Psychological Methods
1997, Vol. 2, No. 1, 3–19


1082-989X/97/$3.00

# Optimal Design in Psychological Research

Gary H. McClelland
University of Colorado at Boulder

Psychologists often do not consider the optimality of their research designs. However, increasing costs of using inefficient designs requires psychologists to adopt more efficient designs and to use more powerful analysis strategies. Common designs with many factor levels and equal allocations of observations are often inefficient for the specific questions most psychologists want to answer. Happenstance allocations determined by random sampling are usually even more inefficient and some common analysis strategies can exacerbate the inefficiency. By selecting treatment levels and allocating observations optimally, psychologists can greatly increase the efficiency and statistical power of their research designs. A few heuristic design principles can produce much more efficient designs than are often used.

Experimental researchers outside psychology often carefully consider the efficiency of their research designs. For example, the high costs of conducting large-scale experiments with industrial processes has motivated the search for designs that are optimally efficient. As a consequence, a substantial literature on optimal design has developed outside psychology. In contrast, psychologists have not been as constrained by costs, so their research designs have been based on tradition and computational ease. Most psychologists are unaware of the literature on optimal research design; this topic receives little or no attention in popular textbooks on methods and statistics in psychology. Experimental design textbooks offer little if any advice on how many levels of the independent variables to use or on how to allocate observations across those levels to obtain optimal efficiency. When advice is offered, it is usually based on heuristics derived from experience rather than statistical principles.

The purpose of this article is to review the basic concepts of optimal design and to illustrate how a few simple changes in research design and analysis strategy can greatly improve the efficiency and statistical power of psychological research. The intent is to expand on the very brief treatments of optimality issues written for social scientists. For example, Kraemer and Thiemann (1987) and Lipsey (1990) in their discussions of maximizing power considered optimality issues but for only one special case. Estes (1991) illustrated the use of expected mean squares for deciding between alternative designs, but he does not provide guidance on determining optimal designs. Optimal design has received somewhat more attention in economics (Aigner & Morris, 1979) and marketing (Kuhfeld, Tobias, & Garratt, 1994).

The disadvantage to psychologists of using nonoptimal designs is either (a) increased subject costs to compensate for design inefficiencies or (b) reduced statistical power for detecting the effects of greatest interest. Both situations are increasingly unacceptable in psychology. Subject costs should be minimized for both ethical and practical reasons. Using a nonoptimal design that requires more animal subjects than does the optimal design is inconsistent with the ethics of animal welfare in experimentation. The imposition of experiment participation requirements on students in introductory psychology classes is difficult to justify if subject hours are used inefficiently. Reduced budgets from funding agencies also constrain the number of subjects to be used. In short, on many fronts there is increasing pressure on psychological experimenters to get more bang for the buck. The use of optimal, or at least more efficient, designs is an important tool for achieving that goal.

The motivation for this article arose from helpful electronic mail discussions with Don Burill. Drafts benefited from suggestions from Sara Culhane, David Howell, Charles Judd, Warren Kuhfeld, Lou McClelland, William Oliver, and especially Carol Nickerson.

Correspondence concerning this article should be addressed to Gary H. McClelland, Department of Psychology, Campus Box 345, University of Colorado, Boulder, Colorado 80309-0345. Electronic mail may be sent via Internet to gary.mcclelland@colorado.edu.

Inadequate statistical power continues to plague psychological research (Cohen, 1962, 1988, 1990, 1992; Lipsey, 1990; Sedlmeier & Gigerenzer, 1989). Journal editors and grant review panels are increasingly concerned about the statistical power of studies submitted for publication or proposed. By and large, the only strategy that psychologists have used for improving power is augmenting the number of observations. However, at least three other less costly strategies are available. One is the use of more sophisticated research designs (e.g., within- vs. between-subjects designs and the addition of covariates). The consideration of these design issues is beyond the scope of this article and is covered elsewhere (Judd & McClelland, 1989; Maxwell & Delaney, 1990). Two is the more efficient allocation of observations in whatever research design is chosen; this is the focus of this article. Three is the use of specific, focused, one-degree-of-freedom hypotheses rather than the usual omnibus tests, which aggregate not only the effects of interest but also a large number of effects that are neither hypothesized nor of interest. As shall be seen, the efficient allocation of observations to conditions and the use of one-degree-of-freedom tests can be used in tandem to improve power without increasing the number of observations.

The most striking difference between traditional experimental designs and optimal designs is that the former usually allocate equal numbers of observations to each level of the independent variable (or variables), whereas optimal designs frequently allocate unequal numbers of observations across levels. The emphasis on equal *n*s in psychological research appears to be due to the relative ease of analyzing data from such designs with desk calculator formulas and of interpreting parameter estimates. This ease makes them suitable for textbook examples that are then emulated. Today, however, the power and ubiquity of modern computing makes computational ease an irrelevant concern for the choice of experimental design.

One appeal of nonoptimal designs with equal allocations of observations to conditions is that such designs are more robust against violations of statistical assumptions, particularly the homogeneity of variance assumption. However, this protection is not strong and, as shown here, may sometimes have a high cost. A better strategy may be to use optimal designs while being vigilant for any violations of the assumption of equal variances. Any violations detected can either be remediated by means of transformations (Judd, McClelland, & Culhane, 1995) or be protected against by more robust comparison methods (Wilcox, 1996). In any case, it is necessary to consider issues of optimality to determine what price in terms of inefficiency is being paid for the weak protection against assumption violations offered by equal allocations of observations.

Psychologists, even if they know about optimal design, may also be reluctant to embrace optimal designs because such designs require the specification of a particular effect or effects to be optimized. That is, should the design be optimal for detecting main effects or interactions or both? Should the design be optimal for detecting linear effects or quadratic effects or both? Similarly, there is reluctance to use focused, one-degree-of-freedom tests because such tests require researchers to specify in some detail what they are expecting to find. Atkinson (1985) noted that optimal design is "distinct from that of classical experimental design in requiring the specification of a model" (p. 466). The tradition has been, instead, to use omnibus, multiple-degree-of-freedom tests to determine whether there are any overall effects and then to follow up with multiple comparison tests to try to determine the specific nature of those effects. This article explores the implications for experimental design of the more modern approach to data analysis that emphasizes focused one-degree-of-freedom hypothesis tests (Estes, 1991; Harris, 1994; Judd & McClelland, 1989; Judd et al., 1995; Keppel & Zedeck, 1991; Lunneborg, 1994; Rosenthal & Rosnow, 1985).

The literature on optimal design is complex and technical (for readable overviews, see Aigner, 1979; Atkinson, 1985, 1988; Atkinson & Donev, 1992; and Mead, 1988). However, without pursuing the technicalities of optimal design, psychologists can greatly improve the efficiency of their research designs simply by considering the variance of their independent variables. Maximizing the variance of independent variables improves efficiency and statistical power.

To demonstrate the importance of the variance of independent variables for improving efficiency and statistical power, one begins by considering a two-variable linear model:

$$Y_i = a + bX_i + cZ_i + e_i. \quad (1)$$

$X$ and $Z$ may be continuous predictors, or they may be codes (e.g., dummy, effect, or contrast) for categories or groups (two coded variables are sufficient to represent three groups). For considering the effect or statistical significance of $X$ or the power of that test, the

coefficient $b$ (or its estimate) and $V(b)$, the variance of the estimate, play important roles. The test of the null hypothesis that $b = 0$ is usually evaluated using either

$$F^*_{1,N-3} = \frac{\hat{b}^2}{\hat{V}(\hat{b})} \tag{2}$$

or

$$t^*_{N-3} = \frac{\hat{b}}{\sqrt{\hat{V}(\hat{b})}}, \tag{3}$$

where the circumflexes indicate sample estimates and where $N$ is the total number of observations (in the sequel, lowercase $n$ represents the number of observations within a condition or subgroup). With appropriate assumptions about the distribution of the errors, $e$, $F^*$ and $t^*$ can be compared with critical values of the $F$ distribution and Student's $t$ distribution, respectively. The statistical power (i.e., the probability of rejecting the null hypothesis given that it is false) is based on the noncentral $F$ distribution:

$$P\,(F^* \geq F_{1,N-p;.05;\delta}), \tag{4}$$

where the noncentrality parameter is defined as

$$\delta = \frac{b^2}{V(\hat{b})}. \tag{5}$$

A 100 (1—α)% confidence interval for $b$ is given by

$$\hat{b} \pm t_{N-3;\alpha/2}\sqrt{\hat{V}(\hat{b})}. \tag{6}$$

The smaller the estimated variance of the estimate, the smaller the confidence interval, and hence the more precise the estimate of the regression coefficient $b$. The standardized effect size (the coefficient of partial determination or the squared partial correlation or the proportional reduction in error produced by $X$ over and above $Z$) is given by,[1]

$$PRE = r^2_{YX.Z} = \frac{\hat{b}^2}{\hat{b}^2 + (N-3)\hat{V}(\hat{b})}. \tag{7}$$

For all these statistical expressions, one is better off (i.e., larger test statistic, more statistical power, smaller confidence interval, and larger effect size) as $b$ (or its estimate) increases and as $V(b)$ (or its estimate) decreases.[2] The regression coefficient $b$ is determined, up to a linear transformation, by nature. The key term in all the previous statistical expressions that can be affected by design is $V(b)$, which is estimated by

$$\hat{V}(\hat{b}) = \frac{\hat{V}(e)}{NV(X.Z)} = \frac{\hat{V}(e)}{NV(X)(1-r^2_{XZ})} = \frac{\hat{V}(e)(VIF)}{NV(X)}. \tag{8}$$

The square root of this variance is frequently reported by regression programs as the "standard error of the estimate." $\hat{V}(e)$ is the estimated variance of the error term in Equation 1; it is frequently reported in an analysis of variance (ANOVA) table in regression programs as the mean squared error or *MSE*. It is simply the variance of the differences between the estimated and predicted values for $Y$. $V(X.Z)$ is the variance of $X$ after controlling for $Z$. If $X$ and $Z$ are independent, then $V(X.Z) = V(X)$. However, if $X$ and $Z$ are correlated, then the variance of $X$ is reduced by $(1 - r^2)$, often reported as the "tolerance." The correlation reduces $V(X)$ to $V(X.Z)$, which in turn increases $V(b)$; hence $1/(1 - r^2)$ is sometimes referred to as the *variance inflation factor* or *VIF*.

To understand how research design affects statistical tests, power, confidence intervals, and effect sizes,

[1] Test statistics, like $F$, can be decomposed into a product of a function of effect size and a function of study size (Rosenthal, 1987, pp. 106–107). This expression for the standardized effect size *PRE* unfortunately suggests that effect size can be increased by increasing the sample size $N$. This is not the case because, as one sees, $N$ is also a component of $V(b)$. Because standardized effect sizes such as squared correlations are biased estimates of the true effect size, there *is* a slight effect of the sample size. However, that bias is negligible for my purposes here.

[2] The equations make it clear that if optimal designs increase the variance of $X$, then they will also increase effect sizes. One might also ask what the optimal designs are for measuring an effect size as precisely as possible, that is, having a small confidence interval around the effect size. The size of this confidence interval, however, is not directly related to the variance of $X$. Rather, as in Cohen and Cohen (1983, p. 111), the width of the confidence interval for effect size is determined only by the magnitude of the effect size estimate itself (larger effect sizes inherently have smaller confidence intervals as a consequence of Fisher's $Z$ transformation) and by the number of observations (the variance of the transformed estimate is a function of $N - p - 2$). Except for its effect on the magnitude of the effect size, varying the allocation of observations across conditions has no impact on the width of the confidence interval.

one needs only to consider how design affects the components of $V(b)$. I refer to all these statistical properties with the short-hand terms of *efficiency* or *precision*. The precision of the estimate of $b$ increases as $V(b)$ decreases. Whether one uses classical significance tests, confidence intervals, or standardized effect sizes to make research decisions, more precise estimates of $b$ improve the quality of those decisions. Thus, one wants to consider how design can change the components so as to decrease $V(b)$. Equation 8 identifies three main design strategies available to researchers for improving precision: (a) increasing the sample size $N$, (b) decreasing the error variance $V(e)$, and (c) increasing $V(X.Z)$, the residual variance of $X$. All three strategies are useful, but psychologists have too often relied only on the first strategy of increasing sample size, which can be costly. Note that there is an exact trade-off among the three strategies. For example, doubling the residual variance of $X$ has exactly the same effect on precision as doubling the sample size $N$, which in turn has exactly the same effect on precision as halving the error variance. The error variance can often be substantially reduced by using different research designs such as adding covariates or using within-subject designs (Judd & McClelland, 1989; Maxwell & Delaney, 1990). While the statistical power benefits of better research designs are important and themselves not used often enough, the remainder of this article focuses on the third strategy of increasing the residual variance of $X$. As one shall see, the residual variance of $X$ can easily be altered by changing the allocation of observations across levels of the independent variable. Appropriate allocations can substantially increase statistical precision.

Consider a quantitative independent variable $X$ with five different approximately equally spaced levels arbitrarily (without loss of generality) assigned the values of $-1$, $-1/2$, 0, $1/2$, and 1. Categorical independent variables are considered later. In the following, independent variables in a linear model of the dependent variable will always be either the numerical values themselves or values of contrast codes to be applied to the cell means computed for any particular design.

If one assumes constant $N$ and constant $V(e)$, the efficiency or relative variance (designated $RV$) of a particular design for estimating the parameter $b$ for variable $X$ is the ratio of the residual variance of $X$ to its maximum possible variance. The trade-off between the residual variance of $X$ and $N$ in the denominator of Equation 8 indicates that the optimal design with $N$ observations has the same precision as a nonoptimal design with $n/RV$ observations or, equivalently, that a nonoptimal design with $N$ observations has the same precision as the optimal design with only $(RV\ n)$ observations. The confidence interval for the nonoptimal design is $\sqrt{1/RV(X)}$ wider than the confidence interval for the optimal design.

The proportional distribution of observations across the five levels of $X$ can be represented by $(p, q, r, s, t)$ with $p + q + r + s + t = 1$. The goal is to determine the allocation of observations across the five levels that maximizes $RV$ for the particular effects of interest. After the optimal allocation is determined, it can be used as a basis for comparing the relative efficiencies of the traditional equal $n$ and other common allocations, including happenstance allocations that occur when researchers sample randomly rather than control the levels of their independent variables.

## Linear Effects

Consider the statistical model

$$Y_i = a + bX_i + e_i. \tag{9}$$

The linear effect represented in this model is of primary importance in many psychological studies. The consequences of different allocations of observations to the levels of $X$ are discussed first. Then the consequences of adopting different analysis strategies for testing this model are examined.

### *Designs for Linear Effects*

In the Appendix (see also Atkinson & Donev, 1992, and Mead, 1988) it is demonstrated that the optimal design for the linear effect, the one that yields the maximum variance for $X$, is (1/2, 0, 0, 0, 1/2). That is, a linear effect is detected or estimated most precisely by testing the difference between the means of the two most extreme levels of the independent variable, where one-half of the observations are allocated to each extreme.

Rather than using this optimal design for estimating the linear effect, many researchers allocate observations equally across all five levels of the independent variable. The equal-$n$ allocation has a relative variance of only $RV(X) = 1/2$. That is, the equal-$n$ design requires $1/(1/2) = 2$ times as many observations as the optimal design for comparable precision in estimating the linear effect. A doubling of costs is not trivial, regardless of whether those costs are measured in terms of sacrificed lab rats, experimenter time, human subject time, or test materials. But using the

equal-$n$ design without doubling the observations risks Type II errors because of the reduced ability to detect the linear effect. The confidence interval for $b$ estimated from the equal-$n$ design is $\sqrt{2} = 1.41$ wider than is the confidence interval for the optimal design.

The "linear efficiency" column of Table 1 presents the relative variances of a variety of other allocations that might be used to test for the linear effect of $X$. Note especially that designs with unequal $n$s are reasonably efficient as long as all the observations are at extreme levels. For example, for the 3:1 ratio (3/4, 0, 0, 0, 1/4), $RV(X) = .75$, so that only 1.33 times as many observations are needed as for the optimal design (1/2, 0, 0, 0, 1/2). This is still considerably more efficient than the equal-$n$ distribution (1/5, 1/5, 1/5, 1/5, 1/5). In fact, unless the ratio of the number of observations at the two extreme levels exceeds 5.8:1, any allocation with all observations at the extremes is at least as efficient as an equal-$n$ design in detecting a linear effect.

Some researchers are unable to control the allocation of observations to the levels of $X$ and must instead use whatever allocation that random sampling gives them. It is instructive to examine the relative efficiencies of such happenstance allocations. For example, the peaked, normallike distribution (1/15, 3/15, 7/15, 3/15, 1/15) has a relative variance of only .23. Thus, to have the same precision for estimating the linear effect, the researcher obtaining this distribution needs $1/.23 = 4.35$ times as many observations as does the experimenter using the optimal design. The 15:1 distribution (.9375, 0, 0, 0, .0625) yields comparable efficiency even though all observations are at the extremes. Note that the bimodal, U-shaped distribution of (.37, .11, .03, .11, .37), which many field researchers would consider undesirable, has a much higher linear efficiency than does the peaked distribution (.80 vs. .23) because a high proportion of its observations are at the extremes.

Table 1
*Relative Linear and Quadratic Efficiencies for Various Allocations of Observations Across Five Levels of the Independent Variable X*

| Design | Proportional allocation | Linear efficiency $RV(X)$ | Quadratic efficiency $RV(X^2)$ |
|---|---|---|---|
| Extreme | (.5, 0, 0, 0, .5) | 1.00 | 0 |
| Center ends | (.25, 0, .5, 0, .25) | .50 | 1.00 |
| Uniform | (.2, .2, .2, .2, .2) | .50 | .70 |
| Every other | (.33, 0, .33, 0, .33) | .67 | .89 |
| Compromise | (.375, 0, .25, 0, .375) | .75 | .75 |
| Peaked | (.07, .20, .47, .20, .07) | .23 | .42 |
| Skewed | (.20, .47, .20, .07, .07) | .29 | .56 |
| | | (.27)[a] | (.53)[a] |
| 2:1 | (.67, 0, 0, 0, .33) | .89 | 0 |
| 3:1 | (.75, 0, 0, 0, .25) | .75 | 0 |
| 5:8:1 | (.85, 0, 0, 0, .15) | .50 | 0 |
| 15:1 | (.94, 0, 0, 0, .06) | .23 | 0 |
| U shaped | (.37, .11, .03, .11, .37) | .80 | .47 |

[a] The first number is the efficiency when that effect is tested alone; the second number (in parentheses) is the efficiency after controlling for the other effect.

Some regression textbooks inappropriately attribute reduced efficiency for detecting a linear effect to "restriction in the range of $X$." However, the real problem is reduced relative variance, which does not necessarily result from a restriction on the range of $X$. For example, the allocation (0, 1/2, 0, 1/2, 0) has a relative efficiency of .25, but the peaked allocation (1/15, 3/15, 7/15, 3/15, 1/15) has a smaller relative efficiency of only .23, even though it has a wider range.

It is also important to note that the large differences in relative efficiencies across designs means that it is inappropriate to compare most effect size measures from studies with different designs. If the *identical* linear effect is estimated using different designs, the *standardized* effect sizes are very different (see also Cohen & Cohen, 1983, pp. 187 and 211). Combining Equations 7 and 8 for the special case of the linear model in Equation 9 yields

$$PRE = r^2_{YX} = \frac{\hat{b}^2}{\hat{b}^2 + \dfrac{(N-2)\hat{V}(e)}{NV(X)}} \qquad (10)$$

where $b$ is the parameter for the linear effect. If instead of using the estimated variance of the errors adjusted for degrees of freedom, one uses $V(e) = \text{SSE}/N$, then Equation 10 becomes

$$PRE = r^2_{YX} = \frac{1}{1 + \dfrac{V(\hat{e})}{\hat{b}^2 V(X)}}. \qquad (11)$$

This and most other standardized effect sizes are maximized by increasing the variance of $X$. Thus, even for the same linear effect, those designs with larger relative variances necessarily yield larger values of $PRE$. Note that $N$ does not appear in Equation 11; therefore, in estimating effect size, increasing sample size does not compensate for an inefficient design. In contrast, the comparison of unstandardized

effect sizes, such as the parameter estimate $b$, does not depend on having the same relative variances.

*Analysis for Linear Effects*

An inappropriate analysis strategy can exacerbate the inefficiency of a nonoptimal research design. For example, consider the plight of the field researcher who has obtained a random allocation across the five levels of $X$ similar to the peaked, normallike distribution of (1/15, 3/15, 7/15, 3/15, 1/15). Combining Equations 2 and 8 yields

$$F^*_{1,N-3} = \frac{\mathrm{V(X)Nb^2}}{\hat{\mathrm{V}}(\mathrm{e})} = \frac{\mathrm{SSR}}{\mathrm{MSE}}, \qquad (12)$$

where *SSR* is the usual sum of squares reduced and *MSE* is the mean squared error. Then for a simple linear regression, $SSR = (7/30)Nb^2 = .23Nb^2$. An alternative analysis strategy is to use a one-way ANOVA to assess possible differences among the five means. Assuming that the higher order trend effects are negligible, this strategy yields a mean squares between (*MSB*) equal to the same *SSR* divided by 4, the degrees of freedom for the five means. That is, $MSB = (.23/4)Nb^2 = .058Nb^2$. This *MSB*, smaller than the *SSR* for the simple linear regression, is compared with the same *MSE*. Obviously, the omnibus, multiple-degree-of-freedom test in the one-way ANOVA is more likely to miss detecting the linear effect. Also note that this random allocation from a happenstance design and the analysis strategy combine to produce an efficiency that is only 5.8% of that which would be obtained with an optimal design. Compensating for this reduction in efficiency requires (1/.058) = 17 times as many observations. Very few researchers can afford the cost of that many additional observations.

As was noted earlier, the use of contrast codes is increasingly advocated as an analysis strategy. However, it may not always be appropriate for quantitative variables and the random allocations resulting from happenstance designs. The *SSR* for a contrast is given by (Judd & McClelland, 1989, p. 292),

$$SSR_{contrast} = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k \lambda_k^2 / n_k}, \qquad (13)$$

where $k$ identifies the level of $X$ and where the $\lambda_k$, which must sum to zero across levels, represents the codes for the contrast variable. The obtained *SSR* is then compared with the *MSE*. As is shown in the Appendix, for the case of the linear contrast codes and the peaked distribution, the $SSR = (5/26)Nb^2 = .19Nb^2$, which is slightly less than the $SSR = (7/30)Nb^2 = .23Nb^2$ obtained by the simple linear regression.[3] The reduction is due to the redundancy among contrast-coded predictors; although the codes themselves are orthogonal, the unequal distribution of observations across the five levels induces a correlation. Even if the analysis does not explicitly include the redundant codes, the standard formula for computing the *SSR* for contrasts compensates for the anticipated redundancy. If the researcher wishes to test for only the linear effect, then using contrast codes exacts a penalty for the irrelevant redundancies; thus, simple linear regression (or, equivalently but more complicatedly, weighted contrast codes) is a more appropriate analysis strategy.

Another inadvisable analysis strategy is to regress $Y$ on the polynomial terms $X$, $X^2$, $X^3$, and $X^4$. Often the goal of this strategy is to estimate the linear effect while simultaneously testing for deviations from linearity. This strategy also substantially reduces the unique *SSR* for $X$ due to an induced redundancy among the predictors (primarily between $X$ and $X^3$ and between $X^2$ and $X^4$). For the case of the peaked distribution, the parameter estimate for the linear effect is the same, but the unique *SSR*, reduced by the redundancy (the squared correlation when $X$ is regressed on the other polynomial terms), is only

$$\begin{aligned} SSR &= V(X.X^2,X^3,X^4)Nb^2 \\ &= V(X)(1 - R^2_{X.X^2,X^3,X^4})Nb^2 \\ &= \frac{7}{30}\left\langle\frac{108}{469}\right\rangle Nb^2 = .054Nb^2. \end{aligned} \qquad (14)$$

In other words, 1/.054 = 18.6 times as many observations as the optimal design are necessary to com-

[3] These results assume the use of unweighted contrast codes; that is, the codes are not weighted by the number of observations at each level. Unweighted contrast codes are the norm in most psychological studies and textbooks because they can always be represented as the difference between group means. But unless the researcher's hypothesis concerns differences among all group means, a heavy price is paid. Contrast codes weighted by number of observations at each level will produce results equivalent to ordinary regression. In this case, weighted contrast codes or the equivalent regression analysis is preferred to the usual unweighted contrast codes.

pensate for the combination of a nonoptimal design *and* an inadvisable analysis strategy. The problem is that estimating the linear effect of $X$ by itself is not the same as estimating the linear effect of $X$ in the context of a polynomial model. By itself, the parameter for $X$ estimates the average linear slope across all observations; in a polynomial model the parameter for $X$ estimates the linear slope when $X = 0$. The two estimates are equal *when the distribution is symmetric about 0* because the average value of $X$ is then 0, but the questions underlying the estimates are still different. It is not surprising that one would have more confidence in an estimate of the average slope across all observations than in an estimate of the instantaneous slope at a particular point.

A final analysis strategy that unfortunately is used much too often is to split the observations into two groups based on the median of $X$. The result has the *appearance* of an optimal design; however, simply recoding observations from the middle levels to the extreme levels does not of course really make them extreme. For the peaked distribution, the resulting *SSR* equals only $(1/9)Nb^2$, about 48% of the *SSR* from the simple linear regression for the same data. Furthermore, the *MSE* is increased by splitting the observations into two groups. In short, using a median split for the peaked distribution is equivalent to discarding at least half of the observations.

For the same reasons, it is not appropriate for researchers to measure $X$ with fewer levels. For example, consider two clinical researchers using frequency statements to measure the degree of sadness. The first researcher uses five response categories: "always sad," "almost always sad," "sometimes sad," "almost never sad", and "never sad." The second researcher uses only two categories: "often sad" and "rarely sad." Presumably, those responding "always," "almost always," and about half of those responding "sometimes" in the first study would select the "often" category in the second; the remaining respondents would select "rarely." In effect, the second researcher has simply required the respondents to perform the median split; therefore, the prior arguments against median splits apply in this situation as well.

Error in the measurement of $X$ can cause the parameter estimates to be biased, although if the measurement error is small in relation to $V(X)$, the bias is likely to be inconsequential (Davies & Hutton, 1975; Seber, 1977). However, procedures that greatly increase measurement error, such as median splits or the use of only a few response categories, or that decrease the variance of $X$ can greatly exacerbate the bias in parameter estimates.

In summary, if only a linear effect is expected, then the optimal design that allocates one half of the observations to each extreme level should be used. With such a design, analysis strategy is not an issue because a simple linear regression, an independent $t$ test, or a one-way ANOVA with two levels all yield exactly the same conclusion. If for whatever reasons a nonoptimal design is used, the inefficiency of that design should not be compounded by using an inappropriate analysis strategy. The only choice is simple linear regression.

## Designs for Quadratic Effects

Some researchers may object to the extreme design (1/2, 0, 0, 0, 1/2) because it does not allow the detection of nonlinear or polynomial effects. An equal-$n$ design across five levels allows the estimation of polynomial trends up to the quartic. However, if there are no theoretical reasons for expecting a quartic effect, then looking for such an effect increases the risk of making Type I errors. Even when higher order effects such as a quartic effect are detected, they are often difficult to interpret without a firm theoretical basis for anticipating them. Furthermore, dispersing scarce resources over many levels increases the likelihood of Type II errors for the effects of greatest interest.

There are, nevertheless, many situations in which simpler nonlinear effects such as a quadratic effect are expected. Tests for quadratic effects are appropriate when the maximum or minimum values of $Y$ are expected to occur at intermediate values of $X$ or when values of $Y$ begin to peak or bottom out as the extreme values of $X$ are approached. The relative efficiency of a design for detecting or estimating a quadratic effect is $RV(X^2)$, the relative residual variance of $X^2$.

In the Appendix it is demonstrated (see also Atkinson & Donev, 1992) that the optimal design for detecting or estimating a quadratic effect is (1/4, 0, 1/2, 0, 1/4). This design can also be justified intuitively. If the relationship is linear, then the mean of the observations at the two extreme levels equals the mean of the observations at the middle level; the test for a quadratic effect simply compares these two means. As before, the maximum variance is obtained by allocating one half of the observations to each mean being compared. The observations at the extreme levels are

equally divided; otherwise the test of the quadratic effect would be partially redundant with the test of the linear effect. If there is a quadratic effect, the mean of the observations at the middle level will be smaller than (U shape) or greater than (inverted U shape) the mean of the observations at the extreme levels.

The maximum variance of $X^2$ for the optimal design (1/4, 0, 1/2, 0, 1/4) is 1/4. However, this variance should *not* be compared with the maximum variance of $X$ relevant for the linear effect, because such comparisons depend on the location of the origin of the scale. Psychologists seldom assume that their scales have more than interval properties, so a change of origin is always allowed. It *is* appropriate, however, to compare the quadratic variances of other designs with the variance of the optimal design for testing the quadratic effect. Thus, for the allocation (1/4, 0, 1/2, 0, 1/4), $RV(X^2) = 1$.

The *Quadratic efficiency* column of Table 1 gives $RV(X^2)$ for a variety of designs. The extreme design, of course, has $RV(X^2) = 0$. The traditional equal-$n$ design, with $RV(X^2) = .7$, requires $1/.7 = 1.43$ times as many observations as the optimal design to have the same precision for estimating a quadratic effect. The every-other design (1/3, 0, 1/3, 0, 1/3) fares better with $RV(X^2) = .89$. Finally, note that the peaked normallike distribution has a relative quadratic efficiency of only .42.

## Designs for Linear and Quadratic Effects

Unfortunately, the design that is optimal for the linear effect is not necessarily very good for the quadratic effect, and vice versa. Figure 1 displays the relative linear and quadratic efficiencies for a number of possible designs. The linear efficiency for the best quadratic design is only .5, and the quadratic efficiency for the best linear design is 0. Also note that the traditional equal-$n$ design has a relative linear efficiency of .5 and a relative quadratic efficiency of .7. Greater relative efficiency for quadratic effects than for linear effects is probably *not* what most researchers intend when they choose the equal-$n$ design.

If detection of a quadratic effect is crucial for testing a theory and if a linear effect would be problematic to interpret in the presence of a quadratic effect, then the optimal quadratic design should be chosen. If researchers want to hedge their bets so that they would still have reasonable efficiency for estimating a linear effect should the expected quadratic effect not be found, a similar symmetrical design with observations at three levels should be used; however, the proportion of observations at the middle level should be reduced. If $r$ is the proportion of observations at the middle level, then the set of possible combinations of linear and quadratic relative efficiencies is, as shown in the Appendix,

$$[(1 - r), 4r(1 - r)]. \quad (15)$$

Considering confidence intervals for each effect suggests how to choose values for $r$ to hedge one's bets.

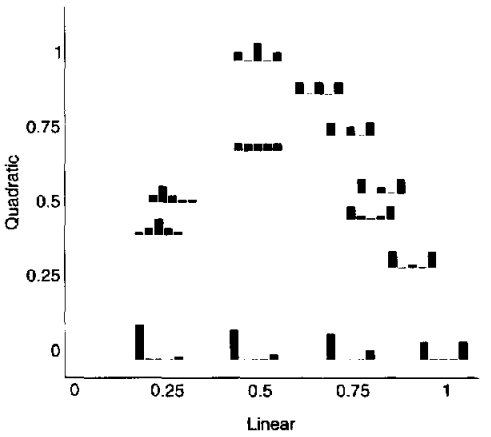


*Figure 1.* Linear versus quadratic relative efficiency for selected allocations of observations across five levels of the independent variable $X$.

### *Separate Confidence Intervals*

Neter, Wasserman, and Kutner (1990, p. 150) recommended constructing separate confidence intervals for each effect; in fact, to do otherwise is complicated with most regression programs. Comparable confidence intervals can be created by selecting a design with equal relative efficiencies for the linear and quadratic effects. For example, setting $r = 1/4$ yields the compromise design (3/8, 0, 1/4, 0, 3/8) shown in Table 1. The relative linear and quadratic efficiencies for this design both equal .75. This design is a good compromise in that it provides reasonable relative efficiency for a test of the quadratic effect and equally good relative efficiency for a backup test of the linear effect. Alternatively, efficiency equal to the respective optimal designs for each effect can be obtained by using $1/.75 = 4/3$ times as many observations.

Note that equal *relative* efficiencies for linear and quadratic effects does not imply equal *absolute* efficiencies. For the symmetric design with proportion $r$ observations at the middle level, the ratio of $V(X)$ to

$V(X^2)$ is $1/r$. The absolute linear and quadratic variances would be equal only when $r = 1$, which has the undesirable effect of making both variances zero. Thus, the confidence interval for the quadratic effect, given the scaling assumptions in this section, is necessarily wider than is the confidence interval for the linear effect, regardless of the design.

Mead (1988) and Atkinson and Donev (1992) also considered the problem of minimizing the less frequently used confidence ellipsoid, and Stigler (1971) and Studden (1982) considered how to choose $r$ to maintain a minimum specified power for detecting, respectively, a quadratic or higher order polynomial component.

### *Example*

It is instructive to consider a hypothetical example from Estes (1991) in terms of the trade-off in optimality between tests of linear and nonlinear effects. Estes based his example on the Sternberg (1966) paradigm in which a "subject is presented with a small set of items, typically randomly selected digits, letters, or short words, then is presented with a test item and responds yes or no as quickly as possible, yes indicating that the test item was in the set of items presented (the memory set) and no indicating that it was not" (Estes, 1991, p. 3). The dependent variable is reaction time (RT), measured in milliseconds. In the example, 25 subjects are allocated equally across five levels of set size, ranging from 1 to 5 items. The mean RTs are presented in Table 2.

According to theory, RT should increase linearly with set size. The *MSE* (within-set variance) is 617.5 for these data; thus the test statistic for the linear trend is $F(1,20) = 146.26$ ($PRE = .88$), which is clearly significant ($p < .0001$). The estimate of the linear trend is 42.5 ms per item, and the 95% confidence interval is [35.6, 49.4]. To test the adequacy of the linear model, Estes (1991) tested deviations from linearity as a set, as recommended by many textbooks, and concluded that

> a test of the nonlinear component yields an $F$ of only 2.99, which falls slightly short of the 5% level. Thus, although the deviations of the data points from [a] straight line . . . may look systematic to the eye, we have no statistical justification for going from the linear to any more complex function to describe the data. We conclude that, within the limits of this study, the theoretical hypothesis represented by the linear function with an upward slope is supported over any alternative model. (p. 80)

However, if detecting nonlinearity is important, then an optimal design and analysis is required. The suggestion of nonlinearity is strong: (a) the nonlinear model reduces error by 31% over the linear model, an average of 10% reduction in error per extra model parameter or degree of freedom, and (b) the obtained $F(3,20) = 2.99$ has a probability under the null hypothesis of .055, which just misses the conventional, but arbitrary, .05 cut-off (Cowles & Davis, 1982).

Either a more focused analysis or a more efficient design easily detects the nonlinearity in these data. The omnibus test of nonlinearity implicitly aggregates tests of the quadratic, cubic, and quartic trends. Although one might expect quadratic or possibly cubic trends, it seems unlikely that there would be quartic and higher order trends. Including all the nonlinear trends in a single test allows lower order trends to be washed out in the omnibus test. With fixed $N$, adding even more levels to the design makes nonlinearity more difficult to detect. In the present case, more focused tests of orthogonal trends yield, with the conventional .05 criterion, near significance for the quadratic trend, $F(1,20) = 4.08$, $p = .054$, $PRE = .17$, significance for the cubic trend, $F(1,20) = 4.55$, $p = .045$, $PRE = .19$, and no evidence for a quartic trend, $F(1,20) = 0.23$, $p = .64$, $PRE = .01$. Focused tests of trend components detect the nonlinearity, but the omnibus test, because it aggregates a test of the quartic trend with the lower order trends, does not.

Even better than a more focused analysis would be to use a more efficient and focused design for testing the quadratic trend, the simplest and most likely nonlinear trend. It is interesting to consider what would have been the result if the same means and *MSE* were obtained in a study with a more efficient design. An $N$ of 25 observations does not divide evenly into the three levels of the optimal design (1/4, 0, 1/2, 0, 1/4) for testing a quadratic effect, so only 24 subjects are

Table 2
*Mean Reaction Times (in Milliseconds) and Alternative Designs for the Hypothetical Memory Search Experiment From Estes (1991)*

| Set size | 1 | 2 | 3 | 4 | 5 | Total $n$ |
|---|---|---|---|---|---|---|
| $M$ | 420 | 500 | 540 | 555 | 605 | — |
| Uniform | 5 | 5 | 5 | 5 | 5 | 25 |
| Optimal for quadratic effect | 6 | 0 | 12 | 0 | 6 | 24 |
| Compromise | 9 | 0 | 6 | 0 | 9 | 24 |

used, as is shown in Table 2. This design is as linearly efficient as the original design with equal number of observations across five levels. The linear trend is still significant, of course, with $F(1,21) = 166.3$ ($PRE = .89$), $p < .0001$. Now, however, the quadratic trend is also significant with $F(1,21) = 7.35$ ($PRE = .26$), $p = .013$.

In this context, the traditional test for a quadratic trend is more appropriately described as a test of determining whether the mean of the observations at the middle level equals the expected mean (i.e., the mean of the observations at the two extreme levels) if the relationship between set size and RT is linear. A higher order effect other than the quadratic (most likely, the cubic for these data given the results of the one-degree-of-freedom tests) may be responsible for the mean at the middle level being other than that expected for a linear relationship. It is also possible, but unlikely, that a combination of higher order effects could yield a mean consistent with a linear model. Thus, as always, failure to reject a null hypothesis should not be considered as confirmation. However, in this example, the test reveals that the mean at the middle level is *not* consistent with a linear relationship between set size and RT. So, the nonoptimal, equal-$n$ design fails to detect a theoretically important deviation from nonlinearity that an optimal design, assuming the same means and *MSE*, can detect easily. Moreover, an optimal design with only 16 (rather than 24) observations also detects a statistically significant deviation from linearity with $F(1,13) = 4.89, p = .046$, and necessarily yields a more precise estimate and narrower confidence interval.

If linearity is expected but, as a hedge, a test of deviation from linearity is desired, the best design for this example is the compromise design (3/8, 0, 1/4, 0, 3/8), shown in the last row of Table 2. Even with only 1/4 of the subjects allocated to the middle level, nonlinearity is detected with $F(1,21) = 5.51$, $p = .029$.

The optimal design for the quadratic test necessarily yields the maximum effect size or *PRE*. Thus, for the same means and same *MSE*, there is no allocation of 24 observations across these five levels of set size that yields an effect size greater than $PRE = .26$. Knowing the maximum possible effect size may be useful in judging the substantive and theoretical significance of the deviation from linearity. For example, a failure to find a significant quadratic effect is not as good a support for the linearity hypothesis as is a demonstration that the quadratic effect, even with an optimal design, trivially reduces the proportion of error in relation to a linear model.

## Nonordinal Effects (Categorical Variables)

In many psychology experiments the levels of the independent variable are simply categories; neither the ordering nor the spacing of the levels is known. Estes (1991, p. 42) used as an illustration a categorization experiment with three different types of instructions: normal (N) instructions and two types of enhanced instructions, one emphasizing attention to relationships (R) and the other emphasizing attention to similarities (S). Determining the best allocation of observations to the three instruction conditions requires consideration of how the data are to be analyzed. As was noted earlier, an increasing number of contemporary textbooks advocate focused one-degree-of-freedom tests that use contrast codes. The use of contrast codes requires the researcher to be specific about the questions to be asked of the data *before* the analysis is performed. Theory is usually the source for the questions to be asked. No theory is at hand for the present example, but questions that might be asked about the three types of instructions spring readily to mind. Once the contrast codes corresponding to the questions are determined, the optimal design can be selected by identifying the allocation of observations that maximizes the variance of the codes.

Suppose the researcher wants to know (a) whether there is a difference between normal and enhanced instructions (N vs. R and S) and (b) whether there is a difference between the two types of enhanced instructions (R vs. S). The following set of contrast codes corresponds to these two questions:

| Code | N | R | S |
|---|---|---|---|
| C1 | −2 | 1 | 1 |
| C2 | 0 | −1 | 1 |

Code C1 compares the mean of the normal instruction condition to the mean of the means of the two enhanced instruction conditions. Code C2, which is orthogonal to C1, compares the means of the two enhanced instruction conditions. If the instruction categories R, N, and S are assigned the values −1, 0, and 1, respectively, then the questions corresponding to C2 and C1 are analogous to tests of the linear and quadratic effects, respectively, discussed above. Thus, while we may not want to refer to C1 as the "quadratic effect," we can nevertheless use the results outlined previously to select an optimal design for this experiment.

The researcher must decide the relative importance

of the questions corresponding to each code. If C1 is critically important, then the optimal design, in the order (N, R, S), is (1/2, 1/4, 1/4), but the relative efficiency for the test of C2 is only .5. At the other extreme, if C2 is critically important, then the optimal allocation is (0, 1/2, 1/2), in which case the C1 question is not asked. As was the case with tests of linear and quadratic effects, the allocation to the "middle" category (N, in this case) can be varied to change the relative importance of each of the two questions. If equal relative efficiency for testing the two questions is desired, then the design (2/8, 3/8, 3/8) is optimal because it yields relative efficiencies of .75 for both questions. The traditional equal-*n* design (1/3, 1/3, 1/3) yields a higher relative efficiency (.89) for question C1 than for C2 (.67). This may not be what the researcher intends.

Theory might suggest other codes. For example, if it is hypothesized that subjects adopt a similarities perspective even in the absence of similarity instructions and that, therefore, only the relationship instructions should have an effect on performance, then the appropriate codes are

| Code | N | R | S |
|---|---|---|---|
| C3 | −1 | 2 | −1 |
| C4 | −1 | 0 | 1. |

C3 tests for a difference between the mean of R and the mean of the means of N and S. C4 asks, as somewhat of an afterthought, whether there is a difference between the means of N and S. Here, C3 seems to be the much more important question, so the design (1/4, 1/2, 1/4) may be most appropriate.

## Higher Order Polynomial Trends or Effects

The principles discussed above can be used to determine optimal designs for detecting cubic and quartic trends. However, there are few instances in psychology for which a cubic or a quartic trend is expected on the basis of theory. Therefore optimal designs for higher order polynomial trends are not described in detail here. In general, however, the optimal allocations to each level are proportional to the absolute values of the orthogonal polynomial contrast codes for the appropriate trend when that trend is the highest order trend possible in the design. For the cubic trend for four equally spaced levels of $X$ (−1, −1/3, 1/3, 1), the contrast code is (−1, 3, −3, 1), therefore the optimal design for detecting a cubic trend is (1/8, 3/8, 3/8, 1/8). Similarly, the optimal design for detecting a quartic trend is (1/16, 4/16, 6/16, 4/16, 1/16) for five equally spaced levels of $X$. For nonordinal or categorical variables with many levels, the optimal design for a specific contrast of means can be determined by allocating an equal number of observations to each side of the contrast and then dividing each side's allocations equally between the levels in that side. For example, the optimal design for comparing three means to two other means is (1/6, 1/6, 1/6, 1/4, 1/4).

Note that a model including all polynomial trends up to a power of 4 is equivalent to allowing *any* pattern of means for the five levels of $X$. If a researcher is really interested in detecting any pattern of means, then the traditional equal-*n* design is optimal. However, if there are not good reasons for expecting a large number of possible patterns, the equal-*n* design will produce a large number of Type I errors. Moreover, the inefficiency caused by squandering observations in a nonoptimal design will produce a large number of Type II errors for the hypotheses of greatest interest.

## Linear × Linear Interactions

Now assume a second quantitative variable $Z$, which has five different equally spaced levels, assigned the same numerical values as $X$. The Linear × Linear interaction (the interaction of greatest interest to most researchers) is represented by the $XZ$ product, if $X$ and $Z$ are also included as terms in the model. McClelland and Judd (1993) derived a formula for the relative variance of the residual product, $RV(XZ)$, controlling for $X$ and $Z$. They also compared the relative efficiencies of a variety of designs for detecting the Linear × Linear interaction. The following discussion is based on those findings.

If $X$ and $Z$ are uncorrelated, then the relative efficiency of a design for detecting the Linear × Linear interaction is simply the product of the separate relative linear efficiencies of that design for $X$ and $Z$. Thus, the relative efficiency for the Linear × Linear interaction is at best equal to the lower of the linear efficiencies of the two variables in the interaction. If the relative linear efficiencies for both $X$ and $Z$ are less than 1.0, then the relative efficiency for the Linear × Linear interaction will be much lower than the relative linear efficiency for either of the two component variables.

Maximum residual variance for $XZ$, and hence greatest efficiency for the Linear × Linear interaction,

is obtained from the four-corner design that allocates 1/4 of the observations to each extreme, represented by the ($X$, $Z$) values (−1, −1), (−1, 1), (1, −1), and (1, 1). Note that even if there is no interaction, the four-corner design still has optimal efficiency for estimating the linear effects of $X$ and $Z$.

Instead of the four-corner design, some researchers hedge their bets by using an equal-$n$ design that allocates observations equally across the cells of, say, a 5 × 5 matrix. This design has the advantage of allowing the detection of any possible interaction, including, for example, the Cubic × Quartic interaction. Not only may it be costly to create and to manage 25 combinations of factors, but more important, the relative efficiency of the equal-$n$ design for detecting the Linear × Linear interaction is only .25. In other words, the equal-$n$ design requires four times as many observations as does the optimal four-corner design to have equal efficiency in detecting a Linear × Linear interaction. Moreover, the relative efficiency of the equal-$n$ design for detecting a linear effect of $X$ or $Z$ is only 0.5. Thus, the ability to detect higher order interactions is costly in terms of the ability to detect those effects that are usually of greatest interest.

A more modest hedge is a design that allocates observations equally to the nine cells defined by $X$ and $Z$ values of −1, 0, and 1; this design allows detection of the Linear × Quadratic, Quadratic × Linear, and Quadratic × Quadratic interactions as well as the Linear × Linear interaction. Few psychological theories predict interactions that are more complicated than these. However, the relative efficiency of this design for detecting the Linear × Linear interaction is only .44. Thus, this design should be chosen only if there is good reason to expect a Quadratic × Quadratic interaction or if a Quadratic × Quadratic interaction could be interpreted if found.

A useful alternative to the optimal four-corner design is one that allocates 1/4 of the observations to the center cell (0, 0) and divides the remaining 3/4 of the observations equally between the four corner cells (hence, 3/16 in each corner cell). This design has equal relative efficiencies (.75) for detecting the linear effects of $X$, the linear effects of $Z$, and the $XZ$ interaction. Few psychological theories require a more sophisticated design. This design has the additional advantage of leaving one degree of freedom for a lack-of-fit test comparing the actual mean of the center cell with the mean predicted by the Linear × Linear interaction and the separate linear effects. A significant deviation of the predicted and actual means for the center cell signals the presence of polynomial effects for $X$ and $Z$ and higher order interactions. If the lack-of-fit test suggests the presence of such effects, then a more complete design is appropriate in a follow-up experiment. Using equal $n$ in this five-point design is not very deleterious. This alternative design is easily generalized to more than two variables and has much to recommend it as a standard design for experimental research in psychology. Sall and Lehman (1996) made a similar suggestion for using the above generic, five-point design.

McClelland and Judd (1993) demonstrated how deleterious random or happenstance allocations (e.g., those resulting from surveys) can be for the detection of interactions. They provided an example of a bivariate normallike distribution that has a relative efficiency of only .06 for detecting the $XZ$ interaction. A field study expecting this distribution requires about 17 times as many observations as the optimal design to have comparable relative efficiency for estimating the Linear × Linear interaction.

## Discussion

Most psychologists use traditional equal-$n$ experimental designs even though those designs are not optimal for the questions that they want to ask of their data. This is not sensible scientifically. Design inefficiencies can always be offset by increasing the number of observations. However, ethical concerns, financial costs, and time constraints preclude increasing the number of observations as a general solution to design inefficiency. Many studies could use fewer subjects in optimal designs and still have the same or greater statistical power as equal-$n$ designs for detecting the effects of greatest interest. In the past, psychologists may have avoided designs with unequal $n$s because of computational complexities. However, modern computing makes computational issues irrelevant for the choice of experimental design.

Throughout this article it has been assumed that the cost of each observation is equal. This may not be the case. For example, in a biopsychological drug experiment, Level 0 might correspond to a control condition, Level 1 might correspond to an injection to increase the natural level of the drug in the body, and Level −1 might correspond to an injection of a blocking agent to reduce the effects of the natural level of the drug. The costs of the drug and the blocking agent may be very different from each other and considerably larger than the cost of the vehicle injection for

the control condition. However, the same efficiency principles can be used to determine an optimal design for a fixed overall cost. Note that allocation of subjects to conditions affects the maximum possible $N$ for a fixed overall cost. Because $N$ and residual variance trade off, the goal is to find the allocation that maximizes $N\ V(X)$ for the effects of interest.

Another assumption throughout this article has been that $V(e)$, the residual variance, remains constant, independent of the design. This is equivalent to the usual homogeneity of variance assumption. Violations of this assumption can be more problematic with unequal-$n$ designs, especially when the group with the larger variance has the smaller $n$ (Wilcox, 1996, p. 131). If heterogeneity of variance is expected, then additional considerations for optimal design apply, such as allocating more observations to those levels or groups expected to be most variable (Kish, 1965).

If there are multiple questions of interest, there is no simple recommendation for an optimal design. The relative importance of the various questions must be considered and a design chosen so that those questions of greatest interest have the greatest relative efficiencies. Usually, it is possible to identify the effect of greatest interest and then include one or two additional effects as a safeguard. In such cases, the optimal design is usually easy to determine. The recommendation to consider the relative importance of the questions being asked when designing an experiment is similar to, but may make unnecessary, the recommendations of Rosenthal and Rubin (1984) and Tukey (1991) to allocate differentially the Type I error rate to the important comparisons when making multiple comparisons. However, if for whatever reasons a nonoptimal design is used, their recommendations should be heeded.

Cox (1958, see especially pp. 137–142) considered more informally some of the issues addressed in this article. His results and conclusions are generally consistent with those reported here. For example, Cox (1958) concluded that "both slope [the linear effect] and curvature [the quadratic effect] are more precisely estimated from three equally spaced levels than from four or more equally spaced levels with the same extreme points" (p. 141). However, at a time when experiments were perhaps less costly and computations were definitely more expensive, Cox was not convinced that the increase in precision from unequal-$n$ designs was worth the increase in computational complexity. For example, he judged the increased sensitivity of the optimal (1/4, 1/2, 1/4) design to be too small to be worth the trouble. Now that computational costs are trivial, a different conclusion might be reached. Cox was also doubtful that researchers would be able to specify their relative interests in the linear and quadratic effects. However, in testing psychological theories the presence or not of a quadratic effect is sometimes crucial. A researcher who wants to be convinced that deviations from linearity are not substantial must use the most powerful test possible, and that requires either an optimal design (1/4, 1/2, 1/4) or 12.5% more observations to compensate for the inefficiency of an equal-$n$ design. On the other hand, if the researcher is most interested in estimating the linear effect and wants only a safeguard test for the quadratic effect, then the design (3/8, 1/4, 3/8) is adequate and requires 12.5% fewer observations than does the equal-$n$ design with the same linear efficiency. If a 12.5% difference in subject costs is not problematical, then equal-$n$ designs might as well be used. However, if a 12.5% increase in costs is cause for concern, then the most appropriate optimal design ought to be used.

Determining the optimal design for complex experiments involving many questions can be difficult. The technical literature on optimal design is so forbidding that psychologists are as unlikely to consult that literature as they have been to heed the sound admonitions to consult power tables and power curves to assess the statistical power of their designs. However, psychologists who are comfortable with the technical literature on power tables and curves would profit by reading the quite accessible overviews of the optimal design literature by Atkinson and Donev (1992) and Mead (1988). Accessing the technical literature on optimal design is, fortunately, not crucial, because adhering to the following general principles can yield a substantially improved design in relation to traditional designs.

First, an optimal design allocates observations to the same number of variable levels as there are parameters in the model. For a linear model, two levels should be used to estimate the two parameters, the intercept and the slope; for a quadratic model, three levels should be used; and so on. Thinking about an appropriate model for the data will therefore often suggest the appropriate number of levels and the corresponding optimal design. Psychologists often seem reluctant to use the relatively small number of levels that their models actually require. Mead (1988) countered this concern:

> The argument for dispersing the fixed total resources among many excess levels to detect any of a large number of possible discontinuities of the model is faulty because there will not be adequate precision at any point to detect a discontinuity, while the dispersal of resources reduces precision for the major criteria. (p. 533)

Second, if a test of lack of fit against the expected model is desirable as a safeguard, only a single level should be added to allow for that test. As Mead (1988, p. 533) stated, "The extra level provides both for a conflict of interests between the competing criteria, and some protection against major failure of the model."

Third, linear effects and Linear × Linear interactions are of primary interest in many psychological experiments. If thinking about a model for the data does not suggest the optimal design, then it is reasonable to consider an "off-the-shelf" design strategy that follows from the second recommendation. That strategy is to test efficiently for linear effects but to safeguard against nonlinearity. For a study with one independent variable, then, the recommendation is to allocate observations in the proportion (3/8, 1/4, 3/8) across three equally spaced levels with the end levels being as extreme as is realistic and feasible. For two independent variables, 3/16 of the observations should be allocated to each corner cell, with the levels of these cells being as extreme as is realistic and feasible, and the remaining 1/4 of the observations should be allocated to the center cell. Generalizations to more than two independent variables are obvious. Such designs provide efficient tests for linearity hypotheses (including the linearity interactions in designs with multiple variables) while still allowing a safeguard test against major violations of those hypotheses. If the safeguard test indicates a significant violation of linearity, then a more complex design is appropriate in a follow-up experiment.

Fourth, if the levels of the independent variable (or variables) are neither selected nor controlled but dictated by random sampling, then it is crucial not to compound the inefficiency of this random allocation of observations by using an inappropriate analysis strategy. Studies with quantitative independent variables should be analyzed with multiple regression with no more polynomial and interaction product terms than are theoretically reasonable. Studies with categorical independent variables should be analyzed with focused one-degree-of-freedom contrasts. Studies with observations allocated by random sampling should also estimate the maximum value of *PRE* that would be obtained with an optimal design.

## References


Aigner, D. J. (1979). A brief introduction to the methodology of optimal experimental design. *Journal of Econometrics, 11,* 7–26.

Aigner, D. J., & Morris, C. N. (Eds.). (1979). *Experimental design in econometrics.* Amsterdam: North-Holland.

Atkinson, A. C. (1985). An introduction to the optimum design of experiments. In A. C. Atkinson & S. E. Fienberg (Eds.), *A celebration of statistics: The ISI centenary volume* (pp. 465–474). New York: Springer-Verlag.

Atkinson, A. C. (1988). Recent developments in the methods of optimum and related experimental design. *International Statistical Review, 56,* 99–115.

Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs.* Oxford, England: Clarendon Press.

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

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

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

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist, 37,* 553–558.

Cox, D. R. (1958). *Planning of experiments.* New York: Wiley.

Davies, R. B., & Hutton, B. (1975). The effects of errors in the independent variables in linear regression. *Biometrika, 62,* 383–391.

Estes, W. K. (1991). *Statistical models in behavioral research.* Hillsdale, NJ: Erlbaum.

Harris, R. J. (1994). *ANOVA: An analysis of variance primer.* Itasca, IL: F.E. Peacock.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach.* San Diego, CA: Harcourt, Brace, Jovanovich.

Judd, C. M., McClelland, G. H., & Culhane, S. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. *Annual Review of Psychology, 46,* 433–465.

Keppel, G., & Zedeck, S. (1991). *Data analysis for research*

*designs: Analysis of variance and multiple regression/ correlation approaches.* New York: W.H. Freeman.

Kish, L. (1965). *Survey sampling.* New York: Wiley.

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

Kuhfeld, W. F., Tobias, R. D., & Garratt, M. (1994). Efficient experimental design with marketing applications. *Journal of Marketing Research, 31,* 545–557.

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

Lunneborg, C. E. (1994). *Modeling experimental and observational data.* Belmont, CA: Wadsworth.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Belmont, CA: Wadsworth.

McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114,* 376–390.

Mead, R. (1988). *The design of experiments: Statistical principles for practical application.* Cambridge, England: Cambridge University Press.

Neter, J., Wasserman, W., & Kutner, M. H. (1990). *Applied linear statistical models* (3rd ed.). Homewood, IL: Richard D. Irwin.

Rosenthal, R. (1987). *Judgment studies: Design, analysis, and meta-analysis.* Cambridge, MA: Harvard University Press.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance.* Cambridge, England: Cambridge University Press.

Rosenthal, R., & Rubin, D. B. (1984). Multiple contrasts and ordered Bonferroni procedures. *Journal of Educational Psychology, 76,* 1028–1034.

Sall, J., & Lehman, A. (1996). *JMP® start statistics: A guide to statistical and data analysis using JMP® and JMP IN® software.* Belmont, CA: Duxbury Press.

Seber, G. A. F. (1977). *Linear regression analysis.* New York: Wiley.

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

Sternberg, S. (1966). High-speed scanning in human memory. *Science, 153,* 652–654.

Stigler, S. M. (1971). Optimal experimental design for polynomial regression. *Journal of the American Statistical Association, 66,* 311–318.

Studden, W. J. (1982). Some robust-type D-optimal designs in polynomial regression. *Journal of the American Statistical Association, 77,* 916–921.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6,* 100–116.

Wilcox, R. R. (1996). *Statistics for the social sciences.* San Diego, CA: Academic Press.

*(Appendix follows on next page)*

# Appendix

## *RV(X)*

Assume that the probability allocation over the five levels of $X$ (−1, −1/2, 0, 1/2, 1) is given by ($p$, $q$, 1-$p$-$q$-$s$-$t$, $s$, $t$) such that $p + q + s + t <= 1$. The mean is then

$$-p - q/2 + s/2 + t \tag{A1}$$

and the variance $V(X)$ is

$$p[-1-(-p-q/2+s/2+t)]^2 + q[-1/2-(-p-q/2+s/2+t)]^2 + (1-p-q-s-t)[-0-(-p-q/2+s/2+t)]^2 + s[(1/2-(-p-q/2+s/2+t)]^2 + t[1-(-p-q/2+s/2+t)]^2, \tag{A2}$$

which when expanded becomes

$$p - p^2 + q/4 - pq - q^2/4 + s/4 + ps + qs/2 - s^2/4 + t + 2pt + qt - st - t^2. \tag{A3}$$

In the symmetric case for which $s = q$ and $t = p$, this variance reduces to

$$2p + q/2. \tag{A4}$$

Clearly, any allocation of probability to $q$ reduces its effect by half, whereas any allocation of probability to $p$ increases its effect by 2. Thus, the maximum variance for a symmetric allocation is obtained when $p = 1/2$, which is equivalent to the distribution (1/2, 0, 0, 0, 1/2). This maximum variance equals 2(1/2) = 1; hence, $V(X) = RV(X)$ for any probability allocation.

## $V(X^2)$

If the distribution over the levels (−1, −1/2, 0, 1/2, 1) is ($p$, $q$, 1-$p$-$q$-$s$-$t$, $s$, $t$), then the levels of $X^2$ are (0, 1/4, 1) with respective proportions (1-$p$-$q$-$s$-$t$, $q + s$, $p + t$). The mean is obviously

$$\frac{q+s}{4} + p + t. \tag{A5}$$

The variance is then

$$(1-p-q-s-t)[0-(q+s)/4-p-t]^2 + (q+s)[1/4-(q+s)/4-p-t]^2 + (p+t)[1-(q+s)/4-p-t]^2, \tag{A6}$$

which when expanded becomes

$$(16p - 16p^2 + q - 8pq - q^2 + s - 8ps - 2qs - s^2 + 16t - 32pt - 8qt - 8st - 16t^2)/16. \tag{A7}$$

For a symmetric distribution ($p$, $q$, $1 - 2p - 2q$, $q$, $p$), this reduces to

$$(16p - 32p^2 + q - 16pq - 2q^2)/8. \tag{A8}$$

These results can be used to calculate any of the variances mentioned in the text. The maximum variance over the squared levels (0, 1/4, 1) obtains, according to the previous argument for $V(X)$, when one half of the observations are at each extreme, that is, when $1 - p - t = p + t$, which implies $p + t = 1/2$. Unless the probability allocation is symmetric so that $p = t$, $X$ and $X^2$ are correlated and the tolerance is less than 1, which reduces the residual variance. Hence, the maximum variance occurs when $p = t = 1/4$, yielding the design over levels of $X$ of (1/4, 0, 1/2, 0, 1/4) and $V(X^2) = 1/4$. The relative variance of any other allocation is given by its residual variance multiplied by 4.

## *SSR* for Linear Contrast

According to Judd and McClelland (1989), the sum of squares reduced or sum of squares regression for a simple linear model is given by

$$SSR = \sum_{i=1}^{n} (\bar{Y} - a - bX_i)^2. \tag{A9}$$

If the $X$ observations are centered by subtracting the mean of $X$, then the intercept $a$ equals the mean of $Y$. That is,

$$SSR = \sum_{i=1}^{n} [\bar{Y} - \bar{Y} - b(X_i - M)]^2 = b^2 \sum_{i=1}^{n} (X_i - M)^2 = Nb^2 \frac{\sum_{i=1}^{n} (X_i - M)^2}{N} = Nb^2 V(X). \tag{A10}$$

According to Judd and McClelland (1989), the *SSR* for a contrast is given by

$$SSR_{contrast} = \frac{\left(\sum_k \lambda_k \bar{Y}_k\right)^2}{\sum_k \lambda_k^2 / n_k}. \tag{A11}$$

If one uses the linear contrast (−1, −1/2, 0, 1/2, 1) and assumes that the linear model is correct so that the expected mean for each cell is given by

$$\bar{Y}_j = a + bX_j, \tag{A12}$$

then for the peaked distribution (1/15, 3/15, 7/15, 3/15, 1/15), the numerator of the *SSR* for the linear contrast is

$$\begin{aligned} &\{-1[a + b(-1)] - \tfrac{1}{2}[a + b(-\tfrac{1}{2})] + 0[a + b(0)] \\ &+ \tfrac{1}{2}[a + b(\tfrac{1}{2})] + 1[a + b(1)]\}^2 \\ &= (b + \tfrac{1}{4}b + \tfrac{1}{4}b + b)^2 \\ &= \tfrac{25}{4}b^2. \end{aligned} \tag{A13}$$

The denominator of the *SSR* for the linear contrast is

$$\frac{(-1)^2}{\frac{N}{15}} + \frac{(\frac{-1}{2})^2}{\frac{3N}{15}} + \frac{0^2}{\frac{7N}{15}} + \frac{(\frac{1}{2})^2}{\frac{3N}{15}} + \frac{1^2}{\frac{N}{15}}$$

$$= \frac{3 + \frac{1}{4} + \frac{1}{4} + 3}{\frac{3N}{15}} = \frac{\frac{13}{2}}{\frac{N}{5}} = \frac{65}{2N}. \quad \text{(A14)}$$

The ratio of the numerator and denominator yields

$$\frac{25b^2}{4}\frac{2N}{65} = \frac{5}{26}Nb^2. \quad \text{(A15)}$$

### Linear–Quadratic Trade-Off

Assume a symmetric probability allocation $[(1 - p)/2, p, (1 - p)/2]$ over the levels $(-1, 0, 1)$. The mean is zero and the variance is

$$\frac{1-p}{2}(-1)^2 + p(0) + \frac{1-p}{2}(1) = 1 - p. \quad \text{(A16)}$$

Over the squared levels (0, 1), the allocation would be $(p, 1 - p)$ with a mean of $1 - p$. The variance is thus

$$\begin{aligned} &p(0 - 1 + p)^2 + (1 - p)(1 - 1 + p)^2 \\ &= p(1 - p)^2 + (1 - p)p^2 \\ &= (1 - p)[p(1 - p) + p^2] \\ &= (1 - p)p. \end{aligned} \quad \text{(A17)}$$

The variance relative to the optimal allocation is $4p(1 - p)$. Therefore,

$$[1 - p, 4p(1 - p)] \quad \text{(A18)}$$

describes the set of relative linear and quadratic efficiencies, obtainable with $p$ observations in the center cell and the remaining $1 - p$ observations divided equally among the extremes.

Received June 28, 1996
Revision received October 3, 1996
Accepted October 3, 1996 ■