Psychological Methods
2004, Vol. 9, No. 2, 147–163

1082-989X/04/$12.00 DOI: 10.1037/1082-989X.9.2.147

The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies

Scott E. Maxwell
University of Notre Dame

Underpowered studies persist in the psychological literature. This article examines reasons for their persistence and the effects on efforts to create a cumulative science. The "curse of multiplicities" plays a central role in the presentation. Most psychologists realize that testing multiple hypotheses in a single study affects the Type I error rate, but corresponding implications for power have largely been ignored. The presence of multiple hypothesis tests leads to 3 different conceptualizations of power. Implications of these 3 conceptualizations are discussed from the perspective of the individual researcher and from the perspective of developing a coherent literature. Supplementing significance tests with effect size measures and confidence intervals is shown to address some but not necessarily all problems associated with multiple testing.

The primary purpose of this article is to examine the importance of statistical power for the formulation of a coherent body of scientific literature. The article addresses this goal through consideration of four interrelated subtopics: (a) why underpowered studies persist, (b) the undesirable consequences of underpowered studies, (c) the extent to which effect size measures and confidence intervals successfully address problems associated with multiple testing, and (d) ways in which designing more powerful studies as well as other possible approaches can contribute toward developing a cumulative science of psychology.

The first section of the article proposes that underpowered studies persist in part because most studies involve tests of multiple hypotheses. The article shows that it is entirely possible that the power to test any specific hypothesis is very low by conventional standards while at the same time the power to detect at least one effect may be quite large. It is well-known that the power to detect an effect often depends on the context in which an effect is to be tested. For example, Cohen's f^2 effect size measure in a regression analysis can be computed for a single predictor, for a subset of predictors, or for all predictors simultaneously. Similarly, in structural equation modeling, Satorra and Saris (1985) presented methods for computing the power associated with a specific parameter, whereas MacCallum, Browne, and Sugawara (1996) presented methods for computing the power associated with the overall test of a model. Thus, the idea that statistical power depends on the specific test of interest is generally well-known. However, the perspective taken in the present article is to consider the ramifications of performing multiple tests in a single study, where each of these tests has its own level of statistical power. The issues to be considered here are relevant whether the individual tests are focused tests such as those associated with a single predictor variable or are global tests of an overall model. The key point is to consider the mere fact that many studies involve multiple tests of one type or another.

The remaining sections of the article consider additional aspects of multiple statistical tests. The second section shows that when power is low for any specific hypothesis but high for the collection of tests, researchers will usually be able to obtain statistically significant results, but which specific effects are statistically significant will tend to vary greatly from one sample to another, producing a pattern of apparent contradictions in the published literature. The third section examines the potential benefits as well as possible limitations of effect size measures and confidence intervals. The fourth section briefly describes various alternative strategies for developing a cumulative science of psychology in the face of challenges presented by low statistical power.

Editor's Note. Patrick J. Curran served as action editor for this article.—SGW

I thank David A. Cole, George S. Howard, and David A. Smith for their helpful comments based on an earlier version of this article.

Correspondence concerning this article should be addressed to Scott E. Maxwell, Department of Psychology, University of Notre Dame, Notre Dame, IN 46556. E-mail: smaxwell@nd.edu

The Persistence of Underpowered Studies

The past 15 years have seen increasing attention given to the role of statistical power in psychological research. Not only have numerous journal articles and book chapters been

written on the topic of statistical power (e.g., Allison, Allison, Faith, Paultre, & Pi-Sunyer, 1997; Cohen, 1992; Green, 1991; Kraemer, 1991; McClelland, 1997; O'Brien & Muller, 1993) but in addition several entire books devoted to the topic have been published (e.g., Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990; Murphy & Myors, 1998) and statistical packages devoted solely to power analysis and sample size determination have emerged (Borenstein, Rothstein, & Cohen, 1997; Elashoff, 1999; Hintze, 1996; O'Brien, 1998; Thomas & Krebs, 1997). Classic reviews of published literature in psychology (Rossi, 1990; Sedlmeier & Gigerenzer, 1989) have documented the fact that underpowered studies tended to dominate the literature in the late 1980s. Despite ever-increasing attention, more recent reviews of published literature in psychology have largely continued to show that the majority of these studies lack appropriate statistical power. For example, Bezeau and Graves (2001), Clark-Carter (1997), Kosciulek and Szymanski (1993), and Mone, Mueller, and Mauland (1996) have shown that the power to detect a medium effect remains very close to the inadequate level of power originally detected by Cohen (1962) 40 years ago in such diverse areas as clinical neuropsychology, articles published in *British Journal of Psychology*, rehabilitation counseling research, and management.

An exception to this continuing trend was provided by Maddock and Rossi (2001), who showed that research in three health-related journals generally has adequate power to detect not only large but also medium effects. This finding could reflect the fact that new methodological perspectives simply take time to influence researchers' behavior. Thus, Maddock and Rossi's findings may continue to be reflected in future literature reviews. Another possibility is that research in the health-related journals tends to be federally funded, and federal funding agencies may be likely to require evidence of sufficient statistical power before deciding to fund a proposal. There are undoubtedly multiple reasons why some literatures may continue to show lack of progress in the design of studies with adequate power whereas other areas, such as health-related research, may have begun to show such progress. The primary purpose of the current article is to examine why underpowered studies may continue to persist in some areas of psychology despite real pressures to obtain statistical significance in order to maximize one's opportunity of publishability.

The paradox of increased attention combined with continuing evidence that underpowered studies persist leads to a natural question of why researchers have often not followed methodologists' recommendations to design studies with sufficient power. This paradox becomes even more puzzling in light of typical editorial practices virtually requiring statistical significance as a prerequisite for publication. Although this practice is questionable on many grounds, nevertheless it might be expected that it would at least lead to studies with sufficient power. Especially over time, researchers would presumably learn that it was necessary to design studies with sufficient power or else publication would be unlikely. Researchers who failed to recognize these contingencies would all too often find themselves in a position in which their results were not publishable. It is very likely that even beginning researchers are aware of these contingencies. Why, then, do underpowered studies persist?

Cohen (1992) expressed puzzlement over this state of affairs 30 years after the appearance of his initial article: "It is not at all clear why researchers continue to ignore power analysis" (p. 155). There are undoubtedly many explanations for the persistence of underpowered studies. The motivation of Cohen's (1992) article was a belief by an associate editor of *Psychological Bulletin* that researchers find power analysis to be too complicated. Cohen (1992) himself speculated that "at least part of the reason may be the low level of consciousness about effect size" (p. 155). This article presents another possible explanation for why researchers continue to design studies lacking in power despite apparent countervailing publication pressures. This reason springs from an entirely different perspective, namely, that most studies involve tests of multiple hypotheses, creating a gap between the power for any single test and the power for the collection of tests. Without my claiming that this is the only reason psychologists continue to design underpowered studies, consideration of this issue suggests novel implications for the consequences of underpowered studies as well as for developing a cumulative psychological science.

Reviews of psychological literature have shown that studies tend to be underpowered in the sense that tests of any specific hypothesis tend to lack adequate power. On the surface, it might seem to follow that the probability of obtaining a statistically significant result in the study would be low, thus jeopardizing the opportunity for publication. However, it is entirely possible that the power of any specific test might be low and yet the probability of obtaining a statistically significant result somewhere in the study could be substantial. The explanation of this apparent contradiction is that most studies involve tests of multiple hypotheses. As a result, the probability of rejecting at least one hypothesis in the collection of tests will clearly exceed the probability that any specific hypothesis is rejected.

For example, Cohen's (1962) original survey was based on 70 studies. Yet embedded in these 70 studies were 4,820 statistical tests, or an average of nearly 70 tests per study. Despite there being generally low power for any single test, Cohen (1962) noted that "with few exceptions, the 70 studies *did* have significant results" (p. 151). Cohen (1962) then went on to say

> This may suggest that perhaps the definitions of size of effect were too severe, or perhaps, accepting the definitions, one might seek to conclude that the investigators were operating under circumstances wherein the effects were actually large, hence their success. (p. 151)

However, most likely of all would seem to be the fact that these studies typically contained so many statistical tests that an appreciable number would be statistically significant even if the power of any single test was inadequate. Even if all effects being tested were small, Cohen's (1962) estimated power of .18 would imply more than 12 statistically significant results per study on average.[1]

Wilkinson and the Task Force on Statistical Inference (1999) went so far as to state the following: "Multiplicities are the curse of the social sciences. In many areas of psychology, we cannot do research on important problems without encountering multiplicity" (p. 599). By *multiplicity*, they simply meant the fact that many studies involve multiple hypotheses and thus multiple hypothesis tests. Any time multiple tests are conducted, one can distinguish between error rates associated with a single test and error rates associated with a collection of tests.

Most experimental design books have discussed multiple comparisons in the context of the difference between the per-comparison alpha rate and the familywise (or experimentwise) alpha rate. Any time multiple tests are conducted, these two alpha rates will differ from one another. Typically, attention focuses on the implications of these multiple tests for the possible inflation in error rate. However, much less attention has been devoted to implications of multiplicity for power. It is important to realize that multiplicity has implications for power regardless of how one chooses to deal with its implications for the Type I error rate. Even though a few authors (e.g., Cohen, 1994) have pointed out that invoking such procedures as the Bonferroni adjustment will lower power, there has been much less awareness of the broader implications of multiplicity for power.

One partial exception to this lack of awareness has been in the analysis of variance (ANOVA) literature. Specifically, in the ANOVA tradition of pairwise comparisons, a distinction is sometimes made between (a) the power for a specific comparison, (b) any-pairs power, and (c) all-pairs power. The *power for a specific comparison* is simply the probability that this specific comparison will be declared statistically significant. *Any-pairs power* is the probability that at least one pairwise comparison will be declared statistically significant. Finally, *all-pairs power* is the probability that all pairs that are truly different from one another will be declared statistically significant. As an illustration of the distinction among these three definitions, consider a three-group design in which we suppose that all three population means differ from one another. An example of the power for a specific comparison would be the probability that the test comparing Groups 1 and 2 is statistically significant. In contrast, any-pairs power would be the probability that at least one of the three pairwise comparisons (i.e., between Groups 1 and 2, 1 and 3, or 2 and 3) is statistically significant. Finally, all-pairs power would be the probability that all three pairwise comparisons are statistically significant.

The distinction among these definitions of power has received much less attention in ANOVA than the similar distinction among Type I error rates. Outside the ANOVA framework, the distinction has received even less attention. However, the distinction applies any time multiple hypotheses are tested in any study, as long as we realize that the specific use of "pairs" may no longer be applicable. To the extent that most psychological studies test multiple hypotheses, this distinction is relevant for much of the psychological literature. Furthermore, the distinction may have important implications for considering the role of power in psychological research.

For example, consider the case of a 2×2 between-subjects factorial design. Data analysis in this design will typically consist of three omnibus tests, namely, the row main effect, the column main effect, and the row by column interaction. The probability that at least one of these tests will be statistically significant is different from the probability that any specific test will be statistically significant. Yet different from both of these probabilities is the probability that all three effects will be statistically significant. How different these three probabilities are from one another will depend on the effect size of each effect, sample size, and whether cell sizes are equal or unequal.

To illustrate the differences in the three conceptualizations of power, suppose that each main effect, as well as the interaction, is nonzero in the population. Specifically, suppose that each effect corresponds to Cohen's (1988) definition of a *medium effect size*. In addition, suppose that cell sizes are equal to one another and that each test is conducted with an alpha level of .05. Table 1 shows values of the three types of power for cell sizes ranging from 10 to 40 (notice that the corresponding total sample sizes range from 40 to 160).[2] To understand the meaning of the probabilities

[1] It does not necessarily follow that the typical study would have more than 12 statistically significant results, because the number of statistical tests was undoubtedly not uniformly distributed over studies. However, the basic point here is that Cohen's (1962) finding that most studies contained statistically significant results should come as no surprise given the number of tests performed.

[2] The probabilities shown in Table 1 were derived by first calculating the power of any specific effect and then using the binomial distribution to find the probabilities of at least one success as well as of three successes in three trials. It should be noted that the binomial distribution assumes that trials are independent of

Table 1
Values of Three Types of Statistical Power in a 2 × 2 Design as a Function of Cell Size

Type of power	Cell size			
	$n = 10$	$n = 20$	$n = 30$	$n = 40$
Any single prespecified effect	.35	.59	.79	.88
At least one effect	.71	.93	.99	>.99
All effects	.04	.21	.47	.69

Note. Both main effects as well as the interaction are presumed to have medium effect sizes. Power values are calculated using a binomial approximation.

shown in the table, first consider the column for a cell size of 10. Historically, a cell size of 10 has often been recommended as a rule of thumb for determining sample size in factorial designs. How well does that rule perform when all three effects are medium? The top row of the table shows that the probability that any specific effect is statistically significant with 10 participants per cell is only .35, much below any recommended level of statistical power. Why then, was this rule not extinguished long ago as researchers learned that it failed to provide adequate power? Of course, one answer could be that most researchers study only large effects, but this "Lake Wobegon" explanation that most effects are larger than medium flies in the face of Cohen's (1988) review of the psychological literature. Instead, the second row of Table 1 provides a more likely explanation. Although the power for any specific effect is inadequate, the power to detect at least one effect is a reasonably respectable .71.

Thus, a researcher who designs a 2 × 2 study with 10 participants per cell has a 71% chance of obtaining at least one statistically significant result if the three effects he or she tests all reflect medium effect sizes. Of course, in reality, some effects will often be smaller and others will be larger, but the general point here is that the probability of being able to find something statistically significant and thus potentially publishable may be adequate while at the same time the probability associated with any specific test may be much lower. Thus, from the perspective of a researcher who aspires to obtain at least one statistically significant result, 10 participants per cell may be sufficient, despite the fact that a methodological evaluation would declare the study to be underpowered because the power for any single hypothesis is only .35.

A later section of the article returns to this distinction and its implications, but for the moment, it is important to realize that there is yet a third type of power still to be considered. Still looking at the $n = 10$ column of Table 1, the third row shows that the probability that all three effects are statistically significant is only .04. Thus, in this scenario, only very rarely will the row main effect, the column main effect, and the interaction all be statistically significant. Notice that ideally all three of these effects should be declared nonzero, because all three are truly nonzero in the population. However, when all three effects exhibit medium effect sizes and there are 10 participants per cell, the probability that all three effects are detected is actually less than the alpha level of .05 established for each test. Stated another way, the probability of making at least one Type II error in this scenario is a shocking .96.

Comparing the three probabilities of .35, .71, and .04 makes it all too clear that whether 10 per cell is an adequate sample size depends greatly on how we conceptualize power. Even if we accept .80 as a standard for desired power, how large our cell size needs to be in a 2 × 2 design depends strongly on whether this .80 value applies to the power of a specific test, the power that we obtain at least one statistically significant result somewhere in the study, or the power that all nonzero effects are detected.

What sample size is needed for a 2 × 2 factorial design? Not only does the answer depend on anticipated or minimally important effect sizes, but Table 1 shows that the answer also depends on which conceptualization of power is deemed most relevant. For example, suppose that all three effects are expected to be medium or that this is the magnitude of effect deemed to be important. Further suppose that the desired level of power is chosen to be .80. How large should the sample be? Table 1 shows that if power is conceptualized in terms of a specific test, a cell size of approximately 30 is appropriate. However, if power is conceptualized in terms of the probability of obtaining at least one statistically significant result, a cell size only about half this large is likely to be sufficient. On the other hand, the probability of detecting all nonzero effects even with $n = 30$ is below .50 in this scenario. In order to reach a power of .80 to detect all nonzero effects here, a cell size of approximately 48 is required, a 60% increase over the number needed for any specific test to have a power of .80 and approximately 3 times as many research participants as required to have a power of .80 to detect at least one effect.

To what extent do the discrepancies shown in Table 1 depict a worst case scenario? In one respect, the probabilities do in fact represent a worst case scenario because all three effects were assumed equal to one another. Other patterns of effect size would result in different patterns of probabilities. For example, if only one of the three effects is

one another, which is only approximately true in factorial designs. Even though the effects themselves are independent, their tests are not completely independent because they all use mean square within as a common error term. Thus, the probabilities shown in Table 1 are only approximate. Nevertheless, the important point here is the pattern of probabilities, which would be much the same even if dependencies were taken into account.

truly nonzero, all three conceptualizations of power become equivalent to one another, and as a consequence the second and third rows of Table 1 would become identical to the first row. Thus, in this case the debate over different conceptualizations of power becomes moot.

However, in another respect, the discrepancies shown in the table may not reflect a worst case scenario. For example, in many studies more than three effects will be tested. Another complication is that in many studies the multiple effects will be related to one another, unlike the effects shown in Table 1, which are orthogonal because of the equal cell sizes. The implications of the relation among effects depend on whether the effects are positively correlated or negatively correlated.

To examine the influence of correlations among effects, it is helpful to establish a baseline where effects are independent. For simplicity, suppose that only two effects are to be tested. Further suppose that the power to detect each effect individually is .50. The first column of values in Table 2 shows three probabilities: (a) the power for a specific effect, (b) the power that at least one of the two effects is detected, and (c) the power that both effects are detected.[3] The pattern of values shown here is similar to that seen earlier in Table 1 in that the discrepancies among the three types of power are sizable.

Next consider a case in which the effects are positively related to one another. A typical example of such positive effects often occurs in multivariate ANOVA where multiple dependent measures are frequently positively correlated with one another. In particular, suppose the degree of correlation between variables is such that the conditional probability that either variable is statistically significant is .80 given that the other variable is significant. As in the baseline condition of uncorrelated effects, suppose that the power of each individual effect is .50. The second column of values in Table 2 shows that the discrepancies among the types of power are much less than in the uncorrelated condition. As the effects become more highly correlated with one another, the distinctions among types of power become less. In particular, in the limit where the correlation equals 1.0, multiple tests become equivalent to literally the same test done multiple times.

Table 2
Influence of Correlated Effects on the Three Types of Power

	Type of effects		
Type of power	Orthogonal	Positively correlated	Negatively correlated
Any single prespecified effect	.50	.50	.50
At least one effect	.75	.60	.90
All effects	.25	.40	.10

Note. The power to detect each effect individually is presumed to be .50. Additional power values are calculated as a function of conditional probabilities.

Finally, consider a case in which the effects are negatively related to one another. At first glance, this might seem to be a rare situation. However, in reality, it is likely to be very common. For example, in multiple regression analysis, when predictors are positively correlated with one another, their corresponding regression weights are generally negatively correlated (Rozeboom, 1966, pp. 507–509). Indeed, this is the statistical basis of the multicollinearity problem. Thus, in the simplest case of two positively correlated predictor variables, the weights associated with each predictor will correlate negatively.[4] In particular, suppose that the degree of correlation between the predictors is such that the conditional probability that either variable is statistically significant is .20 given that the other variable is significant. As in the baseline condition of uncorrelated effects, suppose that the power of each individual effect is .50. The third column of values in Table 2 shows that the discrepancies among the types of power are much greater than in the uncorrelated condition. As the effects become more negatively correlated with one another, the distinctions among types of power become larger.

Although the impact of multiple tests has generally received the most attention within the ANOVA tradition of psychological research, Table 2 suggests that the actual impact may be greater in correlational studies using multiple regression (and associated methods such as structural

[3] The probabilities shown in Table 2 can be found by forming a 2×2 contingency table based on marginal and conditional probabilities. The entries in Table 2 then follow directly from the cells of the 2×2 contingency table.

[4] It may seem counterintuitive that positively correlated predictors in multiple regression have negatively correlated regression weights. To understand this phenomenon, consider a simple example in which two parallel measures of the same construct are used to predict some outcome measure. What would one expect to happen across replications of such a study? When predictors correlate positively with one another (and the sign of the correlation with the outcome variable is the same for both predictors), the sum of their regression weights tends to vary relatively little from one replication to another. For example, if the population value of the sum is .80, sample values of the sum of the weights will also be close to .80. However, how this sum of .80 is split between the predictors may vary greatly. Thus, for example, in one replication we might find that the first predictor receives a weight of .60, in which case the second predictor will tend to receive a weight of around .20. In another replication the first predictor might receive a weight of around .30, which means that the second predictor will tend to receive a weight of around .50. The crucial point here is that when the sample estimate for one predictor is high, the sample estimate for the other predictor will tend to be low. However, it is precisely this pattern that implies that the two regression weights will correlate negatively with one another.

equation modeling), where the simultaneous inclusion of positively correlated predictors in a model leads to negatively correlated effects. Thus, in this respect, the discrepancies shown in Table 1 for a 2 × 2 ANOVA design may underestimate discrepancies likely to be found for the three types of power in multiple regression analysis.

To examine possible discrepancies among the three types of power in multiple regression analysis, consider the case of a regression design with five predictor variables. Implicit in Table 2 is the fact that the magnitude of the discrepancies among the types of power will depend at least in part on how highly correlated the predictors are with one another. As an example, suppose that each and every pair of predictors is correlated at a medium level according to Cohen's (1988) definition. In other words, all zero-order correlations between predictors equal .30. Of course, the magnitude of correlations between each predictor and the outcome variable are also crucial in determining each type of power. For example, suppose that each predictor has a medium correlation of .30 with the outcome variable.

Table 3 shows values of the three types of power under this scenario for a variety of sample sizes, assuming that each individual test is conducted with an alpha level of .05.[5] This table shows that a researcher who follows the traditional 10:1 ratio of sample size to number of predictors avails himself or herself of just over a 50% chance of obtaining at least one statistically significant regression coefficient under this scenario. The table shows that increasing the sample to 100 increases the comparable probability to .84. At a sample size of around 100, diminishing returns set in, so doubling the sample size provides only a modest increase in the probability of obtaining a statistically significant result. Thus, from the perspective of hoping to obtain at least one statistically significant result to report, a sample size of 50 is not completely unreasonable and a sample size of 100 is likely to be judged as entirely sufficient.

From one perspective, a sample size of 50 to 100 for a multiple regression analysis with five predictors may be adequate; however, from other perspectives it may be inadequate. Table 3 shows that the probability that any specific predictor is statistically significant in this scenario is only .26 even with 100 participants. To the extent that there is agreement that power should be at least .80, the table shows that at least 400 participants are needed if power refers to the statistical significance of a specific predictor. Thus, from this perspective, sample size in this scenario needs to be much larger than the 50 to 100 value suggested by the first perspective.

Table 3 also shows that even with a sample of 400, the probability that all five predictors are statistically significant under this scenario is only .22. Thus, even with a sample this large, the probability of making at least one Type II error is .78. A researcher who wanted to be careful to avoid any Type II errors would need an enormous sample size in this situation.

Thus, how large a sample needs to be in a factorial ANOVA design or in a multiple regression study depends greatly on how one conceptualizes statistical power. Although the 10:1 rule of thumb may offer reasonable power to find some effect in both cases, it is likely to be woefully inadequate from the perspective of providing sufficient power to detect a specific effect, much less all nonzero effects. This general principle applies any time multiple hypotheses are tested, adding force to the statement that "multiplicities are the curse of the social sciences" (Wilkinson and the Task Force on Statistical Inference, 1999, p. 599).

Table 3
Values of Three Types of Statistical Power in a Multiple Regression Study as a Function of Sample Size

Type of power	Sample size			
	$n = 50$	$n = 100$	$n = 200$	$n = 400$
Any single prespecified effect	.15	.26	.48	.78
At least one effect	.57	.84	.99	>.99
All effects	<.01	<.01	.01	.22

Note. Power values shown were obtained from 10,000 replications of a simulation with five predictor variables. All zero-order bivariate correlations were presumed to be .30 (a medium effect size) in the population.

Consequences of Underpowered Studies

Much has been written over the past 2 decades about the extent to which psychology has succeeded in developing a coherent literature of research findings (e.g., Hedges, 1987; Meehl, 1978). One undisputed point is that apparent differences in findings are at least partly due to mere artifacts. Prominent among these artifacts is the simple existence of sampling error (Schmidt, 1996). The concept of a sampling distribution is arguably the most fundamental idea in inferential statistics, but its ultimate importance for understanding multiple tests in a single research study, as well as results from a collection of studies, may not be obvious.

For example, consider the three regression tables shown in Table 4. Each panel displays standardized regression coefficients (i.e., beta weights) and associated additional

[5] The power values shown in Table 3, as well as those in other examples involving multiple regression, are based on the assumption that the test for each predictor variable pertains to its unique contribution to a model while controlling for all other predictors. Different results would occur for other types of multiple regression, such as hierarchical regression analysis, where statistical significance might be based on the first step in which a variable entered the model.

Table 4
Regression Data From Three Studies

Depression predictor	β	*SE*	*t*	$p(>\|t\|)$
Replication 1				
Academic competence	.2031	.1070	1.8981	.0608
Appearance	.2129	.1041	2.0454	.0436*
Athletic competence	.1616	.1173	1.3780	.1715
Behavioral competence	−.0659	.1158	−0.5690	.5707
Social competence	.2398	.1109	2.1615	.0332*
Replication 2				
Academic competence	.0753	.1095	0.6873	.4936
Appearance	.1216	.1234	0.9856	.3269
Athletic competence	.2153	.1151	1.8713	.0644
Behavioral competence	.0592	.1183	0.5006	.6178
Social competence	.2216	.1019	2.1753	.0321*
Replication 3				
Academic competence	.1969	.0975	2.0196	.0463*
Appearance	.1627	.0991	1.6422	.1039
Athletic competence	.1781	.0991	1.7977	.0754
Behavioral competence	.0309	.1078	0.2864	.7752
Social competence	−.0244	.1066	−0.2293	.8191

* $p < .05$.

statistics obtained using multiple regression analysis to examine the relationship between depression (the dependent variable) and five predictors, each of which represents self-perceived competence in a specific domain. (The variables are scaled in such a way that positive regression coefficients imply that higher levels of perceived competence correspond to lower levels of depression.) How might one interpret this collection of studies? From a vote-counting perspective (Hedges & Olkin, 1980), the primary impression of these results may be confusion. Social competence was statistically significant twice, whereas academic competence and appearance were each significant once. Neither athletic nor behavioral competence reached the .05 level of significance. Faced with such results, one might at least be tempted to conclude that athletic and behavioral domains can safely be ignored, while arguing about the inconsistent results for the other three domains.

In reality, the replications shown in Table 4 are the first three results obtained from a larger simulation. In all three cases, data were generated from a model in which all variables (i.e., predictors as well as the outcome variable) had a medium correlation of .30 with one another. Thus, any differences in the three replications reflect nothing more than sampling error. Furthermore, any differences between the predictor variables, either within a study or across studies, also reflect only sampling error. Thus, any decision to drop athletic and behavioral domains from further consideration in future studies would clearly be misguided. In fact, in this situation any inference that one domain is more important than another would be erroneous, because the data were sampled from a population in which all domains are known to be equally important.

What factors are responsible for the apparent inconsistencies shown in Table 4 and the likely problem that any single study may mislead as much as it reveals? Two factors conspire to create this problem. First, each regression analysis in Table 4 was based on a sample size of 100. In all three cases, data were generated from a model in which the predictor variables and the criterion all had the same medium correlation with one another. Table 3 showed that in this situation the probability of obtaining at least one statistically significant effect is .84. Thus, it is not surprising that each replication shown in Table 4 contains at least one statistically significant effect. However, Table 3 also showed that the probability that any specific effect is statistically significant is only .26. Thus, we would expect about four statistically significant results among tests of 15 regression coefficients, and in this particular set of replications that is exactly what happened. The point here is that although power is sufficient for obtaining statistical significance somewhere, it is not sufficient for any specific effect. Thus, one culprit making it difficult to interpret the results of any single study properly is the inadequate power for testing any specific effect. As implied in Table 3, apparent inconsistencies and apparent null results would be greatly lessened if the sample size were 400 instead of 100.

A second contributing factor is the misinterpretation of a nonsignificant test. Although the methodological literature is replete with warnings about the dangers of attempting to confirm the null hypothesis, for most mere mortals this temptation is difficult to resist. Attention tends to focus immediately on the column of probability values and their attendant asterisks or lack thereof. As Schmidt (1992, 1996) and others have pointed out, this problem can be minimized by supplementing significance tests with confidence intervals. A later section of the article focuses on the possible benefits of confidence intervals as an adjunct to or replacement for significance tests.

Table 5 continues the theme developed in Table 4. Specifically, Table 5 summarizes the results of performing 10,000 replications of the same regression study based on various sample sizes. As in Table 4, data were drawn from a population in which all zero-order correlations were medium. Thus, the population regression coefficient for each domain is nonzero, and all five population values are equal to one another. Table 5 shows that when sample size is small, the most likely result is that only one predictor will be statistically significant as happened in the second and third replications in Table 4. Under this scenario, which of the five predictors is the one declared to be significant varies randomly from sample to sample. For example, when the sample size is 50, only the coefficient for academic compe-

Table 5
Patterns of Statistically Significant Regression Coefficients With Exchangeable Medium Zero-Order Correlations and Five Predictors for a Range of Sample Sizes

No. of significant predictors	Sample size			
	$n = 50$	$n = 100$	$n = 200$	$n = 400$
No significant predictor	.43	.16	.01	<.01
Exactly one significant predictor	.43	.45	.13	<.01
Exactly two significant predictors	.12	.32	.40	.04
Exactly three significant predictors	.01	.07	.36	.26
Four or more significant predictors	<.01	<.01	.09	.70
Pattern				
Any specific predictor (alone or in combination)	.15	.26	.48	.78
Any specific predictor by itself	.09	.09	.03	<.01
Any specific pair of predictors	.01	.03	.04	<.01

Note. Power values shown were obtained from 10,000 replications of a simulation. All zero-order bivariate correlations were presumed to be .30 (a medium effect size) in the population. Each entry in the table depicts the probability that a pattern of statistically significant results will occur with the specified sample size. The sum of the first five probabilities within each column equals 1.0, except for rounding error.

tence will be statistically significant 9% of the time, only the coefficient for appearance 9% of the time, and so forth. As we have seen, the result is that the specific pattern of results in any single study is idiosyncratic and varies greatly from study to study. In contrast, when $n = 400$, the most typical result is that either four or five predictors will be statistically significant, producing a much more consistent pattern of findings across studies.

One potentially confusing aspect of Table 5 may deserve further comment, especially because similar patterns will appear in other tables to be presented shortly. The power values in three rows of the table (viz., the rows corresponding to exactly one, two, or three significant predictors) initially increase as sample size increases but then begin to decrease at some point. To understand this apparent anomaly, focus on the row entitled "Exactly two significant predictors." As the sample size increases from 50 to 100 or from 100 to 200, the probability of exactly two significant predictors increases. However, as the sample size increases from 200 to 400, this probability decreases dramatically. The reason for this pattern is that when sample size is low, the most likely result is either no significant predictor at all or at most one significant predictor. Thus, increasing sample size from 50 to 100 or from 100 to 200 has the effect of making it more likely to obtain exactly two significant predictors instead of only one or even none. However, increasing sample size from 200 to 400 has the effect of making it less likely to obtain exactly two significant predictors because now it is much more likely to obtain either exactly three significant predictors or four or more significant predictors. Another useful perspective is to consider the probability of obtaining two or more significant predictors. Summing the relevant rows in Table 5 reveals that this probability increases from .13 for $n = 50$ to .39 for $n = 100$ to .85 for $n = 200$ and to 1.00 (rounded off to two decimal places) for $n = 400$. This perspective underscores the point that larger sample sizes lead to a higher probability of obtaining at least two significant predictors, just as would be expected.

Table 6 shows the patterns of statistically significant predictors for the subset of studies in which at least one predictor was statistically significant. To the extent that statistical significance of at least one predictor is a prerequisite for publication, the pattern of results in the published literature would follow the probabilities shown in Table 6. Although the absolute magnitudes of values in Table 6 are higher than the corresponding values in Table 5, the general pattern of results is much the same. In particular, smaller sample sizes yield less consistent results.

The relationship between sample size and pattern of results shown in Tables 5 and 6 is not restricted to multiple regression. Table 7 displays similar findings for a 2×2 factorial design analyzed with ANOVA. Table 8 shows corresponding values for the subset of studies in which at least one effect was statistically significant. As was true in the generation of Table 1, it is once again assumed that each main effect as well as the interaction corresponds to a medium effect size. Tables 7 and 8 show the pattern of results that will occur as a function of cell size. As was true in the regression results depicted in Tables 5 and 6, small samples produce unstable results in the 2×2 ANOVA. Under the stated conditions, cell sizes of 10 lead to a situation in which not only will one or more true effects almost certainly go undetected but in addition the pattern of statistically significant results may well be idiosyncratic to the sampling error present in this particular study. On the other hand, cell sizes of 40 begin to produce a very different pattern. Not only is there now a much more reasonable

Table 6
Patterns of Statistically Significant Regression Coefficients With Exchangeable Medium Zero-Order Correlations and Five Predictors for a Range of Sample Sizes for the Subset of Studies With at Least One Statistically Significant Predictor

No. of significant predictors	Sample size			
	$n = 50$	$n = 100$	$n = 200$	$n = 400$
Exactly one significant predictor	.76	.54	.13	<.01
Exactly two significant predictors	.21	.38	.41	.04
Exactly three significant predictors	.03	.08	.37	.26
Four or more significant predictors	<.01	<.01	.10	.70
Pattern				
Any specific predictor (alone or in combination)	.25	.31	.49	.78
Any specific predictor by itself	.15	.11	.03	<.01
Any specific pair of predictors	.02	.04	.04	<.01

Note. Power values shown here were obtained from a subset of the 10,000 replications performed to obtain the values in Table 5. In all cases the power values in Table 6 are based on more than 5,500 replications. Each entry in the table depicts the probability that a pattern of statistically significant results will occur with the specified sample size. The sum of the first four probabilities within each column equals 1.0, except for rounding error.

prospect of detecting all true effects, but the pattern of results is likely to be much more stable from study to study.

Another typical situation involving sample size and pattern of results occurs when researchers compare groups on more than one dependent variable. As a simple illustration, suppose that two groups are being compared on four dependent variables. Table 9 shows the patterns of results that will occur if in reality the group difference is medium on each dependent variable and if the dependent variables correlate at the medium level (a correlation of .30) with one another.[6] A sample size of 25 individuals per group produces at least one statistically significant result 78% of the time. From this perspective, 25 participants per group might be viewed as sufficient. However, the power for any single dependent variable in this situation is only .41. Furthermore, Table 10 shows that among studies with at least one statistically significant result (i.e., roughly among published studies), two thirds of the time statistically significant results will be obtained for only one or two of the dependent variables when $n = 25$. Thus, with this sample size, the published literature is likely to display notable inconsistencies as to which variables the groups truly differ on. In fact, Table 10 shows that any specific variable has almost exactly a 50–50 chance of being deemed "significant" in any single study with at least one statistically significant result. For a larger sample size of 50 per group, the statistical power for any single variable becomes .69 (see Table 9), but even here the probability that true group differences are revealed on all four dependent variables is only one third. To have a power of .80 for detecting all four group differences in this situation requires a sample size of approximately 100 individuals per group. The point here is not necessarily that 100 is the "correct" sample size but instead that as in the 2 × 2 ANOVA and in multiple regression, small sample sizes lead to a published research literature that is virtually guaranteed to contain numerous inconsistencies about what is statistically significant and what is not.

Table 7
Pattern of Statistically Significant Results in 2 × 2 Analysis of Variance When All Effects Are Medium

Significant effect	Cell size			
	$n = 10$	$n = 20$	$n = 30$	$n = 40$
No effect	.29	.07	.01	<.01
A alone	.15	.10	.04	.01
B alone	.15	.10	.04	.01
A × B alone	.15	.10	.04	.01
A and B (but not A × B)	.08	.14	.14	.09
A and A × B (but not B)	.08	.14	.14	.09
B and A × B (but not A)	.08	.14	.14	.09
A and B and A × B	.04	.21	.47	.69

Note. Both main effects as well as the interaction are presumed to have medium effect sizes. Table entries are probabilities calculated using a binomial approximation. Column totals sum to 1.0, except for rounding error.

The preceding examples illustrate that the tendency to conduct underpowered studies will tend to produce an inconsistent body of literature. Rossi (1997) underscored the potential importance of the hypothetical examples presented here by providing a compelling case study of how lack of power did in fact lead to an inconsistent set of studies examining an actual phenomenon of considerable interest. Specifically, Rossi (1997) described a historical controversy

[6] The values reported in Tables 9 and 10, as well as in Figures 2 and 3, were obtained through simulation using SAS PROC IML. All reported values are based on a minimum of 5,000 replications.

Table 8
Pattern of Statistically Significant Results in 2 × 2 Analysis of Variance When All Effects Are Medium for Subset of Studies With at Least One Statistically Significant Effect

Significant effect	Cell size			
	$n = 10$	$n = 20$	$n = 30$	$n = 40$
A alone	.21	.10	.04	.01
B alone	.21	.10	.04	.01
A × B alone	.21	.10	.04	.01
A and B (but not A × B)	.11	.15	.14	.09
A and A × B (but not B)	.11	.15	.14	.09
B and A × B (but not A)	.11	.15	.14	.09
A and B and A × B	.05	.23	.47	.69

Note. Both main effects as well as the interaction are presumed to have medium effect sizes. Table entries are probabilities calculated using a binomial approximation. Column totals sum to 1.0, except for rounding error.

regarding the existence of spontaneous recovery of verbal associations and showed that this controversy can be understood in terms of inconsistent results as a consequence of underpowered studies.

Effect Sizes and Confidence Intervals

As mentioned earlier, contributing factors to the problems associated with small sample sizes are the overreliance on and misinterpretation of significance tests. Wilkinson and the Task Force on Statistical Inference (1999) recommended that significance tests be accompanied by effect size measures and, ideally, confidence intervals to better inform readers. This section of the article revisits the multiple regression example and the multiple dependent variable example to consider the potential benefits as well as possible limitations of effect sizes and confidence intervals.

Table 9
Patterns of Statistically Significant Results for Four Dependent Variables (DVs) All With Medium Effect Size

No. of significant DVs	Sample size per group		
	$n = 25$	$n = 50$	$n = 100$
No significant DVs	.22	.04	<.01
Exactly one significant DV	.28	.11	.01
Exactly two significant DVs	.25	.21	.03
Exactly three significant DVs	.17	.31	.16
Exactly four significant DVs	.08	.33	.81
Pattern			
Any specific DV (alone or in combination)	.41	.69	.94

Note. Each entry in the table depicts the probability that a specific pattern of statistically significant results will occur with the specified sample size. The sum of the first five probabilities within each column equals 1.0, except for rounding error. A minimum of 5,000 replications were performed for each sample size.

Table 10
Patterns of Statistically Significant Results for Four Dependent Variables (DVs) All With Medium Effect Size for the Subset of Studies With at Least One Statistically Significant Result

No. of significant DVs	Sample size per group		
	$n = 25$	$n = 50$	$n = 100$
Exactly one significant DV	.35	.11	.01
Exactly two significant DVs	.32	.22	.03
Exactly three significant DVs	.22	.32	.16
Exactly four significant DVs	.11	.35	.81
Pattern			
Any specific DV (alone or in combination)	.52	.72	.94

Note. Each entry in the table depicts the probability that a specific pattern of statistically significant results will occur with the specified sample size. The sum of the first four probabilities within each column equals 1.0, except for rounding error. A minimum of 5,000 replications were performed for each sample size.

A convenient effect size measure in multiple regression is the standardized regression coefficient. Thus, instead of simply reporting a variable to be a significant or nonsignificant predictor, a much better strategy may be to report a confidence interval for the standardized regression coefficient for each predictor. For example, it would be possible to form a 95% confidence interval for each regression coefficient shown in Table 4.

Figure 1 displays the result of forming such confidence intervals. An interval for each predictor in Replication 1 is shown at the left of the figure. Moving to the right, the middle portion of the figure shows comparable intervals for each of the five predictors in Replication 2. Finally, the five intervals appearing at the right of the figure are the intervals obtained in Replication 3.

Figure 1 shows that in these studies all 15 confidence intervals would overlap with one another. This has three important implications. First, although the presence or absence of asterisks is often inconsistent from study to study, confidence intervals show that for these three studies, the collection of results is in fact consistent. For example, the three 95% intervals for the academic domain are (–0.01, 0.41), (–0.14, 0.29), and (0.01, 0.39). Any belief that the results of the third study are somehow truly different from those of the first two simply because academic domain was a statistically significant predictor in the third study but not in the first two immediately vanishes when the three confidence intervals are compared with one another. Second, whereas the presence or absence of asterisks tends to convey an air of finality that an effect exists or does not exist, the confidence intervals tend to convey an attitude that considerable uncertainty remains about the true population

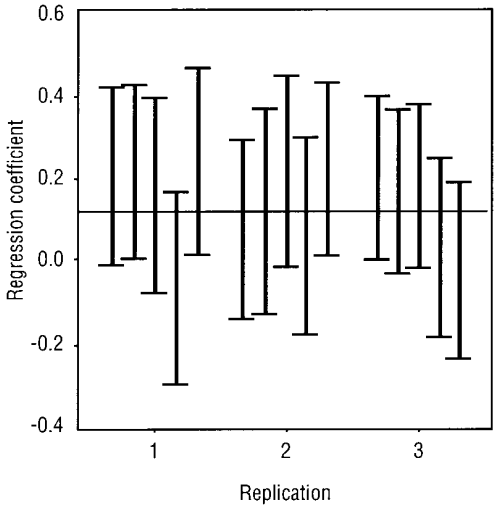

Figure 1. Confidence intervals for regression coefficients of five predictors across three replications. The first five bars represent 95% confidence intervals for each of the five predictor variables as obtained in the first replication. The second five bars are for the second replication, and the final five bars are for the third replication. The horizontal line depicts the true population value of all regression coefficients.

value of this coefficient based on any single study. As Cohen (1994) suggested, the very fact that confidence intervals reveal such uncertainty may partly explain why researchers have been reluctant to report them. A third implication of reporting a confidence interval is that it effectively emphasizes that the sample size of 100 may have been too small here. In particular, confidence intervals based on this sample size reveal all too clearly that any single study has hardly pinpointed the precise true population value of the regression coefficient associated with each domain. Given the population from which these data were drawn, the actual population value for each regression coefficient is 0.14 as shown by the horizontal line in the figure. However, a sample size of 100 produces an interval whose width is approximately .40, thus showing that more data are needed to obtain a precise estimate of the true population value.

Figure 1 suggests that confidence intervals may be less prone to misinterpretation than significance tests. However, even proper interpretation of confidence intervals can become tricky when multiple intervals are constructed in a single study. To put this remark in context, it is important to realize that the same multiple comparison procedures (e.g., Bonferroni, Tukey, Scheffé) that are often used to control familywise Type I error rates can usually be used to produce simultaneous confidence intervals. Thus, researchers who believe it to be important to produce intervals that provide desired coverage probabilities across a set of intervals typically have a straightforward method available for accomplishing this goal. Less obvious, however, are other potential perils associated with interpreting intervals across multiple variables or effects. For example, Schenker and Gentleman (2001) showed that the common practice of examining whether confidence intervals overlap is necessarily conservative and lacks power when used as a basis for deciding whether two independent point estimates are statistically significantly different from one another.

Proper interpretation of even a single interval can be problematic when that interval has been selected from among a group of several intervals. For example, consider once again a situation in which two groups are being compared on four dependent variables. As before, for simplicity we assume a medium population effect size for each dependent variable, and we also assume that variables correlate at a medium level with one another. What should we expect effect sizes and confidence intervals to reveal in such a situation? In particular, one question of interest is what might be observed for the dependent variable with the largest effect size, because all other things being equal, this is the variable an investigator might choose to emphasize in his or her interpretation of group differences.

Figure 2 provides a graphical answer to the question of what results can be expected for the dependent variable with the largest effect size in the sample. Specifically, Figure 2 shows the average *d* value as well as the average upper and lower limits of a 95% confidence interval for *d*, for group sample sizes of 25, 50, and 100 among studies with at least

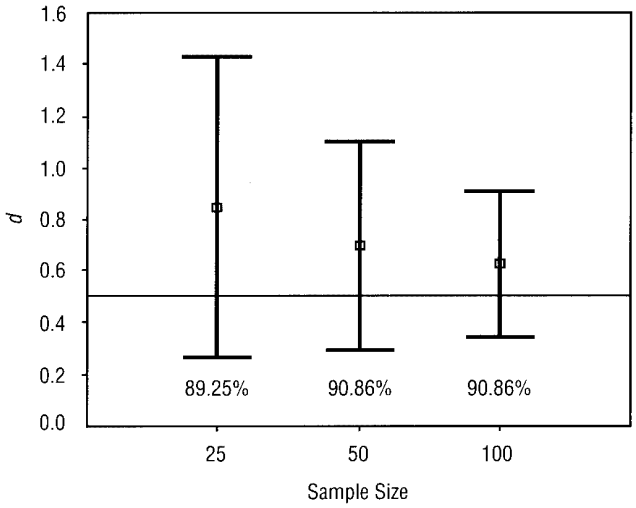

Figure 2. Average confidence interval for *d* based on the dependent variable with the largest sample difference among studies with at least one statistically significant result when population *d* value is 0.5. The horizontal line depicts the true population effect size value of 0.5. The percentages under each bar show the actual percentage of coverage of the nominal 95% confidence interval.

one statistically significant group difference. For simplicity, we assume that the published literature contains all such studies and that studies failing to find at least one statistically significant difference do not appear in the literature. To understand the implications of the figure, consider the results shown for a sample size of 25 per group. In this situation, the average d value obtained for the dependent variable with the largest effect in the sample is 0.85. Even though every dependent variable truly has a medium effect size of 0.50 in the population, the average study will show one dependent variable with a d value slightly above Cohen's (1988) definition of a large effect. Thus, there is an extreme temptation to conclude that the group difference is truly large for at least one of the variables included in the study when in reality the population group difference is only medium. Fortunately, this temptation can be at least partially addressed by requiring researchers to form a confidence interval. In this case, the average interval would stretch from 0.26 to 1.42 for the variable with the largest sample d. Even though this interval is still centered around a large effect, it is much less tempting to infer a strong belief that the true population value of d is large, because the interval is so wide.

Unfortunately, forming a confidence interval does not fully solve the underlying problem. It turns out that only 89.25% of intervals formed for the variable with the largest sample value of d contain the true population value of 0.50. Worse yet, 10.75% of intervals have a lower limit above 0.50. This means that slightly more than 10% of reported intervals will fail to contain the true population value of 0.50 and will instead imply that the true population value of d is above 0.50 with apparent confidence of 95%. This is more than a fourfold increase over the 2.5% figure an unsuspecting reader or researcher would expect based on the nominal confidence level of 95%. Four factors contribute to this distortion. First, the focus here is on the variable with the largest sample effect. Obviously, variables with smaller sample effects will produce smaller values of d, and intervals less likely to overestimate d. However, this is small consolation for researchers who understandably may be most focused on interpreting their largest effects. Second, the intervals shown here are not simultaneous 95% confidence intervals. Requiring simultaneous confidence of 95% would widen every interval and thus reduce the 10.75% figure. However, little emphasis has been paid to simultaneous confidence intervals, except in some very specific ANOVA situations (see Maxwell & Delaney, 2004, for more information about simultaneous confidence intervals in this context). Third, these intervals reflect only those studies with at least one statistically significant result. However, studies without a statistically significant result are usually unlikely to be published, so in this respect the results shown here can be expected to more closely resemble the published literature. Fourth, a major reason for the bias found here is the small sample size.

The middle and right-most intervals in Figure 2 show that the degree of bias and distortion is less for group sample sizes of 50 and 100. The midpoints of the intervals are considerably closer to the true population value of 0.50 than was the case for the smaller sample size of 25. Of course, the intervals themselves are also narrower, although even with a total of 200 participants, the average interval has a width of more than 0.50, which underscores that even with this sample size considerable uncertainty remains about the population value of d. Furthermore, the coverage probability of 90.86% for a 95% interval is only very slightly improved. Almost 10% of intervals will still have a lower limit above the true population value of 0.50 even with 200 participants.

Some traditionalists might suggest that part of the problem shown in Figure 2 reflects capitalization on chance that could be reduced or even eliminated by requiring a statistically significant multivariate test. Figure 3 shows the result of adding this requirement. Although fewer studies will meet this additional criterion, the smaller subset of studies that would now presumably appear in the literature are even more biased than the studies depicted in Figure 2. For example, with $n = 25$, the average value of d has risen from 0.85 to 0.96. Correspondingly, the true coverage probability of a 95% interval is now 81.66%, even smaller than before. As a result, nearly one in every five studies would show a

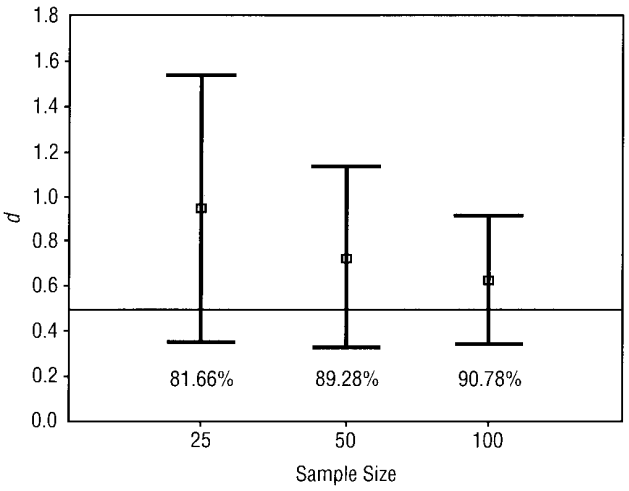

Figure 3. Average confidence interval for d based on the dependent variable with the largest sample difference among studies with at least one statistically significant result and statistically significant multivariate test when population d value is 0.5. The horizontal line depicts the true population effect size value of 0.5. The percentages under each bar show the actual percentage of coverage of the nominal 95% confidence interval.

variable whose lower limit exceeds 0.50, mistakenly implying that with 95% confidence the true population effect is larger than medium.

It is important to emphasize that the distortion seen in Figures 2 and 3 can potentially be addressed by a careful interpretation of results accumulated over multiple studies. In particular, the effect size and accompanying interval for any specific dependent variable will tend to overestimate the true population value one half of the time and underestimate the true population value the other half of the time. As long as investigators focus on a specific variable, multiple studies will eliminate any apparent bias. Multiple studies would fail to reveal this bias only in the case in which researchers focus their interpretation in each study on whichever dependent variable happens to show the largest effect in that individual study. A worst case scenario might exist when each study reports the results only for the variable with the largest effect and yet a meta-analyst reading the collection of studies might reasonably regard the various dependent variables reported across studies as being comparable to one another, thus seeming to justify accumulating results across studies even though the results are not uniformly reported for exactly the same dependent measure.

Tables 9 and 10, along with Figures 2 and 3, illustrate complications that can emerge when multiple dependent variables are combined in a single study, especially with small sample sizes. However, similar complications can exist with regard to multiple independent variables. For example, consider the following statement from Aronson, Wilson, and Brewer (1998):

> One of the most frequently misunderstood aspects of experimentation is the amount of pretesting that is often required to make sure that the independent variable is having the desired impact. When students read published experiments in psychological journals, they often have the impression that the researchers had an idea, designed a study, collected the data in a few weeks, analyzed the data, and presto, found exactly what they predicted. Little do they know that in most cases the experiment was preceded by a good deal of pretesting, whereby different versions of the independent variable were "tried out.". . . This might seem to be misleading, in that the researchers ended up reporting only the version of the independent variable that had the desired effect. (p. 117)

The effects of such pretesting on a body of literature depend on exactly how the pretesting is conducted. One possibility is that after discovering a version of the independent variable that produces the desired effect, an investigator begins his or her study from scratch, collecting brand new data. In this case, the only real harm is that some forms of the independent variable that are truly effective might have gone undetected because of low power. However, another possibility is that the researcher reports the data obtained during pretesting or, at the very least, analyzes data in which additional participants have been added to the pretesting participants. This practice is analogous to selecting the dependent variable with the largest effect, as shown in Figures 2 and 3, and will tend to produce biased estimates of effect sizes.

Further Considerations and Possible Remedies

The previous sections have offered explanations for why underpowered studies persist and why they may be harmful to the development of a coherent scientific literature. An argument could be advanced that consistency of results is less important in the initial stages of research, in which the major goal is to explore a variety of potential relationships in the hope of identifying those are most deserving of subsequent confirmatory studies. Although this argument may have some merit, nevertheless it is important that researchers appreciate the fact that many actual nonzero relationships will go undetected and hence be dropped from further consideration to the extent that initial exploratory studies are underpowered. Even worse, legitimate exploration may slip into what Kerr (1998) labeled as "HARKing," (hypothesizing after the results are known), which he defined as "presenting a post hoc hypothesis in the introduction of a research report as if it were an a priori hypothesis" (p. 197). Kerr reported the results of a survey suggesting that HARKing is widespread in psychological research and has been encouraged by the suggestion of some eminent researchers such as Bem (1987) that

> the data may be strong enough to justify recentering your article around the new findings and subordinating or even ignoring your original hypotheses.. . . If your results suggest a compelling framework for their presentation, adopt it and make the most instructive finding your centerpiece. (p. 173)

The results presented in Tables 1 through 10 show that a researcher adopting such a strategy may have a reasonable probability of discovering apparent justification for recentering his or her article around a new finding. Unfortunately, however, this recentering may simply reflect sampling error given the sample sizes typical of most psychological research. Similarly, the effect sizes shown in Figures 2 and 3 show that this strategy will inevitably produce positively biased estimates of effect sizes, accompanied by apparent 95% confidence intervals whose lower limit may fail to contain the value of the true population parameter 10% to 20% of the time. In any case, notice that no bias results if recentering involves only varying the emphasis of what is written in the discussion section of an article. However, recentering can create a serious bias if it involves selective presentation of results.

A related problem is that even a literal replication in a situation such as this would be expected to reveal smaller effect sizes than those originally reported. The large bias shown in Figures 2 and 3 suggests that the magnitude of effect sizes found in attempts to replicate can be much

smaller than those originally reported, especially when the original research is based on small samples. A less obvious implication of Figures 2 and 3 is that these smaller effect sizes might not even appear in the literature because attempts to replicate may result in nonsignificant results. For example, suppose a researcher decides to replicate a study that reports a *d* value of 0.96 for a comparison of two independent groups based on a sample size of 25 per group. A power analysis would reveal that only 19 participants per group are required to have a power of .80 to detect a difference corresponding to a population *d* value of 0.96. However, suppose that the original *d* value of 0.96 was in fact obtained as the largest difference among four dependent variables. Figure 3 shows that a *d* value of 0.96 is the average maximum sample *d* value observed in a study with 25 participants per group when four dependent variables each have a true population *d* value of 0.50 and correlate .30 with one another. The actual power with 19 participants per group for a single dependent variable in this situation will be only 0.32. Thus, the investigator has only a one third chance of obtaining a statistically significant result in this case. The original sample size of 25 per group raises the power only to 0.41. The end result is that there is clearly a high probability of failing to replicate the original finding, primarily because the original *d* value reported in the literature is in this case a badly biased estimate of the true underlying population *d* value.

Of course, the most obvious solution to this problem is simply to use larger samples. As Cohen (1962) said, "Since power is a direct monotonic function of sample size, it is recommended that investigators use larger samples than they customarily do" (p. 153). Cohen (1962) based this recommendation on the principle that "unless one is to increase the significance level (i.e., increase the risk of Type I errors) or use directional tests (e.g., a one-sided test for *t*) power can generally be increased only by an increase in sample size" (pp. 151–152). From this perspective, there is one and only one solution to the problem of underpowered studies, namely, to increase sample size. Indeed, with an increase in attention to the importance of power, statistics books are increasingly likely to emphasize the role of sample size in influencing power. However, researchers who are faced with practical limitations of sample size may feel that there is little point in conducting a formal power analysis because they may have little ultimate ability to acquire as large a sample as the power analysis might suggest. Thus, researchers may be inclined to do the best they can within reasonable limits in obtaining a reasonable sample size and then simply hope for the best. Unfortunately, this attitude ignores the increasing realization that sample size is not the only factor influencing statistical power. In fact, Cohen (1962) himself alluded to such possibilities more than 40 years ago in a footnote, where he briefly included "improving experimental design efficiency and/or experimental control" (p. 152) as two possible methods of increasing power. Methodologists have developed a number of such methods for improving efficiency and control since the time of Cohen's (1962) original article on power, yet most researchers still seem to equate power with sample size. A number of recent sources (Dennis, Lennox, & Williams, 1997; Hansen & Collins, 1994; Higginbotham, West, & Forsyth, 1988; Lipsey, 1997; Shadish, Cook, & Campbell, 2002; West, Biesanz, & Pitts, 2000) present methods for increasing statistical power without increasing sample size.

Of course, it would be equally shortsighted to believe that simply adopting some of these methods will single-handedly lead to cumulative knowledge in psychology. As Schmidt (1996) pointed out, in many situations it may be the case that the sample size required to have adequate power is still beyond the resources of a single investigator and a single study. As he and others have suggested, meta-analysis provides one method for developing cumulative knowledge over and above single studies. In particular, as Cohn and Becker (2003) showed, meta-analysis can increase power to detect an effect by providing a more precise estimate of a population effect size than would be available from a single study. However, a question can be raised about the extent to which meta-analysis is likely to provide an unbiased estimate of this true underlying effect size. For example, Kraemer, Gardner, Brooks, and Yesavage (1998) showed that including underpowered studies in meta-analyses leads to biased estimates of effect size whenever accessibility of studies depends at least in part on the presence of statistically significant results. Ironically, to the extent that multiple tests are conducted in most studies, the problem identified by Kraemer et al. may be less severe. Presumably many published studies contain a mix of statistically significant as well as nonsignificant results. Even if statistical significance somewhere in the collection of tests is a virtual prerequisite for publication, it may nevertheless be the case that published literature contains tests of specific hypotheses whose statistical test was nonsignificant. Even so, bias will be absent only to the extent that researchers fully report all results, not just those that are statistically significant.

In any event, Kraemer et al. (1998) recommended that underpowered studies be excluded from meta-analyses in order to obtain more nearly unbiased estimates of effect sizes. In fact, this perspective has led some medical researchers to maintain that conducting underpowered clinical trials is unethical to research participants (Halpern, Karlawish, & Berlin, 2002; but for contrary views see Janosky, 2002; Lilford & Stevens, 2002). Thus, even those who argue that meta-analysis is necessary to make up for the lack of power in individual studies may need to realize that the individual studies themselves must have adequate power or else the results of the meta-analysis are likely to be biased. Indeed, this view receives some support from health re-

searchers who have compared the results of meta-analyses to large-scale clinical trials. Although there is some evidence suggesting that results of large-scale trials tend to fall within the boundaries implied by a random effects meta-analytic model (e.g., Berry, 2000), other studies suggest that meta-analytic effect sizes tend to overestimate effects found in large-scale studies, just as would be expected to the extent that underpowered studies are included in the meta-analyses (e.g., Chalmers et al., 1987; LeLorier, Gregoire, Benhaddad, Lapierre, & Derderian, 1997; Villar, Carroll, & Belizan, 1995). Even so, an alternative to excluding underpowered studies from meta-analyses is to include all studies but take into account possible effects of publication bias. For example, Sterne, Egger, and Davey Smith (2001) described a variety of graphical methods to detect the presence of publication bias as well as statistical models intended to adjust for the effects of possible publication bias (see also Hedges & Vevea, 1996; Vevea & Hedges, 1995).

Psychology might borrow yet one other perspective from health researchers. When researchers find themselves facing Schmidt's (1996) concern that requiring a power of .80 implies a sample so large as to "make it impossible for most studies ever to be conducted" (p. 123), an alternative is to consider a collaborative multisite study. Health researchers seem to have recognized the benefits of such designs and implemented support structures to encourage such studies beyond the historical norm in the behavioral sciences. Ironically, more than 50 years ago Toops made a very similar suggestion in advocating the "standard million," whereby each of 1,000 psychologists would obtain data on 1,000 individuals (Widaman, 2000). Thankfully, samples this large are unnecessary even to detect minuscule effect sizes, but it may be an appropriate time to reconsider the merit of Toops's idea, albeit on a much smaller scale. In fact, Howard, Maxwell, and Fleming (2000) recently illustrated how Bayesian methods and meta-analysis can be used as a primary data analytic method when primary data are collected over multiple studies. In principle, another solution is for psychology as a discipline to change publication practices so that studies with nonsignificant results are as likely to be accepted for publication as are studies with statistically significant results. Although such a change would have much to recommend itself in theory, Kraemer et al. (1998) pointed out that in practice such a dramatic change in editorial policy, reviewers' judgments, and authors' expectations is unlikely in the near future. Of course, even if such a policy were adopted, it is still the case that studies with greater power and precision are more informative than their underpowered counterparts, all else being equal.

Conclusion

Unless psychologists begin to incorporate methods for increasing the power of their studies, the published literature is likely to contain a mixture of apparent results buzzing with confusion. Increased reporting of effect sizes and confidence intervals will not by itself increase the consistency of the literature, although it may motivate more powerful studies by highlighting a major source of likely confusion. Not only do underpowered studies lead to a confusing literature but they also create a literature that contains biased estimates of effect sizes. Furthermore, as implied in the curse of multiplicity, researchers may have felt little pressure to increase the power of their studies, because by testing multiple hypotheses, they often assured themselves of a reasonable probability of achieving a goal of obtaining at least one statistically significant result. However, the fact that most methodological reviews have continued to show that studies are underpowered implies that tests of individual hypotheses more often than not lack sufficient power, even though adequate power exists for detecting an effect somewhere in the collection of tests. This discrepancy in conceptualizations of power has almost certainly contributed to a literature that not only is inconsistent but also overestimates actual values of effect sizes.

If psychology is to continue to develop a coherent and accurate body of scientific literature, it is imperative that further attention be given to the role of power in designing studies and interpreting results. In particular, it would seem advisable to require that a priori power calculations be performed and reported routinely in empirical research. Ideally, such calculations would reveal not only the probability that any specific single test is statistically significant but also the probabilities of obtaining at least one statistically significant result as well as the probability that all hypotheses to be tested will be rejected. Of course, one hope in requiring such power calculations is that this would effectively force researchers to design more powerful studies. Fortunately, an assumption that the only way to increase power is to increase sample size is almost always wrong. Psychologists are encouraged to familiarize themselves with additional methods for increasing power. Unfortunately, this step by itself is unlikely to be sufficient.

In most situations, it is simply unrealistic to believe that a single study will provide a definitive answer to the most important question of interest, much less all questions of interest, in part because precise estimates require such large samples but also because of the idiosyncratic nature of many single studies (Wilson & Lipsey, 2001). Thus, either multicenter studies or meta-analysis will often be necessary. In any case, several additional steps, all of which have been suggested previously, seem called for: (a) complete reporting of results, including nonsignificant as well as significant findings, a policy made much more feasible with the advent of electronic presentation of results; (b) the archiving of raw data; (c) the registration of studies prior to carrying them out, so that studies on a given topic are equally accessible regardless of their findings; and (d) the presentation of

confidence intervals even when their major contribution is to reveal the lack of precision with which important parameters are being estimated. Although these policies are clearly not a panacea, their adoption would contribute to a cumulative science of psychology.

References

Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods, 2,* 20–33.

Aronson, E., Wilson, T. D., & Brewer, M. B. (1998). Experimentation in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (Vol. 1, 4th ed., 99–142). New York: McGraw-Hill.

Bem, D. J. (1987). Writing the empirical journal article. In M. Zanna & J. Darley (Eds.), *The compleat academic: A practical guide for the beginning social scientist* (pp. 171–201). New York: Random House.

Berry, S. M. (2000). Meta-analysis versus large trials: Resolving the controversy. In D. K. Stangl & D. A. Berry (Eds.), *Meta-analysis in medicine and health policy* (pp. 65–81). New York: Dekker.

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology, 23,* 399–406.

Borenstein, M., Rothstein, H., & Cohen, J. (1997). *Power and precision: A computer program for statistical power analysis and confidence intervals*. Teaneck, NJ: Biostat.

Chalmers, T. C., Levin, H., Sacks, H. S., Reitmen, D., Berrier, J., & Nagalingam, R. (1987). Meta-analysis of clinical trials as a scientific discipline: I. Control of bias and comparison with large co-operative trials. *Statistics in Medicine, 6,* 315–325.

Clark-Carter, D. (1997). The account taken of statistical power in research published in the *British Journal of Psychology*. *British Journal of Psychology, 88,* 71–83.

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

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

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8,* 243–253.

Dennis, M. L., Lennox, R. D., & Williams, R. (1997). Practical power analysis. In K. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 367–404). Washington, DC: American Psychological Association.

Elashoff, J. D. (1999). *NQuery Advisor* (Version 3.0) [Computer software and manual]. Los Angeles: Statistical Solutions.

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research, 26,* 499–510.

Halpern, S. D., Karlawish, J. H. T., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Journal of the American Medical Association, 288,* 358–362.

Hansen, W. B., & Collins, L. M. (1994). Seven ways to increase power without increasing *N*. In L. M. Collins & L. A. Seitz (Eds.), *Advances in data analysis for prevention intervention research* (NIDA Research Monograph 142, NIH Publication No. 94-3599, pp. 184–195). Rockville, MD: National Institutes of Health.

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

Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin, 88,* 359–369.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21,* 299–332.

Higginbotham, H. N., West, S. G., & Forsyth, D. R. (1988). *Psychotherapy and behavior change: Social, cultural, and methodological perspectives*. New York: Pergamon Press.

Hintze, J. L. (1996). *PASS* (Version 6.0) [Computer software and manual]. Kaysville, UT: NCSS.

Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods, 5,* 315–332.

Janosky, J. E. (2002). The ethics of underpowered clinical trials. *Journal of the American Medical Association, 288,* 2118.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2,* 196–217.

Kosciulek, J. F., & Szymanski, E. M. (1993). Statistical power analysis of rehabilitation counseling research. *Rehabilitation Counseling Bulletin, 36,* 212–219.

Kraemer, H. C. (1991). To increase power in randomized clinical trials without increasing sample size. *Psychopharmacology Bulletin, 27,* 217–224.

Kraemer, H. C., Gardner, C., Brooks, J. O., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23–31.

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

LeLorier, J., Gregoire, G., Benhaddad, A., Lapierre, J., & Derderian, F. (1997). Discrepancies between meta-analysis and subsequent large randomized, controlled trials. *New England Journal of Medicine, 337,* 536–542.

Lilford, R., & Stevens, A. J. (2002). Underpowered studies. *British Journal of Surgery, 89,* 129–131.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage.

Lipsey, M. W. (1997). Design sensitivity: Statistical power for applied experimental research. In L. Bickman & D. Rog (Eds.), *Handbook of applied social research methods* (pp. 39–68). Thousand Oaks, CA: Sage.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130–149.

Maddock, J. E., & Rossi, J. S. (2001). Statistical power of articles published in three health psychology-related journals. *Health Psychology, 20,* 76–78.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Erlbaum.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2,* 3–19.

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

Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology, 49,* 103–120.

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

O'Brien, R. G. (1998). A tour of UnifyPow: A SAS module/macro for sample-size analysis. In *Proceedings of the 23rd SAS Users Group International Conference* (pp. 1346–1355). Cary, NC: SAS Institute.

O'Brien, R. G., & Muller, K. E. (1993). Unified power analysis for *t*-tests through multivariate hypotheses. In L. K. Edwards (Ed.), *Applied analysis of variance in the behavioral sciences* (pp. 297–344). New York: Dekker.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646–656.

Rossi, J. S. (1997). A case study in the failure of psychology as a cumulative science: The spontaneous recovery of verbal learning. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 175–197). Mahwah, NJ: Erlbaum.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction*. Homewood, IL: Dorsey Press.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika, 50,* 83–90.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician, 55,* 182–186.

Schmidt, F. L. (1992). What do 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.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309–316.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Sterne, J. A. C., Egger, M., & Davey Smith, G. (2001). Investigating and dealing with publication and other biases. In M. Egger, G. Davey Smith, & D. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (2nd ed., pp. 189–208). London: BMJ Books.

Thomas, L., & Krebs, C. J. (1997). A review of statistical power analysis software. *Bulletin of the Ecological Society of America, 78,* 126–139.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60,* 419–435.

Villar, J., Carroll, G., & Belizan, J. M. (1995). Predictive ability of meta-analyses of randomized controlled trials. *Lancet, 345,* 772–776.

West, S. G., Biesanz, J., & Pitts, S. C. (2000). Causal inference and generalization in field settings: Experimental and quasi-experimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social psychology* (pp. 40–84). New York: Cambridge University Press.

Widaman, K. F. (2000, October). *Scaling manifest and latent variables to promote a progressive stance in research*. Paper presented at the annual meeting of the Society of Multivariate Experimental Psychology, Saratoga Springs, NY.

Wilkinson, L., and the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wilson, D. B., & Lipsey, M. W. (2001). The role of method in treatment effectiveness research: Evidence from meta-analysis. *Psychological Methods, 6,* 413–429.

Received August 31, 2001
Revision received December 16, 2003
Accepted January 26, 2004 ■