Sample Size to Detect a Planned Contrast and a One Degree-of-Freedom Interaction Effect

By: Douglas Wahlsten

Wahlsten, D. Sample size to detect a planned contrast and a one degree-of-freedom interaction effect. Psychological Bulletin, 1991, 110, 587-595.

Abstract:

A simple method is described for estimating the sample size per group required for specified power to detect a linear contrast among *J* group means. This allows comparison of sample sizes to detect main effects with those needed to detect several realistic kinds of interaction in 2 × 2 and 2 × 2 × 2 designs with a fixed-effects model. For example, when 2 factors are multiplicative, the sample size required to detect the presence of nonadditivity is 7 to 9 times as large as that needed to detect main effects with the same degree of power. In certain other situations, effect sizes for the main effects and interaction may be identical, in which case power and necessary sample sizes to detect the effects will be the same. The method can also be used to find sample size for a complex contrast in a nonfactorial design.

Article:

The power of the analysis of variance (ANOVA) to detect statistical interaction in a factorial experiment is substantially less than the power to detect main effects in many situations (Neyman, 1935; Rodger, 1974; Traxler, 1976; Wahlsten, 1990). For example, with a 2 × 2 design, power to detect the multiplicative type of interaction is 16% when power to detect a main effect is 87% (Wahlsten, 1990); hence the ANOVA will frequently point to additivity of effects when in fact they are multiplicative.

This is not a universal property of the ANOVA method, however. As shown in this article, the degree of the discrepancy in power depends strongly on the specific kind of interaction that is present in the data. The power function depends on the size of an effect and the degrees of freedom. If two effects have the same size and degrees of freedom, power must necessarily be the same.

If the presence or absence of interaction is important for a theory it is imperative that the test of interaction have a power of 80% or, preferably, 90%. This sometimes requires larger samples than are customarily used to detect main effects with ANOVA. But how much larger must they be? Relative sample sizes may give a better impression of the insensitivity of ANOVA to interaction than do relative power values. The power to detect a particular kind of interaction may be half the power to detect a main effect in that circumstance, yet the sample size needed to detect the interaction may be far more than double the sample size needed to detect the main effect with the same degree of power. This larger sample size provides a direct indicator of the additional research subjects, technician time, and grant funds required for the desired sensitivity of the statistical test.

This article presents a convenient way to determine appropriate sample sizes for factorial designs in which there are two levels of each factor. The method can be adapted to many other designs, provided the investigator has a good idea about the specific kind of interaction effect of theoretical interest. The basic idea is that any completely randomized design with fixed effects can be considered a one-way design with *J* groups and then analyzed with a series of orthogonal contrasts. In a 2 × 2 or 2 × 2 × 2 design, for example, the contrasts for main effects and interactions have identical degrees of freedom in both numerator and denominator; only the sizes of the effects themselves differ. Planned contrasts offer many advantages over omnibus tests in designs with more than two levels per factor (e.g., Rosnow & Rosenthal, 1989).

Finding the Sample Size for a Linear Contrast

In the next sections, it is assumed that observations in each population are normally distributed and that samples of equal size are drawn randomly. Consider first a simple comparison of two groups with true means μ_1 and μ_2 having standard deviations equal to σ, which yields a population effect size $\delta = |\mu_1 - \mu_2|/\sigma$. If the probability of a Type I error is set at α and the desired power of the test is $1 - \beta$, then the exact solution for the necessary sample size (n) per group is

$$n = \frac{2(z_\alpha - z_{1-\beta})^2}{\delta^2} \qquad (1)$$

when one uses a one-tailed z test of the hypothesis that $\delta = 0$ and the standard deviations are known and equal. Some variation of this formula is presented in almost every introductory text on statistical inference. Generally, σ is unknown and is estimated from the data. Then a t test having $n_1 + n_2 - 2$ degrees of freedom is used, and Formula 1 is approximate. Because a linear contrast among J group means will also have a t distribution when the normal assumptions hold, it should be possible to generalize Equation 1.

For J groups with means $\mu 1, \mu 2 \cdots \mu_J$ a linear contrast among the means is given by

$$\Psi_c = c_1\mu_1 + c_2\mu_2 + \cdots c_J\mu_J, \qquad (2)$$

where $\Sigma c_j = 0$, and effect size $\delta_c = \Psi_c/\sigma$. If each group has variance σ^2 and samples of n observations are taken from each population, the variance of the sample contrast $\hat{\Psi}_c$ is (Hays, 1988)

$$\text{Var}(\hat{\Psi}_c) = \frac{\sigma^2}{n} \Sigma c_j^2. \qquad (3)$$

Combining the ideas in Equations 1 and 3, the sample size per group to detect the contrast when σ is known is exactly

$$n = (z_\alpha - z_{1-\beta})^2 \sigma^2 \frac{\Sigma c_j^2}{\Psi_c^2}. \qquad (4)$$

If σ^2 is estimated from the sample variance within groups (degrees of freedom $Jn - J$), this equation is an approximation. It is essentially the equation presented by Levin (1975) and applied by Lachenbruch (1988) to determine sample size to detect an interaction in a 2 × 2 design.

However, empirical testing (demonstrated below) reveals that adding 2 to Equation 4 is appropriate when σ is estimated from the data. I propose here that a more satisfactory approximation for a contrast of J means is

$$n = \frac{(z_\alpha - z_{1-\beta})^2 \Sigma c_j^2}{\delta_c^2} + 2, \qquad (5)$$

or, when $\Sigma c_j^2 = 1$,

$$n = \frac{(z_\alpha - z_{1-\beta})^2}{\delta_c^2} + 2. \qquad (6)$$

It might be helpful also to specify the relation in terms of a proportion of variance attributable to the particular effect or contrast, η_c^2 or η_j^2, which is a partial correlation ratio (Maxwell, Camp, & Arvey,1981; see also Glass & Hakstian, 1969). When $\Sigma c_j^2 = 1$ for a particular contrast, the partial correlation ratio is

$$\eta_c^2 = \frac{\delta_c^2}{\delta_c^2 + J}, \text{ and} \qquad (7)$$

$$n = \frac{(z_\alpha - z_{1-\beta})^2 (1 - \eta_c^2)}{J\eta_c^2} + 2. \qquad (8)$$

How good is the approximation proposed here? Kraemer and Thiemann (1987) prepared a master table that is based on a central t approximation to the noncentral t distribution that is very convenient for determining sample size for a comparison of two groups. The table is entered with the critical effect size A, which is equivalent to $\sqrt{\eta^2}$, and the result is degrees of freedom v for the two groups combined. In light of Equation 7, it is possible to generalize their approach to a linear contrast Ψ_c. When $\Sigma c_j^2 = 1$, the critical effect size is

$$\Delta_c = \delta_c / (\delta_c^2 + J)^{1/2}, \qquad (9)$$

where Δ_c equals η_c. The total sample for the J groups is then $N = \nu + J$ where ν is obtained from Kraemer and Thiemann's master table, and sample size per group is $n = (\nu + J)/J$. One point of caution is necessary here. Converting the effect size Δ_c to the equivalent δ_c for insertion in Equation 6 uses Equation 9, which assumes $\Sigma c_j^2 = 1$. For two groups, this equivalent δ_c is not the familiar $\delta_c = |\mu_1 - \mu_2|/\sigma$ given in Kraemer and Thiemann (1987) because that expression yields $\Sigma c_j^2 = 2$. The c_j values must be $\pm 1/\sqrt{J}$ when Equations 6 and 9 are used. Otherwise, Equation 9 should entail J^2 rather than J in the denominator if c_i values are ±1.

Table 1
Comparison of Sample Sizes Calculated to Yield Power of 90% for a Two-Tailed Test of a Contrast With α = .05 When Effect Size Δ_c = .20

		Results				
		Kraemer & Thiemann			Equation 6	
J	δ_c	ν	$(\nu + J)/J$	n	n	Rounded
2	0.408	257	129.5	130	128.1	129
3	0.612	257	86.7	87	86.1	87
4	0.816	257	65.2	66	65.1	66
5	1.021	257	52.4	53	52.4	53
6	1.225	257	43.8	44	44.0	44
8	1.633	257	33.1	34	33.5	34
10	2.041	257	26.7	27	27.2	28

Note. J = number of groups in the experiment. Results are according to the method of Kraemer and Thiemann (1987) or Equation 6 in the present article.

Suppose one wishes to compare four groups with a contrast using α = .05 with a nondirectional test of the null hypothesis $\delta_c = 0$, and the desired degree of power is 90%. For effect sizes ranging from 0.10 to 0.60, the required n from Kraemer and Thiemann (1987) is no more than 0.2 unit higher than the result from Equation 6, and the values of n are identical when rounded upwards. Table 1 compares results from the master table with those of Equation 6 for different numbers of groups. Equation 6 yields a slight underestimate of n when there are $J = 2$ groups and a slight overestimate for more than 6 groups.

Results from Equation 6 or 8 can also be compared with those from Cohen's (1988) sample size tables. For $J > 2$, a linear contrast translates into a partial η_c by means of Cohen's Equation 8.2.19. For a contrast effect in a 2 × 2 design with α = .05 (two-tailed) and power of 90%, when Cohen's Table 8.4.4 and Equation 8.4.4 yield $n = 43.0$, Equation 8 requires $n = 44.0$. Only when $J = 2$ groups can the methods of Cohen (1988), Kraemer and Thiemann (1987) and Equation 5 be compared directly. Equation 5 was not derived with a two-group t test in mind, but a contrast between two means amounts to the same thing. Cohen's tables for sample size use Effect Size f or d, whereas the Kraemer and Thiemann master table uses Δ. A tabled value in one can be compared with the equivalent index of the other using interpolation, which introduces small errors. However, Equation 5 can be used with any value of δ. Table 2 compares n to achieve 90% power, first for Equation 5 versus the master table and then for Equation 5 versus Cohen's (1988) Tables 8.4.1, 8.4.4, and 8.4.7. All noninteger results are rounded up. Equation 5 consistently calls for one or two more observations than do Cohen's tables and one or two fewer than does the master table when $J = 2$.

Table 2
Sample Sizes Required to Detect a Difference Between J = 2 Group Means[a]

	K & T vs Eqn 5				Cohen vs Eqn 5				
	α = .05		α = .01			α = .05		α = .01	
Δ	K & T	Eqn 5	K & T	Eqn 5	f	Cohen	Eqn 5	Cohen	Eqn 5
.7	9	8	12	10	.8	9	11	13	14
.6	13	12	17	16	.6	16	17	22	23
.5	19	18	27	25	.4	34	35	48	49
.2	130	129	183	181	.2	132	134	188	188

Note. Values of n are rounded up to the nearest integer.
[a] With a power of 90% and a two-tailed test using the methods of Cohen (1988), Kraemer and Thiemann (K & T; 1987), and Equation 5 (Eqn 5) in the present article.

The principal advantages of Equation 5 over the methods of Kraemer and Thiemann (1987) and Cohen (1988) are that (a) no tables are needed apart from the ubiquitous cumulative normal table whose critical values can be memorized, (b) any magnitude of effect size may be used without interpolation, and (c) power for complex contrasts that do not fit into the standard factorial ANOVA scheme may be readily evaluated. If *n* is found for each of a series of contrasts in the same data set, it would be wise to use the largest value of *n* required to test effects of particular interest. Several recent textbooks on statistical analysis in the behavioral sciences (Hays, 1988; Marascuilo & Serlin, 1988; Maxwell & Delaney 1990) emphasize the testing of linear contrasts in one-way designs. The preceding equations should be helpful to researchers who intend to apply their methods of analysis. It would be pointless to plan a clever experiment and analyze results with sophisticated instruments yet adopt a sample size that renders the tests insensitive to all but the crudest effects.

Sample size for experiments with many groups can be found quickly with the method of Cohen (1988) if the design is factorial, but complex experiments do not always lend themselves to this approach. For example, reciprocal crossbreeding of animals can be used to demonstrate maternal environment effects on rate of development (Wainwright, 1980), behavior (Bauer & Sokolowski, 1988) or brain size (Wahlsten,1983), and they may also reveal the importance of the cell cytoplasm for individual differences in brain structure (Wimer & Wimer, 1989). One logical method of analyzing such data is a series of orthogonal contrasts. The experimental design entails 16 groups, consisting of 2 parent strains, 2 reciprocal F_1 hybrids, 8 reciprocal backcrosses of an F_1 hybrid to one parent, and 4 reciprocal F_2 hybrids. Table 3 contains an experimental design and hypothesized means for the number of granule cells (in thousands) in the dentate gyrus of the hippocampus of male mice, basing the parent strain means (high and low) and standard deviation within groups (σ = 45.0) on data reported by Wimer and Wimer (1989). The model asserts that the midparent mean of 425 is augmented to varying degrees by high-strain autosomes, Y chromosome, cytoplasm, and maternal environment and is decremented by similar amounts by the low-strain counterparts. As is evident from the table, small samples would be quite adequate to detect the large parent-strain difference, but much larger samples would be required to detect the small Y chromosome effect. The prudent decision would be to choose a sample size that would allow one to detect the smallest effect that is of major importance for the research program. Different sample sizes might be contemplated for different groups, but this would violate the orthogonality of certain comparisons. Alternatively, the groups involved in a comparison of particular interest but with small effect size could be tested with large sample sizes, and then those groups could be analyzed with a separate oneway ANOVA.

Relative Sample Sizes for Specific Cases

For any particular test, Equation 5 or 6 gives the results, once $\delta_c = \Psi_c/\sigma$ is specified. This can be done separately for main effects and interaction, but an expression can be derived for the ratio of sample sizes needed to detect an interaction and a main effect. If n_A is the sample size needed to detect contrast Ψ_A with power 1 — β, and $n_{A\times B}$ is the sample size required to detect contrast $\Psi_{A\times B}$, also with power 1 — β, then from Equation 5 it follows that

$$\frac{n_{A\times B}-2}{n_A-2}=\frac{\Psi_A^2}{\Psi_{A\times B}^2}=\frac{\delta_A^2}{\delta_{A\times B}^2}. \qquad (10)$$

The ratio of sample sizes is unrelated to the chosen values of α or β and is determined primarily by the specific kind of interaction present in the data. The principle embodied in Equation l0 applies to any two contrasts in which the Σc_j^2 values are the same. When both values of *n* are reasonably large, the constants can safely be omitted, and relative sample sizes are inversely proportional to relative squared effect sizes.

Let us examine several situations that commonly occur in psychological research to find out the relative magnitudes of interaction and main effects. Suppose one conducts a 2 × 2 experiment in which Factor A has levels A_1 and A_2 and Factor B has levels B_1 and B_2. Arranging the four groups in the order A_1B_1, A_1B_2, A_2B_1, and A_2B_2 and using coefficients (c_j) of l or -1, the coefficients are (-1, -1,1,1), (-1,1, -1,1), and (-1,1, 1, -1) for the A main effect, B main effect, and A × B interaction, respectively.

Patterns of results are portrayed in Figure 1. Many others might be imagined, of course. The scales for the measures *Y* are arbitrary and can be adapted to any comparable circumstance by linear transformation. The cases may be summarized as follows. The first six cases all have $\mu_{21} - \mu_{11} = 1$, and the last three have $\mu_{21} = \mu_{11}$.

Case 1. A and B effects are equal and additive.

Case 2. A and B are multiplicative, so that the effect of B is twice as large for A_2 as for A_1.

Case 3. There is no effect of B on A_1, but there is a clear effect of B on A_2.

Case 4. There are opposite effects of B on A_1 and A_2.

Case 5. There is no effect of B on A_1 but a large effect on A_2, resulting in reversal of ranks.

Case 6. There are opposite effects of B on A_1 and A_2 but no main effects at all.

Case 7. There is no effect of A at B_1, and there is a greater effect of B on A_2 than A_1.

Case 8. There is no effect of A at B_1, and there is no B effect on A_1.

Case 9. There is no effect of A at B_1, and there are opposite effects of B on A_1 and A_2.

Table 3

Means for Crosses Between Low-Scoring (L) and High-Scoring (H) Strains Analyzed by Orthogonal Contrasts

Cross	Mother	Father	*M*	Inbred parents	Male Y chromosome	Egg cytoplasm	Hybrid mat env
1	L	L	300	−1			
2	H	H	550	1			
3	L	H	390				
4	H	L	460				
5	L	L × H	350		1		−1
6	L	H × L	340		−1		−1
7	H	L × H	510		1		−1
8	H	H × L	500		−1		−1
9	L × H	L	415			−1	1
10	H × L	L	445			1	1
11	L × H	H	505			−1	1
12	H × L	H	535			1	1
13	L × H	L × H	465		1	−1	
14	L × H	H × L	455		−1	−1	
15	H × L	L × H	495		1	1	
16	H × L	H × L	485		−1	1	
Ψ_c				150	40	120	200
δ_c				3.333	0.889	2.667	4.444
Σc_j^2				2	8	8	8
n				4	89	12	4

Note. Means are numbers ($\times 10^3$) of dentate granule cells in the hippocampus of male mice. Standard deviation within a group is 45 ($\times 10^3$), which is based on Wimer and Wimer (1989). The model used to generate the expected means asserts that the mean of the two parent strains (425) is increased by the H strain autosomes by 80, Y chromosome by 5, cytoplasm by 15, and maternal environment (mat env) by 25; whereas the L strain counterparts decrease the mean by the same amounts. Also, a hybrid maternal environment increases the mean by 50 units. Ψ_c = contrast; δ_c = effect size; *n* = sample size per group needed to detect a contrast with power of 90% when the Type I error level is .05, one-tailed.

The values of Ψ_c for each contrast are shown in Table 4. These make it possible to compare sample sizes for the three effects within a particular case, but the arbitrary scales of measurement in the graphs make comparisons between cases difficult. Suppose that in each case the four group means and the variance within groups are such that $\eta^2 = 20$ or Cohen's $f = .50$. According to Table 8.3.14 of Cohen (1988), to have a power of 90% when $\alpha = .05$ for the overall *F* test of significance, there should be $n = 15$ observations in each of the four groups.

Knowing the value of η^2 for the entire experiment and the relative values of Ψ_c for all J — 1 contrasts, sample size can be ascertained without specifying the actual group means or standard deviation within groups. This makes it convenient to calculate and comprehend the relative sizes of all three effects and the sample sizes needed to detect them. A simple way to compute η_c^2 or η_j^2 for each contrast is needed. It can be shown that

$$\frac{\eta^2}{1-\eta^2} = \sum_{j=1}^{j=J-1} \eta_j^2 / (1 - \eta_j^2). \qquad (11)$$

The sum of squares between the J groups must equal the total SS_c for the J-1 contrasts if they are orthogonal (Hays, 1988). Provided Σc_j^2 has the same value for each contrast, the proportion of the SS between groups attributable to a contrast effect is

$$P_j = \frac{\Psi_j^2}{\sum_{j=1}^{j=J-1} \Psi_j^2} = \frac{SS_j}{SS_{\text{between}}}. \qquad (12)$$

From Equations 11 and 12,

$$\eta_j^2 = \frac{P_j}{P_j + \frac{(1-\eta^2)}{\eta^2}} \qquad (13)$$

When $\eta^2 = .20$, $\eta_c^2 = P_c/P_c + 4)$. Consider Case 1. The sum of squared contrasts is 4 + 4 = 8, and for the A main effect, $P_c = 4/8 = 0.5$. Therefore, $\eta_c^2 = .11$, and from Equation 8, 23 observations are needed to detect either main effect with power of 90% when α = .05 and a two-tailed test is used. In Case 2 for the multiplicative A × B interaction, $P_c = 1/19$, $\eta_c^2 = .013$, and n should be 199.5, which is nearly nine times as large as the n needed for the main effects! Table 4 reveals that multiplicative interaction in a 2 × 2 design will be particularly difficult to detect, as will the patterns in Cases 3 and 7. Even in Case 4, in which the lines diverge, the interaction is substantially smaller than the A main effect. Case 8 is interesting because all three effects are equal.

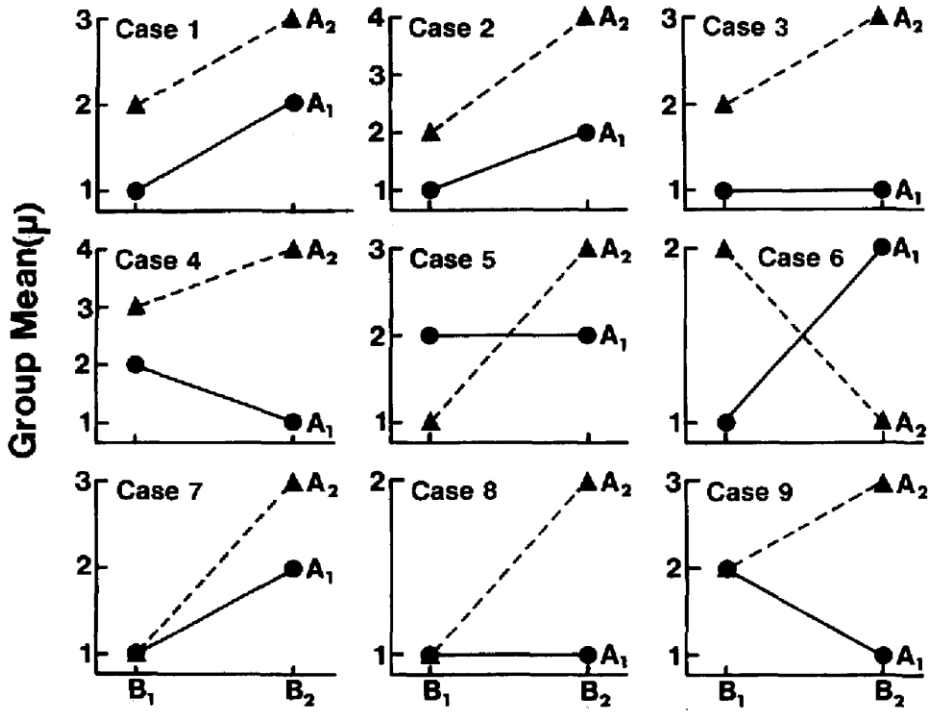

Figure 1. Population means for four groups in a 2 × 2 factorial design with Factors A and B. The nine cases represent different possible patterns of group means. Only Case 1 indicates true additivity of the two factors. Contrast values and sample sizes to detect the interaction are listed in Table 4.

Several caveats are necessary:

1. Certain of the cases in Figure 1 involve a complete absence of one or more effects. Because these null hypotheses are presumed to be true, it would make no sense to do a power or sample size calculation for a nonexistent contrast. However, Equations 11 and 13 hold even when one or two of the contrast effects are zero.

2. Several interesting cases in which interaction effects exceed main effects are not shown in Figure 1. The general conditions when this will occur are derived in the next section.

3. It may seem counterintuitive that $n = 15$ is sufficient for the overall F test, yet n must be considerably greater to detect many of the one degree-of-freedom contrasts, such as the A or B main effect in Case 1 in which additivity obtains. Yet, when the effect size that is equivalent to $\eta_c^2 = 0.111$ for the A main effect is inserted into Equation 5, the n needed for a t test comparing two group means is approximately twice the n needed for an equivalent contrast among $J = 4$ groups. That is the total number of observations required is about the same in both situations. if a single contrast does account for a large proportion of the total $SS_{between}$, then indeed the n will be less than the 15 required for the global F test.

Table 4
Contrast Values for Hypothetical Results From 2 × 2 Designs Shown in Figure 1 and Sample Sizes Needed to Detect Each Contrast[a]

	Case								
Variable	1	2	3	4	5	6	7	8	9
Contrast									
A	2	3	3	4	0	0	1	1	2
B	2	3	1	0	2	0	3	1	0
A × B	0	1	1	2	2	2	1	1	2
Sample size									
n_A	23	25	13	14	—	—	116	32	23
n_B	23	25	116	—	23	—	13	32	—
$n_{A\times B}$	—	200	116	53	23	11	116	32	23

Note. Values of n are rounded up to the nearest integer. Dashes indicate value cannot be calculated because effect size is 0.
[a] With a power of 90% when $\alpha = .05$, two-tailed, and the value of $\eta^2 = .20$ for the one-way analysis of variance on four groups. Under these conditions, $n = 15$ observations per group is sufficient to detect the overall difference between the groups with power of 90%.

This approach can be extended to a 2 × 2 × 2 design or any higher order design with only two levels per factor. Figure 2 illustrates several possible outcomes for a three-way design, and Table 5 describes the sample sizes to detect main effects and interactions. Eight cases are considered, although other interesting patterns are certainly possible. As with the 2 × 2 design, suppose the difference among the eight group means is such that $\eta^2 = .20$. According to Tables 8.3.15 and 8.3.16 of Cohen (1988), n must be about 10 to yield power of 90% when $\alpha = .05$. The value of η_c^2 for each contrast is computed according to *Equation* 13. Table 5 shows that the n to detect a specific contrast is considerably greater than 10 unless that contrast accounts for a large proportion of the total variance between groups. Indeed, only the A main effects in Cases 15 and 16 require fewer than 10 observations. Equal sample sizes are sufficient for the A × B × C interaction and A main effect in Cases 14 and 17, but for Cases 11, 12, and 13 the samples needed to detect the interactions are rather large. It is somewhat disturbing to learn that it will be most unlikely to find a multiplicative second-order interaction (Case 11) with a reasonable degree of power, given the research budgets of most psychologists.

The 2 × 2 in General

The conditions under which the necessary sample sizes for main effects and interaction in a 2 × 2 design are generally the same or different can be determined by supposing that membership in Group A_1 or A_2 determines the parameters of a linear equation and that Treatment B determines the value of X, as shown in Fig. 3. It is assumed that X_2 always exceeds X_1. To simplify presentation of results, let the differences in values be symbolized as $D_x = X_2 - X_1$, $D_a = a_2 - a_1$, and $D_b = b_2 - b_1$. The contrasts then have the expected values $\Psi_A = 2 D_a + (X_1 + X_2)D_b$, $\Psi_B = D_x(b_1 + b_2)$, and $\Psi_{A\times B} = -D_xD_b$.

Sample size to detect a main effect, as compared with that needed for the interaction, can be evaluated by the difference in squared effect sizes. A ratio could also be used, but it is cumbersome for this purpose. Comparing the A main effect to interaction,

$$\Psi_A^2 - \Psi_{A\times B}^2 = 4 (D_a + D_bX_1)(D_a + D_bX_2). \quad (14)$$

Whether the effect for A and interaction are the same or different depends on the choice of treatments B_1 and B_2. Several other generalizations are possible. The A and interaction effect sizes can be the same if and only if

either X_1 or $X_2 = -(a_2 - a_1)/(b_2 - b_1)$. When the intercepts are the same ($a_1 = a_2$), either X_1 or X_2 must be zero if the A and interaction effects are to be the same. That is, if the lines meet at the origin but do not cross, the effect sizes will be the same, no matter what the slopes of the two lines. Because adding a constant to all X values and another constant to all Y values merely shifts the axes and will not affect the results of an ANOVA, it follows that the A main effect will always (a) exceed the interaction effect if the lines do not touch, (b) equal the interaction effect if the lines meet at either X_1 or X_2, and (c) be less than the interaction effect if the lines cross between X_1 and X_2. Note that these conclusions in no way depend on the range from X_1 to X_2. The difference $X_2 - X_1$ will, of course, determine the degree to which effect sizes differ but not the sign of the difference.

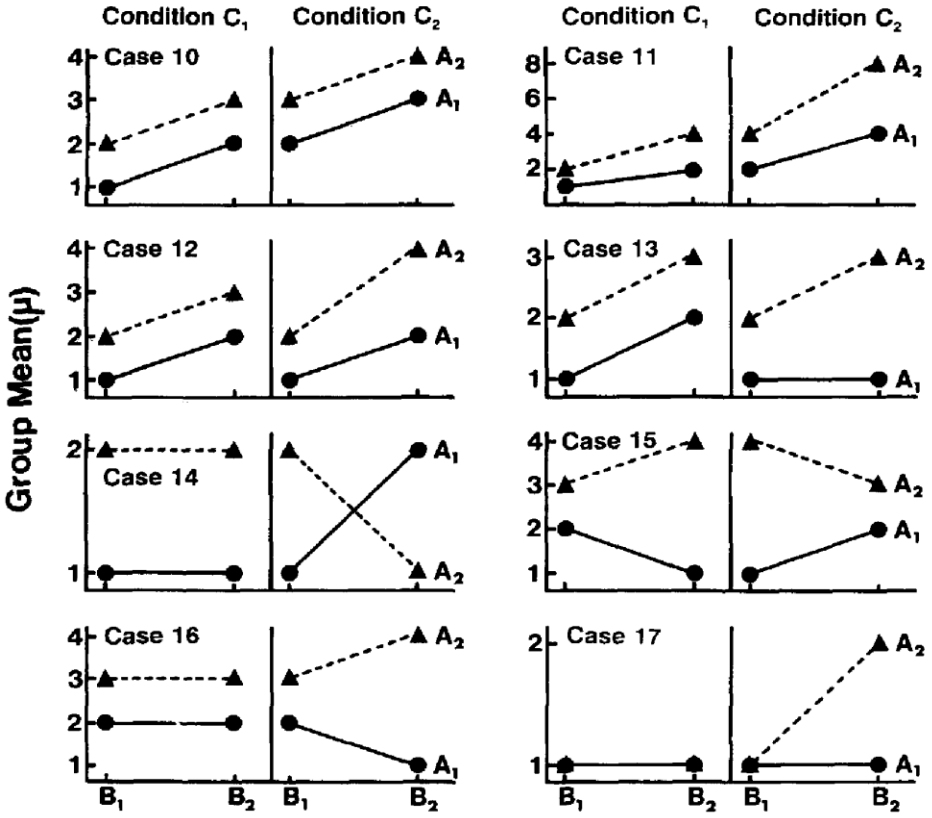

Figure 2. Population means for eight groups in a 2 × 2 × 2 factorial design with Factors A, B, and C. Only Case 10 should yield no significant interaction. Contrast values and sample sizes are listed in Table 5.

Comparing the B main effect with interaction,

$$\Psi_B^2 - \Psi_{A\times B}^2 = D_x^2 4 b_1 b_2. \qquad (15)$$

Table 5
Sample Sizes Needed to Detect Contrasts for Main Effects and Interactions in a 2 × 2 × 2 Design for the Cases Shown in Figure 2 With a Power of 90% When α = .05, Two-Tailed

	Case							
Contrast	10	11	12	13	14	15	16	17
A	18	20	14	11	23	9	9	47
B	18	20	14	25	—	—	—	47
C	18	20	291	207	—	—	—	47
A × B	—	160	291	207	23	—	18	47
A × C	—	160	291	207	—	—	18	47
B × C	—	160	291	207	23	—	—	47
A × B × C	—	1,426	291	207	23	29	18	47

Note. For each case, the differences among group means are such that $\eta^2 = .20$ and the sample size needed for a power of 90% for the overall F test is 10.
Dashes indicate value cannot be calculated because effect size is 0.

Four conclusions follow from this expression: (a) If either slope b_1 or b_2 is zero, the effect sizes and sample sizes are always equal for the B main effect and interaction. (b) If the signs of the slopes b_1 and b_2 are the same, the B main effect is always larger than the interaction. (c) If the signs of slopes b_1 and b_2 are opposite, the interaction effect is always larger than the B main effect. (d) None of these *conclusions* depend *on the location* or range of X values, although the degree of difference in sample sizes will depend on D_x, as shown in Equation l 5. The conclusion that the interaction effect will always exceed the main effect when the lines cross applies to the A main effect but not to the B main effect. Likewise, the conclusion that the interaction effect will always exceed the main effect when slopes have opposite signs applies to the B but not the A main effect.

Limitations

Equation 5 is a normal approximation, but Tables 1 and 2 indicate that it is a reasonably good approximation. When two observations are added to the usual n from the exact expression for a normal distribution, the result is very close indeed to a more elaborate calculation that is based on a central t approximation to the noncentral t distribution (Kraemer & Thiernann, 1987). Several approximations of the noncentral t distribution have been devised (Tiku, 1966), and a simple normal approximation is respectable in this company. The excellence of the approximation to several decimal places is not a major issue for sample size and power calculations, because n is always rounded upwards to an integer and power is of interest only to the nearest 5%. When sample size is calculated to yield desired power, the investigator typically tests a few more subjects than dictated by the algebra, just in case there is a subject misbehavior, apparatus failure, or other unexpected loss of data.

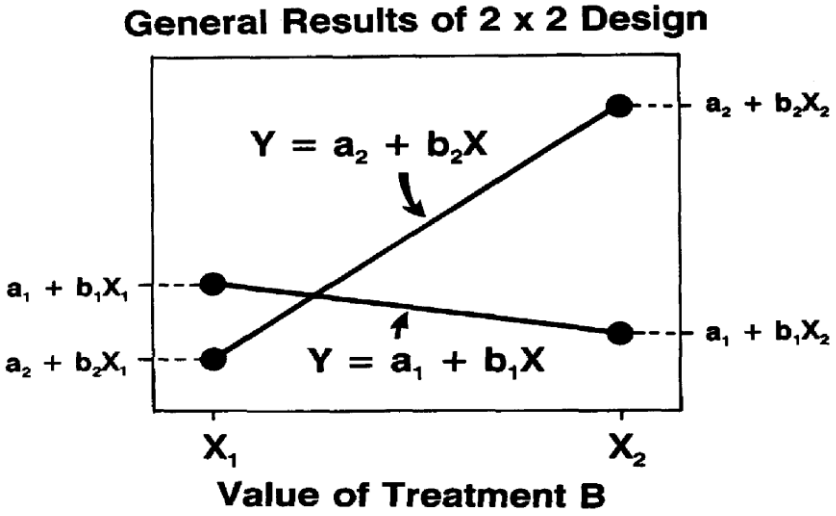

Figure 3. Expected mean scores for four groups in a 2 × 2 design in which group membership (A_1 or A_2) determines parameters of a linear equation and Treatment B determines the value of X.

Computers can facilitate certain kinds of power and sample size calculations. For example, the program by Borenstein and Cohen (1988) can determine n for a one-way design with J groups, given either group means and standard deviations or effect size, but it cannot do this for a contrast among J means, nor can it compute power or n for interaction in a two-way or three-way design using group means. The expert system by Brent, Scott, and Spencer (1988) can determine n for four levels of power and three effect sizes with a one-way design but not a factorial design, and it cannot work with contrasts. The SAS program can determine power, given degrees of freedom and the noncentrality parameter for the noncentral χ^2, F, or t distributions (Hewitt & Heath,1988), and the noncentrality parameter can be determined from effect size as indicated in Wahlsten (1990). Of course, it would be simple to implement Equations 2, 5, and 8 in a program.

Throughout this article, I have assumed that the model involves fixed effects rather than random effects. Power and sample size can be readily determined for a random-effects ANOVA model (Koele, 1982), if this is appropriate for the data in question, but the normal approximation in Equation 5 would not be suitable.

The method advocated in this article appears to give good results even when sample sizes are relatively small. Of course, these calculations will be credible only when the distributions of observations within each group are

normal or nearly so. The consequences of departures from normality for power and sample size related to tests of interaction and complex contrasts remain to be explored. A Monte Carlo approach (e.g., Soper, Cicchetti, Satz, Light, & Orsini, 1988) might be useful for this purpose. Tests that are based on the normal distribution generally have slightly greater power than the so-called "nonparametric" tests only when the normal assumptions hold. Recent investigations demonstrate that relative power of alternative tests depends strongly on the shape of the distribution (Blair & Higgins, 1985). Bootstrap methods have been advocated for computing confidence intervals when the shapes of the underlying distributions are unknown (Efron, 1988; see also Rasmussen, 1988; Strube, 1988). However, the bootstrap, which is applied to the facts observed, provides no direct aid when sample size must be chosen before data collection.

When many tests of significance are done in the course of a large study it would be wise to use a Bonferroni adjustment for the criterion for a Type I error. Equation 5 can still be used to estimate sample size, although *n* will be larger when α is reduced. The ratio of sample sizes in Equation 10, however, will not be changed.

If the *t* or *F* test is good enough for the data collected, then choice of sample size using Equation 5 should be reasonable. I hope that the simplicity of this approach will encourage more investigators to embark on a quest for interaction or other interesting contrast effects with sufficient power provided by an adequate sample. Despite well-established principles of sample size determination and the availability of works making these comprehensible to the users of statistical analysis, Rosnow and Rosenthal (1989) as well as Cohen (1990) observed recently that many psychologists continue to ignore the question of power.

References

Bauer, S. J., & Sokolowski, M. B. (1988). Autosomal and maternal effects on pupation behavior *in Drosophila melanogaster. Behavior Genetics, 18,* 81-97.

Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin, 97,* 119-128.

Borenstein, M., & Cohen, J. (1988). *Statistical power analysis: A computer program.* Hillsdale, NJ: Erlbaum.

Brent, E. E., Jr., Scott, J. K., & Spencer, *J.* C. (1988). *EX-SAMPLE*^{Tm}. An *Expert System to Assist in Designing Sampling Plans.* Columbia, MO: Idea Works.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,1304-1312.*

Efron, B. (1988). Bootstrap confidence intervals: Good or *bad? Psychological Bulletin, 104,* 293-296.

Glass, G. V, & Hakstian, A. R. (1969). Measures of association in comparative experiments: Their development and interpretation. *American Education Research Journal, 6,* 403-414.

Hays, W. L. (1988). *Statistics* (4th ed). *New* York: Holt, Rinehart *&* Winston.

Hewitt, J. K., & Heath, A. C. (1988). A note on computing the chi-square noncentrality parameter for power analysis. *Behavior Genetics, 18,* 105-108.

Koele, P. (1982). Calculating power in analysis of variance. *Psychological Bulletin, 92,* 513-516.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research.* Newbury Park, CA: Sage.

Lachenbruch, P. A. (1988). A note on sample size computation for testing interactions. *Statistics in Medicine,* 7,467-469.

Levin, J. R. (1975). Determining sample size for planned and post hoc analysis of variance comparisons. *Journal of Education Measurement, 12,* 99-108.

Marascuilo, L. A, & Serlin, R. C. (1988). *Statistical methods for the social and behavioral sciences.* New York: W H. Freeman

Maxwell, S. E., Camp, C. J., & Arvey, R. D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology,* 66,525-534,

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data. A model comparison approach.* Belmont, CA: Wadsworth.

Neyman, J. (1935). Comments on Mr. Yates' paper. *Journal of the Royal Statistical Society,* (Suppl. 2), 235-241.

Rasmussen, J. L. (1988). "Bootstrap confidence intervals: Good or bad": Comments on Efron (1988) and Strube (l988) and further evaluation. *Psychological Bulletin, 104,* 297-299.
Rodger, R. S. (1974). Multiple contrasts, factors, error rate and power. *British Journal of Mathematical and Statistical Psychology, 27,* 179 - 198.
Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 49,*1276-1284.
Soper, H. V, Cicchetti, D. V, Satz, P., Light, R, & Orsini, D. L. (1988). Null hypothesis disrespect in neuropsychology: Dangers of alpha and beta errors. *Journal of Clinical and Experimental Neuropsychology, 10,* 255-270.
Strube, M. J. (1988). Bootstrap Type I error rates for the correlation coefficient: An examination of alternate procedures. *Psychological Bulletin, 104,* 290-292.
Tiku, M. L. (1966). A note on approximating to the noncentral *F* distribution. *Biometrika, 53.* 606-610.
Traxler, R. H. (1976). A snag in the history of factorial experiments. In D. B. Owen (Ed.), *On the history of statistics and probability* (pp. 283-295). New York: Marcel Dekker.
Wahlsten, D, (1983). Maternal effects on mouse brain weight. *Developmental Brain Research, 9.* 215-221.
Wahlsten, D. (1990). Insensitivity of the analysis of variance to heredity-environment interaction. *Behavioral and Brain Sciences, 13* , 109- 120.
Wainwright, P (1980). Relative effects of maternal and pup heredity on postnatal mouse development. *Developmental Psychobiology, 13,* 493-498.
Wimer, C. C., & Wimer, R. E. (1989). On the sources of strain and sex differences in granule cell number in the dentate area of house mice. *Developmental Brain Research, 48,* 167-176.