

Running head: SOME IMPLICATIONS OF FACTORING COST

Some Implications of Factoring Cost into Statistical Power Calculations

James L. Dannemiller  
Rice University

Ronald C. Serlin  
University of Wisconsin - Madison

## Abstract

The costs of making Type I and Type II errors were factored into power and sample size calculations for testing a simple directional hypothesis regarding the population mean. These costs were expressed relative to the cost of testing a single subject with any startup costs folded into this unit subject cost. The expected monetary loss from an experiment then involves the cumulative cost of testing the subjects as well as the cost of each error weighted by its probability of occurrence. Type I errors lead to a linear loss function with increases in sample size. Type II errors lead to a nonlinear loss function with a minimum expected loss; this implies that there will be an optimum sample size. This optimum sample size can be calculated with numerical approximation to the derivative of the loss function. The power and minimum expected loss can be calculated for this optimum sample size. Testing fewer or more than this optimum number of subjects results in experiments that are less cost effective than is theoretically possible. Alternatively, this approach also shows that if the cost of a Type II error is taken to be very much larger than the cost of testing a single subject, then this optimum sample size vanishes and an experimenter should always test as many subjects as possible. A distinction is made, therefore, between the costs of experiments designed primarily to yield new knowledge, and experiments designed to determine the benefits of a new therapy or drug relative to the best existing therapy or drug. Several limitations of the current analysis are discussed.

### Some Implications of Factoring Cost into Power Calculations

Most experimental psychologists have had the following intuition at some point or another in the course of doing their research: “I know that I could gain power by testing additional subjects, but testing additional subjects has real costs. There must be a sample size at which the gain in power from testing an additional subject can no longer be justified given the cost of testing that subject.” Having spent time looking at curves of power versus sample size, the investigator knows that the marginal gain in power diminishes with increasing sample size to the point that adding more power by testing more subjects is no longer cost effective. But how does the investigator determine this break-even sample size? Intuitively, if the experimenter could determine the sample size at which the marginal gain in power from testing another subject drops below the marginal cost of testing that subject, then at least in purely economic terms, s/he would have found the optimal sample size.

What is required for the concept of an optimal sample size to make sense? In addition to knowing the unit cost of testing a single subject and the fixed startup costs of designing and implementing the experimental protocol, there are two other critical pieces of information: the costs of Type I and Type II errors. In experimental psychology, we typically do not explicitly put a monetary value on these errors when calculating power. There is nothing, however, that prevents us from doing so. It will be easier in some cases than in others to specify these costs. Our purpose in writing this article is to show that when the costs of these two errors are taken into account, the concept of an optimal sample size emerges. This exercise also shows the conditions under which it makes sense simply to test as many subjects as possible.

Sample sizes typically are determined in Null Hypothesis Significance Testing (NHST) by specifying a desired power, typically 0.80 or more, and using power tables to find the sample size necessary to achieve that power given an effect of a fixed size and a Type I error rate. One

could argue that the costs of Type I and Type II errors are implicitly taken into account in this procedure by choosing the probabilities for these errors. In other words, if it is extremely important to find a predicted effect or the gain in knowledge from finding such an effect is considered to be very valuable, then the power will probably be set at a higher level than in an experiment in which finding a predicted effect is not as important either theoretically or practically. All other things being equal, the most straightforward way to increase power is by increasing the sample size. Larger sample sizes will be necessary when the value of correctly rejecting a false null hypothesis is considered to be high. Larger sample sizes are also costlier than smaller sample sizes if the unit cost of testing a subject remains the same independently of sample size. But now the experimenter faces the dilemma described above: S/he knows that testing additional subjects brings with it diminishing returns in terms of additional power, so at what sample size will the marginal gain from testing one more subject drop below the cost of testing that one additional subject? Is there a sample size beyond which the increase in power is not worth the additional cost of testing those subjects, and if so, then how is this sample size related to the costs of Type I and Type II errors?

Loss and gain can be explicitly factored into the design of experiments when power and sample size are taken into consideration provided that monetary values can be placed on Type I and Type II errors. The three costs treated here are a) the unit cost of testing a subject (assumed to be constant and independent of sample size, b) the cost of making a Type I error, and c) the cost of making a Type II error. Alternatively, it is also possible to consider the complements of the last two costs: the gains from correct decisions. The analysis below shows that if values can be assigned to these costs or gains, and if the experimenter takes as one of his/her objectives to maximize the expected monetary gain from the experiment or to minimize the expected monetary loss, then there will be an optimum sample size that will accomplish these goals.

Testing fewer subjects than this optimum cannot be justified because doing so will lead to less than the maximum expected gain from the experiment. In this case the investigator will have stopped at a sample size for which the marginal gain from testing one more subject remains above the marginal cost of testing that subject. Alternatively, and perhaps more importantly, testing more than this optimum sample size cannot be justified because the marginal gain from testing one more subject will be less than the marginal cost of testing that one additional subject. A rational experimenter would not proceed beyond this optimal sample size because it would literally be a waste of money to do so. This implies that power calculations that do not take into account the absolute or at least the relative costs of making Type I and Type II errors or the gains from avoiding those errors will lead to experiments that are not as cost effective as would be theoretically possible.

Nam (1973) has shown that there is an optimum ratio of sample sizes that maximizes the power of a test comparing treatment and control groups when the costs of treatment differ for the two groups. In this case the optimum ratio of sample sizes can be computed under the constraint that the total cost is some fixed value. Nam's approach did not explicitly take into account the costs of making Type I and Type II errors, but it is illustrative of the fact that when cost is considered in conjunction with power, optima on sample size can result. Maurice (1959) and Colton (1963) showed that there was an optimum sample size when the objective was to pick the better of two treatments based on random assignment to these treatments of  $2n$  of the subjects from the total available pool of subjects,  $N$ . In this case, after the  $2n$  subjects have been tested, and the experimenter has chosen the better of the two treatments based on the data from that sample, the remaining subjects in the pool are assigned to the chosen treatment. If the loss from assigning a subject to the inferior treatment is proportional to the absolute difference between the two treatment population means, then there will be some optimum sample size,  $2n < N$ , that will

minimize the total expected loss. The key to understanding this result is that there will always remain a finite probability that the inferior treatment will be chosen as the superior treatment when the sample sizes are finite. The experimenter chooses the sample size for the original test knowing that there is some nonzero chance that s/he could choose the inferior treatment. A similar situation arises when one considers statistical decisions from a Bayesian point of view. In this case, it is natural to think about minimizing the average or maximum risk of various decisions by taking into account the losses associated with the various decisions and their associated probabilities (Haussler & Opper, 1997).

Jerrell (1988) showed that when costs are placed on the two types of errors relative to the cost of testing a single subject, and a cost function is computed that weights the costs of the two error types by their respective probabilities of occurrence, then clear optima emerge on sample size. The experimenter's objective was taken by Jerrell to be that of minimizing the cost function. This is precisely the approach that we have taken here. The purpose of Jerrell's paper was to present a brief computer program that students could use to explore the relations between sample size and cost. We expand Jerrell's treatment by providing the mathematical development of the cost function and by pointing out the conditions under which an optimal sample size will exist. In doing so, we are forced to consider the value of rejecting a false null hypothesis in various types of research endeavors.

The costs of incorrect decisions have been treated previously in domains such as deferred decision making and sequential testing with optional stopping (Busemeyer & Rapoport, 1988; Lewis & Berry, 1994; Grundy, Healy, & Rees, 1956; Rapoport & Burkheimer, 1971). In the sequential testing literature, the final expected loss from a clinical trial is taken to be the cumulative cost of testing the patients enrolled in the study plus the cost of making an incorrect decision weighted by the probability of that decision. The same approach can be used in a more

classical hypothesis testing framework, and this is illustrated below for the case of a single sample test of the population mean. This is not intended to be an exhaustive treatment of errors, power and cost. Rather, we use this easily understood example to draw out the implications of considering the various costs involved in testing a simple hypothesis.

### Analysis

Consider the simple case of an experimenter who wishes to test a one-tailed alternative hypothesis that the population mean,  $\mu$ , is greater than 0. The null and alternative hypotheses for this situation are:

$$H_0: \mu \leq 0 \quad (1)$$

$$H_1: \mu > 0 \quad (2)$$

To calculate the power for testing this statistical hypothesis, the experimenter needs to fix the Type I error rate,  $\alpha$ . The choice of  $\alpha$ , of course, will depend on how costly the experimenter considers a Type I error. This will be discussed in more detail below when costs are explicitly factored into the design of the experiment.

Once  $\alpha$  is fixed, the experimenter needs to assume an effect size to be able to set the desired power and calculate the required sample size for achieving that power. Power is just the complement of the Type II error rate,  $\beta$ , so once again the experimenter presumably implicitly sets the desired power depending on how costly s/he considers a Type II error. Assuming a population standard deviation of  $\sigma$ , the effect size can be represented as

$$e = \mu_1/\sigma \quad (3)$$

where  $\mu_1$  represents one of the range of population means consistent with  $H_1$ .

To simplify the exposition, we may assume that  $\sigma$  equals 1. In this case, the critical value of the  $z_\alpha$  statistic required for rejecting  $H_0$  in favor of  $H_1$  and the critical value of the sample mean determined by  $\alpha$  are related simply by

$$\bar{X}_\alpha = \frac{Z_\alpha}{\sqrt{n}} \quad (4)$$

where  $n$  is the sample size. Additionally,  $\mu_1$  is equal to the effect size,  $e$ , under these circumstances. The probability of a Type II error,  $\beta$ , in this case is simply the proportion of a normal distribution with mean and standard deviation of  $e$  and  $\sigma/\sqrt{n}$ , respectively, below  $\bar{X}_\alpha$ , and the power is the complement of this proportion. Call these two proportions,  $\beta$  and  $1 - \beta$ , respectively.  $\beta$  is the probability of a Type II error given the effect size,  $e$ , the sample size,  $n$ , and  $\alpha$ . The experimenter usually sets  $1 - \beta$  at some minimally acceptable level or chooses it to reflect the cost associated with making a Type II error and calculates the required sample size.

Now there are two possible states for the null hypothesis,  $H_0$ ; either  $H_0$  is true or  $H_0$  is false. We will ignore in this note the argument that  $H_0$  is never strictly true (e.g., Cohen, 1990). What is the expected loss when  $H_0$  is true? The only error possible when  $H_0$  is true is a Type I error, so it is easy to calculate the expected loss in this case. If we assume that the unit cost of testing a subject is  $c$ , and the cost of making a Type I error is  $d_1$ , then the expected loss is simply

$$L_1 = \alpha \cdot d_1 + n \cdot c \quad (5)$$

The expected loss involves the cost of a Type I error weighted by the probability of its occurrence as well as the cumulative cost of testing the subjects. When  $H_0$  is true, then the expected loss is directly proportional to the sample size and the expected loss function is linear in  $n$  because  $\alpha$  is fixed by the experimenter and does not depend on sample size. This makes intuitive sense because when  $H_0$  is true, then testing the first subject results in an immediate monetary loss, and the loss is directly increased as more subjects are added to the sample. Of course, the experimenter cannot know that  $H_0$  is true, so this loss is unavoidable under these circumstances.

Now consider the second possibility for  $H_0$ ; it is false. In this case, the only error that the experimenter can make is a Type II error of failing to reject the false null hypothesis. The expected loss in this case is

$$L_2 = \beta \cdot d_2 + n \cdot c \quad (6)$$

At first, this looks like the expected loss function when  $H_0$  is true, but  $L_2$  is more complicated than  $L_1$ .  $\beta$  is a nonlinear function of the sample size,  $n$ . The first part of the expected loss function to the left of the addition sign is a nonlinear function of  $n$ , and the second part is once again linear in  $n$  because adding more subjects to the sample increases the cost directly in proportion to the increase in sample size. Thus, the expected loss when Type II error is considered is a nonlinear function of the sample size.

What does this expected loss function,  $L_2$ , look like? Assume that the unit cost of testing a subject is 1, the cost of a Type II error is 100 (i.e., equivalent to testing 100 subjects; this will be discussed below), the effect size is 0.50, and  $\alpha$  is set at .05. Figure 1 shows the expected loss function when  $H_0$  is false (dashed line). Notice that this function has a minimum. This means that there is an optimum sample size for minimizing the expected loss. If the experimenter tests more or fewer than this optimum number of subjects, then s/he can expect to incur greater costs in the long run than are absolutely necessary.

-----  
 insert Figure 1 about here  
 -----

For the case shown in Figure 1, testing 30 to 31 subjects will minimize the expected loss given that a Type II error is 100 times as costly as testing a single subject. Also shown in Figure 1 (dashed line) is the expected loss function when  $H_0$  is true and the cost of a Type I error, 100, is equal to the cost of a Type II error. As noted above, the  $L_1$  function is linear with sample size.

Why does the expected loss function when  $H_0$  is false depend nonmonotonically on the sample size? The portion of the function to the left of the optimum sample size can be understood as follows. With sample sizes this small and a moderate effect size, the probability of a Type II error,  $\beta$ , is reasonably large. Because  $L_2$  depends on the cost of a Type II error, losses with small samples are dominated by the Type II error costs rather than by the cumulative cost of testing these few subjects. Adding more subjects decreases  $\beta$  and its resulting cost more rapidly than the additional cost from adding those subjects, so the net expected loss decreases.

Next consider the portion of the  $L_2$  function to the right of the optimum. This part of the function depends linearly on the sample size. Why? Because with large sample sizes,  $\beta$  becomes nearly constant and approaches 0. Adding more subjects when  $n$  is already large only infinitesimally decreases  $\beta$ , so the expected loss simply reflects the added cost of testing more subjects.

The interesting result from the perspective of experimental design is that when the cost of a Type II error is explicitly factored into the picture, it forces the experimenter to consider a second objective: that of minimizing the expected monetary loss (equivalently, increasing the expected gain)<sup>1</sup>. The first goal, of course, is to maximize the power of the statistical test for rejecting  $H_0$ . This is usually done by setting the desired power at some minimum acceptable level (e.g., 0.80) and calculating the required sample size. If this sample size is prohibitively large, then power might have to be traded off against cost because resources are limited. Alternatively, if the cost of testing the required number of subjects is well within the resources of the project, then the experimenter might decide to purchase more power by adding subjects. At some point, the experimenter will encounter the fact that additional subjects provide only vanishingly small changes in the expected loss, so relative to the cost of making a Type II error, these additional subjects are not cost effective.

The advantage of making this Type II error cost explicit is that it is then easy to see that adding more subjects to the design once the minimum expected loss has been achieved is clearly wasteful of scarce resources – time and money. Experienced experimenters know this qualitatively; as noted above, at some point adding more subjects to the design is not worth the additional small power that those subjects add to the statistical test. The derivation above shows that if one is willing to place a relative value on the cost of a Type II error, then this sample size beyond which the experimenter begins to waste resources can be calculated. Experimenters have an ethical obligation to conduct an experiment that uses the available resources as effectively as possible, especially if the money for the research is being provided by fellow citizens. The approach presented here can help the experimenter in achieving this ethical and practical goal.

Because the experimenter cannot know whether s/he is making a Type I or a Type II error in a given experiment, the total expected loss from the experiment will be the sum of the costs of the two errors weighted by their probabilities of occurrence ( $\alpha$  and  $\beta$ , respectively) plus the cumulative cost of testing  $n$  subjects. This can be represented as

$$L = \alpha \cdot d_1 + \beta \cdot d_2 + n \cdot c \quad (7)$$

How does one calculate the optimum number of subjects given the cost of a Type II error and the other parameters associated with a given situation? Differentiating  $L$  with respect to the sample size,  $n$ , setting this derivative equal to 0, and solving for  $n$  will produce the desired result. More formally, the solution to

$$\frac{dL(n)}{dn} = 0 \quad (8)$$

provides the sample size required to minimize the expected loss. Because the cumulative normal density function contributes to  $L$  and because the minimization is done with respect to  $n$ , numerical methods are necessary to estimate the optimum sample size<sup>2</sup>.

Table 1 shows the power at the optimum sample sizes for a range of Type II error costs with  $\alpha$  set at .05 and an effect sizes of 0.3 and 0.8. Because the product,  $\alpha \cdot d_I$ , in Equation 7 does not depend on the sample size,  $n$ , and because this term simply adds a constant to the expected loss, the cost of a Type I error does not affect the location of the sample size at the minimum expected loss in Equation 8. Also shown in this table are the minimum expected losses at these optimum sample sizes. This latter value just represents the expected loss at the peak of the expected loss function. This is the minimum cost that the experimenter could expect to incur from conducting the different experiments with variation in the probabilities and costs of Type II errors. This table is only meant to be illustrative of the sample sizes and associated power levels that result from putting a relative value on the cost of making a Type II error.

-----  
 insert Table 1 about here  
 -----

Table 1 shows as expected that when the cost of a Type II error increases, the sample size required to minimize the expected loss increases. Hence, the power increases, and the minimum expected loss at this optimum sample size also increases. Notice that the power levels associated with the optimal sample sizes reach levels, ~98%, rarely found in experimental social science research.

An interesting result shown in Table 1 is that the minimum expected loss from fixing power at 0.80 without regard to the cost of a Type II error increases substantially as the cost of that error increases. This makes sense because fixing the power and hence the sample size

without regard to the cost of a Type II error means that one is essentially making experimental design decisions that are insensitive to the real costs of making Type II errors. As these errors become more costly, the fixed-power design becomes much less optimal and much more costly.

Considering a non-zero gain from a correct rejection of  $H_0$ .

Equation 6 above considered only the loss from a failure to reject a false null hypothesis. The argument was that the gain from a correct rejection of this null hypothesis could be ignored because all that mattered from the point of view of minimizing the expected loss was the difference between the gain and the loss. This allowed the simplification of setting the gain from a correct rejection to 0. Next, we consider the implications of considering this gain to be non-zero, and by some arguments enormous relative to the loss from failing to reject a false  $H_0$  and to the cost of testing a single subject.

Suppose that the gain from a correct rejection of a false  $H_0$  is  $g$  (again, relative to the cost of testing a single subject). Equation 6 for the expected gain/loss now becomes

$$L_2 = (1 - \beta) \cdot g + \beta \cdot d_2 + n \cdot c \quad (8)$$

The quantity  $(1 - \beta)$  is just the power: the probability of correctly rejecting a false  $H_0$ . This simplifies to

$$L_2 = g - \beta \cdot (g - d_2) + n \cdot c \quad (9)$$

As noted above, the only effect of including the gain from a correct rejection of a false  $H_0$  is to make the expected loss a function of the difference between the gain from a correct rejection of a false  $H_0$ ,  $g$ , and the loss from a failure to reject a false  $H_0$ ,  $d_2$  as well as to offset the expected loss function by the gain from a correct decision. This latter constant,  $g$ , in Equation 9 has no effect on the optimum sample size; it simply shifts the expected loss function vertically along the y-axis. If  $g$  is not zero, Equation 9 can be made identical to Equation 6 simply by setting  $g$  to zero and increasing the cost of a Type II error,  $d_2$ , by the corresponding amount. But increasing the

cost of a Type II error as shown in Table 1 increases the optimum sample size, so when the gain from a correct decision is factored into the expected loss function, and the cost of a Type II error is held constant, its effect is primarily to increase the optimum sample size, and the power at that optimum sample size.

Consider the behavior of Equation 9 as the sample size,  $n$ , gets very large. When this happens,  $\beta$  approaches 0, and Equation 9 reduces to

$$L_2 = g + n \cdot c \quad \text{for large } n \quad (10)$$

Recall that  $c$  is negative. Again, this implies that there is an optimum sample size because when  $\beta$  has effectively gone to 0, and a correct decision and its associated gain,  $g$ , are effectively guaranteed, adding more subjects will only reduce that expected gain.

This result depends on the assumption that  $g$ ,  $d_2$  and  $c$  are *commensurate*. By commensurate we mean that the gains and losses from correct and incorrect decisions can be quantified, and that they can be reasonably expressed as being multiples of the cost of testing a single subject. An argument can be made, however, that the gain from a correct rejection of a false  $H_0$  should be considered very large with respect both to  $d_2$  and to  $c$ . By the falsification account of scientific inference, we can only learn something when we reject a false theory (Miller, 1994; Popper, 1959). On the rare occasion that we can reject a false theory, what we have learned is then considered enormously valuable relative to the cost a testing a single subject and even to the loss from a failure to reject that false theory. In other words, it is hard to put a price on new knowledge.

By this reasoning, we should consider  $g \gg d_2 \ \& \ c$ . In this case we consider  $g$  to be *incommensurate* with  $d_2$  and  $c$ . If this is the case, then consider the behavior of Equation 8 for this condition:

$$L_2 = (1 - \beta) \cdot g \quad \text{for } g \gg d_2 \ \& \ c \quad (11)$$

There are two very important implications of considering the gain from a correct rejection of a false  $H_0$  to be very large with respect both to the loss from a Type II error and to the cost of testing a single subject. First, *there is no longer an optimum sample size*. As  $n$  increases, the expected loss (gain),  $L_2$ , continues to increase, albeit slowly for large  $n$ . Second, *the expected loss (gain) is directly proportional to the power*. Stated in another way, the experimenter should always test as many subjects as are affordable because this will increase his/her expected gain. The experimenter does not need to worry about wasting scarce resources by testing “too many” subjects when the gain from a correct rejection of a false null hypothesis is taken to be very, very large with respect to the cost of testing a single subject.

#### Limitations of the Current Analysis

There are several limitations of the approach advocated here. These limitations involve a) a reluctance to permit experimental design decisions be determined strongly by economic considerations, b) the uncertainty associated with the effect size estimate, c) the problem of deciding on values for the costs of Type I and Type II errors, and d) whether quantifying the gain from a correct rejection of a false  $H_0$  even makes sense in psychological, theory-driven research. Consider these limitations in turn.

Why do experimenters rarely explicitly consider the cost-effectiveness of their designs? One reason could be that this is seen as compromising the rigor of the science. This could be seen as opening the door to allowing numerous, extra-scientific factors to influence what is ultimately a scientific enterprise in the case of theoretically-driven research. But surely these extra-scientific factors do play a role, even if it is an implicit one, in influencing how research is designed. To take but one example: for each hour that I make a subject sit through my experiment, that subject is incurring an opportunity cost. The opportunity cost is the difference between what that subject gains from sitting through my experiment relative to what s/he could

gain from engaging in some other foregone activity (e.g., studying for an exam; reading a class assignment). Presumably, the experimenter thinks that the ultimate benefits to himself or herself, to the subject, and to society from conducting this experiment outweigh the opportunity costs for the subjects and the other costs associated with the experiment, otherwise it would not make sense to do the experiment. So every time that a decision is made to go forward with an experiment using human subjects, the cost-effectiveness of the design has either implicitly or explicitly been considered.

The second problem with this approach is that it requires a point specification of the effect size. This limitation is not specific to the case of factoring costs into power calculations, but rather it is involved whenever sample size and power are considered. The problem is that the actual effect size could differ from the one used in the power and cost calculations. As a result, the loss function,  $L_2$ , is conditional on the specific effect size employed. If there were a range of effect sizes possible, then it becomes difficult to specify an “optimum” sample size. One suggestion would be to consider the largest effect size that one could expect, and the smallest, and then to proceed to calculate the optimum sample sizes for this range. This leaves an element of uncertainty in the calculation of the cost-effectiveness of the design, but it could be argued that this is still preferable to deciding on some arbitrary, or minimally acceptable power level without taking into account the cost of a Type II error.

The last, and perhaps more obvious, problem with this analysis is that it is not clear how to put an actual value on the costs of Type I and Type II errors. This problem was finessed here by simply referring these costs to the cost of testing a single subject. The actual cost was left unspecified and set for convenience to 1. Reflection on the  $L_2$  function above shows that if both the unit cost of testing a subject and the cost of a Type II error are increased by the same factor, then the only thing that changes is the minimum expected loss at the optimum sample size.

Neither this optimum sample size nor the power associated with it change when these two costs maintain a fixed ratio. This suggests that one does not need to put an actual dollar value on these costs to conduct this analysis. Rather, expressing the cost of Type I and Type II errors in units of the cost of testing a single subject will be sufficient. This is similar to the approach used in evaluating the costs of sequential clinical trials in which the costs of Type I and Type II errors are explicitly interpreted as indicating the importance of a correct decision relative to the cost of testing a single subject (Lewis & Berry, 1994). In the Lewis and Berry study, the cost of incorrect decisions in their sequential clinical trials procedure was considered to be in the range from 2,000 to 12,000 times the unit cost of testing a single subject.

It is interesting to note that the 200:1 increase in the relative cost of a Type II error in Table 1 results in only a 2:1 to a 2.7:1 increase in the optimal sample size for effect sizes of 0.8 and 0.3, respectively, when the cost of a Type I error is held constant at 100 times the cost of a single subject. If an experimenter were uncertain about how to put a value on the loss from a Type II error, this analysis suggest that setting that value very high still results in an optimal sample size that is not unreasonably large. If instead of using this strategy, the experimenter employs the standard 0.80 power rule-of-thumb for determining sample size, Table 1 (effect size = 0.8) shows that a constant and very suboptimal sample size would be adopted ( $n = 10$ ) that increases the expected loss from the experiment by a factor of approximately 3 at a minimum.

There is one way in which the unit of cost in an experiment can always be referred to the cost of testing a single subject. If all of the experimenter's time in designing the experiment, all of the time involved in testing subjects, all of the monetary resources in terms of equipment and depreciation of that equipment used to conduct the experiment, etc. can be aggregated and expressed in monetary terms, then that total cost can be divided by the number of subjects in the experiment to get the cost per subject ( $c$ ). This then makes it possible to think about the cost of a

Type II error as the cost of the one or more future experiments that would result from failing to detect the effect that was predicted and of the cost of having expended those resources in the current, failed experiment. In other words, if I conduct an experiment, and I make a Type II error in that experiment, then I lose the total cost of that experiment as well as any costs associated with future experiments wrongly entailed by the current failed experiment. If my original experiment used 50 subjects, and if all costs have been expressed in units of dollars per subject ( $c$ ), then I have lost  $\$50c$  from that experiment. If I now mistakenly conduct two similar follow-up experiments that would not have been conducted had I not made a Type II error in the first experiment, then the cost of making that Type II error eventually reaches  $3 \times \$50c$  or 150 times the cost of testing a single subject. This simple analysis does not include the lost benefits that might have accrued from finding the predicted effect in the first experiment – in some ways, the most substantial cost of making a Type II error.

Even this solution, however, still leaves the experimenter with the problem of trying to decide just how costly these errors really are. In more clinically oriented research where the costs of missing a novel, effective treatment could be quantified more precisely in terms of continued illness and its associated medical resources, lost productivity, etc., this step in the analysis would be somewhat easier. In theoretically-driven research, however, putting values on these costs is less straightforward. It is beyond the scope of this paper to propose a precise method of estimating these error costs. Rather, the approach here shows that in principle once such costs are specified, they have important and somewhat surprising implications for the cost effectiveness of experimental research designs.

Finally, consider the argument that the gain from rejecting a false null hypothesis should be considered large enough to swamp the cost of a Type II error and the cost of testing a single subject. Null hypothesis significance testing was adopted early in fields like agriculture and

industrial manufacturing (Gower, 1988). In these fields, it is relatively easy to quantify the economic gain from picking the heartier of two crops or the less defective of two manufacturing processes. Indeed, in some treatments of this problem, the gain from a correct rejection of the null hypothesis is set proportional to the effect size, the standardized difference between the means of two populations between which the experimenter is trying to choose (Colton, 1963; Maurice, 1959). When this inferential machinery is imported into the arena of testing theory-driven psychological hypotheses, can these gains and losses be so easily quantified? What price can be put on learning that one's theory of some psychological process was wrong? Does it make sense to quantify the cost of knowledge in this way?

It is not necessary to argue that an actual value could be placed on the gain in knowledge that results from correctly rejecting a false null hypothesis. Rather, one could argue as was done in a previous section that if this gain is very, very large relative to the two costs,  $d_2$  and  $c$ , then it is always to the experimenter's advantage to test as many subjects as possible. Given finite resources for a project, as many of those resources as possible should be devoted to maximizing the power of the statistical test because this increases the expected gain. Perhaps this is one reason that in practice many experimenters simply concentrate on the probability  $1 - \beta$ ; they are effectively operating as if the gain from a correct rejection of a false null hypothesis far outweighs any consideration of the cost of making a Type II error and the cost of testing a single subject.

## References

- Bussemeyer, J. R., & Rapoport, A. (1988). Psychological models of deferred decision making. *Journal of Mathematical Psychology*, 32, 91-134.
- Cohen, J. (1990). Things I have learned so far. *American Psychologist*, 45, 1304-1312.
- Colton, T. (1963). A model for selecting one of two medical treatments. *Journal of the American Statistical Association*, 58, 388-400.
- Gower, J. C. (1988). Statistics and agriculture. *Journal of the Royal Statistical Society*, 151, 179-200.
- Grundy, P., Healy, M., & Rees, D. (1956). Economic choice of the amount of experimentation. *Journal of the Royal Statistical Society, Series A*, 18, 32-48.
- Haussler, D., & Opper, M. (1997). Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25, 2451-2492.
- Lewis, R. J., & Berry, D. A. (1994). Group sequential clinical trials: A classical evaluation of Bayesian decision-theoretic designs. *Journal of the American Statistical Association*, 89, 1528-1534.
- Maurice, R. (1959). A different loss function for the choice between two populations. *Journal of the Royal Statistical Society*, 21, 203-213.
- Miller, D. W. (1994). *Critical Rationalism: A Restatement and Defence*. Chicago: Open Court.
- Nam, J.-M. (1973). Optimum sample sizes for the comparison of the control and treatment. *Biometrics*, 29, 101-108.
- Popper, Karl R. (1959). *Logic of Scientific Discovery*. London: Hutchinson.
- Rapoport, A., & Burkheimer, G. J. (1971). Models for deferred decision making. *Journal of Mathematical Psychology*, 8, 508-538.

Samuels, S. M. (1985). A best-choice problem with linear travel cost. *Journal of the American Statistical Association*, 80, 461-464.

#### Author Note

James L. Dannemiller is in the Department of Psychology at Rice University, and Ronald C. Serlin is in the Department of Educational Psychology at the University of Wisconsin - Madison. This manuscript was started while JLD was at the University of Wisconsin – Madison and completed after he moved to Rice University. This research was supported by NICHD R01 HD32927 to JLD. We thank Jeremy Biesanz, Rick Jenison and David Lane for helpful comments on an earlier draft of this paper. Correspondence concerning this article should be addressed to James L. Dannemiller, Psychology Department-MS 25, Rice University, PO Box 1892, Houston, Texas 77251-1892. Electronic mail may be sent via the Internet to [dannemil@rice.edu](mailto:dannemil@rice.edu).

## Footnotes

<sup>1</sup>Implicit in this treatment of the expected losses from Type I and Type II errors is the idea that the expected gain from correct decisions is 0. When the objective is to find the sample size that minimizes the expected loss, all that matters is the *difference* between the gain from a correct decision and the loss from an incorrect decision. Hence, the gains from correct decisions can be set to 0, and only the costs of errors,  $d_1$  and  $d_2$ , need to be considered.

<sup>2</sup>The *normcdf* function in *Matlab* was used to generate the results. This function returns the cumulative probability below a given value in a normal distribution with parameters  $\mu$  and  $\sigma$ .

<sup>3</sup>Maurice (1959) pointed out that if the loss from an incorrect decision is proportional to the absolute value of the difference between two population means, and the experimenter is trying to choose between the two processes generating the sample observations, then there is no sample size that will minimize the loss when the absolute difference between the two population means is infinite. In this case, the expected loss is infinite as long as there is even a small probability of choosing the incorrect population. The experimenter can never test enough subjects to minimize the expected loss. This is something of a paradox because an infinite effect size would appear to be the most favorable case for deciding between two populations, but mathematically, the loss function is not bounded when the loss is set proportional to the absolute difference between the two population means.

Table 1

*Optimum sample size, power at optimum sample size and minimum expected loss at the optimum sample size for various costs of a Type II error with  $\alpha = .05$  and effect sizes of  $e=0.3$  and  $e = 0.8$ .*

<i>Effect Size, <math>e = 0.30</math></i>	<u>Cost of a Type II error*</u>			
	<u>500</u>	<u>1000</u>	<u>10000</u>	<u>100000</u>
Optimum sample size, $n_{opt}$	106	132	212	285
Power at $n_{opt}$	0.926	0.964	0.997	~1.00
Minimum expected loss at $n_{opt}$	-148	-173	-249	-321
Sample size for 0.80 Power	69	69	69	69
Expected loss with 0.80 Power	-173	-272	-2059	-19920
<i>Effect Size, <math>e = 0.80</math></i>				
Optimum sample size, $n_{opt}$	24	27	38	48
Power at $n_{opt}$	0.989	0.994	~1.00	~1.00
Minimum expected loss at $n_{opt}$	-35	-38	-48	-59
Sample size for 0.80 Power	10	10	10	10
Expected loss with 0.80 Power	-109	-203	-1896	-18820

\*Unit cost of testing a subject,  $c = 1$ ; these error costs are expressed as multiples of  $c$ . The cost of a Type I error was fixed at 100.

## Figure Captions

Figure 1. Expected loss as a function of sample size for a Type I error (solid line) and a Type II error (dashed line). The unit cost of testing a subject is set to 1, the cost both of Type 1 and Type II errors has been arbitrarily set to 100 times the unit cost of testing a single subject, the Type I error probability has been set to 0.05, and an effect size of 0.50 has been used.

