (e.g., $p = .06$ rather than $p > .05$). That effort, in turn, will make it easier for informed consumers and meta-analysts to evaluate the implications of a given *p* value or, more usefully, of that *p* value with its associated effect size.

Returning to Table 1.2, suppose we were confronted with a "nonsignificant" *p* and a "large" effect size – what should this tell us? Were we simply to conclude on the basis of the significance level that "nothing happened," we might be making a serious mistake. A small sample size may have led to failure to detect the true effect, in which case, we should continue this line of investigation with a larger sample size before embracing the null as approximately true.

On the other hand, suppose we obtain a "significant" *p* and a "small" effect size – what should this tell us? The answer depends both on the sample size and on what we consider the practical importance of the estimated effect size. With a large sample size, we may mistake a result that is merely "very significant" for one that is of practical importance. In the next chapter, we describe convenient formulas for computing informative effect sizes when comparing two groups. We also show how to calculate counternull values, that is, effect sizes that are just as well supported by the data as the null hypothesis. Chapter 3 describes extensions of the procedures in Chapter 2 for studies employing more than two groups. Chapter 4 presents procedures applicable when it is useful to conceptualize the data as made up of more than one factor. Chapter 5 describes repeated measures designs in which each subject contributes two or more measurements (e.g., under different experimental conditions). Chapter 6, the concluding chapter, describes procedures employed when multiple contrasts are computed, combined, and compared.

One aspect of our presentation is especially distinctive. Essentially all of our effect size measures are correlations of one kind or another. We feel that using correlations is particularly appropriate in the behavioral and social sciences, where the same conceptual outcome variable (e.g., improvement of mental health) can be measured in a wide variety of ways (e.g., test scores, ratings on a scale, rehospitalizations). In such situations, regression coefficients, for instance, are not directly comparable across different studies even in the same research area, whereas correlation coefficients can be sensibly compared. The outcome variables in much of the social sciences are typically not like those in medicine (e.g., death, blood pressure), chemistry (temperature, pressure), physics (acceleration, velocity), or even some of the social sciences (e.g., money in economics), where the outcome variables have intrinsic meaning.

In the area of interpersonal expectancy effects, for example, dozens of different outcome variables have been employed, including maze-learning and Skinner-box-learning scores earned by rats, pupils' intellectual performance, responses to inkblot tests, reaction times, psychophysical judgments, interview behavior, and person perception tasks. In each of these areas, many different specific measures can be employed, and yet all these measures, of all these outcome variables, are subsumed under the single conceptual outcome variable of interpersonal expectancy effects, the degree to which one person's expectation for another's behavior comes to serve as a self-fulfilling prophecy (Rosenthal, 1966, 1976, 1994; Rosenthal & Rubin, 1978; Rosnow & Rosenthal, 1997). Clearly, the enormous variety of specific measures employed by researchers points to the greater utility of correlation coefficients than of regression coefficients as general indices of effect size.

## REVIEW QUESTIONS

1. A widely used publication manual for psychologists urged that effect sizes be routinely reported and added that "in most cases such measures are readily obtainable whenever the omnibus test statistics (e.g., $t$ and $F$) . . . are reported." What is wrong with the quoted statement?

   **Answer:** The $t$ test is not an omnibus test but a focused test, and not all $F$ tests are omnibus tests (only those with numerator $df > 1$). Furthermore, effect sizes reported for omnibus $F$ tests are typically uninterpretable.

2. Computing the analysis of variance on results of a five-group randomized experiment in which the subjects have been assigned to one of five levels of the independent variable, the researchers find $F(4, 95) = 1.24$, $p = .30$, and report that there was "no effect" of the independent variable. What is wrong with their report?

   **Answer:** Their report is of an omnibus effect that is unlikely to be of any real scientific interest. In addition, an important contrast may be hidden within the omnibus test with $p$ as low as .028.

3. Is it possible for a large effect to escape detection by a significance test, and for a small effect to be statistically significant?

**Answer:** The answer is yes to both questions. A large effect might escape detection if there were too few subjects. A small effect would be detectable if the sample were large enough.

4. In discussing our hypothetical example with group means of 25, 30, 40, 50, and 55, we noted an alternative prediction using lambda weights of $-1$, 0, 0, 0, $+1$. What does the alerting $r$ on these data tell us, particularly in comparison with the alerting $r$ of .99 with the weights of $-2$, $-1$, 0, $+1$, $+2$ that was discussed previously?

   **Answer:** Correlating the alternative lambda weights with the group means gives alerting $r = .83$, which is another strong signal of a predicted relationship of interest. Later, we will see that squaring the alerting $r$ tells us the proportion of the between-conditions sum of squares accounted for by the set of lambda weights. In this case, squaring .83 tells us that the alternative prediction with weights of $-1$, 0, 0, 0, $+1$ accounts for over two thirds of the between-conditions sum of squares. Squaring $r = .99$, the original prediction using weights of $-2$, $-1$, 0, $+1$, $+2$ accounted for 98% of the between-conditions sum of squares. Both hypotheses fare well, but the original one is a better predictor of the children's performance.

5. Another hypothesis mentioned is that economic incentive improves the productivity of work groups but that a boomerang effect can result from excessive reward. Given five increasing levels of economic incentive, this hypothesis was described as an "upside-down ∪" with weights of $-2$, $+1$, $+2$, $+1$, $-2$. Imagine two alternative experiments, with condition means of 50, 53, 56, 56, 43 in Study A, and with condition means of 56, 49, 41, 47, 55 in Study B. What can alerting $r$s computed on these results tell us?

   **Answer:** The alerting $r$s are .86 for Study A and $-.96$ for Study B. Study A's hypothesis is a good predictor, as it can account for 74% of the between-conditions sum of squares. Study B's lambda weights, although they account for 92% of the between-conditions sum of squares, are in the opposite direction of the obtained pattern. That is, an upside-down ∪ was predicted, but the group means are ∪-shaped in Study B.

6. Researcher Smith assigns a total of forty subjects at random to either an experimental or a control condition. She hypothesizes that subjects in the experimental condition will score higher on the dependent variable than will subjects in the control condition. She reports $F(1, 38) = 4.71$, $p = .036$, and concludes that her hypothesis was statistically supported. Researcher Jones repeats Smith's study using a total of twenty subjects, finds the same direction of difference, but reports $F(1, 18) = 2.22$, $p = .15$. Disappointed by this result, he questions the validity of Smith's hypothesis, reporting that he failed to replicate her result. Is he correct?

   **Answer:** Readers who already know how to calculate an effect size based on a focused test (reviewed in the next chapter) will have figured out that Jones was confused. Had he calculated the effect size instead of fixating on the obtained $p$ level, he would have immediately seen that his findings replicated Smith's almost perfectly. The reason why Jones failed to replicate Smith's $p$ value was that he was operating with half as many subjects and therefore less power.