Journal of Applied Psychology
1999, Vol. 84, No. 2, 234-248

0021-9010/99/$3.00

Testing the Hypothesis That Treatments Have Negligible Effects: Minimum-Effect Tests in the General Linear Model

Kevin R. Murphy
Colorado State University

Brett Myors
Macquarie University

Researchers are often interested in testing the hypothesis that the effects of treatments, interventions, and so on are negligibly small rather than testing the hypothesis that treatments have no effect whatsoever. A number of procedures for conducting such tests have been suggested but have yet to be widely adopted. In this article, simple methods of testing such minimum-effect hypotheses are illustrated in a variety of applications of the general linear model. Tables and computational routines that can be used in conjunction with the familiar F test to evaluate the hypothesis that the effects of treatments or interventions exceed some minimum level are also provided.

One of the most common statistical procedures in the behavioral and social sciences is to test the hypothesis that treatments or interventions have no effect, or that the correlation between two variables is equal to zero, and so on. Cohen (1994) referred to these procedures as "nil hypothesis" tests, a label that differentiates them from the more general category of null hypothesis tests (which allow researchers to test the hypothesis that the difference between two treatments is equal to any specific figure, including but not limited to zero) and that makes it explicit that this particular class of tests is used to evaluate the plausibility of the hypothesis that treatments or interventions have no true effect whatsoever. Although nil hypothesis tests are extremely common, there is a substantial controversy about their value and meaning (Chow, 1988; Cohen, 1994; Cortina & Dunlap, 1997; Cowles, 1989; Harlow, Mulaik, & Steiger, 1997; Meehl, 1978; Morrison & Henkel, 1970; Murphy, 1990; Schmidt, 1992, 1996). One of the many criticisms of nil hypothesis tests is that they ask the wrong question. That is, these tests ask whether treatments, interventions, and so forth have *any* effect. The question many researchers (especially those interested in the application of science to solve practical problems) want to ask is whether the effects of treatments are large enough to make a real difference. The statistical tests most frequently encountered in the social and behavioral sciences do not directly address this question.

Several authors (e.g., Schmidt, 1996) have suggested that significance testing is unnecessary and that researchers should instead focus on estimates of effect size. A number of factors, including the increasing use of meta-analysis (Hunter & Hirsh, 1987; Lipsey & Wilson, 1993; Schmidt, 1992), increasing appreciation of the importance of statistical power (Cohen, 1988; Lipsey, 1990; Murphy & Myors, 1998), and the editorial policies of many leading journals (Thompson, in press), have helped to increase the likelihood that effect size estimates will indeed be presented in research reports. However, even when effect size estimates are available, tests of null hypotheses can still play a useful role in evaluating research results.

Effect size estimates, like all sample statistics, are fallible, and it is possible to find what appear to be meaningful effects in a sample even though population effects are either nonexistent or trivially small. Nil hypothesis tests provide a useful supplement to effect size estimates by helping investigators evaluate the likelihood that an effect as large as one observed in a particular sample could plausibly have been obtained from a population in which there is no true effect.[1] Thus, nil hypothesis tests provide useful tools for evaluating

Editor's Note. Nambury Raju served as the action editor for this article.—KRM

Kevin R. Murphy, Department of Psychology, Colorado State University; Brett Myors, Department of Psychology, Macquarie University, Sydney, New South Wales, Australia.

Correspondence concerning this article should be addressed to Kevin R. Murphy, Department of Psychology, Colorado State University, Fort Collins, Colorado 80523-1876. Electronic mail may be sent to krmurphy@lamar.colostate.edu.

[1] As sample sizes increase, the value of significance tests of any kind tends to decrease, and sample-based effect size estimates become increasingly precise. Tests of the null hypothesis (both nil hypothesis tests and the minimum-effect tests described in this article) are most informative when samples are not large enough to yield precise parameter estimates.

the strength of the evidence produced by a study (Frick, 1996). Although significance tests are not the optimal tool for evaluating the credibility of sample-based statistics (e.g., confidence intervals are often better suited to this purpose; Cohen, 1994; Fleishman, 1980; Thompson, 1997), nil hypothesis testing continues to be the most common method for addressing the fact that sample-based effect size estimates may not provide sufficient evidence to conclude that a treatment or intervention has an effect in the population. Given the continuing reliance on nil hypothesis tests, despite decades of criticism of these procedures (Thompson, 1997), it seems unlikely that these procedures will disappear in the foreseeable future. Rather than abandoning nil hypothesis testing, it might be preferable to concentrate on applications of well-known statistical methods that overcome many of the limitations of these tests.

Alternatives to Nil Hypothesis Tests

Rather than testing the hypothesis that treatments, interventions, correlates, and so on have no effect whatsoever, it is often useful to test the hypothesis that the effect of treatments is so small that it can be labeled "trivial" or "negligibly small." For example, suppose that researchers determine that a quality improvement program must produce a 2% decrease in rejected products to justify its cost. Nil hypothesis tests ask whether this program has any effect; it would be more useful to determine whether the researcher can be confident that it has at least the minimum effect needed to justify its costs (i.e., at least a 2% decrease in rejected products). Several procedures can be applied to solving problems of this sort.

Serlin and Lapsley (1985, 1993) described methods for testing the hypothesis that statistical effects fall within or outside ranges of values that define a nontrivial effect; these methods are sometimes referred to as tests of the "good-enough hypothesis." Rouanet (1996) showed how Bayesian methods can be used to assert the importance or negligibility of treatment effects. Although these methods hold considerable promise, they have not yet been widely adopted by applied researchers (perhaps because they can require researchers to substantially change the way they conduct their statistical analyses; Rouanet's method involves evaluating a Bayesian posterior probability distribution based on "noninformative" priors). For example, to our knowledge, these methods have not been applied in any of the articles in the *Journal of Applied Psychology* or similar journals (e.g., *Personnel Psychology, Organizational Behavior and Human Decision Processes, Academy of Management Journal*) in the past 5 years.

The purpose of this article is to describe simple methods of testing the hypothesis that effects in various applications of the general linear model either fail to meet (H_0) or exceed (H_1) some minimal threshold. In particular, we show that these methods (a) overcome some of the most critical limitations of traditional nil hypothesis tests, (b) are easily adapted and used by applied researchers whose primary interests and skills are not in the area of methodology, and (c) preserve as many links as possible to familiar statistical procedures. First, we discuss a method for conducting minimum-effect tests based on the widely applicable F statistic and show how a variety of common statistical tests can be easily adapted to this framework. Second, we show how the use of such minimum-effect allows the researcher to address many of the criticisms currently directed at nil hypothesis testing. Finally, we discuss the statistical power of minimum-effect tests and argue that any reductions in power (compared with parallel tests of the nil hypothesis) are more than compensated for by an increase in the meaningfulness of the statistical tests researchers carry out.

The method described in this article tests the hypothesis that in the population, the effect of treatments, interventions, and so forth is equal to or less than some minimal value. As we note later, setting a standard for defining "negligible" effects can be a complex process, but, once this standard is determined, tests of the hypothesis that the results obtained in a study could reasonably be found if the population effect was in the range described as "negligibly small" are easy to perform. For example, if researchers decide in a particular context that treatments that account for 1% or less of the variance have effects too small to justify their use, their task in developing a statistical test is to determine how large the effect in a particular sample would need to be before they could be confident that the variance accounted for in the population was 1% or greater. The tests developed here do just that, by determining the value of the F statistic needed to reject the hypothesis that the effect in the population is equal to or smaller than whatever benchmark is used to describe a negligibly small effect.

The method described here takes advantage of the fact that the F statistic is nondirectional. All other things being equal, a larger F corresponds to a larger effect. Thus, if the observed value of F found in a study is so large that it could not reasonably be obtained in a population for whom treatments of interventions accounted for 1% of the variance (e.g., the probability of finding an F this large given such a population effect is α or lower), researchers can be even more certain that it would not have been found if the true effect of treatments accounted for less than 1% of the variance. Thus, unlike nil hypothesis tests, which test a point hypothesis, the tests described here allow researchers to determine whether the observed effects would be likely to occur for a range of population values (e.g., true effects ranging from nil to 1% of the variance accounted for).

Note that the methods described here are not completely new or original but are based on long-recognized properties of the F distribution (Patnaik, 1949). Fowler (1985) dis-

cussed a method of testing nonzero null hypotheses based on a cube root transformation of the chi-square distribution and showed how this method could be applied to test the hypothesis that the effects of treatments or interventions were trivially small. Fleishman (1980) described a method for determining confidence intervals around effect size estimates. In some designs, this method could be applied to the same problem attacked here (i.e., determining whether one can rule out the hypothesis that the effect of treatments in the population is trivially small). Finally, methods for testing directional hypotheses in specific circumstances (e.g., that the difference between two means is greater than or equal to some value) have also long been available (Hays, 1994). Unlike the methods described here, these tests use the central F distribution and in most cases will have higher power than comparable tests based on the noncentral F. The purpose of this article is not to develop completely new statistical tests but to bring such tests to the attention of applied psychologists and to present them in terms that are as familiar as possible to researchers trained in traditional nil hypothesis testing. We also hope to illustrate their applicability to a wide range of problems encountered in applied research.

Minimum-effect tests are especially useful to applied researchers because they involve (a) articulating standards used to describe a treatment as having a negligible versus a meaningful effect and (b) a direct statement of the confidence a researcher can have that the effects of treatments meet or exceed this threshold. Applied researchers must often decide not only whether a treatment or intervention has any effect but also whether this effect is large enough to conclude that the treatment is worthwhile. The minimum-effect hypothesis tests we describe directly address this important question.

Minimum-Effect Tests

One way to increase the acceptance and use of minimum-effect tests similar to those developed by Serlin and Lapsley (1985, 1993) and Rouanet (1996) is to show how common statistical procedures can be easily adapted to carry out tests of this sort. In this article, we show how tests of minimum-effect hypotheses can be carried out in numerous applications of general linear models (Horton, 1978; Tatsuoka, 1993) simply by developing new tables (or by computing new critical values) for interpreting the widely used F statistic. Traditionally, the F statistic has been used to test the hypothesis that treatments have no effect whatsoever; we show how this same statistic can be used to test the hypothesis that the effects of treatments, interventions, and so on are equal to or smaller than some minimum value.

There are two reasons why we focus on the F statistic. First, the F test is ubiquitous. Most statistical tests carried out in the social and behavioral sciences (e.g., correlations and regressions, t tests, analyses of variance, or ANOVAs, analyses of covariance, or ANCOVAs, multiple regressions, discriminant function analyses) can be thought of as specialized applications of the general linear model (Horton, 1978; Tatsuoka, 1993), and the F statistic is broadly applicable to testing hypotheses involving the general linear model. Even statistical tests that are not generally framed in terms of the F statistic (e.g., testing the hypothesis that two variables are correlated) could easily be examined using this statistic. Second, moving from tests of the nil hypothesis to tests of minimum-effect hypothesis is essentially a matter of moving from the central to the noncentral F distribution in developing statistical tables or critical values for test statistics. As we show in sections that follow, once researchers specify the operational definition of the type of effect they wish to detect or rule out, it is relatively easy to estimate the critical value of the appropriate noncentral F distribution. When framed in terms of the F statistic, the process of testing minimum-effect hypotheses is an extension of the procedures typically used to test the familiar nil hypothesis.

F Tests for Nil Versus Minimum-Effect Hypotheses

The significance of the F statistic is usually assessed by comparing the value of the F obtained in a study to the value listed in an F table. If the obtained F is larger than the tabled value (F is tabled in terms of degrees of freedom $\nu 1$ and $\nu 2$, where $\nu 1$ represents the degrees of freedom for the substantive question, and $\nu 2$ represents the degrees of freedom for the error term), the nil hypothesis is rejected. Tests of minimum-effect hypotheses proceed in exactly the same way, only using a different set of tables or critical values.

The F tables found in the back of most statistics texts are based on the *central F distribution,* or the distribution of the F statistic that would be expected if treatments or interventions had no effect in the population. The shape and location of the central F distribution vary as a function of the degrees of freedom for the hypothesis being tested and the degrees of freedom for the error term used to test that hypothesis (i.e., $\nu 1$ and $\nu 2$).

Tests of minimum-effect hypotheses are based on on the *noncentral F distribution.* The shape and location of the noncentral F distribution is determined by the degrees of freedom $\nu 1$ and $\nu 2$ and by the noncentrality parameter (λ), which is essentially a function of the size of the effect size and sample size. For example, a good estimate of λ in most applications of the general linear model is given by the ratio $(\nu 2 \times \mathrm{PV})/(1 - \mathrm{PV})$, where PV is the percentage of variance in the dependent variable explained by the variable or variables in the linear model. All other things being equal, the larger the value of λ (and therefore the larger the PV),

the larger the F needed to reject the null hypothesis.[2] The central F distribution is a special case of the noncentral distribution, in which it is assumed that treatments have no effect on the dependent variable, which means that PV = 0 and therefore that $\lambda = 0$.

To develop F tables for tests of minimum-effect null hypotheses, researchers must (a) develop an operational definition of the minimum effect they wish to pay attention to (effects smaller than this are labeled "negligible"; cf. Rouanet, 1996); (b) calculate the noncentrality parameter that corresponds to this effect; and (c) tabulate or approximate the corresponding noncentral F distribution. For example, if researchers decide that effects accounting for 1% or less of the variance in outcomes should be labeled *negligible,* it is possible to then develop a set of F tables for testing the hypothesis that in the population, treatment effects fail to meet this threshold (i.e., that they account for 1% or less of the variance in outcomes). If researchers reject this hypothesis, they will be left with the alternative hypothesis that the effects of treatments are large enough to be meaningful (i.e., large enough to exceed their operational definition of a negligible effect).

Constructing Tables for Tests of Minimum-Effect Hypotheses

Tests of the nil hypothesis involve computing the value of F in a study and comparing it with tabled values based on the central F distribution. For example, in testing the nil hypothesis, with degrees of freedom of 2 and 50, the critical value of F ($\alpha = .05$) is 3.18 (this value can be obtained from virtually any statistics text). This indicates that 95% of the values in the central F distribution with 2 and 50 *df*s fall at or below 3.18; if researchers obtained an F greater than 3.18 in their study, they could be confident that the hypothesis that treatments have no effect was wrong. Formally, an F larger than 3.18 indicates that the probability of finding effects as large as those observed in one's study would be less than .05 if the nil hypothesis had been true.

Tables for testing minimum-effect hypotheses are constructed on the basis of the noncentral F distribution, in which the degree of noncentrality depends on the operational definition of a "negligible" effect, described in terms of the PV explained in the dependent variable. There are a number of serviceable approximations of the noncentral F distribution (e.g., Patnaik, 1949; Tiku & Yip, 1978), as well as programs for computing probabilities in that family of distributions (Narula & Weistroffer, 1986), and simple statistical functions that can be used for this purpose. In the appendix we describe a number of simple methods for approximating or computing critical values for relevant noncentral F distributions.

F tables for testing two different minimum-effect hypotheses, that treatments account for 1% or less of the variance in outcomes and that treatments account for 5% or less of the variance in outcomes, are presented in Tables 1 and 2, respectively. These tables present the critical values of F needed to reject minimum-effect null hypotheses (αs = .05 and .01) for PV values of 1% and 5%. That is, these tables can be used to determine whether the F value obtained in a study is large enough to reject the hypotheses that treatments explain 1% or less of the variance in outcomes or that treatments explain 5% or less of the variance in outcomes.

We believe that these particular thresholds are useful in many contexts. For example, treatments that account for 1% of the variance in outcomes correspond to Cohen's (1988) widely cited convention for defining a small effect (in terms of the d statistic, this corresponds to a difference between group means that is 20% as large as the pooled within-groups standard deviation). A treatment that accounts for 5% of the variance in outcomes falls roughly between Cohen's (1988) conventions for defining small and medium effects. As we note below, there are many situations in which researchers may set minimum thresholds that are lower than 1% of the variance or that are higher than 5% of the variance in defining a negligible effect, but these two values may be useful in a wide range of contexts.

Defining a Minimum Effect

The main advantage of the nil hypothesis is that it is simple and objective. If researchers reject the hypothesis that treatments have no effect, they are left with the alternative that the treatments have some effect. On the other hand, testing minimum-effect hypotheses requires a value judgment and requires that some consensus be reached in a particular field of inquiry. For example, the definition of a "negligible effect" might reasonably vary across areas, and there may be no set convention for defining which effects are so small that they can be effectively ignored and which cannot. However, it is possible to offer some broad principles for determining when effects are likely to be judged to be negligible.

First, the importance of an effect might depend substantially on the particular dependent variables involved. For example, in medical research it is common for relatively small effects (in terms of the PV explained) to be viewed as

[2] As λ increases, the mean and variance of the distribution of noncentral F values shift upward, so that a larger F value is needed to reject the null hypothesis at any fixed values for α, $\nu 1$ and $\nu 2$ as lambda increases (and therefore the percentage of variance) increases. For example, if one can reject the hypothesis that treatments account for 1% or less of the variance in outcomes (with a confidence level of .05), one can also reject any less stringent minimum-effect hypothesis (i.e., that treatments account for one half of 1% of the variance in outcomes) with a confidence level of less than or equal to .05.

Table 1
Critical Values of F for Testing the Hypothesis That Treatments Account for 1% or Less of the Variance in Outcomes

		*v*1																
*v*2	α	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
3	.05	10.43	9.70	9.37	9.19	9.07	8.99	8.93	8.88	8.84	8.81	8.77	8.72	8.67	8.63	8.60	8.58	8.55
	.01	35.15	31.28	29.75	28.93	28.41	28.05	27.79	27.59	27.44	27.31	27.12	26.93	26.73	26.53	26.43	26.33	26.23
4	.05	8.02	7.08	6.68	6.45	6.31	6.20	6.13	6.07	6.03	5.99	5.93	5.87	5.81	5.75	5.72	5.69	5.66
	.01	22.05	18.36	16.92	16.14	15.65	15.31	15.06	14.87	14.72	14.60	14.42	14.24	14.05	13.86	13.76	13.66	13.56
5	.05	6.94	5.93	5.50	5.26	5.10	4.99	4.91	4.85	4.80	4.76	4.70	4.63	4.57	4.50	4.47	4.44	4.40
	.01	17.07	13.61	12.26	11.54	11.08	10.76	10.53	10.35	10.21	10.10	9.93	9.75	9.58	9.39	9.30	9.21	9.12
6	.05	6.35	5.30	4.85	4.60	4.44	4.33	4.24	4.18	4.13	4.08	4.02	3.95	3.89	3.82	3.78	3.74	3.71
	.01	14.56	11.25	9.98	9.29	8.85	8.55	8.33	8.16	8.03	7.92	7.76	7.59	7.42	7.24	7.15	7.06	6.97
7	.05	5.98	4.90	4.45	4.19	4.03	3.91	3.82	3.76	3.71	3.66	3.60	3.53	3.46	3.38	3.35	3.31	3.27
	.01	13.09	9.88	8.65	7.98	7.57	7.28	7.06	6.90	6.77	6.67	6.51	6.34	6.18	6.01	5.92	5.83	5.74
8	.05	5.74	4.64	4.18	3.92	3.75	3.63	3.54	3.47	3.42	3.37	3.31	3.24	3.16	3.09	3.05	3.01	2.97
	.01	12.14	8.99	7.79	7.15	6.74	6.46	6.25	6.09	5.96	5.86	5.70	5.54	5.38	5.21	5.13	5.04	4.95
9	.05	5.58	4.45	3.98	3.72	3.54	3.42	3.34	3.27	3.21	3.17	3.10	3.02	2.95	2.87	2.83	2.79	2.75
	.01	11.49	8.38	7.20	6.57	6.17	5.89	5.69	5.53	5.41	5.30	5.15	4.99	4.83	4.66	4.58	4.49	4.40
10	.05	5.46	4.31	3.83	3.57	3.39	3.27	3.18	3.11	3.05	3.01	2.94	2.86	2.79	2.71	2.67	2.63	2.58
	.01	11.02	7.93	6.77	6.14	5.75	5.48	5.28	5.12	5.00	4.90	4.75	4.59	4.43	4.26	4.18	4.09	4.00
11	.05	5.37	4.20	3.72	3.45	3.27	3.15	3.06	2.99	2.93	2.88	2.81	2.74	2.66	2.58	2.54	2.49	2.45
	.01	10.67	7.60	6.44	5.82	5.43	5.16	4.96	4.81	4.69	4.59	4.44	4.28	4.12	3.96	3.87	3.78	3.69
12	.05	5.31	4.12	3.63	3.36	3.18	3.06	2.96	2.89	2.83	2.79	2.71	2.64	2.56	2.48	2.43	2.39	2.34
	.01	10.40	7.34	6.19	5.57	5.19	4.92	4.72	4.57	4.45	4.35	4.20	4.04	3.88	3.72	3.63	3.54	3.45
13	.05	5.27	4.05	3.56	3.28	3.10	2.98	2.88	2.81	2.75	2.71	2.63	2.55	2.47	2.39	2.35	2.30	2.25
	.01	10.20	7.13	5.99	5.37	4.99	4.72	4.52	4.37	4.25	4.15	4.00	3.85	3.69	3.52	3.44	3.35	3.26
14	.05	5.24	4.00	3.50	3.22	3.04	2.91	2.82	2.75	2.69	2.64	2.56	2.49	2.40	2.32	2.27	2.23	2.18
	.01	10.04	6.96	5.82	5.21	4.83	4.56	4.36	4.21	4.09	3.99	3.84	3.69	3.53	3.36	3.28	3.19	3.10
15	.05	5.22	3.96	3.45	3.17	2.99	2.86	2.76	2.69	2.63	2.58	2.51	2.43	2.34	2.26	2.21	2.16	2.12
	.01	9.91	6.82	5.68	5.08	4.69	4.43	4.23	4.08	3.96	3.86	3.71	3.56	3.40	3.23	3.14	3.05	2.96
16	.05	5.20	3.92	3.41	3.13	2.94	2.81	2.72	2.64	2.58	2.53	2.46	2.38	2.29	2.20	2.16	2.11	2.06
	.01	9.81	6.71	5.57	4.96	4.58	4.31	4.12	3.97	3.85	3.75	3.60	3.45	3.29	3.12	3.03	2.94	2.85
17	.05	5.20	3.89	3.38	3.09	2.91	2.77	2.68	2.60	2.54	2.49	2.41	2.33	2.25	2.16	2.11	2.06	2.01
	.01	9.73	6.62	5.47	4.87	4.48	4.22	4.02	3.87	3.75	3.65	3.50	3.35	3.19	3.02	2.93	2.84	2.75
18	.05	5.19	3.87	3.35	3.06	2.87	2.74	2.64	2.57	2.50	2.45	2.38	2.30	2.21	2.12	2.07	2.02	1.97
	.01	9.67	6.54	5.39	4.78	4.40	4.13	3.94	3.79	3.67	3.57	3.42	3.27	3.11	2.94	2.85	2.76	2.66
19	.05	5.20	3.85	3.32	3.03	2.84	2.71	2.61	2.54	2.47	2.42	2.34	2.26	2.18	2.08	2.04	1.98	1.93
	.01	9.62	6.47	5.32	4.71	4.33	4.06	3.87	3.72	3.60	3.50	3.35	3.19	3.03	2.86	2.77	2.68	2.59
20	.05	5.20	3.84	3.30	3.01	2.82	2.69	2.59	2.51	2.45	2.39	2.32	2.23	2.14	2.05	2.00	1.95	1.90
	.01	9.58	6.41	5.26	4.65	4.27	4.00	3.80	3.65	3.53	3.44	3.29	3.13	2.97	2.80	2.71	2.62	2.52
21	.05	5.21	3.83	3.29	2.99	2.80	2.66	2.56	2.48	2.42	2.37	2.29	2.21	2.12	2.02	1.97	1.92	1.87
	.01	9.55	6.36	5.21	4.60	4.21	3.94	3.75	3.60	3.48	3.38	3.23	3.07	2.91	2.74	2.65	2.56	2.46
22	.05	5.23	3.82	3.27	2.97	2.78	2.64	2.54	2.46	2.40	2.35	2.27	2.18	2.09	2.00	1.95	1.90	1.84
	.01	9.53	6.32	5.16	4.55	4.16	3.90	3.70	3.55	3.43	3.33	3.18	3.02	2.86	2.69	2.60	2.50	2.41
23	.05	5.24	3.81	3.26	2.96	2.76	2.62	2.52	2.44	2.38	2.33	2.24	2.16	2.07	1.97	1.92	1.87	1.82
	.01	9.52	6.29	5.12	4.51	4.12	3.85	3.65	3.50	3.38	3.28	3.13	2.98	2.81	2.64	2.55	2.46	2.36
24	.05	5.26	3.80	3.25	2.94	2.75	2.61	2.50	2.42	2.36	2.31	2.23	2.14	2.05	1.95	1.90	1.85	1.79
	.01	9.51	6.25	5.08	4.47	4.08	3.81	3.62	3.46	3.34	3.24	3.09	2.93	2.77	2.60	2.51	2.41	2.31
25	.05	5.27	3.80	3.24	2.93	2.73	2.59	2.49	2.41	2.34	2.29	2.21	2.12	2.03	1.93	1.88	1.83	1.77
	.01	9.51	6.23	5.05	4.43	4.05	3.78	3.58	3.43	3.31	3.21	3.06	2.90	2.73	2.56	2.47	2.37	2.27
26	.05	5.29	3.80	3.23	2.92	2.72	2.58	2.48	2.40	2.33	2.28	2.19	2.11	2.01	1.92	1.86	1.81	1.75
	.01	9.51	6.21	5.03	4.40	4.01	3.75	3.55	3.39	3.27	3.17	3.02	2.86	2.70	2.52	2.43	2.34	2.24
27	.05	5.31	3.80	3.22	2.91	2.71	2.57	2.46	2.38	2.32	2.26	2.18	2.09	2.00	1.90	1.85	1.79	1.73
	.01	9.51	6.19	5.00	4.38	3.99	3.72	3.52	3.37	3.24	3.14	2.99	2.83	2.67	2.49	2.40	2.30	2.20
28	.05	5.33	3.80	3.22	2.90	2.70	2.56	2.45	2.37	2.30	2.25	2.17	2.08	1.98	1.89	1.83	1.78	1.72
	.01	9.52	6.17	4.98	4.35	3.96	3.69	3.49	3.34	3.22	3.12	2.96	2.80	2.64	2.46	2.37	2.27	2.17
29	.05	5.35	3.80	3.21	2.89	2.69	2.55	2.44	2.36	2.29	2.24	2.15	2.07	1.97	1.87	1.82	1.76	1.70
	.01	9.53	6.16	4.96	4.33	3.94	3.67	3.47	3.31	3.19	3.09	2.94	2.78	2.61	2.44	2.34	2.25	2.14
30	.05	5.38	3.80	3.21	2.89	2.68	2.54	2.43	2.35	2.28	2.23	2.14	2.05	1.96	1.86	1.80	1.75	1.69
	.01	9.54	6.15	4.94	4.31	3.92	3.64	3.44	3.29	3.17	3.07	2.91	2.75	2.59	2.41	2.32	2.22	2.12
40	.05	5.64	3.85	3.21	2.86	2.64	2.49	2.38	2.29	2.22	2.16	2.07	1.97	1.87	1.77	1.71	1.65	1.58
	.01	9.75	6.12	4.86	4.20	3.79	3.51	3.30	3.14	3.01	2.91	2.75	2.59	2.42	2.23	2.14	2.03	1.92
50	.05	5.93	3.94	3.24	2.87	2.63	2.47	2.35	2.26	2.19	2.12	2.03	1.93	1.83	1.71	1.65	1.59	1.52
	.01	10.04	6.18	4.85	4.16	3.74	3.44	3.23	3.07	2.94	2.83	2.67	2.50	2.32	2.13	2.03	1.92	1.81
60	.05	6.24	4.04	3.29	2.90	2.64	2.47	2.35	2.25	2.17	2.11	2.01	1.91	1.80	1.68	1.62	1.55	1.47
	.01	10.35	6.28	4.88	4.16	3.72	3.42	3.20	3.03	2.90	2.79	2.62	2.45	2.26	2.07	1.96	1.85	1.73

Table 1 (*continued*)

		*v*1																
*v*2	α	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
70	.05	6.57	4.14	3.35	2.92	2.66	2.48	2.35	2.25	2.17	2.10	2.00	1.89	1.78	1.66	1.59	1.52	1.44
	.01	10.67	6.39	4.93	4.18	3.73	3.41	3.19	3.01	2.87	2.76	2.59	2.41	2.23	2.03	1.92	1.80	1.68
80	.05	6.83	4.26	3.41	2.96	2.69	2.50	2.36	2.26	2.17	2.10	2.00	1.89	1.77	1.64	1.57	1.50	1.42
	.01	10.98	6.51	4.99	4.22	3.74	3.42	3.19	3.01	2.86	2.75	2.57	2.39	2.20	1.99	1.89	1.77	1.64
90	.05	6.97	4.37	3.48	3.00	2.71	2.52	2.38	2.26	2.18	2.11	2.00	1.88	1.76	1.63	1.56	1.48	1.40
	.01	11.29	6.64	5.06	4.26	3.77	3.43	3.19	3.01	2.86	2.74	2.56	2.38	2.18	1.97	1.86	1.74	1.61
100	.05	7.24	4.49	3.55	3.04	2.74	2.54	2.39	2.28	2.19	2.11	2.00	1.88	1.76	1.62	1.55	1.47	1.38
	.01	11.60	6.76	5.13	4.30	3.80	3.45	3.21	3.02	2.87	2.75	2.56	2.37	2.17	1.96	1.84	1.72	1.58
120	.05	7.76	4.74	3.66	3.13	2.81	2.59	2.43	2.31	2.21	2.13	2.01	1.89	1.75	1.61	1.54	1.45	1.36
	.01	12.20	7.02	5.28	4.40	3.86	3.50	3.24	3.04	2.88	2.76	2.56	2.36	2.15	1.93	1.81	1.69	1.55
150	.05	8.61	5.01	3.86	3.28	2.92	2.66	2.49	2.36	2.25	2.17	2.03	1.90	1.76	1.61	1.53	1.44	1.34
	.01	13.04	7.40	5.51	4.56	3.98	3.59	3.31	3.09	2.93	2.79	2.58	2.37	2.15	1.92	1.79	1.66	1.51
200	.05	9.58	5.57	4.22	3.49	3.08	2.81	2.61	2.46	2.33	2.23	2.09	1.93	1.78	1.61	1.52	1.43	1.32
	.01	14.39	8.02	5.90	4.83	4.18	3.75	3.43	3.20	3.01	2.86	2.63	2.40	2.16	1.91	1.78	1.64	1.48
300	.05	11.62	6.54	4.85	3.97	3.43	3.08	2.84	2.65	2.51	2.38	2.20	2.02	1.83	1.64	1.54	1.43	1.31
	.01	16.85	9.18	6.63	5.36	4.59	4.07	3.70	3.42	3.20	3.03	2.76	2.49	2.22	1.93	1.78	1.62	1.45
400	.05	13.49	7.44	5.43	4.42	3.78	3.35	3.07	2.85	2.68	2.54	2.32	2.11	1.90	1.68	1.56	1.44	1.30
	.01	19.02	10.26	7.33	5.87	4.98	4.39	3.97	3.65	3.40	3.20	2.90	2.60	2.29	1.97	1.80	1.63	1.44
500	.05	15.23	8.29	5.98	4.82	4.13	3.65	3.29	3.04	2.84	2.69	2.45	2.21	1.97	1.72	1.59	1.45	1.30
	.01	21.10	11.28	8.00	6.36	5.37	4.70	4.23	3.88	3.60	3.38	3.04	2.71	2.36	2.01	1.83	1.64	1.43
600	.05	16.94	9.11	6.51	5.21	4.43	3.91	3.54	3.24	3.01	2.84	2.57	2.31	2.03	1.76	1.62	1.47	1.31
	.01	23.08	12.25	8.64	6.83	5.74	5.01	4.49	4.10	3.80	3.55	3.18	2.81	2.44	2.06	1.86	1.66	1.44
1,000	.05	23.25	12.26	8.59	6.76	5.66	4.93	4.40	3.99	3.66	3.42	3.05	2.68	2.30	1.93	1.74	1.54	1.34
	.01	30.44	15.89	11.01	8.59	7.13	6.16	5.47	4.95	4.54	4.22	3.73	3.24	2.75	2.25	2.00	1.74	1.46
10,000	.05	135.80	68.43	45.99	34.77	28.04	23.55	20.34	17.94	16.07	14.57	12.33	10.06	7.80	5.56	4.44	3.31	2.19
	.01	152.70	76.89	51.63	38.99	31.39	26.35	22.74	20.04	17.94	16.26	13.73	11.21	8.69	6.16	4.90	3.63	2.36

Note. *v*1 = degrees of freedom for effect; *v*2 = degrees of freedom for error term.

meaningful and important (Rosenthal, 1993). One reason is that the dependent variables in these studies often include quality of life and even survival. A small PV might translate into many lives saved.

Second, decisions about what effects should be deemed negligible might depend on the relative likelihood and relative seriousness of Type I versus Type II errors in a particular area. As we note in a later section, the power of statistical tests in the general linear model decreases as the definition of a negligible effect expands. In any particular study, power is higher for testing the nil hypothesis that treatments have no effect than for testing the hypothesis that they account for 1% or less of the variance in outcomes and higher for tests of the hypothesis that treatments account for 1% or less of the variance than for the hypothesis that treatments account for 5% or less of the variance in outcomes. If Type II errors are seen as being particularly serious in a particular area of research, it might make sense to choose a low figure as the definition of a negligible effect.

On the other hand, there are many areas of inquiry in which numerous well-validated treatments are already available (see Lipsey & Wilson, 1993, for a review of numerous meta-analyses of treatment effects), and, in these areas, it might make sense to "set a higher bar" by testing a more demanding hypothesis. For example, in the area of cognitive ability testing (where the criterion is some measure of performance on the job or in the classroom), it is common to find that tests account for 20%–25% of the variance in the criterion (Hunter & Hirsch, 1987; Hunter & Hunter, 1984). Tests of the traditional nil hypothesis (i.e., that tests have no relationship whatsoever to these criteria) are relatively easy to reject; if $\rho^2 = .25$, a study with a sample size of 28 will have a power of .80 for rejecting the nil hypothesis (Cohen, 1988). Similarly, the hypothesis that tests account for 1% or less of the variance in these criteria is easy to reject; if $\rho^2 = .25$, a study with a sample size of 31 will have a power of .80 for rejecting this minimum-effect hypothesis (Murphy & Myors, 1998). In this context, it might make sense to define a negligible relationship as one in which the tests accounted for 10% or less of the variance in these criteria.

Utility analysis has been used to help determine whether particular treatment have effects that are large enough to warrant attention (Landy, Farr, & Jacobs, 1982; Schmidt, Hunter, McKenzie, & Muldrow, 1979; Schmidt, Mack, & Hunter, 1984). Utility equations suggest another important parameter that is likely to affect the decision of what represents a negligible versus a meaningful effect—that is, the standard deviation of the dependent variable, or SD_y. When there is a substantial and meaningful variance in the outcome variable of interest, a treatment that accounts for a

Table 2
Critical Values of F for Testing the Hypothesis That Treatments Account for 5% or Less of the Variance in Outcomes

		*v*1																
*v*2	α	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
3	.05	11.72	10.30	9.76	9.48	9.30	9.18	9.09	9.02	8.97	8.92	8.86	8.79	8.73	8.66	8.63	8.59	8.56
	.01	39.41	33.23	31.00	29.84	29.13	28.64	28.30	28.03	27.82	27.66	27.41	27.15	26.90	26.64	26.52	26.39	26.26
4	.05	9.31	7.67	7.05	6.72	6.52	6.38	6.28	6.20	6.14	6.09	6.02	5.94	5.86	5.79	5.75	5.71	5.67
	.01	25.49	19.86	17.85	16.81	16.17	15.74	15.42	15.19	15.00	14.85	14.63	14.40	14.17	13.93	13.82	13.70	13.58
5	.05	8.31	6.54	5.88	5.53	5.32	5.17	5.06	4.98	4.91	4.86	4.78	4.70	4.62	4.54	4.49	4.45	4.41
	.01	20.28	14.97	13.10	12.13	11.54	11.14	10.85	10.63	10.45	10.31	10.10	9.89	9.68	9.46	9.35	9.24	9.13
6	.05	7.82	5.94	5.25	4.89	4.66	4.51	4.40	4.31	4.24	4.19	4.11	4.02	3.94	3.85	3.80	3.76	3.71
	.01	17.73	12.58	10.78	9.86	9.29	8.91	8.63	8.42	8.25	8.12	7.92	7.72	7.51	7.30	7.20	7.09	6.99
7	.05	7.56	5.59	4.87	4.49	4.26	4.10	3.99	3.90	3.83	3.77	3.68	3.60	3.51	3.42	3.37	3.32	3.28
	.01	16.29	11.21	9.46	8.55	8.00	7.63	7.36	7.15	6.99	6.86	6.67	6.47	6.27	6.07	5.96	5.86	5.75
8	.05	7.44	5.37	4.62	4.24	3.99	3.83	3.71	3.62	3.55	3.49	3.40	3.31	3.22	3.12	3.07	3.03	2.98
	.01	15.41	10.35	8.61	7.72	7.18	6.81	6.54	6.34	6.19	6.06	5.86	5.67	5.47	5.27	5.17	5.07	4.96
9	.05	7.39	5.23	4.46	4.06	3.81	3.64	3.51	3.42	3.34	3.28	3.19	3.10	3.01	2.91	2.86	2.81	2.76
	.01	14.85	9.78	8.04	7.16	6.62	6.25	5.99	5.79	5.63	5.50	5.31	5.12	4.92	4.72	4.62	4.52	4.42
10	.05	7.39	5.13	4.34	3.93	3.67	3.50	3.37	3.27	3.20	3.13	3.04	2.94	2.85	2.75	2.70	2.64	2.59
	.01	14.48	9.38	7.64	6.75	6.21	5.85	5.58	5.38	5.23	5.10	4.91	4.72	4.52	4.32	4.22	4.12	4.01
11	.05	7.42	5.08	4.26	3.83	3.57	3.39	3.26	3.16	3.08	3.02	2.92	2.82	2.72	2.62	2.57	2.51	2.46
	.01	14.23	9.09	7.34	6.45	5.91	5.55	5.28	5.08	4.93	4.80	4.61	4.41	4.22	4.02	3.92	3.81	3.71
12	.05	7.47	5.04	4.20	3.76	3.49	3.31	3.17	3.07	2.99	2.93	2.83	2.73	2.62	2.52	2.46	2.41	2.35
	.01	14.07	8.88	7.12	6.23	5.68	5.31	5.05	4.85	4.69	4.56	4.37	4.18	3.98	3.78	3.68	3.57	3.47
13	.05	7.54	5.03	4.15	3.70	3.43	3.24	3.10	3.00	2.92	2.85	2.75	2.65	2.54	2.43	2.38	2.32	2.26
	.01	13.98	8.73	6.95	6.05	5.50	5.13	4.86	4.66	4.50	4.38	4.18	3.99	3.79	3.59	3.48	3.38	3.27
14	.05	7.62	5.02	4.13	3.66	3.38	3.19	3.05	2.94	2.86	2.79	2.69	2.58	2.47	2.36	2.31	2.25	2.19
	.01	13.93	8.61	6.82	5.91	5.36	4.98	4.72	4.51	4.35	4.22	4.03	3.83	3.63	3.43	3.33	3.22	3.11
15	.05	7.71	5.03	4.11	3.63	3.34	3.15	3.01	2.90	2.81	2.74	2.64	2.53	2.42	2.31	2.25	2.19	2.13
	.01	13.91	8.53	6.71	5.80	5.24	4.86	4.59	4.39	4.23	4.10	3.90	3.71	3.50	3.30	3.19	3.09	2.98
16	.05	7.81	5.04	4.10	3.61	3.32	3.12	2.97	2.86	2.77	2.70	2.59	2.48	2.37	2.25	2.19	2.13	2.07
	.01	13.91	8.47	6.63	5.71	5.15	4.77	4.49	4.29	4.13	4.00	3.80	3.60	3.39	3.19	3.08	2.97	2.86
17	.05	7.91	5.06	4.09	3.60	3.29	3.09	2.94	2.83	2.74	2.67	2.56	2.44	2.33	2.21	2.15	2.09	2.02
	.01	13.94	8.43	6.57	5.64	5.07	4.69	4.41	4.20	4.04	3.91	3.71	3.51	3.30	3.09	2.99	2.88	2.77
18	.05	8.02	5.09	4.09	3.59	3.28	3.07	2.92	2.80	2.71	2.64	2.52	2.41	2.29	2.17	2.11	2.05	1.98
	.01	13.99	8.40	6.52	5.58	5.00	4.62	4.34	4.13	3.97	3.83	3.63	3.43	3.22	3.01	2.90	2.79	2.68
19	.05	8.13	5.11	4.10	3.58	3.26	3.05	2.90	2.78	2.69	2.61	2.50	2.30	2.26	2.14	2.08	2.01	1.95
	.01	14.04	8.39	6.49	5.53	4.95	4.56	4.28	4.07	3.90	3.77	3.57	3.36	3.15	2.94	2.83	2.72	2.61
20	.05	8.20	5.15	4.11	3.58	3.26	3.04	2.88	2.76	2.67	2.59	2.47	2.36	2.23	2.11	2.05	1.98	1.91
	.01	14.11	8.38	6.46	5.49	4.91	4.51	4.23	4.02	3.85	3.71	3.51	3.30	3.09	2.88	2.77	2.65	2.54
21	.05	8.30	5.18	4.12	3.58	3.25	3.03	2.87	2.74	2.65	2.57	2.45	2.33	2.21	2.08	2.02	1.95	1.88
	.01	14.18	8.39	6.44	5.46	4.87	4.47	4.19	3.97	3.80	3.66	3.46	3.25	3.04	2.82	2.71	2.60	2.48
22	.05	8.41	5.22	4.13	3.58	3.25	3.02	2.86	2.73	2.63	2.56	2.44	2.31	2.19	2.06	1.99	1.92	1.85
	.01	14.26	8.40	6.43	5.44	4.84	4.44	4.15	3.93	3.76	3.62	3.41	3.20	2.99	2.77	2.66	2.54	2.43
23	.05	8.52	5.26	4.15	3.59	3.25	3.02	2.85	2.72	2.62	2.54	2.42	2.30	2.17	2.04	1.97	1.90	1.83
	.01	14.34	8.41	6.42	5.42	4.82	4.41	4.12	3.90	3.73	3.59	3.38	3.16	2.95	2.73	2.61	2.50	2.38
24	.05	8.63	5.30	4.16	3.59	3.25	3.01	2.84	2.71	2.61	2.53	2.41	2.28	2.15	2.02	1.95	1.88	1.81
	.01	14.43	8.43	6.42	5.41	4.80	4.39	4.09	3.87	3.69	3.55	3.34	3.13	2.91	2.68	2.57	2.45	2.33
25	.05	8.74	5.34	4.18	3.60	3.25	3.01	2.84	2.71	2.60	2.52	2.40	2.27	2.14	2.00	1.93	1.86	1.79
	.01	14.53	8.46	6.42	5.40	4.78	4.37	4.07	3.84	3.67	3.53	3.31	3.10	2.87	2.65	2.53	2.42	2.29
26	.05	8.85	5.38	4.20	3.60	3.25	3.01	2.84	2.70	2.60	2.51	2.39	2.26	2.12	1.99	1.92	1.84	1.77
	.01	14.63	8.48	6.43	5.39	4.77	4.35	4.05	3.82	3.64	3.50	3.29	3.07	2.84	2.62	2.50	2.38	2.26
27	.05	8.96	5.43	4.22	3.62	3.26	3.01	2.83	2.70	2.59	2.51	2.38	2.25	2.11	1.97	1.90	1.83	1.75
	.01	14.73	8.51	6.44	5.39	4.76	4.34	4.03	3.80	3.62	3.48	3.26	3.04	2.82	2.59	2.47	2.35	2.22
28	.05	9.07	5.47	4.25	3.64	3.26	3.01	2.83	2.70	2.59	2.50	2.37	2.24	2.10	1.96	1.89	1.81	1.73
	.01	14.83	8.55	6.45	5.39	4.75	4.33	4.02	3.79	3.61	3.46	3.24	3.02	2.79	2.56	2.44	2.32	2.19
29	.05	9.18	5.52	4.27	3.65	3.27	3.02	2.83	2.69	2.59	2.50	2.36	2.23	2.09	1.95	1.87	1.80	1.72
	.01	14.93	8.58	6.46	5.39	4.75	4.32	4.01	3.77	3.59	3.44	3.22	3.00	2.77	2.53	2.41	2.29	2.16
30	.05	9.29	5.57	4.29	3.66	3.28	3.02	2.83	2.69	2.58	2.49	2.36	2.22	2.08	1.94	1.86	1.78	1.70
	.01	15.04	8.62	6.47	5.40	4.75	4.31	4.00	3.76	3.58	3.43	3.20	2.98	2.75	2.51	2.39	2.27	2.14
40	.05	10.44	6.01	4.56	3.83	3.39	3.09	2.88	2.72	2.59	2.49	2.34	2.19	2.03	1.86	1.78	1.69	1.60
	.01	16.14	9.06	6.70	5.51	4.80	4.32	3.98	3.72	3.52	3.35	3.11	2.86	2.61	2.36	2.22	2.09	1.95
50	.05	11.39	6.50	4.84	4.02	3.53	3.20	2.96	2.78	2.64	2.53	2.36	2.19	2.01	1.83	1.74	1.64	1.54
	.01	17.26	9.56	6.98	5.69	4.92	4.40	4.03	3.75	3.53	3.35	3.09	2.82	2.55	2.28	2.14	1.99	1.84
60	.05	12.49	6.94	5.14	4.23	3.68	3.31	3.05	2.86	2.70	2.58	2.39	2.20	2.01	1.82	1.72	1.61	1.50
	.01	18.38	10.06	7.29	5.90	5.07	4.51	4.11	3.81	3.58	3.39	3.11	2.82	2.53	2.24	2.09	1.93	1.77

Table 2 (*continued*)

		*v*1																
*v*2	α	1	2	3	4	5	6	7	8	9	10	12	15	20	30	40	60	120
70	.05	13.34	7.42	5.45	4.43	3.84	3.44	3.16	2.94	2.77	2.64	2.44	2.23	2.03	1.81	1.71	1.59	1.48
	.01	19.46	10.58	7.61	6.13	5.23	4.64	4.21	3.89	3.64	3.44	3.14	2.84	2.53	2.22	2.06	1.89	1.72
80	.05	14.39	7.84	5.71	4.65	4.01	3.58	3.26	3.03	2.85	2.71	2.49	2.27	2.04	1.82	1.70	1.58	1.46
	.01	20.52	11.08	7.93	6.35	5.41	4.77	4.32	3.98	3.72	3.50	3.18	2.86	2.54	2.21	2.04	1.87	1.69
90	.05	15.17	8.31	6.02	4.88	4.15	3.70	3.37	3.12	2.93	2.77	2.54	2.30	2.07	1.83	1.70	1.58	1.44
	.01	21.57	11.59	8.25	6.59	5.58	4.92	4.44	4.08	3.80	3.57	3.24	2.90	2.55	2.21	2.03	1.85	1.66
100	.05	16.18	8.81	6.27	5.05	4.32	3.83	3.49	3.21	3.00	2.84	2.60	2.34	2.09	1.84	1.71	1.57	1.43
	.01	22.59	12.08	8.57	6.82	5.76	5.06	4.56	4.18	3.88	3.65	3.29	2.94	2.58	2.21	2.03	1.84	1.64
120	.05	17.88	9.64	6.89	5.45	4.64	4.09	3.70	3.41	3.17	2.98	2.71	2.43	2.15	1.87	1.72	1.57	1.42
	.01	24.59	13.05	9.20	7.28	6.12	5.35	4.80	4.38	4.06	3.80	3.41	3.02	2.63	2.23	2.03	1.83	1.61
150	.05	20.52	10.86	7.64	6.06	5.11	4.48	4.03	3.69	3.41	3.20	2.88	2.57	2.24	1.92	1.75	1.59	1.41
	.01	27.47	14.46	10.12	7.95	6.65	5.78	5.16	4.69	4.33	4.04	3.60	3.17	2.73	2.28	2.06	1.83	1.59
200	.05	24.55	12.87	8.94	7.02	5.88	5.11	4.56	4.15	3.84	3.56	3.18	2.79	2.40	2.01	1.82	1.62	1.41
	.01	32.04	16.72	11.60	9.04	7.51	6.49	5.76	5.21	4.78	4.44	3.93	3.42	2.90	2.38	2.12	1.85	1.58
300	.05	32.04	16.65	11.52	8.94	7.35	6.32	5.59	5.05	4.62	4.28	3.77	3.25	2.74	2.22	1.97	1.70	1.44
	.01	40.62	20.94	14.40	11.13	9.16	7.86	6.92	6.21	5.67	5.23	4.58	3.92	3.26	2.60	2.27	1.94	1.59
400	.05	39.07	20.15	13.84	10.68	8.79	7.53	6.62	5.95	5.42	4.98	4.33	3.71	3.08	2.44	2.12	1.80	1.48
	.01	48.68	24.94	17.05	13.11	10.74	9.16	8.04	7.19	6.53	6.00	5.21	4.42	3.63	2.84	2.44	2.04	1.63
500	.05	46.31	23.76	16.24	12.45	10.16	8.63	7.57	6.77	6.15	5.65	4.91	4.16	3.40	2.66	2.28	1.90	1.52
	.01	56.40	28.82	19.60	15.02	12.26	10.43	9.11	8.13	7.36	6.75	5.83	4.91	3.99	3.07	2.61	2.14	1.67
600	.05	52.82	26.98	18.38	14.08	11.50	9.78	8.55	7.63	6.91	6.34	5.48	4.59	3.73	2.86	2.44	2.00	1.56
	.01	63.87	32.55	22.11	16.89	13.75	11.65	10.16	9.05	8.18	7.48	6.44	5.39	4.35	3.30	2.78	2.25	1.72
1,000	.05	78.99	40.07	27.09	20.61	16.71	14.12	12.27	10.88	9.79	8.93	7.63	6.32	5.00	3.71	3.06	2.41	1.75
	.01	92.43	46.81	31.60	23.96	19.41	16.37	14.20	12.57	11.30	10.29	8.77	7.25	5.73	4.21	3.45	2.68	1.92
10,000	.05	601.30	301.20	201.20	151.10	121.10	101.10	86.81	76.09	67.76	61.09	51.08	41.08	31.07	21.06	16.05	11.04	6.03
	.01	637.80	319.40	213.30	160.30	128.40	107.20	92.03	80.66	71.81	64.74	54.12	43.51	32.89	22.28	16.97	11.67	6.36

Note. *v*1 = degrees of freedom for effect; *v*2 = degrees of freedom for error term.

relatively small PV might nevertheless lead to practical benefits that far exceed the costs of the treatment.

For example, suppose a training program costs $1,000 per person to administer and that it is proposed as a method of improving performance in a setting in which the current SD_y (i.e., the standard deviation of performance) is $10,000. If the effects of training account for less that 1% of the variation in job performance, one might conclude that the projected cost of training will exceed the projected benefit (based on the equation $\Delta U = r_{xy}x \cdot SD_y - C$, where ΔU is the projected overall benefit, r_{xy} is the relationship between the training and the criterion [in this case, PV = .01 translates into r_{xy} = .10], and C represents the cost), which suggests that PV = .01 represents a sensible definition of the minimum effect needed to label training effects as "nontrivial."

As we show in the Appendix, in situations in which the minimum threshold for defining a negligible effect does not correspond to the 1% or 5% benchmarks used to define Tables 1 and 2, it is necessary to compute one's own critical F value. In the Appendix we describe a number of methods for attacking this problem, some of which (e.g., using functions in SPSS or Excel) are easy to implement regardless of one's level of comfort with mathematical operations.

Illustrative Examples

The implications of the approach described here are best appreciated by illustrating its application in a series of analyses using a range of general linear model test statistics. In the examples that follow, we use the terms *1% minimum* and *5% minimum* to refer to tests of the hypothesis that treatments account for 1% or less of the variance in outcomes and the hypothesis that treatments account for 5% or less of the variance in outcomes, respectively.

t Tests

Perhaps the most familiar statistical problem in the social and behavioral sciences involves using the *t* test to compare the mean scores of participants who receive some treatment with the mean scores of participants who do not. A research design in which participants are randomly assigned to treatment and control conditions is easy to implement and can provide clear-cut and convincing answers to what are sometimes highly complex questions. To illustrate the application of minimum-effect hypothesis testing in this context, we have chosen a study of the effect of regular exercise on placental growth rates in pregnant women (Clapp & Rizk, 1992).

Clapp and Rizk (1992) measured placental volumes in 18

healthy women who maintained a regular routine of exercise during pregnancy. Nine of the women engaged in aerobics, 4 ran, and 5 swam. The control group comprised 16 women who did not engage in such a regimen. Placental volumes were measured using modern ultrasound techniques at 16, 20, and 24 weeks gestation. Key results from this study are presented in Table 3.

Two aspects of that study are especially noteworthy. First, the sample was small (i.e., $N = 34$). In most cases, this would mean low power levels. However, the effects were strong in that study. In the three time periods studied, exercise accounted for between 29% and 47% of the variance in placental volume (*df*s = 1.71–2.41). Because the apparent effects of exercise were substantial, it should be easy to rule out the hypothesis that the treatment has no effect (the nil hypothesis) or even that the true effects of exercise are, at best, small (e.g., tests of a 1% minimum-effect hypothesis).

The t test has 32 *df*s for that study; if it is squared, the statistic is distributed as F with 1 and 32 *df*s. The critical F values for 1 and 32 *df*s when testing the nil hypothesis are 4.14 and 7.51 (αs = .05 and .01, respectively). All of the observed F values in Table 3 were greater than 7.51, so we can reject the hypothesis that exercise has no effect whatsoever at the .01 level and can conclude that regular exercise has some impact on placental volume.

One can conclusively rule out the possibility that exercise has no effect, and the data suggest that the actual effect is large. However, it is always possible that the true effect is small and that the large r^2 and d values observed here represent chance fluctuations in a process that usually produces only a small effect. One can use Tables 1 and 2 to test the hypothesis that the effects of treatments exceed two particular thresholds for defining negligible effects (i.e., tests of the hypothesis that treatments account for 1% or less or 5% or less of the variance in outcomes).

The critical F ($\alpha = .05$) for the tests of hypothesis that treatments account for 5% or less of the variance here is 9.52; this value can be obtained by linear interpolation between values reported in Table 2 for $v2 = 30$ and $v2 = 40$. All F values in Table 3 exceed 9.52, so we can reject the hypothesis that treatments account for 5% or less of the variance in outcomes, with a 95% level of confidence. We can also reject this hypothesis at the .01 level for mean placental volumes at 20 and 24 weeks (critical $F = 15.26$; observed Fs = 15.66 and 29.44 at 20 and 24 weeks, respectively). Furthermore, because we can reject the 5% minimum-effect hypothesis in all of the tests described above, we can also reject the 1% minimum-effect hypothesis for these tests.

Correlation

McDaniel (1988) used large samples to study the validity of measures of preemployment drug use as predictors of job suitability in the military. Validity coefficients for the preemployment use of drugs such as marijuana, cocaine, and various stimulants and depressants were calculated in participant samples ranging in size from 9,224 to 9,355. Several of the validities reported by McDaniel (1988) are shown in Table 4.

Even though drug tests never account for more than one half of 1% of the variance in job suitability, all the correlations shown in Table 4 are "statistically significant." However, the outcome of this significance test clearly says more about the power of the study than about the size of the effect. With samples this large, it is almost impossible not to reject the nil hypothesis; correlations as small as .02 are significantly different from zero.

If we test the hypothesis that drug tests have a negligibly small relationship with measures of job suitability (where a negligible relationship is defined as one in which tests account for 1% or less of the variance in suitability), the large sample does *not* lead one to reject this minimum-effect hypothesis. At 1 and 9,224 *df*s, the critical F ($\alpha = .05$) for rejecting the hypothesis that drug tests account for 1% or less of the variance in job suitability is 126.1, which is well above any of the observed F values in Table 4. In other words, none of the validities falls outside of the range we have designated as *negligible.* Although they are significantly different from zero, it is clear that the relationship between these tests and job suitability is so weak that the

Table 3
Placental Volumes Reported by Clapp and Rizk (1992)

	Control (n = 16)		Treatment (n = 18)					
Week	*M*	*SD*	*M*	*SD*	*t*	*F*	r^2	*d*
16	106	18	141	34	3.68	13.55	.29	1.94
20	186	46	265	67	3.96	15.66	.33	1.71
24	270	58	410	87	5.45	29.66	.47	2.41

Note. Volumes are expressed in cubic centimeters. *d* represents the mean difference divided by the control group's standard deviation. The use of pooled standard deviation values yields somewhat smaller *d* values, but they will nevertheless exceed conventional benchmarks for "large" effects.

Table 4
Predictive Validity of Preemployment Drug Use Reported by McDaniel (1988)

Drug	*N*	Validity	r^2	*t*	*F*
Marijuana	9,355	.07	.004	6.79	46.06
Cocaine	9,224	.04	.001	3.84	14.78
Stimulants	9,286	.07	.004	6.76	45.72
Depressants	9,267	.07	.004	6.76	45.67

variables could be treated as virtually independent. Tests of the nil hypothesis might lead researchers to believe that these correlations are meaningful or important (this is what the term *significant* implies to many readers). Tests of a minimum-effect hypothesis are less likely to be misleading. If the effect is truly negligible, test statistics will not routinely reach statistical significance, no matter how large the sample.

As with all other null hypothesis testing procedures, the necessity of a formal statistical test and the amount of information provided by that test tend to decrease as samples get larger. If the sample is large enough, observed effect size estimates will be precise, and even if the correlation between a test and a criterion is only marginally larger that one's operational definition of a negligible effect (e.g., one decides that correlations of .01 or smaller are negligibly small and finds a correlation of .0100000001), one will reject the null hypothesis. However, unlike tests of the nil hypothesis, where the use of large samples lead one to reject the nil hypothesis regardless of the research question being asked, in minimum-effect testing the use of large samples does not by itself guarantee the outcome of a significance test. If the population effect is truly negligible, large samples increase the likelihood of reaching the correct conclusion (i.e., failing to reject a null hypothesis that is true). We expand on this point in our discussion of the power of minimum-effect tests.

ANOVA

Wickens and Prevett (1995) examined the effects of different types of aircraft navigation displays (e.g., ego-referenced vs. self-referenced) on performance in a simulated flight task. Effect size measures were not included in their article, but an examination of cell means suggests that some of the "significant" effects reported might be so small as to be negligible. Table 5 shows the outcomes of several tests designed to assess the effects of navigation displays on short-term situation awareness.

With one exception, all the results reported in Table 5 are statistically significant (if aircraft navigation displays had no read effect whatsoever, the probability of finding a response time effect for world-referenced questions as large as that reported here would be .06). However, tests of the hypothesis that the effects of navigation displays account for either 5% or less or 1% or less of the variance in outcomes lead to a somewhat different set of conclusions. Comparing the reported *F* values with the critical values shown in Table 1, it is clear that one cannot reject the hypothesis that the effects of navigation displays on response time are negligible (using 1% of the variance or less as the operational definition of negligible); path effects and accuracy effects are large enough to allow one to reject the hypothesis that the true effects of navigational displays are negligible.

None of the effects reported in Table 5 is particularly strong, and if the operational definition of a negligible effect is more stringent, one might not reject the null hypothesis at all. In particular, one cannot reject the hypothesis that the true effects of treatment accounts for 5% or less of the variance in any of the outcome variables shown in Table 5.

How Minimum-Effect Tests Address Criticisms of Traditional Nil Hypothesis Testing

Earlier, we argued that the use of minimum-effect tests would allow researchers to address many of the criticisms of

Table 5
Effects of Aircraft Navigation Displays on Situation Awareness Measures Reported by Wickens and Prevett (1995)

Dependent variable	*F*[a]	Critical *F* traditional[b]	Critical *F* min-eff (1%)[c]	Critical *F* min-eff (5%)
World referenced				
Accuracy	3.21	2.45	3.13	5.45
Response time	2.29	2.45	3.13	5.45
Path effects	4.50	2.68	3.66	6.89
Ego referenced				
Response time	2.88	2.45	3.13	5.45
Path effects	4.04	2.68	3.66	6.89

Note. min-eff = minimum-effect test.
[a] *df*s = 4, 120 for accuracy and response time; *df*s = 3, 120 for tests of path effects. [b] $\alpha = .05$. [c] Test of the hypothesis that treatments account for at least 1% of the variance in the dependent variable.

nil hypothesis testing. For reviews of arguments against traditional nil hypothesis testing, see Chow (1988), Cohen (1994), Cowles (1989), Meehl (1978), Morrison and Henkel (1970) and Murphy (1990). For reviews of arguments in favor of this approach, see Chow (1996), Cortina and Dunlap (1997), and Hagen (1997). Here, we outline the arguments that support this conclusion. To understand our arguments, it is important to consider two important weaknesses of nil hypothesis testing. First, tests of the nil hypothesis are in many senses a trivial exercise because the researcher often already knows the true state of H_0. Second, these tests can end up indicating more about the sensitivity of the study design than about the substantive phenomenon being studied.

Tests of the Traditional Null Are Trivial

In serious tests of the nil hypothesis, the hypothesis that treatments have no effect whatsoever is virtually always wrong, in part because the hypothesis being tested is impossibly precise. In principle it may not be possible to demonstrate that H_0 is in some circumstance correct because the precision of this hypothesis outstrips the potential precision of any plausible method of testing it. This problem is shard by all point hypotheses (e.g., the hypothesis that treatments will have an effect of precisely 1 *SD* is also wrong in a formal sense) and is in no way peculiar to tests of the nil hypothesis. Nevertheless, because the nil is a point hypothesis, it is unlikely that it is true (empirical reviews suggest that treatments in various areas of research rarely have no effect whatsoever; see Haase, Waechter, & Solomon, 1982; Hedges, 1987; Hunter & Hirsh, 1987; Lipsey, 1990; Lipsey & Wilson, 1993; Schmitt, Gooding, Noe, & Kirsch, 1984), and the results of a nil hypothesis test should rarely lead researchers to change their mind about the possibility that treatments have absolutely no effect.

Nil Hypothesis Tests Are Sometimes an Indirect Assessment of Study Sensitivity

Studies of the statistical power of tests of the nil hypothesis (e.g., Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990; Murphy & Myors, 1998) consistently show two things. First, if the sample is small enough, the probability of rejecting the nil hypothesis will be small regardless of the substantive question being investigated. Second, if the sample is large enough, the probability of rejecting the nil hypothesis is virtually 1.00, again regardless of the substantive question being investigated. In other words, it hardly matters what the question is; with a small sample size, researchers will not reject H_0, and with a large sample size, they will. It is hard to escape the conclusion that many tests of the nil hypothesis are little more than indirect counts of the sample size.

There are a number of ways that researchers might increase the sensitivity of their studies besides collecting a large sample (e.g., repeated measures designs), and there is some range of sample sizes in which nil hypothesis tests might be informative. Nevertheless, if it is critical reject H_0, researchers can always accomplish this goal simply by collecting more data. The fact that the outcomes of tests of H_0 can be determined by sample size, independent of the substantive question being asked (i.e., even the most poorly conceived treatment will have an effect that can be reliably detected if the study achieves a sufficient level of sensitivity), suggests that these tests can sometimes tell researchers how many participants showed up, not whether the treatments or interventions have a meaningful effect.

How Minimum-Effect Tests Address These Weaknesses

Tests of the hypothesis that the effects of treatments are so small that they should be described as negligible do not share the flaws of tests of the nil hypothesis outlined earlier. The probability that the nil hypothesis is true is usually so small that tests of this hypothesis are essentially pointless. The same is not true for tests of the hypothesis that treatment effects are negligible. There are treatments that do have negligible effects, and a hypothesis testing procedure that allows researchers to reliably differentiate these treatments from treatments that have some meaningful effect is useful indeed. In particular, if a range of values can be defined that represents a consensus definition of a negligible effect, testing and rejecting the hypothesis that the effects of treatments are at best negligibly small leaves researchers with the alternative hypothesis that the effects of treatments are not negligible but are large enough in magnitude to be described as meaningful effects.

Unlike nil hypothesis tests, researchers cannot be certain of rejecting a minimum-effect null hypothesis simply because the sample is large. The correlations shown in Table 4 were all statistically significant, in the sense that they were different from zero, but if researchers test the hypothesis that these correlations are at best negligible in magnitude (e.g., $r^2 = .01$ or less), they will not reject this hypothesis. No matter how large the sample, we would not make the mistake of labeling these correlations "significant" if we used a minimum-effect test rather than a test of the traditional nil hypothesis. As noted above, the outcome of a nil hypothesis test does not necessarily depend on the substantive question being studied; with a large enough sample size, researchers should always reject the nil hypothesis. The same is not true for tests of minimum-effect hypotheses.

Statistical Power of Minimum-Effect Tests

As statistical tests become more sensitive (usually as the result of increasing the sample size), the power of all tests of the nil hypothesis approaches 1.00 (Cohen, 1988). In tests of minimum-effect hypotheses, the probability that researchers will reject H_0 does not always approach 1.00 as the sample size increases. In particular, if the true effects of treatments or interventions are indeed trivial, the likelihood of rejecting a minimum-effect hypothesis does not increase as the sample size increases. Rather, it (correctly) decreases (Murphy & Myors, 1998). For example, if researchers are testing the hypothesis that treatments account for 5% or less of the variance in outcome, power curves become asymptotic at 1.00 only when treatments do account for more than 5% of the variance in the population. If the true effect of treatments falls below the minimum threshold for defining a negligible effect, the likelihood of rejecting a minimum-effect hypothesis decreases as samples become larger.

It is easier to obtain high levels of power for tests of the nil hypothesis than for minimum-effect tests, at least in part because the nil hypothesis is so easily rejected. Power is always lower for tests of minimum-effect hypotheses than for parallel tests of the nil hypothesis, and power decreases as the hypothesis being tested becomes less restrictive (e.g., for any fixed $v1$ and $v2$ values, power is lower for tests of the hypothesis that treatments account for 5% or less of the variance than for tests of the hypothesis that they account for 1% or less of the variance). One way of illustrating the differences in statistical power for tests of the traditional null versus tests of a minimum-effect hypothesis is to compare the degrees of freedom ($v2$, which is a function of sample size) needed to achieve a particular level of power with each test.

If the true effect of treatments is small (e.g., treatments explain 2% of the variance in outcomes), substantially larger samples are needed to achieve power of .80 in tests of the 1% minimum-effect hypothesis than for tests of the nil hypothesis. If $v1 = 1$, a $v2$ value of 3,225 is needed to achieve power of .80 for this minimum-effect test. In contrast, the $v2$ value needed to achieve this level of power in tests of the nil hypothesis, given the same true effect (i.e., PV = .02), is 385. That is, it takes a large sample to reliably discriminate effects that are close to negligible from those that meet the operational definition of negligible effects, whereas samples do not need to be nearly as large to reliably reject the nil hypothesis. On the other hand, if the true effects of treatments are larger (e.g., 10% of the variance in the population is explained by treatments), the differences in sample sizes needed to achieve power of .80 are considerably smaller (for $v1 = 1$, the $v2$ values needed for power of .80 are 73 and 117 for the nil and the 1% minimum-effect hypothesis, respectively). Murphy and Myors (1998) provided extensive tables for assessing the power of nil hypothesis tests and minimum-effect tests in most general linear model applications.

Tests of minimum-effect hypotheses, although more demanding than tests of the nil hypothesis, are also more informative. If researchers reject the hypothesis that treatment effects fall somewhere in the interval from nil to their operational definition of a negligible effect, they are left with the alternative hypothesis that the effects of treatments are large enough that they cannot be ignored. This might not sound like much, but when it is compared with the alternative hypothesis implied by traditional nil hypothesis tests (i.e., that treatments have some effect, although perhaps only a small one), it is clearly an advance. Nil hypothesis tests merely allow researchers to reject a hypothesis they already know to be false (or at least unlikely), that is, treatments have no effect whatsoever. Tests of minimum-effect hypotheses indicate something that is worth knowing (i.e., whether researchers can state with confidence that the effects of treatments are large enough to be meaningful). These tests require researchers to make an explicit decision about what sorts of effects are likely to be meaningful and what sorts are not, and this decision is not always a simple one. However, researchers interested in the application of their work must often make just this decision, and once a minimum threshold has been set, it is easy to structure null hypothesis testing procedures around this threshold.

Limitations of Minimum-Effect Tests

As we noted at the beginning of this article, many of the statistical procedures used in the social sciences can be thought of as variations of the same general linear model, giving the tests described in this article a wide degree of applicability. However, these methods cannot be applied indiscriminately. First, there is a wide range of modern statistical methods (e.g., structural modeling, confirmatory factor analysis) that makes little use of the F test or of null hypothesis testing of any kind or that rely on trimmed distributions or transformations that make the use of the F test inappropriate. Second, several prominent researchers (e.g., Schmidt, 1992, 1996) have argued that reliance on statistical hypothesis tests has impeded progress in the social and behavioral sciences, and it is not clear whether the methods described here would satisfy their basic criticism. Third, the choice of research designs can lead to nontrivial violations of important assumptions about one's data that affect the structure and accuracy of null hypothesis tests. In particular, research designs that involve obtaining multiple measures from each respondent can lead to violations of assumptions of independence of compound symmetry. Murphy and Myors (1998) discussed methods for dealing with such violations, but the corrections they described are at best approximations.

Another limitation to minimum-effect tests is one shared

with nil hypothesis tests (i.e., they do not indicate enough about the size of the effect). At the beginning of this article, we noted that minimum-effect tests can be a useful supplement to effect size measures, in that they indicate whether such estimates could plausibly come from populations in which treatment effects are trivially small. However, they do not directly indicate whether sample-based effect size estimates are accurate. Procedures described by Fleishman (1980) can be applied to address this problem, but only in a limited range of applications.

Serlin and Lapsley's (1985) approach is similar in many ways to the approach described here, with one critical difference. Their approach is designed to test the hypothesis that a particular parameter (e.g., the difference between two means) is "close enough" to some theoretically specified value to allow researchers to conclude that the theory is right. Our approach tests a traditional null hypothesis (e.g., that the difference between two means is less than or equal to some constant, which would be zero in tests of the nil hypothesis and greater than zero in minimum-effect tests). A key to developing cumulative research knowledge is to obtain and combine the best possible estimates of important parameters (e.g., the correlation between general cognitive ability and success in high school), and Serlin and Lapsley's (1985) approach has the advantage of being directly focused on researchers' ability to make good estimates of such parameters.

Conclusions

Tests of minimum-effect hypotheses may be less familiar than the nil hypothesis test, but virtually everything researchers know about hypothesis testing applies to tests of this sort. In fact, much of what researchers already know about statistical hypothesis testing applies better to tests of minimum-effect hypotheses than to tests of the traditional nil hypothesis. For example, the literature on nil hypothesis testing includes numerous articles dealing with methods of controlling Type I errors (see Zwick, 1993, for a review of several such procedures). If the hypothesis is that treatments have no effect whatsoever, concern over Type I errors is probably misplaced. Type I errors can occur only when H_0 is true, and in tests of the nil hypothesis, this is rarely if ever the case. Minimum-effect null hypotheses might very well be true (i.e., the effect of a particular treatment might very well be negligible), and researchers can bring the existing literature on Type I errors to bear in evaluating the seriousness of this threat and the most appropriate response to the possibility that Type I errors will occur in tests of minimum-effect hypotheses.

The method described here allows researchers to reorient their approach to hypothesis testing without substantially changing the actual procedures used to conduct statistical tests. The only difference between this approach and the traditional approach to null hypothesis testing is that a different set of tables or critical values is used in evaluating the F statistic. Alternatively, users can compute their own critical F values for minimum-effect hypothesis testing using a noncentral F distribution calculator such as the one found in SPSS. Instead of comparing the F researchers obtain in a study with the critical F value found in most statistics textbooks, this approach compares the obtained F with the distribution of F values researchers would expect if the true effects of treatments were at best negligible.

The biggest single obstacle to adopting the approach described here is that consensus must be reached about working definitions for negligible effects. That is, some judgment must be exercised in deciding what hypothesis to actually test. Traditional null hypothesis tests appear to have an advantage in that they are objective: No decisions need be made about what hypothesis should be tested. However, that objectivity is purchased at the price of pursuing an essentially trivial and pointless exercise of testing a hypothesis researchers already know to be false. In our view, the "subjectivity" in calling some effects negligible is a small price to pay for developing a framework for statistical hypothesis testing that is informative and useful.

The tests described here should be especially useful for researchers faced with the problem of determining whether treatments have effects that are large enough to pay attention to. Nil hypothesis tests do not address this question, even when significance tests are accompanied by effect size estimates. For example, suppose that a researcher decided that treatments accounting for less than 1% of the variance in school performance were not worth adopting. He or she tests the nil hypothesis in a study comparing treatments and control conditions and rejects it ($p < .05$), and the sample-based estimate of the PV accounted for by the treatment is .03. This finding does not necessarily tell the researcher whether he or she can have any confidence that the variance accounted for in the population is greater than .01. It is entirely possible that the treatment has some effect (which is what the traditional null hypothesis test is designed to indicate) but that the population effect is smaller than the effect observed in the sample. The minimum-effect tests described in this article would directly answer the question of whether one can be confident that a treatment that appears to be worthwhile in a sample is likely to exceed the threshold for defining negligible effects in the population.

References

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin, 103,* 105–110.

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility.* Thousand Oaks, CA: Sage.

Clapp, J. F., & Rizk, K. H. (1992). Effect of recreational exercise on midtrimester placental growth. *American Journal of Obstetrics and Gynecology, 167,* 1518–1521.

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

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods, 2,* 161–173.

Cowles, M. (1989). *Statistics in psychology: An historical perspective.* Hillsdale, NJ: Erlbaum.

Fleishman, A. I. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement, 40,* 659–670.

Fowler, R. L. (1985). Testing for substantive significance in applied research by specifying nonzero null hypotheses. *Journal of Applied Psychology, 70,* 215–218.

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods, 1,* 379–390.

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology, 29,* 58–65.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Hays, W. L. (1994). *Statistics* (5th ed.). New York: Harcourt Brace.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? *American Psychologist, 42,* 443–455.

Horton, R. L. (1978). *The general linear model: Data analysis in the social and behavioral sciences.* New York: McGraw-Hill.

Hunter, J. E., & Hirsh, H. R. (1987). Applications of meta-analysis. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 321–357). New York: Wiley.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96,* 72–98.

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

Landy, F. J., Farr, J. L., & Jacobs, R. R. (1982). Utility concepts in performance measurement. *Organizational Behavior and Human Performance, 30,* 15–40.

Lipsey, M. W. (1990). *Design sensitivity.* Newbury Park, CA: Sage.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist, 48,* 1181–1209.

McDaniel, M. A. (1988). Does pre-employment drug use predict on-the-job suitability? *Personnel Psychology, 41,* 717–729.

Meehl, P. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy: A reader.* Chicago: Aldine.

Murphy, K. R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist, 45,* 403–404.

Murphy, K. R., & Myors, B. (1998). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests.* Mahwah, NJ: Erlbaum.

Narula, S., & Weistroffer, H. (1986). Computation of probability and noncentrality parameter of a noncentral F distribution. *Communications in Statistics B, 15,* 871–878.

Patnaik, P. B. (1949). The non-central χ^2 and F-distributions and their applications. *Biometrika, 36,* 202–232.

Rosenthal, R. (1993). Cumulating evidence. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 519–559). Hillsdale, NJ: Erlbaum.

Rouanet, H. (1996). Bayesian methods for assessing the importance of effects. *Psychological Bulletin, 119,* 149–158.

Schmidt, F. L. (1992). What do the data really mean? Research findings, meta-analysis and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., Hunter, J. E., McKenzie, R. C., & Muldrow, T. W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 71,* 432–439.

Schmidt, F. L., Mack, M. J., & Hunter, J. E. (1984). Selection utility in the occupation of U.S. park ranger for three modes of test use. *Journal of Applied Psychology, 69,* 490–497.

Schmitt, N., Gooding, R. Z., Noe, R. D., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407–422.

Serlin, R. A., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist, 40,* 73–83.

Serlin, R. A., & Lapsley, D. K. (1993). Rational appraisal of psychological research: The good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Hillsdale, NJ: Erlbaum.

Tatsuoka, M. (1993). Elements of the general linear model. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 3–42). Hillsdale, NJ: Erlbaum.

Thompson, B. (1997, August). *If statistical significance tests are broken/misused, what practices should supplement or replace them?* Paper presented at the 105th Annual Convention of the American Psychological Association, Chicago.

Thompson, B. (in press). Review of "What if there were no significance tests." *Educational and Psychological Measurement.*

Tiku, M. L., & Yip, D. Y. N. (1978). A four-moment approximation based on the F distribution. *Australian Journal of Statistics, 20,* 257–261.

Wickens, C. D., & Prevett, T. T. (1995). Exploring the dimensions of egocentricity in aircraft navigation displays. *Journal of Experimental Psychology: Applied, 1,* 110–135.

Zwick, R. (1993). Pairwise comparison procedures for one-way analysis of variance designs. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical issues* (pp. 43–72). Hillsdale, NJ: Erlbaum.

(Appendix follows)

Appendix

Estimating the Noncentral *F* Distribution

The noncentral *F* is hard to tabulate because it depends on three separate parameters, but there is a reasonably simple approximation based on the central *F* distribution (Horton, 1978; Patnaik, 1949; see also Tiku & Yip, 1978). If one computes

$$g=\frac{(\nu 1+\lambda)^2}{(\nu 1+2\lambda)} \quad \text{(A1)}$$

and

$$k=\frac{(\nu 1+\lambda)}{\nu 1}, \quad \text{(A2)}$$

then

$$Pr(F_{\nu 1,\nu 2,\lambda}>c\cdot k) \text{ is approximately equal to } \Pr(F_{g,\nu 2}>c), \quad \text{(A3)}$$

where *Pr* is the probability, $F_{\nu 1,\nu 2,\lambda}$ is a value from the noncentral *F* distribution, $F_{g,\nu 2}$ is a value from the central *F* distribution, and *c* is any constant. To apply this approximation to the problem of finding the critical value of the noncentral *F* (given α, $\nu 1$, $\nu 2$ and λ), one can simply find the critical value for the central *F* (with degrees of freedom *g* and $\nu 2$) and multiply that value by *k*.

For example, suppose that the working definition of a negligibly small effect is one in which treatments account for 4% of the variance or less (i.e., percentage of variance [PV] = .04), and one collects data in a design in which $\nu 1 = 2$ and $\nu 2 = 120$ (e.g., comparing the means of three treatments, with $N = 124$). As we noted earlier, λ can be estimated on the basis of PV and $\nu 2$, that is, estimated $\lambda = \nu 2 \cdot \text{PV}/(1 - \text{PV})$, which for PV = .04 and $\nu 2 = 120$ yields estimated $\lambda = 5.0$, and this in turn allows one to estimate the appropriate noncentral *F* value. Inserting these $\nu 1$, $\nu 2$, and estimated λ values into Equations A1 and A2 yields $g = 4.08$, $k = 3.5$. The value of the central $F(4, 120)$ needed to reject the null hypothesis ($\alpha = .05$) is 2.45. If one multiplies this value by 3.5 (i.e., *k*), one finds that the approximate *F* value needed to reject the minimum-effect hypothesis that treatments account for 4% or less of the variance in outcomes ($\alpha = .05$) is 8.57. As one should expect, this value falls between the values shown in Tables 1 and 2 of the critical *F* needed to reject a minimum-effect hypothesis when $\nu 1 = 2$ and $\nu 2 = 120$ and when PV = .01 ($F = 4.74$) or PV = .05 ($F = 9.64$).

Another alternative is to use a computer program specifically designed for computing probabilities in the noncentral *F* distribution (e.g., Narula & Weistroffer, 1986). Finally, it is possible to obtain estimates of the appropriate noncentral *F* using functions that are built into a number of widely used computer programs. As noted in the text, SPSS contains a function that can be used for this purpose. The SPSS statement

$$\text{COMPUTE } pF = \text{NCDF.}F[F,\nu 1,\nu 2,\text{PV}\cdot \nu 2/(1-\text{PV})]$$

computes a variable *pF* that gives the probability of *F*, given the noncentral *F* distribution associated with the $\nu 1$, $\nu 2$, and PV values included in that statement. Note that the term $[\text{PV}\cdot \nu 2/(1 - \text{PV})]$ is used to approximate the value of λ. Similar functions can be found in programs such as Excel.

Received June 16, 1997
Revision received July 6, 1998
Accepted July 17, 1998 ■