

*Citation: Stamps, A.E. (2003). Summa contra Pisces: how to fully utilize contemporary statistical protocols. <http://ieq.home.att.net>, downloaded on [insert date you downloaded here].*

## **Summa Contra Pisces: How to fully utilize contemporary statistical protocols**

Arthur E. Stamps III

Institute of Environmental Quality, San Francisco

16 December 2003

Empirical facts are the crucible  
without which your thoughts aren't deducible.  
If you can't contrive  
a level 05,  
your data are irreproducible.

### **Abstract**

In 2001 the standards for reporting statistical findings in the behavioral sciences changed. Under the old system, a finding was considered tenable if it achieved " $p < .05$ ". The new standards require a solid understanding of six concepts: alpha error, beta error, effect size, power analysis, focus, and meta-analysis. Under the new standards, the contribution of a finding for a particular claim is measured in terms of how it changes the collective body of knowledge on that claim. Examples are given showing how the new standards can greatly increase a researcher's individual productivity as well as eliminate major errors from the collective literature.

In 2001 the standards for reporting results in behavioral journals were altered in what amounts to a sea-change in acceptable protocols. It may not be a full paradigm shift, but it is at least a two-cents shift. As a frequent reviewer for journals in the field of environment and behavior (over a score of reviews per year) I have seen dozens of papers that were written and submitted in perfectly good faith, only to meet with serious objections or outright rejection based on the new standards. Since both major revision and the need to start the review process all over with another journal are serious wastes of everyone's time, it would seem useful to explain the current standards so the objections can be eliminated before getting to the review stage.

Based on writing over 250 journal reviews, editing another 250 conference submissions, and doing a literature review of statistical usage in E & B journals covering over 500 published articles (e. g.,  $n > 1000$ ), it seems that the most common research protocol is (a) select examples based on convenience, whether an example seems "typical", or even whether examples will support a hypothesis, (b) obtain large numbers of respondents, (c) have respondents report a

large number of responses, (d) run many statistical tests, and (e) report the existence or non-existence of a finding based on whether " $p \leq .05$ " or " $p > .05$ ". In contrast, the contemporary statistical protocols require use of the following concepts: (a) alpha errors, (b) beta errors, (c) effect sizes, (d) power analysis, (e) the role of multiple degree-of-freedom tests, and (f) meta-analysis. The intent of this position paper is to explain these statistical concepts so that social scientists can work more effectively.

This paper is intended, in general, for those trying to create a reliable body of collective knowledge, and, more specifically, to those who also prefer to answer questions and resolve disputes by obtaining empirical evidence and applying probability theory. Thus readers who rely on personal experience, judgment, authority, or rhetorical deliberation will probably not find this article to be very helpful.

The tone is intentionally informal. The reason is that the formal expositions have been in the literature for decades if not centuries. Most of the ideas presented in this paper are neither new nor original (the connection between Bacon and meta-analysis being the counter-example). However, the thinking is that (a) the ideas have already been presented, earlier and better, in formal terms, (b) the ideas are still not being implemented, and therefore (c) what is needed is a more accessible presentation. The wording and style used in this paper (poems included) have communicated very well to live social science audiences, so there is some evidence to believe it will be useful to distant audiences as well. References to the math are included for readers who prefer formal expositions. It is true that there is madness in this method; hopefully, by the end of the paper, it will become apparent that there is also method in the madness.

The sea-change in acceptable statistical protocols for the behavioral sciences occurred with the 5th edition of the *Publication Manual of the American Psychological Association* (American Psychological Association, 2001). The bulk of the *APA Manual* is concerned with issues best left to professional copy editors or typesetters (indentation, use of italics, tone of language, etc.). None of these copy-editing change reported results; after all, a correlation of  $r = .80$  is larger than a correlation of  $.20$  regardless of whether "r" is in normal or italic font, centered or flushed right, labeled "extremely high" or "not so high", or whether the writing voice is serious or informal. *Section 1.10 Results* of the *Manual* is different. Section 1.10 dryly describes itself as "The Results section summarizes the data collected and the statistical or data analytic treatment used." (American Psychological Association, 2001, p. 20). The heart of the matter does not surface for another page or two, when the *Manual* mentions the need to report effect sizes rather than whether or not a finding were "significant", justify use of multiple degree-of-freedom tests, and reporting data conformable with future meta-analyses. Where did these requirements come from? Why bother with them?

A really short answer is that use of the new standards, in contrast to copy-editing issues, can very well produce conclusions totally different to the inferences drawn using the old standards. This is big thunder. For example, a correlation of  $.80$  can be "non-significant" and a correlation of  $.20$  can be "highly significant". If results are reported only in terms of "significant" or "non-significant", the implication will be that  $.20$  is larger than  $.80$ . Since this is something one would not normally do on purpose, it is quite useful to know how to avoid it. Here's how.

### *The Alpha and Beta of it*

In the usual technical jargon alpha ( $\alpha$ ) and beta ( $\beta$ ) errors are typically described in double or triple negatives. That makes the definitions a bit hard to follow. For instance, it is difficult to decipher the following logic "We failed to reject the null hypothesis therefore....". Here is a more technical explanation of the " $p < .05$ " criterion:

"The level of significance of a statistical test defines the probability level that is to be considered too low to warrant support of the hypothesis being tested. If the probability of the occurrence of observed data (when the hypothesis being tested is true) is smaller than the level of significance, then the data are said to contradict the hypothesis being tested, and a decision is made to reject this hypothesis. Rejection of the hypothesis being tested is equivalent to supporting of the possible alternative hypotheses which are not contradicted.

"The hypothesis being tested will be designated by the symbol  $H_0$ . The set of hypotheses that remain tenable when  $H_0$  is rejected will be called the alternative hypothesis and will be designated by  $H_1$ . The decision rules in a statistical test are with respect to the rejection or nonrejection of  $H_0$ . The rejection of  $H_0$  may be regarded as a decision to accept  $H_1$ ; the nonrejection of  $H_0$  may be regarded as a decision against the acceptance of  $H_1$ . If the decision rules reject  $H_0$  when in fact  $H_0$  is true, the rules lead to an erroneous decision. The probability of making this kind of error is at most equal to the level of significance of the test. Thus the level of significance sets an upper bound on the probability of making a decision to reject  $H_0$  when in fact  $H_0$  is true.

"If the decision rules do not reject  $H_0$ , when in fact one of the alternative hypotheses is true, the rules also leads to an erroneous decision. This type of error is known as a *type 2* error. The potential magnitude of a type 2 error depends in part upon the level of significance and in part upon which one of the possible alternate hypotheses actually is true. Associated with each of the possible alternatives is a type 2 error of a different magnitude. The magnitude of a type 1 error is designated by the symbol  $\alpha$ , and the magnitude of the type 2 error for a specified alternative hypothesis is designated by the symbol  $\beta$ . (Winer, Brown, & Michels, 1991)

In exchange for all that language we would like to propose this:

The alpha error is the probability of reporting noise.

The beta error is the probability of missing something.

For example, suppose you were interested in a burglar alarm. Obviously it would drive you nuts if it went off every time the cat, a fly, or a speck of dust entered the room. Alarms caused by the cat, the fly and the dust are false alarms. They report a burglar when, in fact, there isn't. If we represent the alarm in a typical statistical model it would look something like this:

Sounding (alarm) = burglar + everything else (the cat, fly, dust, etc.).

The "everything else" term is, in a clean experiment, random noise. So, if the alarm goes off and there isn't any burglar, the alarm is registering noise. That is the alpha error: reporting noise. People who like to sleep, which probably includes most of us, will obviously want to change the alarm so it doesn't pick up unwanted events (the cat, etc.). This may work well for a while, but at a certain point, the alarm is so dull that it won't detect a burglar. Now we have another problem: the alarm fails to report the event of interest. That is the classic beta error: missing something.

Conventional values for acceptable alpha and beta errors are .05 and .20 respectively. Those standards say we are willing to put our professional reputations on the line if there is less than a one-in-twenty chance we were fooled by noise, and that there was a 20% chance that we missed for what we were looking. Which precise levels to use will depend on what we are trying to do. If the consequences of accepting noise are high (giving a placebo instead of an effective medication), then we will need to adjust  $\alpha$ . If the consequences of missing something are high (an incoming missile), we will need another value for  $\beta$ . Choices for  $\alpha$  and  $\beta$  have enormous consequences on the required size of an experiment, so just trying to make both  $\alpha$  and  $\beta$  astronomically small is likely to generate an infeasible experiment. Some guidance can be gleaned from the purpose of basic empirical research proposed above: the creation of a reliable body of collective knowledge through empirical evidence and probability theory. The conventional values for  $\alpha$  and  $\beta$  seem to have worked pretty well over the last century or so, and thus we would recommend, barring a cost-benefit analysis of other choices for particular types of research, using the .05 and .20 standards for pure research.

### *Effect sizes*

Now that we have a firm grasp on probabilities of reporting noise (alpha error) and missing something (beta error) we can turn to item #3 on the list of contemporary statistical concepts: effect size. Effect size is the strength of a relationship in a population. For instance, in baseball, a hitter's batting average is the strength of the person's hitting. A player batting .400 is a better hitter than a player batting .010. Here the effect size is a proportion. Another type of effect size is a correlation ( $r$ ). If one were interested in, say, how strongly reading and arithmetic abilities were connected, a reasonable experimental design would be to obtain a sample of students, evaluate them on both abilities, and obtain a correlation over the number of students. A third commonly-used effect size is the standardized mean contrast ( $d$ ). This is simply the contrast between means divided by an estimate of a standard deviation. To continue the student example, if one were interested in whether reading ability improved between grades 6 and 8, one would obtain two samples of students, compute the two averages, divide by the pooled standard deviation, and obtain a value of  $d$ . A fourth type of effect size is proportion of variance accounted for. This will be necessary for hypotheses involving multiple variables and, of course, for variance components.

Effect sizes have several rather important properties. Perhaps the most important property is that significance levels are not effect sizes. The reason is that the significance level is a function of two things: effect size and sample size. Rosenthal and Rosnow (1991) devote virtually a whole book to this very point. The inside cover of their book lists 13 equations

showing how commonly used statistics are related to effect sizes and sample sizes. Here are some examples:

$$\chi^2(1) = \phi^2 \times N,$$

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{df},$$

$$F = \left( \frac{\mu_1 - \mu_2}{\hat{\sigma}} \times \frac{1}{\sqrt{\frac{1}{n_1} - \frac{1}{n_2}}} \right)^2.$$

These equations are explained fully in Rosenthal and Rosnow (1991)<sup>1</sup>; at this time we only wish to point out that the statistic upon which alpha is calculated ( $\chi^2$ , t, F) is a function of the effect size (phi, correlation, standardized mean difference) times the sample size (N, df,  $n_1$ ,  $n_2$ ). Thus significance levels respond to two variables, not one. Trying to estimate an effect size solely from a significance level is analogous to trying to calculate the length of a side of a rectangle given only its area. Can't be done.

The major implication of this is that the alpha level measures something other than the effect size, so it is wrong to conclude that a finding of  $p < .001$  is stronger than a finding of  $p < .05$  or even of  $p > .05$ . The alpha level depends on sample size as well as effect size, so a tiny effect size (e.g., a batting average of .010) can be "highly significant" ( $p < .001$ ), while a large effect size (batting average = .400) can be "non-significant" ( $p > .05$ ) with a sufficiently small sample). If only  $p$  inequalities are reported, then the literature could well give the impression that players batting .010 are better than players batting .400. As mentioned above, this is something none of us would do on purpose. The remedy is to report the effect size in the article. Information on significance is then reported as a confidence interval around the effect size.

Since this mistake is quite common, we would like to repeat the point for emphasis. The inference described above, indicating that results are present or absent depending on the " $p$  level", is not only incorrect but can seriously distort the true findings. That is the reason why the current APA standards require reporting of effect sizes. It is also the reason a lot of submissions are getting summarily sent back for major revision before the reviewer spends time on a detailed reading. Given the lag time for peer review submission cycles (months or even years for behavioral journals) it is clearly in the author's best interest to pre-empt as many objections as possible as soon as possible. Accordingly, authors wishing to expedite the peer review process (not to mention cleaning up the literature) would be well advised to report effect sizes in the first draft.

### *Power analysis*

The next trick is to use power analysis. Why so? Power logically follows the beta error. The beta error, it will hopefully be recalled, is the probability of missing something. A related,

and much more interesting question, is "What is the probability of finding something?". Suppose you were a miner looking for the Mother Lode. You have a choice of picking the digging spot using (a) a divining rod, which had about a 1% chance of finding gold, or (b) ground-penetrating radar, which had about an 80% chance. Clearly you would go with the 80%. The probability of finding something is called the power of a test. To wit:

The power of a test is the probability of finding something.

Conceptually power is just  $1.0 - \beta$ . Statistical power comes into play at least twice: once when calculating how large experiments need to be, and again when deciding whether an effect is so small as to justify the claim that there is no effect. For instance, being able to calculate how large experiments need to be is really, really useful. Unlike speculative discourse, empirical data are going to be expensive, in time, money, or both. Economy matters. If you have an 80% chance of getting an answer with  $n = 100$ , there is precious little reason to squander  $n = 1000$  on the idea.

Power is a function in four parts: alpha, beta, effect size, and sample size. With conventional values of  $\alpha = .05$  and  $\beta = .20$ , power is set to  $.80$  and so we are down to calculating the sample size required to detect an effect size at  $\alpha = .05$ , power =  $.80$ . An example might clarify what this means in practice. We consider the case of measuring effects with standardized mean differences ( $d$ ), and for Cohen's recommendations for small, medium, and large effect ( $d = 0.2, 0.5, 0.8$ ), (Cohen, 1988, pp. 25-26). The results are instructive. From Table 2.3.5 (Cohen, 1988, pp. 36-37) we find that, in order to detect one claim at  $\alpha = .05$ , power =  $.80$ , a large target effect would require  $n = 26$  in each of the groups ( $N = 52$ ), a medium effect would require  $n = 64$ , ( $N = 128$ ), and a small effect would require  $n = 400$  ( $N = 800$ ). We see here the dramatic effect of effect size on sample size. If one is looking for the side of a barn at 3 paces, a small experiment is fine. If one is interested in a needle in a haystack, a large experiment is necessary.

Astronomy provides another useful analogy for power analysis. If one is looking for a moon of Pluto, one would need the services of a large telescope. On the other hand, if one were looking for the Earth's moon, use of the Keck telescope would be a major waste of a precious resource (Keck seconds). Here we have a major consequence of effect size. Smaller effects will require more power to detect; larger effects will require less power. Power analysis enables researchers to spend just the right amount of resources to find a specified target size for specified levels of  $\alpha$  and  $\beta$ . If you are on any sort of budget (time, money, etc.), power analysis will obviously be highly useful to you.

A related use of power analysis is justifying claims of lack of a relationship. Recall that non-significance does not entail a non-relationship, because it may be that the experiment just did not have enough power to detect the relationship. The astronomical example would be a claim that, when viewed with the human eye, Pluto did not have any visible moons, therefore Pluto does not have moons. Such a claim would be summarily rejected because the human eye is just not powerful enough to detect Pluto, let alone its moons.

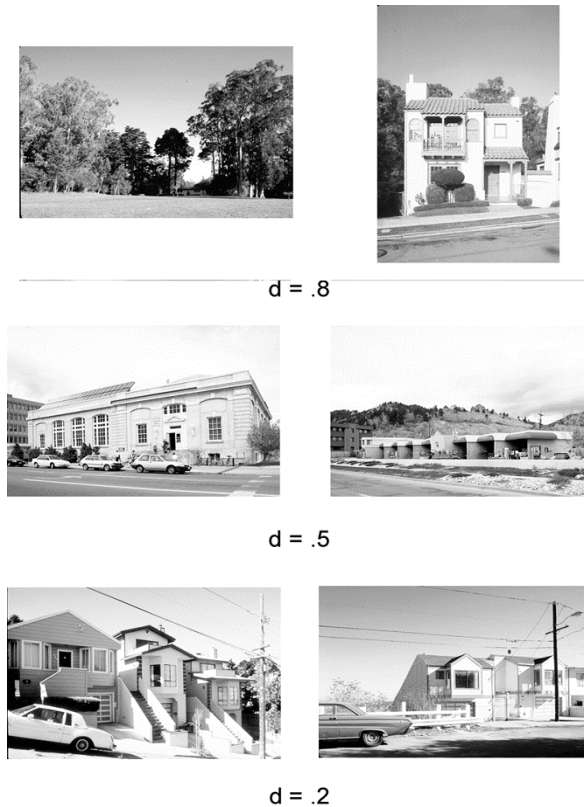


Figure 1. *Illustrations of Cohen's effect sizes.*

Under the new system it is not possible to make a claim for total non-existence of an effect. The reason is that the amount of data needed increases dramatically as the target effect diminishes. Consider, for example, the consequences of trying to detect differences in proportions. (The math is described in Cohen (1988, Chapter 6)). If one wants to detect a difference between .50 and .60 with  $\alpha = .05$  and power = .80, then one would need two samples with  $n = 392$  in each sample. If one is looking for the difference between .50 and .55, the  $n$  per sample goes up to 1570. For the difference between .49 and .51, the  $n$  per sample is 9800. There may be cases where it is necessary to detect the difference between .49 and .51 (trying to predict what is expected to be a close national election, perhaps), but we suspect that most scientists would prefer to settle for a larger target size in exchange for the ability to investigate more questions within their limited resources. The pending question then becomes "How small can an effect be before we consider it to be beneath our consideration?"

One option for basic research is to use Cohen's (1988) threshold for a "small effect". For a proportion, that would be the difference between .50 and .55; for a standardized mean difference,  $|d| < 0.20$ ; for a correlation,  $|r| < .10$ , for percentage of variance, 2%. These are explicitly conventional values, just like the choices of  $\alpha = .05$  and  $\beta = .20$ . Use of Cohen's conventions may or may not lead to counter-productive consequences, depending on what you are trying to do. Rosenthal and Rosnow (1991, p. 43) report a medical study in which the correlation between treatment and no treatment and survival was "only"  $r = .08$ , while the chances of survival given the treatment were twice the survival rate given the placebo (odds ratio = 2.08). Contrary-wise, Cohen cites psychological examples of small differences ( $d < 0.20$ ) such as heights of 15 and 16 year old women and difference in IQ between twins and non-twins. In order to see how Cohn's criteria could be applied to design issues, we collected over 3200 pairs of environmental scenes and calculated values for  $d$ . The purpose of that inquiry was to find out if there were a threshold below which regulation of visual impacts would be a waste of everyone's time (Stamps, 2000). The picture of the results is shown in Figure 1. Cohen's numbers may seem arbitrary, but then again, how many of us would want to spend years of our careers arguing over an effect as small as the difference shown in the bottom line of Figure 1?

Thus, we would suggest the following: under conventional values of  $\alpha = .05$ ,  $\beta = .20$ , and Cohen's thresholds for "small effect", studies can report a lack of a relationship by saying that the experiment had an 80 % chance of detecting a small effect at  $\alpha = .05$ , but no such effect

was detected. Therefore, if present in the population, the strength of the relationship appears to be below the threshold for small.

An implication is that if one is attempting to rule out some idea, then large samples are going to be necessary. Inspection of any of the power tables in Cohen (1988) gives some idea of the task. For a standardized mean difference  $n$  increases from 26 for a large effect up to  $n = 400$  for a small effect. Thus, a submitted paper claiming no difference between two groups of size 25 each would, in a modern review, be summarily dismissed because the  $n$  is not even remotely in the ballpark of what would be required to make a solid claim ( $25 \ll 400$ ). In fact, determining the power for  $n = 25$ ,  $\alpha = .05$ , effect size  $d = .2$  is a simple look-up, and turns out to be 11% (Cohen, 1988, p. 36). The interpretation is that the experiment was designed to have only an 11% chance of finding an answer. Might as well use a divining rod. Contemporary reviewers are more and more likely to do a quick power analyses to see if the size of the experiment is in the ballpark. Prospective authors can avoid the embarrassment of rejection due to lack of power simply by doing the power analysis themselves.

### *Multiple degree-of-freedom tests*

Multiple degree-of-freedom tests arise when we wish to investigate relationships among groups of many variables. Sometimes the logic is clear-cut: there is one dependent variable and there is one claim with  $dfh = 1$ . Examples are a standardized mean difference between two groups or a correlation between two conditions, both of which have one degree of freedom ( $dfh = 1$ ). In the behavioral sciences, it is more likely that  $dfh$  will be more than one. It is exceptionally easy to crank up  $dfh$ . For a numerical variable such as weight of apples,  $dfh = 1$ . The difference between the weights of two samples of apples also has  $dfh = 1$ , as does a correlation or a trend in weight. However, if one codes apple weight into low, medium, and high, then  $dfh$  inflates to 2. If one is interested in an interaction between weight and color and codes color as red, orange, yellow, green, blue, and purple, then  $dfh$  for the interaction becomes  $2 \times 5 = 10$ . A triple interaction of factors with five classifications in each factor would have  $dfh = 64$ . Studies investigating all possible groups of independent variables will have  $dfh(tot) = 2^{dfh(indep)}$ , where " $dfh(tot)$ " is the total hypothesis degrees of freedom and " $dfh(indep)$ " is the hypothesis degrees of freedom for the independent variables. (For the mathematical audience, the cardinality of the question is the cardinality the power set.)

### *The full fish*

All this assumes that one is interested in a single dependent variable, say degree of taste. Often there is more than one dependent variable. An inquiry may, for instance, contain data not only on preference but also on liking, pleasure, good feeling, etc. Then  $dfh(tot)$  will increase enormously. (The cardinality is the product of two power sets; even worse than one power set.) Here we have the logical interpretation of the "fishing trip". The claim is that some or all of the dependent variables might be related to some or all of the independent variables. For example, suppose we are interested in the following idea:

Preference, pleasure and liking = building type(5) + social group(5) + interactions + residual.

Here we have three dependent variables and two independent factors, each with  $dfh = 4$ . Full and complete analysis would entail running models for preference, pleasure, liking, preference x pleasure, preference x liking, and pleasure x liking for each of the possible subsets of the dependent and independent variables. The consequence is that

$$dfhtot = 2^{dfh(dep)} \times 2^{dfh(indep)} = 2^{(dfh(dep) + dfh(indep))}$$

Here we have the full fish: Run all tests among all subsets of variables, whether dependent or independent. It is really easy to run up  $dfhtot$ . For this example,  $dfhtot = 2^{(3+8)} = 2048$ . So what is wrong with that? Well, from a statistical point of view, plenty. Recall that the probability of reporting noise is the alpha error. It turns out that the alpha error is heavily dependent on the number of claims made on the same data. Consider a day at the race track. The fellow sitting next to you wins in the first race, then in the second, and so on all day. You would think him a most wonderful bookie. Then you learned that he bet on every horse in every race....

The relevant math is the calculation of a simultaneous alpha level. For  $m$  tests, each of which has its own individual alpha level, the overall alpha level is  $\alpha_{\text{simultaneous}} = 1 - (1-\alpha)^m$  (Winer et al., 1991, p. 153). Thus, if we have one test and think we are taking a 5% chance of being fooled by noise, that is indeed our risk. However, if we run five tests, each at  $\alpha = .05$ , the chances of fooling ourselves becomes 22%; with 10 tests we will have  $\alpha = 40\%$ , and with 2048 tests (easy to do with multiple dependent factors and comprehensive testing),  $\alpha$  is just about 1.00, in which case it is virtually certain that the results are indistinguishable from pure noise. If we patch this problem by accepting only claims for which  $\alpha < .05/m$ , then, for  $dfh = 2048$ , we would consider only findings with  $p < .00002$ . That would miss a lot of findings — the classic  $\beta$  error. Consequently, the intent of being thorough and complete and running all possible tests is certainly laudable in the abstract but not in the statistics.

Increasing  $dfh$  has drastic effects on experimental power. Let us consider the case of regression (Cohen, 1988, Chapter 9). Tables 9.3.1 and 9.3.2 list power for different combinations of  $dfh$  and  $dfe$  (called "u" and "v" in the book), for  $\alpha = .01$  and  $\alpha = .05$ . If you have one claim and wish to have an overall alpha of .05, then Table 9.3.2 (for  $\alpha = .05$ ) is appropriate; if you have five claims and want an overall alpha of .05 then Table 9.3.1 is the way to go. (If you have more than five claims, you will probably need to use custom software to do the power analysis. The equations are available in Johnson, Kotz, and Balakrishnan (1995). The required sample sizes are literally off the charts, which says something useful regarding how much data would be needed to justify more than five or so claims based on a single set of data).

For simple regressions (Cohen, 1988, p. 445)  $N$  is a function of the non-centrality parameter of the F distribution ( $\lambda$ ) and the target effect size ( $f^2$ ):  $N = \lambda / f^2$ . Consequently, for any given  $f^2$ , the size of the experiment will be a direct function of  $\lambda$ . So let us look at the practical implications of Tables 9.3.1 and 9.3.2. First, as one might well expect, decreasing alpha increases the required sample size. For power = .80,  $dfh = 1$ ,  $dfe = 20$ ,  $\alpha = .05$ ,  $\lambda \approx 8$ , while for  $\alpha = .01$ ,  $\lambda \approx 14$ . Thus, the required sample size will increase by about 75% if one wishes to make five tests rather than one. Second, as one well might not expect, changing  $dfh$  is

not the same as changing  $dfe$ . The range of  $dfh$  in Table 9.3.1 goes from 1 to 120;  $dfe$  is given as 20, 60, 120, and infinity. It turns out that increasing  $dfe$  has very little effect on  $\lambda$  after 60 or so. For example, for  $dfh = 1$  and  $\lambda = 14$ , power at  $dfe = 60$  is .86, power at  $dfe = 120$  is .87, and power at  $dfe = \infty$  is .88. Thus, any  $dfe$  over 60 is probably going to be a waste of resources, in terms of power. On the other hand, for  $dfe = 60$ ,  $\lambda = 14$ , power at  $dfh = 1$  is .80; for  $dfh = 20$  power is .11, and for  $dfh = 60$  power = .03. Increasing the size of the hypotheses ( $dfh$ ) has a huge effect on power: from .80 down to .03. That is a change from an 80% chance of finding something all the way down to a 3% chance. In the maximum tabled entry ( $dfh = 120$ ,  $\lambda = 40$ , a really huge effect) even an infinite  $dfe$  will provide power of only .51. No matter how large the experiment, if the size of the hypotheses is large, the size of the experiment is pretty much irrelevant, as far as power goes. The practical implication is this: Increasing the number of cases in an experiment is unlikely to compensate for increasing the size of the hypotheses. Increases in the size of the hypotheses will greatly increase the chances of missing something, and calling for more data is just not going to help.

For the simple regression model there is only one dependent variable. We once reviewed a paper with 50 dependent variables and one independent variable. The math says nothing about how many of each kind of variable is acceptable. If there is more than one such variable, then set correlation (Cohen, 1988, Chapter 10) is often applicable. The math here goes beyond the scope of this paper. Instead of calculating  $R^2$ , statistics are based on ratio's of determinants of correlation matrixes (Cohen, 1988, p. 468) and significance tests use Wilks  $\Lambda$  or Rao F. Without the math, we can only make a general statement about multivariate models: all the problems with  $dfh$  and  $dfe$  present in the univariate models are present, only more so. Researchers wishing to use multivariate models will probably be expected to include a power analysis in the first draft to make sure the experiment is sized correctly.

There are some other noteworthy consequences of using multiple degree-of-freedom tests. Consider the following example. Suppose one were interested in predicting visual complexity. Inspection of the literature generated seven possible design features that could be influential, and so an experiment is conducted on 10 street scenes. Suppose the result were  $R = .90$ , a very impressive figure. Now suppose some fool takes it into his head to attempt replication using different variables such as number of oranges grown from 1988 to 1997, acreage of zoos in the United States, number of new Broadway productions, 1959 - 1968, zip codes of military bases, January rainfall in 10 cities, age of United States Presidents at time of inauguration, and time of Venus rising, first listed day of the month, January - October 2000. It turns out that these variables (all selected at random from *The World Almanac* (Famighetti, 1999) also predicted the findings for visual complexity rather impressively ( $R = .85$ ). Inclusion of one more nonsense variable (magnitudes of major earthquakes, 1963 - 1989) bumped  $R$  up to .90 — the same result reported in the original study. What happened?

The expected value for  $R$  increases as the number of variables increases, regardless of what the variables are. Testing  $dfh = 7$  on 10 cases is going to result in claims such as visual complexity is related to zip codes of military bases. (That is, in fact, another example of a classic alpha error.) One solution is to use a shrunken  $R$ ; another solution is to focus the experiment by designing it explicitly to test for a known number of claims, desired levels of  $\alpha$  and  $\beta$ , and an explicit effect size. Cohen and Cohen (1993, pp. 169-171) provide some relevant

advice in their *Section 4.6.2 Less is More*. We can probably all guess what that section's conclusion is.

Focusing has another useful consequence: removing redundancies. Take the case of studying effects on the three responses of pleasure, preference, and liking. We have already seen that using three dependent variables instead of one is going to cost (a) a lot more work, and (b) a much greater chance of reporting noise. The relevant question then becomes "When will we need multiple measures?".

Here is one way to think about it. If we are attempting to measure one thing with multiple measures, then the measures should be highly correlated, else they would be measuring different things. For instance, ratings of pleasure, preference, and liking typically correlate above  $r = .90$ . Multiple measures are often used to increase the reliability of the whole set of measures. The Spearman-Brown formula is applicable (Rosenthal, 1987, pp. 10-11). Let's look at Table 1. If the average correlation between measures is .90, then the effective reliability of two such measures is .95; of 4 measures; .97; of 8 measures, .98; of 16 measures, .98. There is (a) little increase in effective reliability and (b) the marginal improvement is diminishing. On the other hand, the required sample size, even if we are so generous as to look for a medium sized effect, a simple model with only two independent variables, and not even think about interactions between or among the variables, all that being forgiven, the required sample size still keeps going up. So there would have to be a really strong case for deciding that raising the reliability of the measures would be worth the increase in experimental effort.

Table 1. Consequences of using multiple measures on reliability and sample size

Number of Measures	Collective reliability ( $\bar{r} = .90$ )	Collective reliability ( $\bar{r} = .30$ )	N
1	.90	.30	80
2	.95	.46	93
4	.97	.63	146
8	.98	.77	186
16	.98	.87	266

$\bar{r}$  is the average correlation between measures.

N is the approximate total number of cases for  $\alpha = .01$ ,  $\beta = .20$ , from Cohen (1988, Chapter 9)

On another hand, if the reliability between the measures were .30, then the effective reliability of multiple measures will increase (from .30 to .87 for this example). Now, however, there is a logical problem. If the measures are not highly correlated, they are indicating different things, and so it would be confusing to lump them into the same concept. There might be cases where one might go ahead and do this. For instance, if one is attempting to predict a measure that won't be available until later (acceptance into graduate school, job performance), then it might be worthwhile to assemble a large quantity of loosely related predictors. For inquiry designed to support theory though, the implication seems to be that when multiple measures are highly correlated they will be logically useful but will have negative impacts on the size of the experiment for very little gain in reliability, and, conversely, when the measures are not highly correlated, they will help statistically but confuse the logic. Thus, for the purpose of creating a collective body of knowledge, either way, one is probably better off by staying away from multiple measures if at all possible.

In some forms of knowledge, such as narratives, more and more detail is considered to be a good thing. Research suggests that amount of detail is one of the factors people use to ascertain believability. Completeness is another factor (Pennington & Hastie, 1993). Perhaps the greatest advocate for copious detail was Edmund Husserl (1962/1913). Among his prolific writings was the idea of "adequacy", which, if we understand it, seemed to be that sheer quantity of descriptors and views was good, regardless of whether they made any sense or were internally coherent (Husserl, 1962/1913, p. 48). A classic simple example is the Necker Cube (Figure 2), which can be interpreted as facing upwards, or downwards, or as a flat snowflake, the top (or bottom) of an umbrella, etc. (Ihde, 1986). The principle seems to be, following May West, "Richer is better".

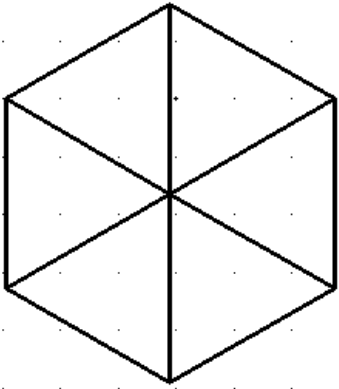


Figure 2. A Necker cube.

Is richer better? Perhaps yes, perhaps no. Readers with backgrounds in anthropology will probably be familiar with "adequacy" under labels such as "dense" or "thick" descriptions. For some purposes (small talk, understanding world views of other cultures, trying to reach the perfect, complete description of an individual object or event), richness may be appropriate. However, anyone who has soldiered his or her way through even a fraction of Husserl's 40,000+ pages of dense, thick prose will probably wonder if there might be a point where all this richness begins to be a bad thing. For those who use probability theory to make decisions (which includes everyone in the intended audience for this paper), the answer is yes, early and often. While massive detail is good for a story, it is not good for a theory

and is certainly not good in sizing experiments. If one is making decisions based on empirical data and probability theory, "adequacy" means redundancy at best. More typically, richness generates very large increases in  $\alpha$  errors,  $\beta$  errors, and logical confusion.

If one is trying to synthesize a reliable collective body of knowledge that applies to more than one instance, then multiple, overlapping measures will greatly complicate the tasks of backtracking and replicating studies. (The math behind backtracking is given in Lawvere and Schanuel (2000)). More mathematically inclined readers will also be familiar with the problems of multicollinearity, ill-conditioning, and singular matrixes (Nering, 1974). These are all manifestations of the same issue: redundancy. For readers versed in information theory (Cover & Thomas, 1991), the amount of information communicated by a redundant signal is zero. Take, for example, the task of trying to identify a person. If one knows the person has very large shoes, that knowledge will help to distinguish the person from other people. However, additional knowledge that the person also has very large socks won't make the discrimination much easier, because large socks nearly always go with large shoes. Shoe and sock size are highly correlated; either one is useful; both together are redundant. The amount of work required for data collection, writing, reading, and storage has doubled and the number of possible hypotheses has increased with the power set of number of variables, all with virtually no benefit in terms of identifying the person. Mathematically one not only does not need multiple measures, but is actually much better off without them.

There is a choice here that is covered up by the jargon "multiple degrees-of-freedom tests". On one hand, there is richness, completeness, thickness, density. On the other hand, there is, to quote Ockham "It is futile to do with more what can be done with fewer". (Adams, 1987, p. 156). Or, to focus on the essentials, the choice is:

### Richness or Parsimony?

#### *Experimental design*

Parenthetically, many papers sent to me to review also have weak experimental designs. This is important because any statistical test, no matter how simple or complex, will be invalid if the experimental design is inappropriate. For instance, a study on perception of the North Star will be invalid, no matter how many subjects or how fancy the computer runs are, if the observations are collected at the South Pole. This may seem trite unless you have recently reviewed a paper that was supposed to be on behavior differences for two sets of political beliefs but the results were reported as differences in income. There is no remedy for such a mismatch between question and experimental design.

One often-used approach to this problem is just not to have an experimental design. Data are collected and elaborate analyses such as clustering, multidimensional scaling, or regressions are run. These methods are exceptionally difficult to defend against alpha errors because of the large number of alternatives investigated. A recent example is a paper attempting to investigate all factors, levels, and interactions of four factors, two of which have two levels and two of which have four levels. The number of possible tests was on the order of  $2^{14}$ , which meant that each effect had to be significant at  $p < .000002$  in order to be safe from reporting noise. That is quite difficult to achieve.

Another commonly-used approach is to increase the number of participants. This has the effect of increasing *dfe*, which in turn increases the power of a statistical test. But, as we have seen above, not by much. Power is influenced much more by number of hypotheses rather than *dfe*, so merely increasing the number of participants quickly reaches its asymptotic utility. In addition, large numbers of participants can be a misleading indicator of experimental quality. For example, suppose a researcher wanted to test the hypothesis that all shrines are white. Accordingly, the researcher goes to the Taj Mahal and finds that it is white. In order to be sure of the hypothesis, another 1000 participants are recruited and asked if the Taj appeared white. The results seem quite solid with  $n > 1000$ . Then the researcher goes to the Viet Nam Memorial. What happened?

Behavioral scientists might be used to thinking of "*n*" as "*nsubj*". But that need not be the case. In statistics, a more informative term for "*n*" is "*dfe*", or the error degrees of freedom. If one is comparing two groups of people, then *dfe* will indeed be a function of number of participants. On the other hand, for repeated measures experiments, *dfe* for differences between stimuli is a function of the product of *nsubj* and *nstim*. Thus, an experiment with *nsubj* = 30 and *nstim* = 30 will have *dfe* on the order of 900, not 30. Also, if one is attempting to generalize over environments, the relevant analysis is variance components, and the crucial quantity is the number of environments, not the number of subjects or number of responses. Consider the

following analogy. Suppose one were interested in the effects of classroom color on cancer rates in children. The research protocol was to select one child from each of four classrooms and have 400 doctors rate the four children on 50 symptoms. Does that protocol sound more solid than having 400 children selected at random from 40 classrooms and running simple contrasts in rates of cancer on-set between types of classrooms? If one were interested in reporting demographic effects in the evaluation of one particular project — say the reconstruction of the World Trade Center in New York — then indeed *nsubj* would be crucial. However, if one is interested in differences due to general categories of environments (color of classroom, etc.), then the bottleneck is likely to be *nstim*. Readers interested in the mathematics of this line of thought are referred to the literature on variance components, both mathematical (Burdick & Graybill, 1992; Searle, Casella, & McCulloch, 1992), and applied (Deming, 1994; Juran, 1992).

### *Meta-analysis*

Meta-analysis is where it all comes together. A simple interpretation of meta-analysis is that it permits us to do statistical tests with data from multiple experiments. Thus, for example, if any particular individual cannot do a definitive experiment because the cost of reaching  $\alpha = .05$  and  $\beta = .80$  is too high or the target effect size is very small, then, through the magic of meta-analysis, the data from one experiment can be combined with data from other experiments so the combined data are definitive. Another example is to test between-experiment factors. We all can do within-subject and between-subject analyses; now we can also do between-experiment analyses. So if there are ten studies in the literature on correlations between reading and math in red classrooms, and another ten studies reporting correlations between reading and math in yellow classrooms, then the effect of color in classroom on the correlation between reading and math can be calculated. These uses are all well-documented in the literature (Hedges & Olkin, 1985) so we will not repeat the details here.

Another way to view meta-analysis is through the framework of Francis Bacon (1605). Bacon was writing at the end of the 16th. century. The topic of interest was how to create a reliable, collective body of knowledge that was not based on authority or personal experience, but rather on empirical evidence and reporting methods of obtaining that evidence well enough so skeptical readers could test their skepticism through replication. After all, if some stranger came up with a nonsense claim, it could be due to all manner of nefarious influences; but if one got the same answer in the privacy of one's own lab, the results would be harder to fault. Bacon's label for a research community that asked questions and accepted answers on the grounds of empirical experiments was the "New Atlantis" (Bacon, 1980).

I would now like to propose a mathematical abstraction of the New Atlantis: a meta-analytic structure. The inputs are (a) a claim for which effect sizes can be calculated, (b) the effect size, and (c) the  $n$  upon which the effect size was calculated. A basic meta-analysis then looks something like Table 2

There are at least two ways in which it makes a great deal of sense to use meta-analysis in the allocation of research resources. The first is choosing a target effect size for an experiment. Please recall that the size of an experiment can be calculated given values for  $\alpha$ ,  $\beta$ , and effect

size. If one is guessing about required sample size, then it might make senses to "be conservative" or "be comprehensive". If, however, one already has some idea of the effect size, then sizing an experiment becomes a simple look-up. A meta-analytic literature review would provide the collective estimate for the target effect size. For example, if the collective estimate were  $r = .10$ , then the size of an experiment would be on the order of  $n = 600$  (Cohen, 1988, p. 87), but if the collective  $r$  were  $.50$ , then the new experiment would need only  $n = 23$ . Table 2 shows how meta-analysis can be helpful in calculating sample sizes for a prospective experiment. In Table 2, each study is represented with  $n$  and an effect size, in this case  $r$ . The collective estimate of the correlation ( $\hat{r}$ ) is shown after each study is incorporated into the collective body of knowledge. Thus, after study 1, the collective  $r$  is  $.92$ ; after study 2, the collective  $r$  is  $.91$ , etc. If one had access to a published meta-analysis such as the one shown in Table 2, the target effect size would be, at a minimum,  $r = .91$ , so power of  $.80$  could be reached with as small an  $n$  as 8. So, it may or may not be the case that large  $n$  is needed. Without meta-analysis we are stuck with trying to be safe and taking the risk of wasting resources; with meta-analysis, we know how many data will be needed for a given experiment and thus can allocate our time and money accordingly.

Table 2. A basic meta-analysis

Study	Individual Results		Collective Results			
	$n$	$r$	$\Sigma n$	$\hat{r}$	ci	$n_{over}$
1	8	.92	8	.92	.61, .98	15
2	8	.89	16	.91	.71, .97	90
3	8	.97	24	.93	.83, .96	315
4	12	.89	36	.92	.53, .97	693
5	21	.97	57	.94	.91, .99	2433
6	900	.99	957	.99	.99, .99	>e 10
7	80	.93	1037	.99	.99, .99	> e10
8	113	.99	1150	.99	.99, .99	> e 10

$n$  and  $r$  are the entries for each experiment.  $n$  is the sample size upon which  $r$  was calculated. The collective results are  $\Sigma n$  (the sum of all sample sizes up to and including the line upon which it is shown,  $\hat{r}$ , the estimate of the collective correlation, ci, which is the confidence interval on the collective correlation, and  $n_{over}$ , which is the sample size for a new experiment that, would a finding of  $r = 0$ , would make the collective ci include zero.

The second way in which meta-analysis makes a great deal of sense is measuring the effect of an individual experiment on the collective body of knowledge. Traditional scientific evaluation would involve inspection of each study and accepting a finding if and only if it achieved  $p < .05$ . With meta-analysis, the focus changes. Instead of asking whether a particular finding for a particular experiment were "significant", the new method is to calculate the effect of the new findings on the collective body of knowledge. Here we have an explicit connection with Lord Bacon: we are all working as a community to create a collective body of knowledge. Formally, the collective body of knowledge is represented by the four columns on the right side of Table 2.  $\Sigma n$  is, as one might expect, is the running total for  $n$ .  $\hat{r}$  is the collective estimate for the correlation, and "ci" is the confidence interval for that collective correlation. Thus, after the first study, the collective  $r$  was  $.92$  and the ci was  $[.61, .98]$ . After study 8, the collective  $r$  was  $.99$  and the ci was  $[.99, .99]$ . The last column in Table 2 is labeled " $n_{over}$ ". It is the amount of new data that would be required to make the whole collective  $r$  non-significant. For example, if

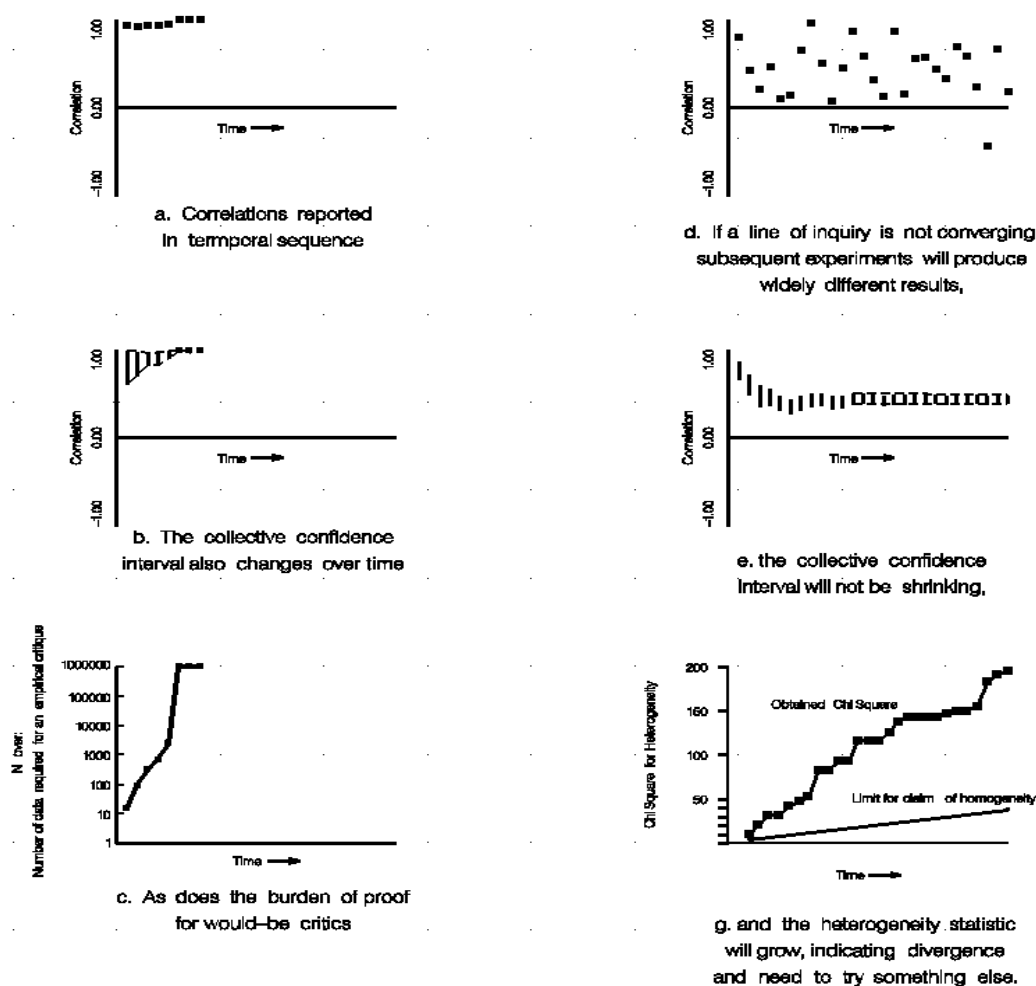


Figure 3. Planning research in terms of a whole line of inquiry instead of deciding if findings from single experiments are "significant" leads to much more information. If (a) effect sizes are reported in temporal sequence, the effect of each study on the collective ci can be calculated (b) as well as the burden of proof for possible critics (c). If a line of inquiry is not panning out (d), that will show up in a non-converging collective ci (e) and increases in heterogeneity (g).

study 2 had reported  $r = 0$  on  $n = 15$ , then the collective ci would include zero and the implication would be that there was not enough data to distinguish the claim from noise. For readers of Bacon, an alternate interpretation is the burden of proof for opposing parties. A valid empirical critique for the example in this table would have to be supported by a study with  $r = 0$  and  $n \gg 1,000,000$ .

Even this simple meta-analysis is enormously useful in planning and reviewing research. Recall that both planning and reviewing involve allocation of resources: time and money for doing the research; pages or money for journal or grant review. For example, Figure 3 shows the numbers in Table 2 in temporal sequence.

In Figure 3a all the findings are quite similar. It's as if you got one result, your buddy got the same result, his or her buddy got the same result, and so on. But in meta-analysis the focus is on the collective results, here represented by the collective  $ci$  (Figure 3b). Here it is pretty clear that the collective  $ci$  is converging nicely on its asymptotic value. What does that mean? Stick a fork in it; it's done. Time to move on to another question. The reason is that, with a tight  $ci$ , the amount of additional data needed to influence the collective  $ci$  is huge. An estimate of that amount of work is given by  $n_{\text{over}}$ . As can be seen in Figure 3c, converging sequences of results means increases in  $n_{\text{over}}$ . The more solid the collective finding, the harder it will be to overturn that finding empirically. (More precise estimates can be obtained by power analysis, but that is beyond the scope of this paper.) Thus, if faced with a choice between doing yet another study on this topic with  $n = 1,000,000$  or working on all the other bright, new shiny ideas you could try with the same effort, which choice seems like the better move?

The same reasoning, of course, applies to journal or grant review. If the current values for a relationship between  $x$  and  $y$  were  $r = .830$  on  $n = 185$ , the  $ci$  would be  $[.779, .870]$  and the  $n_{\text{over}}$  would be just about 12,000. Suppose the reviewer had the capacity to do basic meta-analyses. (It can be done on a spreadsheet, so this supposition is not so far-fetched.) If a submission comes in claiming that there were no relationship between  $x$  and  $y$ , the reviewer would send the draft back asking that the author(s) please provide the experiment with  $n = 12,000$  and  $r = 0$  that would substantiate their claim. If a submission came in reporting that a correlation of  $.80$  was found on  $n = 20$ , the reviewer would combine the new findings with the collective body of knowledge and find that the collective  $r$  would change from  $.830$  to  $.827$  and the  $ci$  would change from  $[.779, .870]$  to  $[.778, .866]$ . Assuming the study were otherwise solid, the reviewer would be likely to suggest that the effect of the submission on the collective body of knowledge was quite small and that precious journal papers would be better allocated if other work were published.

Table 2 represents a researcher's dream sequence. Not only did the researcher get strong results, so did a lot of his friends, and the amount of data needed for a valid empirical critique is so large that no one is likely to try an empirical challenge. But Table 2 is not the only possible temporal sequence. Take a look at Figure 3d. In the top row it is clear that the findings have been quite disparate. With sufficient sample size, each correlation would be "highly significant" and so would be accepted for publication under the old standards. However, Figure 3e shows the effect of each study on the collective  $ci$ . About half-way through this sequence, the  $ci$  reached its asymptotic limit. The last half of the studies had literally no effect on the collective effect size. The  $n_{\text{overs}}$  for this particular set of data were quite large (18,000 by the 10 th. study; over 100,000 by the 26 th. study), so empirical scientists would have trouble maintaining that there was no effect. On the other hand, there is also measure for heterogeneity among findings in meta-analysis. It is a chi square variable and, if significant, suggests that a single overall collective estimate is not appropriate and the line of work should be changed to focus on different conditions (Hedges & Olkin, 1985, p. 122-128). Figure 4e shows the running values for the heterogeneity chi square. It has been clear that, from the beginning, it would have been better if this line of inquiry had been split into more focused studies. In the review context, another paper in this line of work is likely to meet the objection that, even after extensive trials, results for this claim are not converging, and so why should more journal pages be spent on it?

Thus, the temporal sequence shown in Figure 3d represents, to a researcher, a really bad if not the worst scenario. Probably the only thing that could make the scenario worse would be if some politically challenged researcher pointed it out in print. An idea sometimes worked and sometimes didn't work. The limits of the idea were pretty firmly established a dozen studies ago, so all that subsequent energy could have been spent on more promising ideas. Without meta-analysis, researchers can proceed along for decades publishing findings that, individually are "significant", but are, collectively, redundant or discrepant. The advantage of meta-analysis, under these conditions, is that it can be used as a running report card, so if this miserable scenario is developing, the researcher will know about it right away and can take effective action.

Accordingly, meta-analysis makes sense at least two ways: sizing prospective studies, and changing the focus of inquiry from achieving " $p < .05$ " to the statistical synthesis of a collective body of knowledge. And that is why we think this paper is, if not a full pair of dime shift, at least a two-cents shift in thinking about research protocols.

### Summary

For those whose train of thought derailed somewhere back there, here are the highlights:

To fish or not to fish, that is the question.  
Whether't is noblier in the end, to suffer  
the slings and arrows of self-righteous critics,  
or have power against a sea of quibbles  
and by opposing, end them?

1. The protocol of accepting or rejecting claims based on the criterion of " $p < .05$ " leads to major errors and should be replaced with effect sizes and their confidence intervals.
2. Use of effect sizes requires understanding of the basic statistical concepts of alpha error (probability of reporting noise), beta error (probability of missing something), power analysis (probability of finding something), focus (*dfh* very low) and meta-analysis (synthesizing statistical findings over more than one study).
3. Contemporary standards for empirical behavioral science require reporting of effect sizes and power, so the peer review process will be expedited by authors providing those numbers in the first draft.
4. Meta-analysis changes the focus of inquiry from individual, isolated findings to the collective body of knowledge.
5. Use of the contemporary statistical protocols will greatly enhance the productivity of planning, reviewing, and publishing research. Researchers desiring to avail themselves of all the advantages of the contemporary protocols can find the most effective help in three books:

Rosenthal and Rosnow (1991), Cohen (1988), and Hedges and Olkin (1985). Researchers whose basic concepts require complex inter-relationships will also need Cohen and Cohen (1993).

The methods proposed here will not provide totally complete answers to all questions. Rather, they accept limitation on scope in return for reliability of results. Researchers striving for total completeness might be better served by narrative studies or rhetorical discourse. Nor can statistical methods promise perfect predications, only probabilities. Empirical science, at least down at the bench level, is very much like gambling; one gets an idea, one designs an experiment and a test, and one takes one's chances. The advantage of using probability theory is that one has a clear idea of what these chances are, and that is highly useful at both individual and collective levels of science. As Kenny Rogers nearly wrote in "The Researcher":

You gotta  
 Know when you're cold,  
 Know when to shut up.  
 Know when to talk away.  
 Know when you've won.  
 Don't count your claims,  
 while sitting at the keyboard.  
 There'll be time enough for gloating,  
 when the study's done.

### References

- Adams, M. M. (1987). *William Ockham* (Vol. 1). Notre Dame, Indiana: University of Notre Dame Press.
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (Fifth ed.). Washington, D.C.: American Psychological Association.
- Bacon, F. (1605). *Advancement of learning* (1952 ed.). New York: Collier.
- Bacon, F. (1980). *The great instauration and new Atlantis*. Arlington Heights, Illinois: AHM Publishing Corporation.
- Burdick, R. K., & Graybill, F. A. (1992). *Confidence intervals on variance components*. New York: Marcel Dekker.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Cohen, J., & Cohen, P. (1993). *Applied regression/correlation analysis for the behavioral sciences*. Hillsdale, New Jersey: Lawrence Erlbaum.
- Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.
- Deming, W. E. (1994). *The new economics for industry, government, education*. Cambridge, MA: Center for Advanced Engineering Study, MIT.
- Famighetti, R. (Ed.). (1999). *The World Almanac and Book of Facts 2000*. Mahwah, New Jersey: World Almanac Books.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, Florida: Academic Press.

- Husserl, E. (1962/1913). *Ideas: general introduction to pure phenomenology* (B. Gibson, Trans.). London: Collier MacMillan.
- Ihde, D. (1986). *Experimental phenomenology*. Albany, New York: State University of New York.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions Volume 2*. New York: Wiley.
- Juran, J. M. (1992). *Juran on quality by design: the new steps for planning quality into goods and services*. New York: The Free Press.
- Lawvere, F. W., & Schanuel, S. H. (2000). *Conceptual mathematics: a first introduction to categories*. Cambridge: Cambridge University Press.
- Nering, E. D. (1974). *Elementary linear algebra*. Philadelphia: W. B. Saunders.
- Pennington, N., & Hastie, R. (1993). The story model for juror decision making. In R. Hastie (Ed.), *Insider the juror: the psychology of juror decision making*. Cambridge: Cambridge University Press.
- Rosenthal, R. (1987). *Judgment studies: design, analysis, and meta-analysis*. Cambridge: Cambridge University Press.
- Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: methods and data analysis*. New York: McGraw-Hill.
- Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.
- Stamps, A. E. (2000). *Psychology and the aesthetics of the built environment*. Norwell, MA: Kluwer Academic.
- Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design*. New York: Mc Graw-Hill.

---

<sup>1</sup> The third equation uses the identity that  $F = t^2$  for  $dfh = 1$ .