Overcoming Feelings of Powerlessness in "Aging" Researchers: A Primer on Statistical Power in Analysis of Variance Designs

Joel R. Levin
University of Wisconsin—Madison

A general rationale and specific procedures for examining the statistical power characteristics of psychology-of-aging empirical studies are provided. First, 4 basic ingredients of statistical hypothesis testing are reviewed. Then, 2 measures of effect size are introduced (standardized mean differences and the proportion of variation accounted for by the effect of interest), and methods are given for estimating these measures from already-completed studies. Power and sample size formulas, examples, and discussion are provided for common comparison-of-means designs, including independent samples 1-factor and factorial analysis of variance (ANOVA) designs, analysis of covariance designs, repeated measures (correlated samples) ANOVA designs, and split-plot (combined between- and within-subjects) ANOVA designs. Because of past conceptual differences, special attention is given to the power associated with statistical interactions, and cautions about applying the various procedures are indicated. Illustrative power estimations also are applied to a published study from the literature. It is argued that psychology-of-aging researchers will be both better informed consumers of what they read and more "empowered" with respect to what they research by understanding the important roles played by power and sample size in statistical hypothesis testing.

I wish to thank two anonymous reviewers for their extremely beneficial comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Joel R. Levin, Department of Educational Psychology, 1025 West Johnson Street, University of Wisconsin—Madison, Madison, Wisconsin 53706. Electronic mail may be sent via Internet to levin@macc.wisc.edu.

In this article, I use the term effect size in a broad sense to include a variety of measures that reflect the degree of nonoverlap of two or more normal distributions rather than the narrower definition of a standardized difference in means that is used by some. The broader concept, of course, encompasses the narrower.
Power for Basic Comparisons of Means

Example 1: Comparing Two Independent Means in One-Factor ANOVA Designs

I now attempt to make the abstract concrete in the context of a hypothetical study that compares the cognitive performance of older adults who were randomly assigned to two experimental conditions. In Condition 1 (instruction), the participants had received several weeks of special cognitive task instruction before performing the final task; in Condition 2 (no instruction control), participants were simply asked to perform the final cognitive task. Before collecting the data, the researcher is willing to specify a Type I error probability of .05 for an $F$ test (or, equivalently, for a non-directional $t$ test) of the hypothesis that the $K = 2$ population means are equal. In addition, suppose that extensive previous work with the cognitive measure has informed the researcher that a mean difference between instructed and uninstructed adults of at least 6 points is a practically important one to detect. Note that in the current context, the 6-point difference is the “effect size” of which I have been speaking. However, such a raw score difference in means is useful for sample size or power determination only if it is accompanied by the known variability associated with the cognitive task measure in the instructed and uninstructed populations. Thus, to proceed along these lines, one also would need to know or specify the magnitude of the within-populations variability; for the current example, for instance, it might be known that the common standard deviation, $\sigma$, is equal to 10 points.

Effect size as a standardized mean difference. What if—as is true in the more usual situation—one cannot specify the minimum effect of interest in raw score units, or if no measure of within-treatments variability is available? No problem. A researcher simply needs to think of an effect in more global terms, namely, as relative score units, or a standardized mean difference $(\mu_1 - \mu_2)/\sigma$, and to which I refer hereafter as $\psi$ (see also Levin, 1975). For the current example, suppose that after careful consideration of effect sizes based on similar previous work on the topic, the researcher decides that a standardized mean difference between instructed and uninstructed adults of at least $0.6 \ SD$ (i.e., $\psi = .60$) is a practically important one to detect. Moreover, if the assumption of normality is added, the concept of a standardized mean difference can be translated directly into the proportion of distributional overlap (or non-overlap) involving the two populations. With $\psi = .60$, for example, the means of the two experimental conditions would be $0.6 \ SD$ apart, which indicates that (a) the mean (or 50th percentile) of the higher scoring group corresponds to the 73rd percentile of the lower scoring group, which in turn indicates that (b) there is 38% nonoverlap in scores from the two combined distributions (see, e.g., Cohen, 1988). For two-sample comparisons of means, Cohen suggested that a $\psi_e$ (which Cohen calls $d$) of .20 might be considered a “small” effect, .50 a “medium” effect, and .80 a “large” effect. These ballpark values are not universals, however and are likely to vary from discipline to discipline and even from research question to research question within disciplines. Yet, Cohen’s small, medium, and large adjectives are decidedly not arbitrary, having been
derived from extensive surveys of findings in the psychological research literature (see also Glass, 1979). Theoretical, practical, or previous empirical considerations can assist a researcher in deciding on the value of $\psi$, that is useful in each particular power—sample size determination situation. Researchers need not have raw scale effect size information to make meaningful power—sample size calculations; they can simply rethink their effect sizes in relative $\psi$ terms. Alternative relative effect size measures that may prove helpful in certain situations are described shortly.

With these specifications of alpha and effect size for this hypothetical study, along with a decision to include, say, 16 participants per condition (determined on the basis of time, monetary, or participant availability considerations, or simply because 16 participants per condition "sounds good"), the researcher's feelings of powerlessness or lack of sensitivity certainly are justifiable. Why? Let me examine the facts of this case. Given that (a) $\alpha = .05$, (b) the specified minimum effect of interest is given by $\psi = .60$, and (c) $n$, the number of participants per condition, is 16, one can determine (d) the probability that the proposed study will detect a difference between instructed and uninstructed adults (i.e., the study's power). Using available tables (e.g., Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990; Tiku, 1967), charts (e.g., Kirk, 1995, Tables E.12 and E.13), or microcomputer programs (e.g., Elashoff, 1995; Erdfelder, Faul, & Buchner, 1996) or simply performing hand calculations with either exact or approximation formulas (e.g., Levin, 1975; Levin & Serlin, 1981)—which are provided for each of the situations discussed in this article—reveals that the answer to the power problem is .375.

What does this mean? Perhaps surprisingly, based on the characteristics of this particular study ($\alpha = .05$ and $n = 16$ participants per condition), it means that even if a 0.6 SD difference between the two conditions exists, the researcher would have less than a 40% chance of detecting that difference (i.e., the probability is only .375 that the null hypothesis will be rejected). Complementarily, this study's characteristics reveal that the probability is .625 that the researcher will commit a Type II error. If the effect size in fact is greater than that which the researcher designated as useful (i.e., in reality, the difference in means exceeds the specified 0.6 SD difference), then the researcher's power would exceed .375, which is why the minimum effect of interest should be included in the specifications.

For the current example, the researcher may well be justified in formulating a directional (one-tailed) alternative hypothesis, namely, that adults given the special instruction will outperform those in the no-instruction control group. There are positive power consequences of making such a directional prediction, for if the obtained difference in means should turn out to be in the predicted direction (i.e., experimental greater than control), then the researcher is rewarded with an increase in power. Here, that power increase is substantial, from .375 for the two-tailed test to .51 for the one-tailed test. As long as the direction of difference is in fact both justified (i.e., a difference in the opposite direction is not meaningful, definable, etc.) and specified before data collection (i.e., predicted) rather than after ("postdicted"), then testing directional hypotheses clearly is advantageous. Nonetheless, even with a one-tailed test in this situation, power of only .51 to detect the 0.6 SD difference that was considered useful should hardly warm the researcher's heart.

However, how much power should be regarded as acceptable and how can that more pleasant state of affairs be achieved? Concerning the first part of the question, power and Type II error conventions are not as firmly entrenched as are the previously noted Type I error conventions. Clearly, however, power greater than .51 to detect desired effects would be a reasonable goal, and powers in the neighborhood of .75–.90 would undoubtedly be the envy of most researchers. Indeed, Cohen (1988) and other applied statisticians typically tout powers of .80 as respectable for detecting one's specific effects of interest.

In this article, most of the examples and calculations are based on a researcher's a priori decision to have an 80% chance of detecting a specified effect of interest ($1 - \beta = .80$) while controlling the probability of detecting a chance effect at .05 ($\alpha = .05$).

As an aside, some might wonder why the acceptable risks of Type I and Type II errors are so inequitable, protecting the former more so than the latter (.05 vs. .20, respectively). This typically boils down to a matter of the perceived seriousness of consequences associated with each type of error. In most (but not all) research-based decisions, incorrect rejections are deemed more serious than incorrect nonrejections: It is better to miss something that is there (Type II error) than to "find" something that is not there (Type I error). I strongly endorse this hypothesis-testing philosophy and have previously argued (e.g., Levin, 1985, pp. 225–226), in accord with Greenland (1975) and others, that Type I errors are the root of much of what is evil about current publication practices: Type I error "findings" are published, are elevated to the status of truths, lead other researchers on wild goose chases and failures to replicate, and take many years and dollars to undo. Type II errors, on the other hand, typically lead to more circumspect conclusions, promote alternative and more efficient approaches for examining a problem, and (at least according to optimist philosophy) will "improve" treatments to such a degree that rightful effects will ultimately be uncovered—to science's benefit. Of course, in the best of all worlds, both $\alpha$ and $\beta$ should be set comparably low. In the real world, however, a researcher's resource limitations (including psychological ones, such as patience and stamina) invariably preclude that. For example, to detect a 0.5 SD difference between two means ($\psi = .50$) based on a nondirectional test with $\alpha = \beta = .05$ requires a total of 210 participants. How many researchers' budgets and psyches could stand that?

Now, concerning the second part of the preceding question, although there are a number of power-achieving options (which I consider throughout this article), one such option is to respond to the earlier sample size question. Applying any of the previously indicated tools (i.e., tables, charts, computer programs, 

---

3 Another option for increasing an empirical study's statistical power is to increase the dependent variable's reliability (say, e.g., by increasing the number of test items or measures or by eliminating poorly functioning ones). That psychometric tactic is separate and distinct from the focus here and is therefore not considered. The interested reader should consult sources such as Cleary and Linn (1969), Cohen (1988), and Lipsey (1990).
formulas) provides the answer to that question: With $\alpha = .05$ and, say, power of .80 to detect the 0.6 SD difference of interest in the current example, one would need to include 45 or 36 adults in each experimental condition for a two-tailed or a one-tailed test, respectively (corresponding to total sample sizes of 90 and 72, respectively). As such, it is easy to see why the originally considered study with a mere 16 participants per condition meant "no chance of success," at least in the mind of the hypothetical reviewer or statistical consultant. Alternatively, a reviewer’s reaction that the study’s actual sample size was inadequate might justifiably elicit an "oughta quit" reaction on the part of the researcher.

It also is easy to see why often-heard claims by researchers (e.g., "The power for this study was .80") are semantically empty ones: Power to detect what sized effect, and with what Type I error probability, given, the particular sample size? That is, the correct (and complete) claim for the current example needs to include the following: "The power for this study is .80 to detect a 0.6 within-conditions standard deviation difference between means based on a two-tailed $\alpha$ of .05 and 45 participants in each condition."

Useful, and versatile, approximation formulas for yielding the preceding results are as follows:

$$z_2 = z_1 - (n^*/2)^{1/2}(|\mu _1 - \mu _2|/\sigma ),$$

where $n^* = n - 2$

and

$$n^* = 2(z_1 - z_2)/[(\mu _1 - \mu _2)/\sigma ]^2,$$

which, because $(\mu _1 - \mu _2)/\sigma = \psi_2$, may be equivalently written, respectively, as the power equation

$$z_2 = z_1 - [(n^*/2)^{1/2}|\psi_2|],$$

where $n^* = n - 2$  \hspace{1cm} \text{(1)}

and the sample size equation

$$n^* = 2(z_1 - z_2)^2/\psi_2^2,$$

where $n^* = n - 2$  \hspace{1cm} \text{(2)}

where $z_1$ is the score in a standard normal distribution corresponding to either the 100$[1 - (\alpha/2)]$ or 100$(1 - \alpha )$ percentile for two-tailed and one-tailed tests, respectively, $z_2$ is the score in a standard normal distribution corresponding to the 100$\beta$ percentile, and $n$ is the number of participants in each experimental condition (see Levin & Serlin, 1981). With $n^* = n - 2$ in Equation 1 or $n = n^* + 2$ in Equation 2, the approximations are appropriately conservative and reasonably close to what would be obtained through exact calculations.

I now illustrate the use of each of these equations for the Example 1 specifications. To determine power with $\alpha = .05$ and, say, power of .80, one has $z_2 = z(.20) = -0.84$, with the same other specifications, and by Equation 2:

$$n^* = 2[(1.96 - (-.84))^2/(.60)^2] = 43.56,$$

which, because these are human beings I am talking about, needs to be rounded (always up) to the nearest integer, or $n^* = 44$, with the augmentation of 2, this becomes $n = 46$ (compared with the exact calculation of $n = 45$).

As an alternative to conducting approximate power and sample size analyses via Equations 1 and 2, one can refer to Lipsey’s (1990, pp. 90–96) helpful approximation charts. In addition, two recent exact sample size and power determination microcomputer programs, nQuery Advisor (Elashoff, 1995) and GPOWER (Buchner, Faul, & Erdfelder, 1992; Faul & Erdfelder, 1992) are highly recommended. A graph from the GPOWER program, summarizing the power and sample size adventures related to the current example, is shown in Figure 1.

A directional (one-tailed) test of the difference between the two condition means, which might well be justified here, one would replace $z_1 = 1.96$ with $z_1 = z(.95) = 1.65$ in the above calculations. The corresponding approximate power and sample size values are $1 - \beta = .48$ (exact = .51) for 16 participants per condition and $n = 37$ (exact = .36) for power of .80, respectively.

**Effect size as a proportion of explained variation.** A second effect size measure that I use in the examples in this article is the proportion of total variance in the dependent measure that can be "explained" or "accounted for" by the independent or manipulated variable, or

$$\sigma_\text{explained}/\sigma_\text{total} = \sigma_\text{explained}/(\sigma_\text{explained} + \sigma_\text{error}).$$

In the current example, this would be the proportion of variance in adults’ cognitive task performance that can be accounted

---

*Although power also can be determined for unequal sample size situations (Cohen, 1988), I do not do so here. With "near-equal" sample sizes, approximate power can be obtained from the formulas provided in this article by replacing the common $n$ with the harmonic mean of the sample sizes.

* I am grateful to Janet Parks of Bowling Green University for bringing the GPOWER power-sample size programs to my attention. Although I consider only comparison-of-means situations in this article, GPOWER also provides power and sample size computations underlying tests of correlation and multiple regression coefficients. nQuery Advisor promises the same (although they are not part of Version 1.0) and additionally includes power—sample size procedures for survival analysis and interval estimation, and it can handle unequal $n$ situations. Both programs provide procedures for tests of proportions. GPOWER (both the MS-DOS and Macintosh versions) may be obtained from the authors at no charge according to the instructions provided by Erdfelder, Faul, and Buehner (1996). For more information about GPOWER, contact either Franz Faul, Department of Psychology, University of Kiel, Olshausenstr. 40, D-24098 Kiel, Germany (electronic mail address: gpo12@rz.uni-kiel.de), or Edgar Erdfelder, Department of Psychology, University of Bonn, Roemerstr. 164, D-53117 Bonn, Germany (electronic mail address: erdfelder@uni-bon.de). For information about nQuery Advisor (for Windows 3.1 or higher), contact Statistical Solutions Ltd., 60 State Street, Suite 700, Boston, Massachusetts 02109 (electronic mail address: info@statsolusa.com).
for by their instructional condition. In comparison-of-means contexts, proportion of variation accounted for, or "strength of relationship," measures include $\eta^2$, $\omega^2$, and $e^2$. I use $\omega^2$ for the current power and sample size considerations because its sample estimator is associated with only a slight degree of bias in estimating its population analog (Kirk, 1995, pp. 177–180; Maxwell, Camp, & Arvey, 1981). As indicated previously, substantial nonoverlap between two score distributions implies a large standardized mean difference, as represented by a large value of $\psi$, which in turn implies a strong relationship between the independent variable (e.g., conditions) and the dependent variable (e.g., performance), as represented by a large value of $\omega^2$. Cohen (1988, pp. 280–284) defined $f^2$ as the ratio of explained variance to unexplained variance, or

$$f^2 = \frac{\sigma_{explained}^2}{\sigma_{error}^2} = \frac{(\sigma_{explained}^2/\sigma_{total}^2)}{(\sigma_{error}^2/\sigma_{total}^2)} = \omega^2/(1 - \omega^2).$$

(3)

The complementary relationship is given by

$$\omega^2 = f^2/(1 + f^2).$$

(4)

For the two-sample problem, $f$ can be shown to be simply half of $\psi$, or $f = \psi/2$.

Returning to the two-group cognitive instruction example, then, one sees that the effect size given by $\psi = .60$ corresponds to $f = .30$. When substituted into Equation 4, this is $\omega^2 = .083$. Thus, a 0.6 $SD$ difference in means is equivalent to instructional conditions accounting for about 8% of the total cognitive task variation. Note that in explained variation terms, $\psi = .20$ (Cohen's small effect) is equivalent to $\omega^2 = .01$, $\psi = .50$ (medium effect) is equivalent to $\omega^2 = .06$, and $\psi = .80$ (large effect) is equivalent to $\omega^2 = .14$. In addition, for $\psi = 1$, $\omega^2 = .20$; for $\psi = 2$, $\omega^2 = .50$. Note that like $\psi$, $\omega^2$ is a relative (scale-free) measure of effect size. The two different but transformationally equivalent measures are presented and developed here because of differences in researcher preference in conceptualization. That is, some researchers will prefer to think of effect size as a standardized mean difference, whereas others will prefer to think of it as the proportion of the dependent measure’s variation that is accounted for by the independent variable.

Determining power or sample size according to the alternative $\omega^2$ conceptualization of effect size can be accomplished in the two-sample case either approximately or exactly. Approximating the needed power or sample size is as simple as (a) converting $\omega^2$ to $f$ through Equation 3, (b) dividing $f$ by 2 to obtain $\psi$, and (c) using the $n$-adjusted Equations 1 or 2 to determine power or sample size, respectively. Determining exact power or sample size can be accomplished through tables, charts, or

---

**Figure 1.** Required total sample sizes ($N$) to achieve different powers for detecting a 0.8-within-groups standard deviation difference between two independent means, based on a two-tailed alpha of .05 (from the GPOWER microcomputer program of Buchner, Faul, & Erdfelder, 1992).
microcomputer programs that incorporate an explained variance measure or one of its derivatives presented here. To consult standard tables (Tiku, 1967) or the classic Pearson and Hartley (1951) charts (e.g., Maxwell & Delaney, 1990, Table A.11; Kirk, 1995, Table E.12) requires initial calculation of $\phi$, a function of the $F$ distribution's "noncentrality parameter" and that can be defined in terms of $u$ as

$$\phi = N^2/\omega^2(1 - \omega^2)]^{1/2}$$

(5)

$$= (N^2f^2)^{1/2},$$

(5a)

where $N'$ is equal to $n$ in the two-sample case. More generally, and as will be used from this point on, if $N$ is the total (across groups or conditions) sample size—or the total number of observations in within-subjects designs, to be discussed later—then $N' = N/(\nu_{\text{error}} + 1)$. For the two-sample case, $\nu_{\text{error}} = 1$ and therefore $N' = N/2 = n$. When Equation 5 is used in conjunction with the Pearson-Hartley charts, power can be determined in just one step, whereas sample size determination involves a two- or three-step iterative process. Note also that by substituting previously shown equivalences into Equation 5, one can define a two-sample $\phi$ in terms of the standardized mean difference measure, $\psi_a$, namely, as

$$\phi = n^{1/2}\psi_a/2.$$  

(6)

I now check the previous power calculations when expressed in explained variance terms and incorporated into Kirk's (1995) Tables E.12 and E.13. As shown earlier, for $\psi_a = .60$, $\omega^2 = .083$. With the Pearson and Hartley charts in Table E.12, I first compute $\phi = [16(.083/917)]^{1/2} = 1.20$. Consulting the charts with $\phi = 1.20$, $\nu_1 = \nu_{\text{error}} = 1$, $\nu_2 = \nu_{\text{error}} = N - K = 32 - 2 = 30$, and $\alpha = .05$ (for a two-tailed test), one finds that power is somewhere in the mid-to-high .30s (when I computed it exactly before it was .375). In the current case, one also can determine sample size, in terms of selected values of $\omega^2$, directly from Table E.13 (only for powers of .70, .80, and .90). For these specifications, one finds that for $\omega^2 = .083$, the per condition sample size required to yield power of .80 is about $n = 45$ (which corresponds to the previously calculated exact value).

Estimating effect sizes from previous empirical studies. One of the most difficult things for researchers who are earnestly beginning to think about power and sample size issues is deciding what effect size values (in the current context, values of either $\phi$ or $\omega^2$) to specify as being useful. Short of having firm external criteria on which to base that decision, and foregoing the previously provided Cohen (1988) ballpark adjectives (small, medium, and large), one can make estimates or assessments of effect size magnitudes from either the literature in one's field or one's own prior (including pilot) research (e.g., Glass, 1979). The plethora of meta-analyses conducted today, on virtually every topic from soup to nuts, provides just such effect size magnitudes. Across-study averaged sample $d$s (my $\psi_S$s) or $\omega^2$s are the 'stuff' of meta-analyses, and that stuff may be used as estimates in the preceding power—sample size procedures. In cases in which such effect size measures are not provided directly, one can obtain them through various statistical equivalences. For example, in the two-group $t$ or $F$ test situation, one can estimate $\psi_a$ in a couple of different ways. If means and standard deviations (or variances) are provided, the simplest way to estimate $\psi_a$ is

$$\text{estimated } \psi_a = (M_1 - M_2)/(\text{MSE})^{1/2},$$

(7)

where $M_1$ and $M_2$ are the two sample means, and $\text{MSE}$ is the mean square error (or pooled sample variance). If mean and variability information are not available but $t$ or $F$ statistics are, one can estimate $\psi_a$ as

$$\text{estimated } \psi_a = t(2/n)^{1/2} = (2F/n)^{1/2}.$$  

(8)

Similarly (see Kirk, 1995, p. 178), $\omega^2$ can be estimated as

$$\text{estimated } \omega^2 = (t^2 - 1)/[(t^2 - 1) + N]$$

$$= (F - 1)/[(F - 1) + N].$$

(9)

from a $t$ test or an ANOVA based on the $K = 2$ experimental conditions. Equivalently, $\omega^2$ can be estimated from the ANOVA sum of squares between conditions ($SSB$), sum of squares total ($SST$), and $MSE$ as

$$\text{estimated } \omega^2 = (SSB - MSE)/(SST + MSE).$$

(10)

An estimated $\omega^2$ that is negative (which is possible when calculated on the basis of sample data) should be regarded as zero.

Thus, effect sizes are "out there" in the literature, waiting for the diligent sample size determiner to uncover. It is important to mention, however, the following: (a) The effect sizes out there may or may not be of the magnitude that a particular researcher will consider to be useful when specifying a minimum effect of interest in his or her own sample size calculations. (b) Effect sizes derived from published studies are likely to be inflated or biased against the null hypothesis (e.g., Greenwald, 1975) estimates, as a result of such studies having made it to the stage of publication in the first place (see also Frick, 1995; Greenwald et al., 1995; Rosenthal, 1979). (c) The two alternative sample-based formulas provided here for estimating effect sizes ($\psi_a$ and $\omega^2$, respectively) generally will not produce the same sample size and power estimates. They are alternative ways of conceptualizing an effect size based on sample estimators that have differing degrees of bias associated with them (more and less for estimated $\psi_a$ and $\omega^2$, respectively). If it is important that the two conceptualizations correspond exactly, then bias-adjusting modifications of the estimated $\psi_a$ measure can be made (see Kirk, 1995, pp. 178–180).

Example 2: Comparing Three or More Independent Means in One-Factor ANOVA Designs

The just-discussed power—sample size concepts and procedures generalize directly to situations in which one wishes to
compare three or more independent means simultaneously in the one-way ANOVA situation. I illustrate by adding a third condition to the hypothetical experiment with older adults. Thus far, I have specified a no-instruction control condition and a cognitive instruction condition. Any positive cognitive outcomes associated with the latter condition might indeed be a consequence of the cognitive instruction received. On the other hand, they may be due in part (or even exclusively) to the adults' exposure to and practice with 'thinking' activity of any kind, social interaction with the experimenter, or any number of other experimental factors. That is, perhaps it is not cognitive instruction that is crucial to improved cognitive task performance. To investigate that possibility, the researcher designs a study with a second instructional condition, "nancognitive," irrelevant, or placebo instruction, which provides participants with practice (comparable in form and amount to that in the cognitive instruction condition) on tasks unrelated in content to those making up the focal criterion task. Thus, there are now three conditions to compare: no instruction (Condition 1), placebo instruction (Condition 2), and cognitive instruction (Condition 3). However, how many participants should be included?

Levin (1975) generalized the just-developed $\psi_0$ construct from the $K = 2$ to $K > 2$ group situation. The details of that generalization are not discussed here, but are summarized. The bottom line is that a standardized difference in two means is a really a special case of a standardized linear contrast

$$\psi_0 = (a_1\mu_1 + a_2\mu_2 + \cdots + a_K\mu_K)/\sigma = \sum_{k=1}^{K} a_k\mu_k/\sigma, \quad (11)$$

where $\sigma$ represents the (assumed common) within-populations standard deviation, and $a_k$ represents the chosen coefficients or weights applied to the set of $K$ means, subject to the restriction that the coefficients sum to zero. For the previous Example 1, the unstandardized mean difference of interest, $\psi = \mu_1 - \mu_2$, can be alternatively expressed using the current contrast coefficient notation as $\psi = (1)\mu_1 + (-1)\mu_2$, which satisfies the required zero-sum-of-coefficients criterion. Moreover, the constant 2 that is apparent in Equations 1 and 2 also is a special case of including only two (out of any $K$) means in the contrast. The generalized ($K > 2$ means in the contrast) term is reflected in the sum of the squared coefficients, $\sum_{k=1}^{K} a_k^2$, which for only two means included equals $(1)^2 + (-1)^2 = 2$ because all other coefficients ($a_k$) would be zero. With a pairwise contrast or comparison defined as a difference involving only two means and a complex contrast defined as a difference involving more than two means with nonzero coefficients (e.g., $\sum [\mu_1 + \mu_2] - \mu_3$), it should be emphasized that the procedures of this section can be applied equally appropriately to either pairwise or complex contrasts. With these generalizations, then, the previously discussed standardized mean difference approach for determining power or sample size is more universally applicable. Two such approaches are now briefly described; for more detailed information, see Levin (1975) and Levin and Serlin (1981).

Specifying the minimum contrast of interest for an omnibus ANOVA F test. Embedding a specified contrast of interest in the context of an omnibus ANOVA F test requires application of the previously mentioned $\phi$-based charts or tables. In that regard, the $K$-group ANOVA generalization of Equation 6 is

$$\phi = [n\psi^2/((\nu_1 + 1) \sum_{k=1}^{K} a_k^2)]^{1/2}, \quad (12)$$

based on $\nu_1 = K - 1$ df and either $\nu_2 = K(n - 1)$ error df for the power problem or $\nu_2 = \infty$ df for the first iteration of the sample size problem; see Kirk (1995, pp. 185–186) for the special case representation of Equation 12 as applied to pairwise comparisons. To illustrate a power calculation for the three-group example, suppose that the researcher will be conducting a one-way ANOVA F test based on $\alpha = .05$, with 16 participants in each of the three experimental conditions (thus, $\nu_1 = 2$ and $\nu_2 = 45$). If a difference of 0.8 of a within-groups standard deviation exists between any two conditions, that is, $\psi_s = [(\mu_1 - \mu_2)/\sigma] = .80$, and—for the worst-case scenario—all other treatment effects are assumed to be equal to zero, how much power would the researcher have to reject the no-difference hypothesis (i.e., to obtain a statistically significant omnibus F test)? Applying Equation 12, one finds that $\phi^2 = 16(.80)^2/(2 + 1)(1^2 + (-1)^2) = 1.71$ and therefore that $\phi = 1.31$. Entering the Pearson-Hartley charts (based on $\alpha = .05$, $\nu_1 = 2$, and $\nu_2 = 45$), one finds that power is equal to about .49. This result was confirmed by Faul and Erdfelder's (1992) GPOWER microcomputer program (obtained by specifying $\mu_1 = -A$, $\mu_2 = 0$, $\mu_3 = A$, and $\sigma = 1$, resulting in $p = .5266$), which also indicates that twice as many participants per condition (i.e., $n = 32$) are required to achieve power of .80 for the current specifications.

To relate $\psi_s$ to Cohen's $f$ directly, note that from Equation 5a

$$f^2 = \phi^2/N'. \quad (5a)$$

Extending this and incorporating the definition of $\phi$ in terms of $\psi_s$ (Equation 12), one has for the one-factor ANOVA model

$$f^2 = [n\psi^2/((\nu_1 + 1) \sum_{k=1}^{K} a_k^2)]^{1/2}, \quad (12)$$

$$f^2 = [n\psi^2/((\nu_1 + 1) \sum_{k=1}^{K} a_k^2)]/N'$$

$$= n\psi^2/N \sum_{k=1}^{K} a_k^2$$

and, because $N/n = K$,

$$f^2 = \psi^2/K \sum_{k=1}^{C} a_k^2$$

or, in more general terms,

$$f^2 = \psi^2/(C \sum_{i=1}^{C} a_i^2), \quad (13)$$

where $C$ is the number of cells in the design. This relationship between $f$ and $\psi_s$ will continue to prove valuable throughout this article and for application to the microcomputer power and sample size packages beyond. For example, for the preceding three-group application to GPOWER, the pairwise $\psi_s$ of .80 had to be converted to $f$, which could be alternatively accomplished via Equation 13 as $f^2 = (.80)^2/3(2) = .1067, \quad f = .3266$. 


One more thing is worth noting in this context. Contrary to what was previously thought (Levin, 1975), calculating the power or sample size associated with $\phi_x$, as was just done, refers only to the probability of rejecting the omnibus $F$ test hypothesis given the specified value of $\psi_x$. It does not refer to the probability of subsequently finding a significant contrast via one of the variety of controlled multiple-comparison procedures at the researcher's disposal (see, e.g., Seaman, Levin, & Serlin, 1991). To approximate multiple-comparison power-sample size associated with the so-called "simultaneous" multiple-comparison procedures (e.g., those developed by Scheffé, Tukey, Dunnett, among others, which, in each case, have critical values that can be reexpressed as a critical $t$-like value), Equations 15 and 16 may be used with the critical $t$-like value substituted for $z_i$ in those equations (e.g., Levin & Serlin, 1981).

Specifying the minimum contrast of interest for a planned $t$ test of that contrast. As I have attempted to argue vigorously in the past (e.g., Levin, 1975, 1985), researchers should abandon omnibus ANOVAs in favor of testing planned contrasts (including directional ones) whenever they are in a position to justify doing so. The consequences of that are almost always increased power, as I now demonstrate. Suppose that the researcher's primary interest is in the comparison between the cognitive instruction condition (Condition 3) and the placebo control condition (Condition 2). If that contrast amounted to at least 0.8 of a within-conditions standard deviation difference, how much power would the researcher have to detect it in the current context based on $K = 3, n = 16$, and $\alpha = .05$? Equation 12 can be readily adapted to answer this question by recomputing $\phi$ as follows:

$$\phi = (n\psi^2_x/2 \sum_{i=1}^{\alpha} a_i^2)^{1/2}. \quad (14)$$

Note that all that has changed is the replacement of $\nu_1 + 1$ in Equation 12 with 2 in Equation 14, with the latter representing 1 (the degrees of freedom associated with any contrast) + 1. Following through for the current example yields $\phi^2 = 16(.80)^2/2(2) = 2.56$ and $\phi = 1.60$. (Note that Equation 5a also may be adapted for use here with $f$ calculated via Equation 13. Thus, again from Equation 13, $f = .3266$ and for a planned comparison, with $\nu_1 = 1$ in Equation 5a, $N = 48/[1 + 1] = 24$, and $\phi = 1.60$.) Correspondingly, the previous charts or tables are now entered with $\nu_1 = 1$ and $\nu_2 = 45$, which reveals power of .60 (both GPOWER and nQuery Advisor confirm this result). For power of .80, one would need to have about 26 participants per condition rather than the original 16. Note that by testing a planned contrast (rather than conducting an omnibus $2$-$df$ $F$ test): (a) power increased for the original specifications with 16 participants (from .49 for an omnibus test to .60 for testing the contrast) and (b) fewer participants were required to attain power of .80 ($n = 26$ for testing the contrast vs. $n = 32$ for the omnibus test).

I now explore two informative changes in the current example. First, there is the problem of alpha, or at least it would be a problem to some researchers or journal reviewers. If one wished to test only the one contrast involving Conditions 2 and 3, there would be no problem. The Type I error probability associated with the study would be given by $\alpha = .05$. If that were the case, however, then why bother to include Condition 1 in the study (i.e., if that condition's mean were not going to be examined in the statistical analyses)? More likely, the researcher wishes to compare each condition with each other, as a set of three contrasts. Or, in some situations, a set of two comparisons might be desired. For example, in the current case, the two orthogonal contrasts, $\psi_1 = (\mu_4 - \mu_1) - \mu_2$ and $\psi_2 = \mu_1 - \mu_2$, would reflect cognitive instruction versus no cognitive instruction and no instruction versus placebo control, respectively. In the case in which all three pairwise comparisons are of interest, to control the overall ("familywise" or "experimentwise") Type I error probability at .05, the researcher might be advised to use the Bonferroni inequality and divide the .05 into three equal portions, thereby testing each contrast at $0.05/3 = .0167$. Unfortunately, the available power-sample size charts and tables cannot accommodate this change in alpha (with few exceptions, only the traditional values of .05 and .01 are provided). What to do, then? Two options are available: (a) Go directly to a more alpha-flexible microcomputer program, such as GPOWER or nQuery Advisor, or (b) try some simple approximation formulas. The latter, proposed by Levin (1975), represent the $K > 2$ extensions of Equations 1 and 2. In particular, the power equation

$$z_2 = z_1 - [(n*/\sum_{k=1}^{\alpha} a_i^2)^{1/2} |\psi_x|], \quad \text{where} \quad n^* = n - 2, \quad (15)$$

and the sample size equation

$$n^* = \sum_{i=1}^{\alpha} a_i^2 (z_1 - z_2)^2/\psi_x^2, \quad \text{with} \quad n = n^* + 2. \quad (16)$$

To illustrate, I first reconsider the exact power and sample size calculations just presented for comparing Conditions 2 and 3. Applying Equation 15, based on $z_1 = 1.96$ for $\alpha = .05$, two-tailed (to be consistent with original calculations), and an adjusted $n^* = 16 - 2 = 14$, one has

$$z_2 = 1.96 - [(14/2)^{1/2}(0.80)] = -.16$$

and power = P(z > z_2) = P(z > -.16) = .56 (compared with the exact .60).

For Equation 16, with $z_2 = -.84$ for power of .80, one has

$$n^* = 2 \{1.96 - (-.84)^2/2.80^2\} = 24.5 \quad \text{or} \quad 25,$$

which, when 2 is added, yields 27 participants per condition (compared with the exact 26). For a directional test of the Condition 3 (cognitive instruction) versus Condition 2 (placebo) contrast, simply replace $z_1 = 1.96$ with $z_2 = 1.65$ in the calculations. The corresponding approximate power and sample size then become $1 - B = .68$ and $n = 22$, respectively.

Now, back to the future. These approximation formulas were introduced not for checking results from exact procedures but for use when exact procedures were either not convenient or not available. Suppose, therefore, that the earlier discussed Bonferroni approach were to be adopted here, for which the researcher is interested in testing the three pairwise comparisons,
each at $\alpha = .0167$. Forget the standard charts and tables in that situation because they will not help. Rather, use of Equation 15 with $z_i = 2.39$ to reflect the two-tailed critical value based $\alpha = .0167$ (see, e.g., Kirk, 1995, Table E.14, with $n_2 = \infty$) yields a $z_i$ of .27 and approximate power of .39 (exact calculations reveal power of .42). If, instead, the researcher planned to test only two of the three pairwise comparisons (based on a per contrast $\alpha = .05/2 = .025$, two-tailed, with associated $z_i = 2.24$), approximate power by Equation 15 would be .45 (exact power = .48).

A second change in the current example provides for the case of complex comparisons, such as the one noted earlier in which the researcher might wish to compare the cognitive instruction condition (Condition 3) with the two control conditions (Conditions 1 and 2) combined. Equations 15 and 16 are equally adept at incorporating more than two-mean pairwise comparisons. Before illustrating that here, however, a note of extreme caution is in order: Do not attempt any of these complex contrast power-sample procedures without first being certain to scale the contrast coefficients properly. What exactly does that mean? It means that for correct power or sample size calculations to result, one must always be careful to be comparing a mean with a mean. To ensure this, the absolute values of the chosen contrast coefficients must sum to 2; that is, $\sum_{i=1}^{K} |a_i| = 2$, in addition to the earlier presented restriction that the contrast coefficients must sum to zero for the contrast to be a "legitimate" one. Thus, to compare the equally weighted combination of $\mu_1$ and $\mu_2$ with $\mu_3$, the following two sets of coefficients are conceptually equivalent:

$$a_1 = \frac{1}{2}, \quad a_2 = \frac{1}{2}, \quad a_3 = -1 \quad \text{(Set 1)}$$

$$a_1 = 1, \quad a_2 = 1, \quad a_3 = -2. \quad \text{(Set 2)}$$

Only the Set 1 coefficients will yield computationally correct powers or sample sizes, however, in that for that set $\sum_{i=1}^{K} |a_i| = \frac{1}{2} + \frac{1}{2} + 1 = 2$. It also should be noted that any "legitimate" contrast may be properly scaled simply by dividing each coefficient by $\frac{1}{2} \sum_{i=1}^{K} |a_i|$. So, for example, dividing each Set 2 coefficient by $\frac{1}{2}(1 + 1 + 2) = 2$ produces the same properly scaled coefficients as in Set 1. This scaling algorithm is important to remember when determining power or sample size in situations in which the "properly scaled" contrast coefficients are not apparent (e.g., factorial ANOVA designs, to be discussed shortly) and when contrasts derived from standard tables are formulated (e.g., tables of orthogonal trend coefficients, such as Kirk’s, 1995, Table E.10).

To illustrate, suppose that one wanted to determine the approximate power associated with a planned contrast (two-tailed, $\alpha = .05$) investigating the difference between the cognitive instruction condition and the two control conditions combined that represents at least a .80 difference. For that contrast, one would incorporate the properly scaled Set 1 coefficients into Equation 15’s power calculations. With coefficients given by $a_1 = \frac{1}{2}, a_2 = \frac{1}{2}, \quad$ and $a_3 = -1$, $\sum_{k=1}^{K} a_i^2$ in the denominator of Equation 15 is equal to $(\frac{1}{2})^2 + (\frac{1}{2})^2 + (-1)^2 = 1.5.$ Accordingly, with the conservative adjusted $\alpha* of 14 in that approximation formula,

$$z_2 = 1.96 - [(14/1.5)^{1/2}(0.80)] = -.48$$

for power equal to .68. Notice that compared with the previously pairwise comparison of the same magnitude, the current complex comparison is associated with greater statistical power (approximate values of .56 vs. .68, respectively). That (i.e., increased power) is invariably the consequence of defining complex, as opposed to pairwise, comparisons if all else is held constant. (Determining the corresponding power via nQuery Advisor and GPOWER yields exact power of .72 for the planned complex comparison.)

Finally, and as was pointed out in the preceding discussion, the $z$ approximation power and sample size Equations 15 and 16 are readily adaptable to directional (one-tailed) hypothesis testing. Thus, if the researcher could reasonably argue that his or her interest was in determining only whether the cognitive instruction condition produced mean performance that is higher than that in the two control conditions combined (based on $\alpha = .05$), then the previous $z_2 = z(975) = 1.96$ is replaced by $z_2 = z(95) = 1.65$, which in turn yields

$$z_2 = 1.65 - [(14/1.5)^{1/2}(0.80)] = -.79,$$

for an increase in approximate power to .79 (exact calculations reveal power of .82).

Effect size as a proportion of explained variation. There is nothing new here. Simply refer to the $\omega^2$ discussion for the two-sample situation and apply either Equation 5 or Equation 5a. For the three-group example, I return to the omnibus test of the .80 pairwise comparison and assess its power. Transforming $\psi = .8$ into a proportion of variance construct follows from the previously given relationship between $\phi$ and $\psi$, provided in Equation 13. Thus, here, $f^2 = .8^2/(3)(2) = .1067$. Accordingly, one sees that with $f^2 = .1067$, $K = 3$, and $N = 48$ (and thus $N' = 48/(2 + 11) = 16$), Equation 5a yields $\delta = [16(.1067)]^{1/2} = 1.31$ and power of .49, thereby confirming the previous calculations.8

Estimating effect sizes from previous empirical studies. Just as in the two-sample case, for the $K$ independent sample problem, $\psi_e$ can be estimated from previous empirical work, including a researcher’s own pilot studies and the published literature. When mean and variability information are available, $\psi_e$ can be estimated as the generalization of Equation 7:

$$\text{estimated } \psi_e = (a_1M_1 + a_2M_2 + \cdots + a_kM_k)/MSE^{1/2} = \sum_{k=1}^{K} a_kM_k/MSE^{1/2}, \quad (17)$$

where $a_k$ are the desired and properly scaled contrast coefficients.

Kirk (1995, Table E.13) also provided a sample size table, adapted from Foster (1993), based on a researcher’s specification of either $f$ or $\omega^2$ directly. Note also that the complex contrast provided here is indeed a "legitimate" contrast according to the zero-sum-of-coefficients criterion, inasmuch as $a_i = \frac{1}{2}, a_2 = \frac{1}{2},$ and $a_3 = -1$, which results in $\frac{1}{2} + \frac{1}{2} + (-1) = 0.
OVERCOMING FEELINGS OF POWERLESSNESS

and $M_s$ and $MSE$ are the sample means and mean square error, respectively. If mean and variability information are not available, but the results of a $t$ (or a $1-df F$) test of the contrast are, then $\psi$, can be estimated as the generalization of Equation 8:

$$\text{estimated } \psi_s = \sqrt{\left( \sum_{k=1}^{K} a_i^2/n \right)^{1/2}} = \sqrt{(F \sum_{k=1}^{K} a_i^2/n)^{1/2}}.$$  \hspace{1cm} (18)

Extending the two-sample Equations 9 and 10 (e.g., Kirk, 1995, p. 178), one can estimate $\omega^2$ for the $K$-sample situation either as

$$\text{estimated } \omega^2 = \nu_i(F - 1)/[\nu_i(F - 1) + N],$$  \hspace{1cm} (19)

where $\nu_i = K - 1$ and $N$ is the total (across-conditions) sample size, or

$$\text{estimated } \omega^2 = (SSB - \nu_iMSE)/(SST + MSE).$$  \hspace{1cm} (20)

Example 3: Comparing Independent Means in Factorial ANOVA Designs

For completely between-subjects factorial ANOVA designs, the previously discussed power and sample size procedures generalize virtually directly. I add "virtually" because users of those procedures must be careful to mind their Ps and Qs, or in this context, their formula as and $n$s. To demonstrate the reality of this virtual, I extend the three-group cognitive instructional example by adding a second factor to the design, age. Thus, suppose one has the previous three treatments (Factor A) that are assigned randomly in equal numbers to both younger and older adults (Factor B) in the structure of a $3 \times 2$ factorial design (hereafter referred to as having $I = 3$ rows and $J = 2$ columns). For traditional equal $n$ factorial designs, three orthogonal hypothesis "families" include the $A$ main effect, the $B$ main effect, and the $A \times B$ interaction, respectively; for each family, one can determine power or sample size based on the researcher's alpha and minimum effect size specifications. One issue to be considered is the likely possibility that sample size calculations for the various specifications will yield different numbers of participants for the different hypothesis families. At least two different alternatives are of course operationally possible for a researcher: (a) Decide which of the three hypothesis families is most important to the study and base sample size (or power) calculations on that family. (b) Determine the required sample size for each of the three hypothesis families and select whichever is largest. For the current example, I proceed according to the second alternative.

Effect size as a standardized mean difference. For each family (A, B, and A X B), the researcher must make the same three specifications as before ($\alpha$, $\psi_s$, and either $n$ or power) to determine the desired missing ingredient (power or $n$). To simplify things here, suppose that (a) for each family, an omnibus $F$ test is to be conducted with $\alpha = .05$; (b) the minimum effect size of interest is specified as $8\sigma$ for either a pairwise difference (for the two main effects) or for a comparably scaled "tetrad" difference (to be discussed shortly) for the interaction; and (c) there are 8 participants in each of the six cells of the design (i.e., a total of 48 participants, consisting of 24 younger adults and 24 older adults, each randomly assigned in equal numbers to the three experimental conditions).

To determine power for the omnibus tests associated with each family, I refer to the labeled means in Table 1 and base all calculations on the associated cell $n$s (i.e., $n = 8$). For the $A$ main effect, a pairwise comparison includes any two row means (e.g., $\mu_1$ vs. $\mu_3$), which are formed by averaging the two corresponding cell means in row 1 ($\mu_{11}$ and $\mu_{12}$) and in row 3 ($\mu_{31}$ and $\mu_{32}$), respectively. To satisfy the sum-of-coefficients scaling criterion discussed previously, I specify the critical $A$ main effect (instructional conditions) contrast of interest, based on the cell means, as

$$\psi_s = [\psi(\mu_{11} + \mu_{12}) - \psi(\mu_{31} + \mu_{32})]/\sigma = .8,$$

for which it can be seen that the sum of the absolute values of the associated contrast coefficients, $a_{11} = 1/2$, $a_{12} = 1/2$, $a_{21} = 0$, $a_{22} = 0$, $a_{31} = -1/2$, and $a_{32} = -1/2$, is equal to the required total of 2. Thus, given these coefficients (for which $\sum_{i=1}^{3} a_i^2 = 1$), along with $\psi_s = .8$, $\nu_i = \nu_A = I - 1 = 2$, and $n = 8$, extending Equation 12 one has

$$\phi = n\psi_s^2/(\nu_i + 1) \sum_{i=1}^{3} a_i^2[1/2] \hspace{1cm} (21)$$

so that

$$\phi = [8(.8)^2/(2 + 1)(1)]^{1/2} = 1.31.$$  

This corresponds exactly to the $\phi$ obtained in Example 2 when the three treatments were cued in a one-way ANOVA layout on the same number of total participants per treatment (16 there vs. 8 younger plus 8 older adults here). Alternatively, Equation 13 may be extended directly to factorial designs as

$$f^2 = \psi_s^2/C \sum_{i=1}^{I} \sum_{j=1}^{J} a_i^2 \hspace{1cm} (22)$$

where $C = IJ$. Thus, here, $C = 3(2) = 6$, $f^2 = (8)^2/6(1) = .1066$, and from Equation 5a with $N' = 48/(2 + 1) = 16$, $\phi = [16(.1066)^2]^{1/2} = 1.31$. From the Pearson and Hartley charts ($\alpha = .05$), there is a $\phi$ of 1.31 based on $\nu_i = 2$ and $\nu_{\text{total}} = IJ(n - 1) = 3(2)(7) = 42$ is associated with power of about .48 (compared with .49 previously). Note that whatever slight

Table 1

<table>
<thead>
<tr>
<th>Instructional condition (A)</th>
<th>Age group (B)</th>
<th>Across ages</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Younger adults</td>
<td>Older adults</td>
</tr>
<tr>
<td>No instruction</td>
<td>$\mu_{11}$</td>
<td>$\mu_{12}$</td>
</tr>
<tr>
<td>Noncognitive</td>
<td>$\mu_{21}$</td>
<td>$\mu_{22}$</td>
</tr>
<tr>
<td>Cognitive</td>
<td>$\mu_{31}$</td>
<td>$\mu_{32}$</td>
</tr>
<tr>
<td>Across conditions</td>
<td>$\mu_{4}$</td>
<td>$\mu_{5}$</td>
</tr>
</tbody>
</table>

Note. There were 8 participants in each cell of the design.
power decrease there is for the A main effect in the factorial ANOVA model compared with that for the same effect in the one-way layout (assuming that all is held constant) is due simply to the fewer error degrees of freedom here (in this case, \( \nu_{error} = 42 \) relative to there (\( \nu_{error} = 45 \)).

I now determine power for the main effect of Factor B (age), given the same specifications. Referring again to Table 1, one can see that when comparing the two means, \( \mu_{11} \) and \( \mu_{22} \), one is comparing the average of \( \mu_{11} \), \( \mu_{21} \), and \( \mu_{31} \), with the average of \( \mu_{12} \), \( \mu_{22} \), and \( \mu_{32} \). The desired 0.8 SD difference between younger (\( B_1 \)) and older (\( B_2 \)) adults may be expressed as

\[
\psi = \left( \frac{1}{3} \left( \mu_{11} + \mu_{22} + \mu_{31} \right) - \frac{1}{3} \left( \mu_{12} + \mu_{22} + \mu_{32} \right) \right) / \sigma = 0.8.
\]

For this properly scaled contrast, there are three coefficients of \( \frac{1}{3} \) and three of \(- \frac{1}{3} \), which results in \( 3 \cdot 3 = 9 \). With \( \nu_1 = \nu_{error} = 42 \), plugging into Equation 21 yields

\[
\phi = \left( \frac{8 \cdot 0.8}{(1 + 1)(0.16)} \right)^{1/2} = 1.96.
\]

Alternatively, from Equations 22 and 5a, respectively,

\[
f^2 = 8^2/16(0.16) = 0.16
\]

and

\[
\phi = \left( \frac{48}{(1 + 1)(0.16)} \right)^{1/2} = 1.96.
\]

From the Pearson-Hartley charts with \( \nu_1 = 1 \) and \( \nu_{error} = 42 \), power is about .77. Note the sensibility of this outcome. For the parabola to those of the main effects (to reflect a comparison of differences to a mean for the former situation (power = .48). For the A main effect there are 3(8) = 24 participants per age group (in essentially a two-sample comparison), whereas for the A main effect there are 2(8) = 16 participants per condition (in essentially a three-sample comparison). The probability of detecting a .8\( \sigma \) pairwise difference in the former situation (power = .77) is considerably greater than the probability of detecting the same difference in the latter situation (power = .48).

“What about the power for the Condition \( \times \) Age interaction?” I can imagine readers asking. If they are not, they should be. To capture the imagination, return to Table 1. A true or “legitimate” interaction contrast consists of a difference in differences to represent the differential effect that an interaction is (see Marascuilo & Levin, 1970, for an extended discussion).

As such, the most basic type of interaction contrast, a tetrad difference, involves four cells of a factorial design, two representing columns in one row (e.g., \( \mu_{11} \) and \( \mu_{22} \)) and another two representing the same two columns in a different row (e.g., \( \mu_{31} \) and \( \mu_{32} \)). In the current example, with only two levels of Factor B, all interaction contrasts must involve columns 1 and 2 but, of course, that will not always be the case. The interaction contrast then compares the mean difference between the two columns in one row (\( \mu_{11} - \mu_{12} \)) with the corresponding mean difference in the other row (\( \mu_{31} - \mu_{32} \), or (\( \mu_{11} - \mu_{12} \)) - (\( \mu_{31} - \mu_{32} \)). (Note that the words row and column can be exchanged in the preceding sentence and the same numerical value of the contrast will result.) With the contrast coefficients scaled comparably to those of the main effects (to reflect a comparison of means rather than sums), a properly scaled standardized A \( \times \) B tetrad interaction contrast (involving, say, rows 1 and 3) may be defined as

\[
\psi = \left( \frac{1}{2} \left( \mu_{11} - \mu_{12} \right) - \frac{1}{2} \left( \mu_{31} - \mu_{32} \right) \right) / \sigma,
\]

for which \( a_{11} = \gamma_1, a_{33} = -\gamma_2, a_{22} = 0, a_{23} = 0, a_{13} = -\gamma_3, \) and \( a_{32} = \gamma_2 \) and, thus, \( \sum \gamma_i = 1, \) just as for the A main effect.

Similarly, as with the A main effect where \( \nu_1 = 2, \) for the A \( \times \) B interaction, \( \nu_2 = (2 - 1)(2 - 1) = 2, \) and once again, \( \nu_{error} = 42. \) Moreover, because (a) all other specifications are the same for the A main effect and the A \( \times \) B interaction, namely, \( n = 8, \alpha = 0.05, \) and minimum \( \psi = 8, \) and (b) the same power formula is applied and charts referred to, might one expect to obtain the same power for the omnibus \( F \) test in the two cases? Determining that the Pearson-Hartley \( \psi \) is equal again is to 1.31, which corresponds to power of about .48, the answer appears to be a resounding yes. If the minimum effect of interest and all other relevant considerations are held constant, then an interaction is associated with the same power as a main effect. This revelation will (perhaps) help to refute commonly heard arguments that interactions are harder to detect than main effects or that one needs larger sample sizes to detect interactions (than main effects).

What such arguments really are arguing is that interactions typically are harder to find than main effects because interactions are typically smaller in magnitude.

The truth of the matter is that if all else is held constant, including particularly the respective sizes of the effect (here, e.g., a .8\( \sigma \) difference in the constituent means in both cases) based on comparable scaling and the same degrees of freedom for the effect (here, \( \nu_1 = 2 \)), then the power (or required sample size) associated with a specified main effect or interaction will be exactly the same; see also Cohen’s (1988, pp. 373–374) discussion in his now-classic “power” book, which represents a fundamental correction of material in the first edition of that reference (Cohen, 1969, pp. 367–369).

To view this argument from a different perspective, I refer to the labeled cell means of the simplest factorial design possible, the 2 \( \times \) 2, as represented here by focusing on the four means in the first two rows of Table 1 (i.e., \( \mu_{11,2}, \mu_{21,3}, \mu_{31,2}, \) and \( \mu_{22,3} \)). Assuming equal cell ns, the A main effect amounts to the difference between the two averaged cell means in row 1 and those in row 2, or \( \frac{1}{2} (\mu_{11} + \mu_{12}) - \frac{1}{2} (\mu_{31} + \mu_{32}) \). Similarly, the B main effect amounts to the difference between the averaged cell means in column 1 and those in column 2, or \( \frac{1}{2} (\mu_{11} + \mu_{12}) - \frac{1}{2} (\mu_{21} + \mu_{22}) \). Finally, the A \( \times \) B interaction, which, as was previously noted, is a difference in differences, can be written as \( (\mu_{11} + \mu_{12}) - (\mu_{31} + \mu_{32}) \). This can be seen to be the same as \( (\mu_{11} + \mu_{31}) - (\mu_{12} + \mu_{32}) \), or the difference between the summed means on one diagonal and those on the other diagonal. To be scaled and interpreted comparably to the main effects (as a mean vs. a mean rather than as a sum vs. a sum), simply divide each implicit coefficient by 2 to obtain \( \frac{1}{2} (\mu_{11} + \mu_{22}) - \frac{1}{2} (\mu_{12} + \mu_{21}) \). It is this final, comparable form of a basic interaction that is the focus of interpretation for the power and sample size procedures that are offered here. To make the argument concrete, suppose that the means in the first two rows of Table 1 were associated with an A main effect represented by \( \psi = .8 \). To have power of .90 for detecting an effect of that size...
(with $\alpha = .05$), 17 participants per cell would be required ($N = 68$). With the same specifications for the B main effect, 17 participants per cell also would be needed ($N = 68$). Finally, with the current comparably scaled half difference-in-difference (to reflect a mean vs. mean) conceptualization, guess how many participants would be required to detect an $A \times B$ interaction representing $.80$ based on $\alpha = .05$ and power of .90? You guessed it: the same 17 participants per cell ($N = 68$).

Two other points are noteworthy here: First, just as was illustrated for the one-way layout, it is possible to determine the power or sample size associated with “complex” contrasts (including interaction contrasts) or those involving more than four cells. As long as such contrasts are “legitimate” main effect or interaction contrasts and they are scaled properly, Equation 21 can be applied directly. Second, if one wishes to determine the power—sample size for detecting 1-df planned contrasts, one simply needs to switch to the factorial-design analogs of Equation 14 (for exact values) and Equations 15 and 16 (for approximate values), being mindful of the contrast and coefficient scaling caveats that were just noted.

**Effect size as a proportion of explained variation.** Determining power and sample size in a factorial design, based on $\omega^2$, follows directly from the one-way application of Equation 5, combined with the immediately preceding discussion. To illustrate sample size determination, for example, suppose that one wants to know how many participants to include in a 3 (cognitive instruction) $\times$ 2 (age) study to have power of .80 to detect an $A \times B$ interaction that accounts for at least 15% of the total variance—assuming that no other effects are present, and that is defined (e.g., Kirk, 1995, pp. 397–398) as a “partial” $\omega^2 = \sigma^2_{\text{effect}}/(\sigma^2_{\text{error}} + \sigma^2_{\text{error}})$—when an $F$ test of that effect is to be conducted with $\alpha = .05$. For this problem, $\omega^2$ is given as .15, which then is incorporated into Equation 5 along with the $B$-to-be-solved-for $N'$. This is referred to the $\phi$ of the Pearson-Hartley charts that corresponds to power of .80, based on $\alpha = .05$, $\nu_1 = 2$, and first-step $\nu_2 = \infty$. From the charts, the indicated value is approximately $\phi = 1.78$, and so

$$\phi^2 = 1.78^2 = N'(.15)/.85.$$  

When solved, this yields $N' = 17.95$, which, when decoded, yields a total $N$ of 17.95 $(2 + 1) = 53.86$, or 54. Coincidentally, $N$ happens to be divisible by the number of cells (6) here, and so $n = 54/6 = 9$. In the second step, I recompute $\phi$ based on $N = 54$, $N' = 54/3 = 18$, and again (but not necessarily) obtain

$$\phi = [18(.15)/.85]^{1/2} = 1.78.$$  

With this hypothetical $n = 9$, the would-be error degrees of freedom are given by $\nu_3 = 3(2)(8) = 48$. Checking this in the Pearson-Hartley charts, one finds that the power associated with $\phi = 1.78$, $\nu_1 = 2$, and $\nu_3 = 48$ is about .76. Thus, $n = 9$ is not enough (for power of .80). Therefore, in Step 3, try $n = 10$, for which $N = 60$ and $N' = 60/3 = 20$. Now,

$$\phi = [20(.15)/.85]^{1/2} = 1.88.$$  

which is based on $\nu_1 = 2$ and $\nu_2 = 54$. With those values in the Pearson-Hartley charts, power to detect an $A \times B$ interaction is in excess of the desired .80—almost .82, in fact, as confirmed by GPOWER. So, go for that interaction with $n = 10$, $N = 60$, and feel powerfully good about it, although the specified minimum $\omega^2$ of .15 for this example represents a sizable, and perhaps unrealistic, effect size to go for. As an aside, with the same specifications for the A main effect ($\alpha = .05$ and $\omega^2 = .15$), the required total sample size for power of .80 would be exactly the same as that just computed: With $N = 60$, $N' = 60/3 = 20$, again it is found that $\phi = 1.88$, for which, with $\nu_1 = 2$ and $\nu_2 = 54$, power is almost .82. This underscores my previous comments about the equivalence of main effect and interaction powers when they are compared and interpreted on a level playing field.

Remember that when determining sample size for specified proportions of variation accounted for in factorial designs, the solution based on the current equations is with respect to the operative $N'$ and not the actual sample size. Thus, the resultant $N'$ must be “decoded” to yield the total sample size, as $N = N'(\nu_1 + 1)$. Moreover, when this total $N$ is rounded up to the nearest integer, it is likely that the resulting value is not one that will be equally divisible among the $U$ cells of the design. Thus, $N$ typically will need to be increased to the nearest integer that is divisible by $U$.

**Estimating effect sizes from previous empirical studies.** As with the one-way layout, factorial ANOVA effect sizes, represented by either $\psi_e$ or $\omega^2$, can be estimated from factorial ANOVA data already in hand. In particular, $\psi_e$ can be estimated for each effect as

$$\psi_e = (a_{11}M_{11} + a_{12}M_{12} + \cdots + a_{ij}M_{ij})/\text{MSE}^{1/2} = \sum_{i=1}^{I} \sum_{j=1}^{J} a_{ij}M_{ij}/\text{MSE}^{1/2}$$

(23) based on properly scaled contrast coefficients. Alternatively, from either $t$ or 1-df $F$ tests of the contrast, $\psi_e$ can be estimated as

$$\psi_e = t(\sum_{i=1}^{I} \sum_{j=1}^{J} a_{ij}/n)^{1/2} = (F \sum_{i=1}^{I} \sum_{j=1}^{J} a_{ij}/n)^{1/2}.$$  

(24)

One can estimate a partial $\omega^2$ for each factorial ANOVA effect either in terms of sum of squares equations or in terms of $F$ statistics (by extending Equation 19) as

$$\omega^2 = \nu_i(F - 1)/[\nu_i(F - 1) + N],$$

(25) where $\nu_i = \nu_{\text{effect}}$ (see Kirk, 1995, pp. 397–398; see also my earlier comments about the differing bias associated with the current sample estimators of $\psi_e$ and $\omega^2$).

**Additional issues.** Before ending my discussion of between-subjects factorial ANOVA situations, I make two final comments. All of the power, sample size, and effect size estimation procedures developed here (including those involving interactions) can be extended directly to encompass higher order factorial ANOVA designs (i.e., designs with more than two factors). Moreover, they can be adapted readily to situations in which tests of “simple effects” (e.g., Kirk, 1995, pp. 377–383), rather than main effects and interactions, are of experimental interest.

To illustrate the latter point, reconsider the current 3 (condi-
More Powerful Experimental Designs

So far, I have emphasized the essential role played by sample size in power determination. In particular, with all else held constant, statistical power increases directly as a function of increases in sample size. What, however, is an "aging" researcher to do if increasing the sample size is not an available option? That is, suppose that resources are not available for the researcher to include more participants in a study. Can power be increased by other means? The answer is again resoundingly affirmative, as I demonstrate throughout the remainder of this article. In fact, that point has already been manifested (but not demonstrated here) in the immediately preceding discussion of factorial ANOVAs. Specifically, a primary function of factorial designs is to examine the effects of two or more independent variables, or factors, simultaneously. When at least one of those factors is a nonexperimental (nonmanipulated) variable, then including that factor in the design (as a so-called "blocking" variable) often serves to reduce the error variability relative to a design in which that factor is not included, which in turn will increase power. For example, suppose that the three-condition study of Example 2 were conducted using adults without regard to their ages (i.e., undifferentiated younger and older adults were recruited and assigned randomly to the three experimental conditions). Assuming that age is related to performance on the outcome measure—a critical assumption, to be elaborated on shortly—then that one-factor design will not be as powerful as the factorial design of Example 3 in which younger and older adults were identified, separated, and then served as levels of a second factor according to which participants were assigned randomly, within the two age groups (blocks), to the three experimental conditions. In this section, I consider alternative designs that are more powerful than basic one-factor independent-samples ANOVA designs, and I demonstrate explicitly how and why they are more powerful (see also Footnote 3).

Example 4: Comparing Independent Means in Analysis of Covariance Designs

Consider the following straightforward addition to a one-factor ANOVA: Participants are again assigned randomly to K independent conditions, but this time a relevant antecedent variable, or covariate, is incorporated into an analysis of mean differences through an analysis of covariance (ANCOVA). In a "true" (Campbell & Stanley, 1966) experiment (i.e., with random assignment to conditions), the extent to which the post-treatment means are "adjusted" for group differences on the covariate is an incidental (and consequent) feature of this design. What is neither incidental nor consequential about conducting an ANCOVA, however, is the increased power that will result relative to an independent-samples ANOVA.\(^8\) Statistical power is increased to the extent that the researcher-selected covariate (X) is correlated with the study's dependent variable (Y) within each of the K samples. The higher the common within-samples correlation between X and Y (\(\rho_{XY|W_i}\)), whether positively or negatively, the greater the power benefits to the researcher. This is because the error variance for an ANCOVA (\(\sigma^2\)) is related to the error variance for the corresponding ANOVA (\(\sigma'^2\)), as follows:

\[
\sigma^2 = \sigma'^2(1 - \rho^2_{X|W}),
\]

which in turn implies that \(\sigma^2 \leq \sigma'^2\). The two will be equal whenever \(\rho_{X|W} = 0\) (i.e., when there is no within-samples correlation between the researcher's selected covariate and the dependent variable) and the ANCOVA's error variance will become increasingly smaller as the correlation between the covariate and the dependent variable gets increasingly larger in absolute value. This relationship between the two error variances guides the power and sample size discussion that follows. From this,

\(^8\) Two traditional functions of the analysis of covariance (ANCOVA) are what have been called the "bias-removing" function (for statistically eliminating preexisting group differences in nonexperimental [nonrandom assignment] studies) and the "variance-reducing" function (for increasing power in both true experimental and nonexperimental studies; see Marescalco & Levin, 1983, p. 119). Because nonexperimental applications of ANCOVA are fraught with interpretive difficulties (e.g., Elashoff, 1969) and I advocate the use of more scientifically credible research methodologies by applied social scientists (e.g., Levin, 1994), I illustrate only the variance-reducing function of ANCOVA in the context of true experimental studies here.
it also can be seen that the researcher's choice of a covariate must be a judicious one. That is, the correlation between X and Y must be sufficiently large to benefit from the use of an ANCOVA, not to mention the potential practical inconveniences involved in identifying and obtaining a "good" covariate. Otherwise, little or no power will be gained in choosing an ANCOVA over an ANOVA.

Reconsider the original two-group comparison of cognitively instructed and uninstructed older adults (Example 1). Again assume that 16 participants were assigned randomly to each condition. Suppose, in addition, that before the study there was an intellectual task (X) administered to all participants, a task similar in many respects to the cognitive skills measured at the end of the experiment (Y) and that is included as a covariate in the statistical analysis. As an alternative covariate, one might have been able to collect information about how many years of formal education each participant had received. I now examine the power characteristics associated with this modified statistical approach.

**Effect size as a standardized mean difference.** Given the preceding discussion, a standardized mean difference in the current randomized ANCOVA context \( \psi' \) may be reconceptualized in relation to that of the corresponding ANOVA standardized mean difference \( \psi \), as

\[
\psi' = \psi / \sigma(1 - \rho_A^2)^{1/2} = \psi / (1 - \rho_A^2)^{1/2}.
\]

That is, exactly the same-sized treatment effect will be functionally larger in the ANCOVA, relative to an ANOVA, by a factor of \( 1/(1 - \rho_A^2)^{1/2} \). With this reconceptualized "effective" effect size notion (and being aware that there is one fewer error degree of freedom in the ANCOVA than in the corresponding ANOVA), one can determine power or sample size in the same fashion as in the independent-samples situation based on either exact Equations 6, 12, and 14 or approximation Equations 1, 2, 15, and 16 (with the subtraction or addition of 2 for the latter four, as discussed previously). For the current example, suppose that one were fortunate enough to identify an intellectual task that was moderately correlated with the dependent measure. According to Equation 2, it then can be used in conjunction with the postinstruction measure—\( r = .6 \) corresponds to \( \omega^2 = .083 \) (actually, .08257) and, thus, according to Equation 27,

\[
\omega^2 = .08257 / (1 - .08257) = .1071.
\]

Plugging this value into Equation 5, with \( n = 16 \), one has

\[
\psi = (16(.1071)/(1 - .1071))\sqrt{1/2} = 1.39,
\]

which corresponds to the just-computed value based on the analogous \( \psi' \) measure. Accordingly, with \( \nu_1 = 1 \) and \( \nu_2 = 29 \), the ANCOVA power is again .47 (compared with .375 in Example 1, where no covariate was included).

**Estimating effect sizes from empirical studies.** The key concept for an applied researcher in ANCOVA situations is how best to estimate the correlation between the covariate and the dependent measure, \( \rho_A \), which, as demonstrated here, serves to reduce within-conditions variability and thereby increases statistical power. A prior question, however, is, What exactly is the appropriate "correlation" that one wishes to estimate? Contrary to what some might think, and as has been indicated (at least implicitly) throughout the preceding discussion, the pooled within-conditions correlation—not the total (or across-conditions) correlation—is the operative power-enhancing construct in ANCOVA situations (see, e.g., Levin, 1985; Marascuilo & Levin, 1983, p. 57). Given the necessary assumption that \( \rho_A \) is of comparable magnitude in each of the \( K \) experimental conditions, the within-conditions correlation coefficient can be estimated from previous research in a couple of different ways.

One approach is for a researcher to rely on previous studies (or to collect pilot data) that included measures of X and Y in the same \( K \) experimental conditions as in the planned study and then to compute the within-conditions sample correlation coefficient, \( r_x \), based on pooling the \( K \) sample variances and covariances associated with the \( N_i \) participants in each condition. Alternatively, a researcher could investigate the linear relationship between X and Y (i.e., compute \( r_x \)) in only one of the experimental conditions. The important point here is that the correlation between X and Y must not be computed for the total sample (i.e., across participants from different experimental conditions); that would be an inappropriate—and potentially confounded—estimate of the true \( \rho_A \). Once \( \rho_A \) has been appropriately estimated from \( r_x \), it then can be used in conjunction with the previously discussed sample estimates of standardized mean differences or proportions of explained variance to estimate the corresponding "inflated" effective measures to be used in power or sample size determination. In particular, for the ANCOVA,

\[
\text{estimated } \sigma^2 = \text{MSE} \left(1 - \rho_A^2\right),
\]

where MSE is the mean square error from an independent sam-
ples ANOVA. More simply, if results from a previous ANCOVA study are available, then

\[ \text{estimated } \sigma^2 = MSE', \]

where \( MSE' \) is the ANCOVA mean square error. These values then can be used in conjunction with Equations 17 and 20 to estimate either \( \psi_1^2 \) or \( \omega^2 \), respectively. Alternatively, if \( F \) statistics are available, \( \psi_1^2 \) and \( \omega^2 \) can be estimated directly from Equations 18 and 19, respectively.

**Example 5: Comparing Means in One-Sample Repeated Measures Designs**

Correlated samples situations are encompassed by what are referred to in traditional terms as *randomized block designs* and *split-plot designs* (Kirk, 1995). In psychological research, randomized block designs typically are manifested by true blocking studies on the one hand (of which the *matched pair design* is the simplest version) and one-sample "repeated measures" studies (e.g., the one-sample pretest-posttest design) on the other. Split-plot designs usually are manifested by multiple-sample repeated measures studies. Both such designs are termed "correlated samples" because between-treatments comparisons are based on either matched or the same participants (in contrast to the previous independent-samples designs, in which between-treatments comparisons involve totally unrelated participants). As with ANCOVA, the statistical consequences of a researcher's informed choice of a correlated samples design over an independent-samples design are—despite the loss in error degrees of freedom—to reduce the error variance that is used for assessing treatment effects. This, in turn, has important practical advantages related to the current power and sample size discussion. In particular, the primary reason for preferring correlated samples over independent-samples designs is because the former are expected to increase statistical power (given a fixed sample size), or, equivalently, because the former are expected to require fewer participants to attain the same power. This primary reason will become apparent in the examples that follow.

Also, as with ANCOVA (e.g., Elbashir, 1969; Huitema, 1980), univariate analyses of repeated measures data (the focus here) are associated with a set of stringent statistical assumptions. An often-ignored critical assumption is that of sphericity (or circularity), which relates to the structure of the variance-covariance matrix underlying the repeated measures (see Kirk, 1995, pp. 274–282). The bottom-line message for researchers is that whenever an Omnibus \( F \) test is conducted on \( K > 2 \) repeated measures (or, more generally, whenever the \( F \) test is based on more than one numerator degree of freedom), there is good reason to believe that the sphericity assumption will be violated, which in turn has Type I error inflationary consequences. In such cases, alternative statistical strategies are needed, including multivariate analyses of repeated measures (e.g., Marascuilo & Levin, 1983, pp. 373–381). Although beyond the scope of this article, the corresponding multivariate power and sample size procedures could and should be applied in such situations (see, e.g., Cohen, 1988, Chapter 10).

Now, back to the item of concern. To paraphrase an all-purpose statistical maxim (reference unknown): If there is a \( \rho_a \), there must be a \( \rho_b \). And, of course, there is, in the context of within-subjects repeated measures designs, the simplest illustration of which is the one-sample pretest–posttest design. In that context, I define \( \rho_b \) as the correlation between participants' performance on the two dependent measures (e.g., pretest and posttest or, more generally, Measure Y and Measure Y'). When there are more than two measures, let \( \rho_B \) be the (assumed common) correlation between all pairs of dependent measures and \( \sigma^2 \) the (assumed common) within-measures variance to simplify the discussion. Given this \( \rho_B \) concept (which is not the same as the just-discussed \( \rho_a \), the correlation between the antecedent variable and the single dependent variable), it can be shown that the error variance of a correlated samples repeated measures design \( (\sigma^2) \) is related to that of the independent-samples ANOVA design \( (\sigma^2) \), as

\[ \sigma^2 = \sigma^2(1 - \rho_B). \]

From this, one can see that if the dependent variable performance of participants on Measures Y and Y' is positively correlated, then an error variance reduction will occur for a repeated measures design (relative to an independent-samples design), with a corresponding power increase. With either little or no correlation or a negative correlation, undesirable power consequences for the repeated measures design will result from a loss of degrees of freedom, along with an increase in error variance in the case of negatively correlated performance measures.

To illustrate the power benefits associated with selecting a repeated measures design, suppose that in the ongoing example, a one-group pretest–posttest design (i.e., a completely within-subjects design with \( K = 2 \) repeated measures) is conducted in which 16 adults are initially given a cognitive skills pretest, then administered the extended cognitive instruction, and finally given a parallel cognitive skills posttest. Assuming that pretest and posttest performance are moderately correlated (say, \( \rho_B = .50 \)), let us look at the power advantages of adopting a one-sample repeated measures (pretest–posttest) design in this context. In doing so, I overlook the methodologically based interpretative problems associated with attributing any pretest–posttest improvement to the cognitive instruction per se, for the current one-group pretest–posttest design is sorely lacking in internal validity (see, e.g., Campbell & Stanley, 1966).

Effect size as a standardized mean difference. Analogous

---

9 This is particularly likely to be true when the repeated measures consist of measures of participants' learning, change, or growth over time.

10 In a one-factor repeated measures design with \( K \) levels of the independent variable, there are only \( (K - 1)/K \) as many error degrees of freedom as there are in the corresponding independent samples design \( ([K - 1]([n - 1]) \text{ vs. } K[n - 1]) \) respectively). It therefore is important that the correlation between measures \( (\rho_B) \) is sufficiently large to compensate for the former's loss of degrees of freedom.

11 As an alternative two-repeated-measures variation of this hypothetical study, a comparison could be made of the instruction adults' posttest performance on two different dependent measures, with one expected to be more positively affected by the instruction than the other (e.g., one task requiring fluid abilities and the other crystallized abilities; see, e.g., Salthouse, 1988).
to what was done in the ANCOVA Example 4, I can define an effective standardized contrast that takes into account the correlation between participants’ performance on the two measures as

\[ \psi_i' = \psi_i / \sigma(1 - \rho)\sqrt{2} = \psi_i / \sqrt{1 - \rho} \]

Thus, a standardized mean difference of 0.6 SD in an independent-samples design would (assuming \( \rho = 0.50 \)) amount to a standardized mean difference of \( \psi_i' = 0.6 / (1 - 0.5) = 0.849 \), which represents a substantial increase in the effective effect size. Indeed, incorporating \( \psi_i' \) and \( n = 16 \) into either Equation 6 or Equation 12 yields \( \psi = 1.70 \), which with \( \alpha = 0.05 \), \( \nu_1 = K - 1 = 1 \) and \( \nu_2 = (K - 1)(n - 1) = 1(15) \), reveals power of about 0.61 (compared with 0.375 for the independent-samples situation). This example underscores the increase in power (or equivalently, the decrease in required sample size) that can be expected when one selects a within-subjects design over a between-subjects ANOVA design, assuming that there is at least a moderate correlation between participants’ performance on the two (or more) measures; see also Maxwell and Delaney’s (1990) Tables 13.9–13.11 for some convenient sample size comparisons in between- and within-subjects designs (although considerable interpolation with respect to researcher-specified effect sizes and power may be required).

**Effect size as a proportion of explained variation.** Adapting Equation 27 to a correlated samples repeated measures context to represent the proportion of total within-subjects variation that is accounted for by the repeated measures factor, one has

\[ \omega^2 = \omega^2 / [\omega^2 + (1 - \rho)(1 - \omega^2)] \]  

(29)

Thus, given that an independent-samples standardized mean difference of \( \psi_i = 0.6 \) corresponds to \( \omega^2 = 0.08257 \) (see Example 4) in a repeated measures design in which \( \rho = 0.50 \), those effects are inflated to \( \psi_i' = 0.849 \) (as just illustrated) and \( \omega^2 = 0.08257 / [0.08257 + (0.50)(0.91743)] = 0.1525 \), respectively. With \( \omega^2 = 0.1525 \) and \( n = 16 \), from Equation 5 one finds that \( \phi = \sqrt{1 - 0.1525 / 0.8475} = 0.70 \) and power of 0.61. Thus, this is a two-factor, 2 (A: time) × 3 (B: subtest), completely within-subjects design. To determine power or sample size in an a priori fashion, for each of the three effects (A, B, and the A × B interaction) one would need to specify effective effect sizes that incorporate the correlation between measures (\( \rho \)), as was done in the one-factor example just presented (i.e., \( \psi_i' \) or \( \omega^2 \)). If the more restrictive assumption of global or omnibus sphericity is believed to hold, then a common correlation might be used for all three effects; instead, if the less restrictive assumption of local sphericity is considered more likely, then separate correlations would be incorporated into the different effects (e.g., Kirk, 1995, pp. 463–464; also see the previous discussion of the sphericity assumption in repeated measures studies, along with Footnote 10). Determining power—sample size in a post hoc fashion, based on sample estimates and use of the common \( MSE' \) for each effect, incorporates the more restrictive global sphericity assumption (Kirk, 1995, pp. 459–461).

I now illustrate the common-error approach for power—sample size determination. Suppose that the instructional intervention.

12 When determining repeated measures power—sample size using the “other ANOVA [analysis of variance] tests” option of the GPOWER program, the requested total \( N \) refers to the total number of observations rather than participants. Version 1.0 of nQuery Advisor does not include procedures for correlated sample situations.

13 I favor the less restrictive separate-error procedure for repeated measures hypothesis testing (as do certain microcomputer statistical packages), which is reflected in the \( F \) statistics in Table 2. Sample size...
tion study, based on 16 adults, $I = 3$ subtests, and $J = 2$ time periods, yielded the results summarized in Table 2. Estimated proportions of variance accounted for by each effect are included in the last column of Table 2. These can be computed, for each effect, using Equation 19. For example, for the subtest main effect,

$$\text{estimated } \omega^2 = \frac{2(5.09 - 1)/[2(5.09 - 1) + 96]}{15} = 8.18/104.18 = .08.$$ 

For determining how much power there would be to detect an effect of that magnitude in an equivalent follow-up study, based on 96 observations, $\alpha = .05$, and a common variance estimate, $\text{MSE}'$ (common), $N' = N/(v_1 + 1) = 96/(2 + 1) = 32$, one has, from Equation 5,

$$\phi = \frac{132(.08)/.92}{11/12} = 1.67,$$

which, with $\nu_1 = 2$ and $\nu_2 = 75$, is associated with power of about .71.

Similar calculations based on the time main effect estimate (with $\omega^2$ estimated from the common $\text{MSE}'$ equal to .11, $N' = 48$, $\nu_1 = 1$, and $\nu_2 = 75$) yield $\phi = 2.44$ and power of about .93. Alternatively, power for the time main effect can be calculated in terms of the estimated $\psi'$ measure (based on Equation 17) as

$$\text{estimated } \psi' = |\frac{1}{2}(15 + 17 + 13) - \frac{\nu}{2}(27 + 33 + 15)|/195^{1/2} = (15 - 25)/13.96 = .72$$

and for which, from Equation 21,

$$\phi = [16(.72)^2]/[1 + (1)(.08)]^{1/2} = 2.49$$

for power of about .94, close to what was obtained using the alternative $\omega^2$ effect size estimate.

Finally, the estimated partial $\omega^2$ for the interaction may be found to be .02, for which (with $N' = 32$, $\nu_1 = 2$, and $\nu_2 = 75$)

$$\phi = [32(.02)/.98]^{1/2} = .81$$

with associated power of only about .21.

A note on true blocking designs. As was mentioned earlier, true blocking designs (in which participants are rank ordered on the basis of some relevant antecedent variable and then are assigned randomly in blocks of $K$ to the $K$ experimental conditions) are encompassed by correlated samples designs and thus can be fit directly into the foregoing repeated measures discussion.\textsuperscript{14} Indeed, in the repeated measures context, each participant represents his or her own "block" from which the $K$ different measures are taken or to which $K$ different treatments are applied. Although blocking designs and the ANCOVA traditionally have been thought of as competing design and analysis strategies in true experimental studies (e.g., Feldt, 1958), recent investigators (e.g., Maxwell & Delaney, 1990) have concluded that the former are neither as powerful nor as practicable as the latter, which is why I have restricted my attention to ANCOVA here.

Moreover, it is possible to combine a blocked random assignment procedure with ANCOVA to produce a particularly powerful design and analysis strategy called alternate ranks ANCOVA (Maxwell, Delaney, & Dill, 1984). However, the alternative ranks design and analysis are not compatible with one of the major practical advantages of pure ANCOVA relative to blocking, namely, that preexperimental covariate information, although available (and a requisite for the proper use of ANCOVA), may not be accessible to a researcher at the time that

\textsuperscript{14} As was alluded to earlier, the most familiar and widely used true blocking design is represented by studies in which participants are assigned randomly to two experimental conditions in matched pairs.

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>$F_s$</th>
<th>$F_c$</th>
<th>Estimated partial $\omega^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time periods (T)</td>
<td>1</td>
<td>2,400</td>
<td>2,400</td>
<td>10.67</td>
<td>12.31</td>
<td>.11</td>
</tr>
<tr>
<td>Subtests (S)</td>
<td>2</td>
<td>1,984</td>
<td>992</td>
<td>3.89</td>
<td>5.09</td>
<td>.08</td>
</tr>
<tr>
<td>T x S</td>
<td>2</td>
<td>852</td>
<td>416</td>
<td>3.47</td>
<td>2.13</td>
<td>.02</td>
</tr>
<tr>
<td>Error (common)</td>
<td>15</td>
<td>3,375</td>
<td>225</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (S)</td>
<td>30</td>
<td>7,650</td>
<td>255</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error (T x S)</td>
<td>30</td>
<td>3,600</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total within subjects</td>
<td>80</td>
<td>19,841</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The means for the pretest verbal, numerical, and reasoning subtests were 15.0, 17.0, and 13.0, respectively. The means for the posttest verbal, numerical, and reasoning subtests were 27.0, 33.0, and 15.0, respectively. Each mean is based on 16 observations. $F_s$ and $F_c$ are the $F$ statistics associated with the separate (assuming local sphericity) and common (assuming global sphericity) error terms, respectively. The partial $\omega^2$ estimates are based on the common error term. ANOVA = analysis of variance.
participants are assigned randomly to experimental conditions. The latter is necessary for blocked random assignment, but not for ANCOVA based on simple random assignment of participants to experimental conditions.

**Example 6: Comparing Means in Combined Between-Subjects and Within-Subjects (Split-Plot) Designs**

Finally, I examine the split-plot situation, in which one factor (with J levels) represents a between-subjects factor and the other factor (with J levels) represents a within-subjects factor. To do so, suppose that 16 older adults are compared with 16 younger adults (i.e., J = 2) on each of J = 2 memory tasks, with one task measuring the participants’ recall and the other their recognition of a list of previously presented objects (see Table 3 for the design and labeled population means).

In such a split-plot design, the main effect of age is assessed by a between-subjects error term, whereas the main effect of task and the Age × Task interaction each are assessed by a within-subjects error term. Although a statistical test of the age main effect is equivalent to conducting a two independent-samples test that compares the two age groups with respect to participants’ average performance on the two tasks, because of the factorial structure of the split-plot design, the power procedures illustrated here are based on the cell ns (as in Example 3). The procedures for each split-plot effect combine the effect size conceptualization of the one-sample repeated measures situation (Example 5) with the factorial design conventions for either the standardized contrast approach or the proportion of explained variation approach (see Example 3).

Specifically, suppose that the researcher is interested in detecting a 0.6 within-population standard deviation difference in mean performance between the two age groups. Before proceeding, however, note that just as the correlation between measures, \( \rho_{B} \), helps to define the “effective” effect size \( \psi'_{B} \) in completely within-subjects designs (see Example 5), it does the same for both between- and within-subjects effects in split-plot designs. For the two within-subjects effects (here, the main effect of task and the Age × Task interaction), \( \rho_{B} \) operates in the same manner as it did in the completely within-subjects design, namely, reducing error variance and increasing power as specified by Equation 28. For the between-subjects main effect, however (here, the age main effect), the magnitude of the nominal effect size, \( \psi_{B} \), is adjusted—typically reduced—as follows (see also Kirk, 1995, p. 520):

\[
\psi'_{B} = \psi_{B}/[1 + (K - 1)\rho_{B}]^{1/2}.
\]

Thus, whenever there is a positive correlation between measures, effect sizes associated with between-subjects factors will appear to be smaller in split-plot designs than they are in independent-samples designs. In the current case, for a comparison of the two age groups averaged across the two tasks, with \( \psi_{B} = 6 \), the previous within-subjects \( \rho_{B} \) of .6, and K = 2 tasks, the effective standardized difference would be reduced to

\[
\psi'_{B} = 6/[1 + (2 - 1)(.5)]^{1/2} = .49.
\]

In terms of the properly scaled cell means that make up the two row means (\( \mu_{11} \) and \( \mu_{21} \)) in Table 3, the contrast representing the age main effect is defined as

\[
\psi = \frac{1}{2}(\mu_{11} + \mu_{12}) - \frac{1}{2}(\mu_{21} + \mu_{22}),
\]

for which the sum of squared coefficients equal to \((\psi)^2 + (\psi')^2 + (-\psi)^2 + (-\psi')^2 = 1 \). With \( \alpha = .05, \nu_{1} = I - 1 = 1, \nu_{2} = 16 \) observations per cell, from Equation 12 one has

\[
\phi = [16(.49)^2]/2(1) = 1.39.
\]

Equivalently, \( f^2 = .49^2/(1) = .06 \), yielding \( \phi = [16(2/2).06]^{1/2} = 1.39 \). The power associated with \( \phi = 1.39 \), based on \( \nu_{1} = 1 \) and \( \nu_{2} = I(n - 1) = 15 \), is equal to .47. Note that compared with Example 1, in which a .6 SD difference between two groups of 16 students each was of interest (with \( \alpha = .05 \)), power was only .375, compared with .47 for apparently the same specific hypotheses (other than the \( \rho_{B} \)-reduced effect size). I say “apparently,” however, because this power is associated with twice as many total observations per mean (namely, \( Jn = 2[16] = 32 \) compared with the case for Example 1. At the same time, it is worth noting that because of the between-tasks correlation of .50 here (along with \( \nu_{2} = 30 \)), the present 2 × 2 split-plot power of .47 is not as great as it would be for a corresponding 2 × 2 independent-samples design also based on a total of 64 observations (for which \( \nu_{2} = 60 \)), in which the power associated with \( \phi_{x} = 6 \) \( f^2 = .09 \) is found to be .66.

I now consider the within-subjects task main effect. In a split-plot design, there are \( \nu_{1} = J - 1 \) and \( \nu_{2} = I(n - 1) \) for the within-subjects main effect. So, combining the Example 5 specifications for the task effect (\( \rho_{B} = .5, \psi = 6 \), and therefore \( \psi' = .849 \)) with the task main effect contrast coefficients based on the individual cell means that make up the two column means, \( \mu_{1} \) and \( \mu_{2} \), in Table 3 (\( \mu_{11} \), \( \mu_{12} \), \( \mu_{21} \), and \( \mu_{22} \), respectively), one finds, again with a sum of squared coefficients equal to 1, that

\[
\phi = [16(.849)^2]/2(1) = 2.40.
\]

\footnote{The seemingly paradoxical implications of this situation are that when one is assessing the effect of a between-subjects factor in a split-plot design, the larger the positive correlation between measures, the less power there is to detect the effect relative to that in an independent samples factorial design, a consequence of the aggregation of “redundant” measures. Yet, the larger the positive correlation among measures, the greater the relative power for assessing the two within-subjects effects, as reflected by differentiating among the redundant measures.}
yields $1/1 = .849$. So that the interaction can be comparably discussed "differential effect" of the factorial design in Example 3. Here, for example, an interaction would indicate that the difference between younger and older adults’ performance was greater on one memory task than the other (e.g., greater on the recall task than on the recognition task). For the split-plot situation, interaction degrees of freedom are given by $\nu_1 = (I - 1)(L - 1)$ and $\nu_2 = I(J - 1)(n - 1)$. Let me put all this together, then, assuming again that $\rho_b = .5$ and $\psi^2 = .6$ (which yields $\psi'_b = .849$). So that the interaction can be comparably interpreted as a comparison of involving means rather than sums (see the previous factorial design discussion), the appropriate coefficients for the interaction are $\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2},$ and $\frac{1}{2}$, respectively, for $\mu_1, \mu_2, \mu_3,$ and $\mu_4$ in Table 3, which, once again, yields a sum of squared coefficients of 1. With this comparable scaling for the interaction, one finds that as with the task main effect, $\phi = [(16(.849)^2)/(1 + 1)(1)]^{1/2} = 2.40$. Moreover, with $\alpha = .05, \nu_1 = 1$, and $\nu_2 = 30$, power is again equal to .91, which once more illustrates that comparably defined and interpreted effects (whether main or interaction) will be associated with the same statistical power if they are based on the same effect size, degrees of freedom, and error variance ($\sigma^2$).

Because the power—sample size procedures based on "proportion of explained variation" effect size measures follow directly from the corresponding material presented in Examples 3 and 5, there is no need to include them here. (The same comment applies to both between- and within-subjects $\psi'_b$ and partial $\psi^2$ effect size estimates.) I do, however, conclude this section with an application of the current split-plot power procedures to an example from the published aging literature.

Example From the Published Literature

To illustrate both a post hoc power calculation and a prospective sample size determination, based on results from an already-completed study, consider one by Wiggs (1993, Experiment 1), which was published in this journal. In an investigation of people’s frequency judgments, 20 older adults and 20 younger adults were assigned to a verbal (English nouns) stimulus condition, and another 20 of each were assigned to an ideogram (Japanese kanji characters) stimulus condition. Thus, there were two 2-level between-subjects factors, age and stimulus condition. In addition, stimulus frequency was manipulated within participants, which subsequently resulted in a three-level factor, frequency difference (between paired items differing in frequency). There was therefore a $2 \times 2 \times 3 (I \times J \times K)$ split-plot design, with the first two factors representing between-subjects variation and the third factor within-subjects variation. The mean proportion of errors, by age level, stimulus condition, and frequency difference, is presented in Table 4.

<table>
<thead>
<tr>
<th>Age level and stimulus condition (between subjects)</th>
<th>Frequency difference (within subjects)</th>
<th>Across frequency differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Older</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>.27</td>
<td>.20</td>
</tr>
<tr>
<td>Ideogram</td>
<td>.38</td>
<td>.37</td>
</tr>
<tr>
<td>Across stimuli</td>
<td>.325</td>
<td>.290</td>
</tr>
<tr>
<td>Younger</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal</td>
<td>.34</td>
<td>.19</td>
</tr>
<tr>
<td>Ideogram</td>
<td>.33</td>
<td>.23</td>
</tr>
<tr>
<td>Across stimuli</td>
<td>.335</td>
<td>.240</td>
</tr>
<tr>
<td>Across ages and stimuli</td>
<td>.330</td>
<td>.270</td>
</tr>
</tbody>
</table>

Note. Data are based on 20 older and 20 younger adults in each stimulus condition (Wiggs, 1993, Experiment 1).  
\[ \text{MSE}_{\text{within}} = .02. \]  
\[ \text{MSE}_{\text{between}} = .04. \]

More precise calculations from the F ratio provided by Wiggs (1993) yield an estimated $\phi$, of .28. I continue to use the value of .30 in this example, however. Note also that the present $\text{MSE}_{\text{between}}$ already incorporates the estimated $\rho_b$, and so there is no need to estimate and include it here.
\[ f^2 = \frac{.30^2}{12(.333)} = .0225. \] Therefore, with \( N' = N/(\nu_1 + 1) = 120/(3)(20) = 2(2)(3)(20)/(1 + 1) = 120 \), from Equation 5, \( \phi = [120(.0225)]^{1/2} = 1.64 \), yielding power of .63. With the observed effect size of 0.3 SD units as an estimate (and even though the interaction effect was found to be statistically significant here, with \( \alpha = .05 \)), a larger sample size would be needed to have reasonable power to detect an effect of that magnitude in a replication study. For power of .80 (based on \( \alpha = .05 \) and the same design specifications), either the iterative application of Equation 21 or microcomputer power analysis reveals that a total of 360 observations, or 30 younger and older adults per stimulus condition (rather than the current 20 adults), would be required.

Incidentally (and in keeping with one of our earlier power-enhancing recommendations) based on Wiggs’s (1993) and others’ findings, a one-tailed test of this contrast might well be justified in a proposed replication study. To incorporate directional hypothesis testing with the available microcomputer programs, one simply has to double the specified alpha (here, from .05 to .10). Doing so for the current estimated Age X Stimulus interaction contrast reveals that a power of .74 is associated with the current sample size of 240 and (b) a total of 276 participants, or 23 older and younger adults per stimulus condition, would be required.

To illustrate an alternative power estimate for the same interaction in terms of the more conservative proportion of variance (partial \( \omega^2 \)) approach, based on a reported \( F \) of 4.58, from Equation 25 and Kirk (1995, p. 519),

\[ \text{estimated } \omega^2 = (4.58 - 1)/((4.58 - 1) + (12)(20)) = .015 \]

and from Equation 5, again based on \( N' = 120 \),

\[ \phi = [120(.015/.985)]^{1/2} = 1.35, \]

which results in power of almost .47 (compared with .63 with the less conservative approach).

I now consider the between-subjects stimulus main effect. In terms of the cell means in Table 3, the mean difference between verbal and ideogram stimuli is given by the following:

\[ \text{estimated } \psi = \frac{1}{2}(.27 + .20 + .15 + .34 + .19 + .15) \]

\[ - \frac{1}{2}(.38 + .38 + .33 + .29 + .23) = -0.113, \]

which indicates that there was an average of .11 more frequency judgment errors with the ideograms than with the verbal stimuli. Incorporating \( MSE_{\text{between}} = .04 \) produces an estimated \( \psi' \) (in absolute value) of .113/\( .04^{1/2} = .565. \) Thus, again with \( \alpha = .05, \nu_1 = 1, \nu_2 = 76, \) and the same sum of squared coefficients,

\[ \phi = [20(.565)^2/(1 + 1)(.333)]^{1/2} = 3.09, \]

or, equivalently, \( f^2 = (.565)^2/12(.333) = .080 \) and \( \phi = [120(.080)]^{1/2} = 3.09, \) for which power is equal to .99. (A one-tailed test of this estimated contrast would be associated with power greater than .99.)

In terms of estimated partial \( \omega^2 \), based on a computed \( F \) of 19.17 (not reported by Wiggs, 1993),

\[ \text{estimated } \omega^2 = (19.17 - 1)/(19.17 - 1 + 12(20)) = .070 \]

and

\[ \phi = [120(.070/.930)]^{1/2} = 3.01, \]

which results in power of .98.

For the between-conditions age main effect (see Table 3),

\[ \text{estimated } \psi = \frac{1}{2}(.27 + .20 + .15 + .38 + .38 + .37) \]

\[ - \frac{1}{2}(.34 + .19 + .15 + .33 + .29 + .23) = .037. \]

With \( MSE_{\text{between}} = .04, \) this is translated into an estimated \( \psi' \) of .037/\( .04^{1/2} = .185 \) and \( \phi = 1.01 \) \( (f^2 = .0086), \) which with \( \alpha = .05, \nu_1 = 1, \nu_2 = 76, \) yields power of only .29. For power of .80 based on the current specifications, a much larger sample, namely, a total of 924 observations, or 77 older and 77 younger adults per stimulus condition (rather than the current 20), would be needed. (A one-tailed test of this contrast reveals power of .41 for the original sample size and that 720 total observations, 60 older and 60 younger adults per stimulus condition, are required for power of .80.)

With respect to the estimated proportion of between-subjects variance accounted for by the age main effect, based on a computed nonsignificant \( F \) of 2.06 (not reported by Wiggs, 1993),

\[ \text{estimated } \omega^2 = [(2.06 - 1)/(2.06 - 1) + (12)(20)] = .004 \]

and

\[ \phi = [120(.004/.996)]^{1/2} = .69, \]

for which power is equal to .16.

I now consider two of the within-subjects effects: the sizable frequency difference main effect and the Age X Frequency Difference interaction. For each of these, based on \( \nu_1 = 2 \) and \( \nu_2 = I(K - 1)(n - 1) = 2(2)(2)(19) = 152, \) I estimate \( \omega^2 \) from the \( F \) ratios (each computed from Table 3, with \( MSE_{\text{within}} = .02, \) and not reported by Wiggs, 1993) and apply Equation 5. For the stimulus main effect, \( F = 11.23 \) and so estimated \( \omega^2 = (3 - 1)(11.23 - 1)/(3 - 1)(11.23 - 1) + 12(20)) = 20.46/260.46 = .079, \)

which, with \( \alpha = .05 \) and \( N' = N/(\nu_1 + 1) = 240/(2 + 1) = 80, \) yields

\[ \phi = [80(.079/.921)]^{1/2} = 2.62 \]

for power of almost .99.

Alternatively, suppose that instead of wanting to assess the frequency difference main effect via an omnibus \( \nu_2 = 2 df F \) test, the researcher preferred to restrict attention to the contrast involving Frequency Difference Levels 1 and 3 (which, in the case of only three equally spaced levels of a quantitative factor, is equivalent to assessing the factor’s linear trend component). In terms of Table 4’s properly scaled cell means, that particular frequency difference contrast is given by

\[ \text{estimated } \psi = \frac{1}{2}(.27 + .38 + .34 + .33) \]

\[ - \frac{1}{2}(.15 + .37 + .15 + .23) = .33 - .23 = .10, \]
with an associated sum of squared coefficients of $4\left(\frac{1}{4}\right)^2 + 4\left(-\frac{1}{4}\right)^2 = .5$. With $MSE_{within} = .02$, this represents an estimated $\psi^{'2}$ of $100/0.02^{1/2} = .707$ and $\phi = 3.16 \left(f^2 = .083\right)$. If examined as a planned 1-df contrast based on $\alpha = .05$, power would exceed .99 (for either a one- or a two-tailed test).

Finally, for the Age × Frequency Difference interaction, with its computed $F$ of 1.71 (and which was found not to be statistically significant by Wiggs, 1993),

estimated $\omega^{2} = \frac{(3 - 1)(1.71 - 1)}{((3 - 1)(1.71 - 1)) + 12(20)} = 1.42/241.42 = .006$,

so that

$\phi = [80(0.006/.994)]^{1/2} = .69$.

The associated power for an interaction effect of that magnitude (with $\alpha = .05$) is only .17. To increase power to .80 in a replication study (based on the current estimate and with all other specifications the same), a considerably larger sample—135 older and 135 younger adults per stimulus condition (rather than the current 20)—would be needed.

The Age × Frequency Difference interaction could be assessed alternatively by the previously given contrast between Levels 1 and 3 of the frequency difference factor. For those two levels, the interaction contrast is defined in terms of Table 4's properly scaled cell means as

estimated $\psi = \frac{1}{2}[0.27 + 0.38] - (0.34 + 0.33)]$

and for which the sum of the squared coefficients again is equal to .5. With $MSE_{within} = .02$, this represents an estimated $\psi^{'2}$ of $-0.04/0.02^{1/2} = .283$ and $\phi = 1.27 \left(f^2 = .013\right)$. If assessed as a planned 1-df contrast based on $\alpha = .05$, power would be equal to .42 (for a two-tailed test) or .55 (for a one-tailed test). To increase power to .80 would require either 51 or 40 older and younger adults per stimulus condition for a two-tailed and a one-tailed test of the contrast, respectively.

### Table 5

Summary of Selected Effect Size and Power–Sample Size Equivalences

<table>
<thead>
<tr>
<th>Measure</th>
<th>Variance proportions ($\omega^2$)</th>
<th>Standardized contrasts ($\phi_c$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A: Effect size equivalences</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohen's $f^2$</td>
<td>$f^2 = \omega^2/(1 - \omega^2)$</td>
<td>$f^2 = \phi^2/c \sum a_i^2$</td>
</tr>
<tr>
<td></td>
<td>$\omega^2 = f^2/(1 + f^2)$</td>
<td></td>
</tr>
<tr>
<td><strong>B: Power–sample size equivalences</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pearson-Hartley $\phi^2$</td>
<td>$\phi^2 = N'\omega^2/(1 - \omega^2)$</td>
<td>$\phi^2 = n b \phi^2/[n(v_1 + 1)] c \sum a_i^2$</td>
</tr>
</tbody>
</table>

Note. $C =$ the number of cells in the design; $a_i =$ the properly scaled contrast coefficients; $N' =$ $N(v_1 + 1); n =$ the number of observations associated with each cell mean.

## Summary and Conclusions

What precisely was the point of this introductory power trip? My major purpose was to help researchers who are investigating the psychology of aging to become more sensitive to the role played by statistical power in hypothesis testing. Included in that increased sensitivity is a greater awareness of both the general issues surrounding statistical power (e.g., problems in interpreting statistically nonsignificant results from studies with inadequate power, the relationship between effect size, and required sample size) and specific procedures for ending up with respectable power or sample sizes for the experimental designs and analyses that are commonly applied by psychology-of-aging researchers. Indeed, as a reviewer of this journal pointed out, evaluation of a study’s power characteristics is already an explicit, and, from my perspective, a welcome addition to certain professional journals’ editorial policies. For example, the journal *Measurement and Evaluation in Counseling and Development* states as a matter of policy that “studies in which statistical significance is not achieved will be seriously considered for publication if power estimates of protection against Type II error are reported and reasonable protection is available.”

Moreover, what I have attempted to provide here was strictly a cook’s power tour rather than a comprehensive coverage of the waterfront, to the extent that the research examples and associated discussion were limited and were restricted only to comparisons of means. (Table 5 includes a summary of certain equivalences uncovered in the current excursions.) So that one will not get the wrong impression, analogous power considerations are relevant, and specific procedures exist, relative to testing hypotheses about other parameters, such as proportions, correlation coefficients, and multiple correlation coefficients (e.g., Cohen, 1988; Gatsonis & Sampson, 1989), among others (see also Footnote 5). In addition, procedures exist for extending the power concepts and considerations discussed here directly to situations in which aggregates of individuals (e.g., dyads, small groups, classrooms, or institutions), rather than individual participants, constitute the experimental “units” to which treatments or interventions are administered (e.g., Barcklowksi, 1981; Levin & Serlin, 1993).

Before concluding this power journey, I throw one additional
monkey wrench into the mix. The concern here was exclusively with a researcher having sufficient statistical power to detect effects that were deemed worthy of detecting. Such emphasis was in response to the more usual research circumstance of a study being found to have not enough statistical power to uncover effects that some would deem interesting or useful. It also should be realized, however, that the opposite research circumstance, but presumably less frequent, also is possible, namely, a large-scale study based on so many participants that it is associated with too much power, in the sense that even tiny, unimportant effects end up being detected. In the latter instance, the statistically significant results that are likely to pop out all over the place are just as likely to be totally uninformative and, at the least, need to be supplemented with “sample size-free” effect size measures that reflect the practical significance of the findings. However, just as with the “too hot”—“too cold” Three Bears metaphor that can be invoked to illustrate the extremes of a phenomenon, some consideration also has been given to the “just-right” aspects of power and sample size determination. Specifically, procedures have been proposed that enable a researcher to have sufficient power for detecting effects of interest, but not excessive power for detecting effects of noninterest. The curious reader is referred to some of those considerations (e.g., Serlin, 1993; Serlin & Lapsley, 1985; Walster & Cleary, 1970).

In some instances, one may apply the technical procedures of this article to assess the power characteristics of an already-conducted study or of a to-be-conducted study for which (because of either physical or fiscal constraints) there is no possibility of recruiting and including additional participants. In other instances, the technical procedures discussed here may be applied to determine how many participants are needed for the primary analyses of a planned study to achieve the desired power for detecting specified effects. The latter (sample size) situation is the situation to which psychology-of-aging researchers should aspire, and should they be fortunate enough to be in that situation, then more power to them! With careful attention to the sample size requisites of a planned for study, “aging” researchers will begin to experience no more feelings of powerlessness. To the contrary, with a firm grasp of the information contained in this article (including the associated references and computer programs), they will begin feeling empowered.

References


Received September 25, 1995
Revision received June 19, 1996
Accepted June 19, 1996