Statistical Difficulties of Detecting Interactions and Moderator Effects

Gary H. McClelland and Charles M. Judd

Although interaction effects are frequently found in experimental studies, field researchers report considerable difficulty in finding theorized moderator effects. Previous discussions of this discrepancy have considered responsible factors including differences in measurement error and use of nonlinear scales. In this article we demonstrate that the differential efficiency of experimental and field tests of interactions is also attributable to the differential residual variances of such interactions once the component main effects have been partialed out. We derive an expression for this residual variance in terms of the joint distribution of the component variables and explore how properties of the distribution affect the efficiency of tests of moderator effects. We show that tests of interactions in field studies will often have less than 20% of the efficiency of optimal experimental tests, and we discuss implications for the design of field studies.

Moderated Multiple Regression

Many theories in psychology posit that one independent variable moderates the relationship between another independent variable and the dependent variable, or, equivalently, that two independent variables interact so that the effect of either one on the dependent variable depends on the level of the other. Statistical tests for interactions between categorical variables are well known in the analysis of variance (ANOVA). Saunders (1955, 1956) was apparently the first to develop a methodology for testing interactions or moderator effects for continuous variables and to refer to this methodology as "moderated multiple regression." Cohen (1978) and Arnold and Evans (1979), among others, formalized and verified Saunders's suggestion; Baron and Kenny (1986) distinguished between the testing of moderator effects and mediating effects; and Jaccard, Turrisi, and Wan (1990) and Aiken and West (1991) provided thorough, modern treatments of testing and interpreting moderator effects.

The increasing use of moderated regression has produced a conundrum: Although experimentalists frequently detect interaction effects, nonexperimentalists conducting field research have found moderator effects to be extremely difficult to detect. We investigate this conundrum by first describing the moderated regression methodology and demonstrating its mathematical equivalence to methodologies for detecting interactions in the ANOVA. We then consider the evidence that moderator effects are difficult to detect in field studies. Next, we demonstrate that differences in the joint distributions of the predictor variables between experiments and field studies are at least partially to blame for the difficulty in detecting moderator effects in field studies. Finally, in light of these issues, we suggest how field researchers might increase the statistical power of their tests of moderator effects.

In moderated multiple regression, standard multiple regression procedures are used to test for the existence of moderator effects by testing the statistical reliability of the product XZ in the following model:

$$Y_i = \beta_0 + \beta_X X_i + \beta_Z Z_i + \beta_{XZ} X_i Z_i + \epsilon_i . \qquad (1)$$

A mathematically equivalent way of implementing this test is to determine whether the residual part of the product XZ is related to the residual part of Y after both X and Z have been used to predict both XZ and Y. All else being equal, the greater the value of the partial regression coefficient[1] β_{XZ}, the greater the

Gary H. McClelland and Charles M. Judd, Department of Psychology, University of Colorado.

This research was partially supported by Grant MH-45049 from the National Institute of Mental Health to Charles M. Judd and Bernadette Park. We are grateful to Richard Jessor for allowing us to use his data on problem behaviors as a case study of the difficulties of detecting interactions in field studies. We also thank Jill Van Den Bos for facilitating our analysis of those data and for suggesting improvements to the article. Helpful suggestions for improving the presentation of the statistical issues were made by Leona Aiken, Charles Berry, Jerry Busemeyer, David Kenny, Terry Moore, Lou McClelland, Carol Nickerson, Ewart Shaw, Juliet Shaffer, and an anonymous reviewer. Any remaining errors are our responsibility.

Correspondence concerning this article may be addressed to either Gary H. McClelland or Charles M. Judd, Center for Research on Judgment and Policy, Department of Psychology, Campus Box 345, University of Colorado, Boulder, Colorado 80309-0345. Electronic mail may be sent to either mcclella@psych.colorado.edu or cjudd@clipr.colorado.edu.

[1] We follow the usual convention of using Greek letters to represent unknown model parameters. These partial regression coefficients should not be confused with the standardized regression coefficients reported as "beta" by some computer programs.

Psychological Bulletin, 1993, Vol. 114, No. 2, 376-390
 0033-2909/93/$3.00

moderating effect of Z on the relationship between X and Y (or, equivalently, the greater the moderating effect of X on the relationship between Z and Y). Thus, the test of the null hypothesis H_0: $\beta_{XZ} = 0$ is a test for the existence of a reliable moderating effect of Z on the $X - Y$ relationship. As Cohen (1978), Arnold and Evans (1979), Cronbach (1987), Aiken and West (1991), and others have noted, it is important that this test be hierarchical, determining whether β_{XZ} is reliably different from zero when controlling for both X and Z. The corresponding squared partial correlation, or proportional reduction in error (PRE; see Judd & McClelland, 1989), describes the model improvement due to adding the product term.[2] The equivalence of the test of H_0: $\beta_{XZ} = 0$ in the moderated multiple regression model of Equation 1 to the test of an interaction in a 2×2 experimental design in an ANOVA is easily demonstrated by using contrast codes of -1 and $+1$ as values of both X and Z to represent the levels of the independent variables. The statistical tests are identical in the two cases. Thus, experimental interaction can be viewed as a moderater effect and either independent variable can be viewed as the moderator.

Difficulty of Detecting Moderator Effects

Experimentalists frequently report finding statistically reliable interactions, and few complain of the difficulty of finding such effects. By contrast, as Morris, Sherman, and Mansfield (1986) noted, despite frequently compelling theoretical reasons for expecting moderator effects and despite the widespread knowledge of how to identify such effects statistically, moderator effects are notoriously difficult to detect in nonexperimental field studies. Many authors have lamented the difficulty of detecting reliable moderator effects in field studies (e.g., Jaccard, Helbig, Wan, Gutman, & Kritz-Silverstein, 1990; Morris et al., 1986; Zedeck, 1971). For example, Zedeck (1971) reported a number of unsuccessful attempts to find moderator variables and concluded that "moderators are as elusive as suppressor variables" (p. 305).

Even when reliable moderator effects are found, the reduction in model error due to adding the product term is often disconcertingly low. Evans (1985), for example, concluded that moderator effects are so difficult to detect that even those explaining as little as 1% of the total variance should be considered important. Champoux and Peters (1987) and Chaplin (1991) reviewed much of the social science literature and reported that field study interactions typically account for about 1%–3% of the variance.[3] We illustrate this problem shortly with an analysis of a case study in which the interaction was significant, the estimated partial regression coefficient was as large as it could theoretically be, and yet the squared partial correlation coefficient was only .01.

Periodically, the difficulty of detecting moderator effects leads investigators to question the appropriateness of assessing these effects by testing the partial regression coefficient of the XZ product as outlined earlier. Those investigators then published recommendations for alternative procedures that were subsequently shown to be flawed. For example, Allison (1977), Cronbach (1987), Dunlap and Kemery (1987), Friedrich (1982), and Wise, Peters, and O'Connor (1984) all demonstrated that various purported alternatives to moderated multiple regression are incorrect. We know of no credible published refutation of the appropriateness of testing the reliability of the partial regression coefficient for the product as a test of moderator effects. Yet, Evans (1991) noted that despite repeated warnings, many investigators are so frustrated by not finding theorized moderator effects that they continue to use and to invent inappropriate statistical procedures. This state of affairs undoubtedly reflects bafflement as to why moderator effects are so difficult to detect in field studies despite compelling theoretical expectations for such effects and despite the apparent ease with which such effects are found in experiments.

Relative Statistical Power

A number of factors accounting for the differential statistical power of experiments and field studies for detecting interactions have been identified. First, overall model error is frequently less in experiments because they are conducted under more controlled laboratory conditions than is possible in most field settings. Less noise means that reliable effects are easier to detect. Second, Busemeyer and Jones (1983) and Aiken and West (1991) discussed the considerable deleterious effect of measurement error on the detection of moderator effects. Errors in measuring X and Z are exacerbated when X and Z are multiplied to form the product variable XZ.[4] Studies conducted by experimentalists, who can assign observations to levels of X and Z and thus control measurement error to a greater extent than can field researchers, are less likely to be affected adversely by measurement error even though they are not immune. Third, theoretical constraints on the nature of the interaction in field studies often restricts the magnitude of the moderator regression coefficient. Experimentalists frequently report disordinal or crossover interactions, but theory often leads field researchers to expect only ordinal or fan-shaped interactions. For example, coping responses and social support are presumed to moderate the relation between stressful life events and adverse outcomes such as depression (e.g., Finney, Mitchell, Cronkite, & Moos, 1984; Pearlin, Menaghan, Lieberman, & Mullan, 1981). However, it may not be reasonable theoretically to presume that coping responses and social support can be so strong as to make stressful life events have an antidepressant effect. The theoretical constraint of ordinal interactions effectively limits the possible magnitude of β_{XZ} and therefore makes such effects more difficult to detect in field studies. Finally, other factors such as the functional form of the interaction (i.e., products of higher order terms) and the nonlinearity of X and Z are known to pose difficulties in the detection of moderator effects (Busemeyer & Jones, 1983; Jaccard et al., 1990; Lubinski

[2] Note, as we discuss more fully shortly, that this is not the same as saying that the multiplicative term accounts for a particular proportion of the variance.

[3] As Aiken and West (1991) noted, it is important to recognize that these authors are making statements about the squared semipartial correlation (i.e., increment in R^2), which is not the most useful effect size index in this context. Nevertheless, the point here is simply that reported effect sizes for moderator terms are often small.

[4] Kenny and Judd (1984) showed that these problems can be ameliorated using structural equation models.

& Humphreys, 1990). These problems pose difficulties for both experiments and field studies, although they may be more problematical for field studies, which generally have more levels of X and Z.

The aforementioned list is incomplete. That is, even if both experiments and field studies are affected to the same degree by these problems, experiments still have a considerable advantage over field studies in the statistical detection of interactions and moderator effects. We demonstrate that the difference in *relative* statistical power between the two types of studies is attributable at least in part to properties of the joint distribution of X and Z, and we quantify that difference in terms of relative efficiency.

Component–Product Covariances

Before considering the properties of the distributions of X and Z that facilitate the detection of interactions in experiments relative to field studies, we first eliminate one property that has caused considerable confusion in the literature. That property is the covariance between X and XZ or between Z and XZ. Some authors (e.g., Althauser, 1971; Morris et al., 1986) have mistakenly attributed the difficulty of detecting moderator effects to the inherent covariance between the individual components and their product. However, Aiken and West (1991), among others, noted that the covariances $C(X, XZ)$ and $C(Z, XZ)$ are almost always dramatically reduced when X and Z are centered before performing the moderated regression analysis. When X and Z are centered and are either jointly symmetric or stochastically independent, $C(X, XZ) = C(Z, XZ) = 0$ (Finney et al., 1984). Furthermore, Friedrich (1982) and Smith and Sasaki (1979) showed that it is *always* possible to change the origins of the X and Z scales (by subtracting appropriate constants) to ensure that $C(X, XZ) = C(Z, XZ) = 0$. Cohen (1978) demonstrated that changes of scale origin do not affect β_{XZ} or its statistical test. Given that a change of origin can always be found to ensure a zero covariance between the product and its components and given that such a change of origin does not alter the moderator statistical test,[5] the covariance, if any, between the components and their product is in principle irrelevant for detecting moderator effects. In practice, it is sometimes useful to transform the origin to reduce or eliminate $C(X, XZ)$ and $C(Z, XZ)$, thus avoiding computational problems in some computer algorithms for regression.

Residual Variance of the Product

At one level, the statistical power issues of testing the hierarchical addition of XZ to a model already containing X and Z are the same as those of testing the addition of any other predictor variable, say W, to the model. One such well-known power issue concerns the variability of W; the effects of variables with restricted ranges or reduced variances are difficult to detect and sizes of those effects are often small. In hierarchical multiple regression it is actually the residual variance in W (or the unique variation in W that is not shared with either X or Z) that determines the statistical power of the test. This suggests that the residual variance of XZ ought to be of concern when testing for moderator effects. Unlike the general situation of adding an arbitrary W to the model, the residual variance of XZ, indeed the complete distribution of XZ, is determined entirely by the joint distribution of X and Z. We provide a formula (derived in the Appendix) for the residual variance of XZ and then use this formula to demonstrate that, relative to experiments, statistical power is inherently low in field studies because of the typical joint distributions of X and Z.

We use the notation $V(XZ.X, Z)$ to refer to the residual variance[6] of the product XZ after controlling for X and Z. Although we prove a more general result in the Appendix, here we make the simplifying assumption that X and Z have already been centered so that their expected values are zero. Then, the residual variance of the product is given by

$$V(XZ.X, Z) = V(X)V(Z) + C(X^2, Z^2) - C^2(X, Z) - \frac{C^2(X^2, Z)V(Z) + C^2(X, Z^2)V(X) - 2C(X, Z)C(X^2, Z)C(X, Z^2)}{V(X)V(Z) - C^2(X, Z)}, \quad (2)$$

where Vs represent variances and Cs represent covariances.

A number of useful insights can be derived from Equation 2. First, Equation 2 demonstrates that the residual variance of the product is completely determined by the properties of the joint distribution of its components. Because the statistical power for detecting a moderator effect and estimates of the size of that effect depend on the variability of the residual product, we can use Equation 2 to examine the properties of the joint distribution of X and Z that are important in determining the residual variance of the product.

Second, Equation 2 shows that the residual variance of the product is simply the product of the component variances adjusted by various covariances. Not surprisingly, whatever restricts the ranges or variances of X and Z must also reduce the range and variance of the residual product. Furthermore, the multiplication of the component variances means that any range or low-variance problems are exacerbated when trying to detect moderator effects.

Third, Equation 2 demonstrates that several covariances provide important adjustments to the product of the component variances. Researchers using multiple regression need to be concerned about the correlation or linear dependence between predictors X and Z, represented in Equation 2 by the covariance $C(X, Z)$. However, Equation 2 implies that researchers using *moderated multiple regression* also need to be concerned about other covariances, namely, the covariances involving squares of the predictors. If X and Z are stochastically independent (this is much more than linear independence), then all of the adjustment covariances in Equation 2 equal zero, so that the residual variance of the product equals the product of the variances exactly. However, except for designed, balanced experiments, it is unlikely that all of the covariances will be zero in practice. Two bivariate distributions of X and Z having the

[5] Changes of origin will, however, alter β_X, β_Z, and their statistical tests. See Aiken and West (1991), Jaccard, Turrisi, and Wan (1990), or Judd and McClelland (1989) for careful instructions for testing and interpreting these coefficients.

[6] The residual variance of XZ should not be confused with the residual or error variance of Y.

same component variances $V(X)$ and $V(Z)$ may yield much different residual variances of the product because of their different covariance structures. Thus, we must consider how the covariances adjust the residual variance of the product.

What do the covariances in Equation 2 reveal about the joint distribution of X and Z and its impact on the residual variance of the product? We answer this question by providing three aids for understanding the representation of the residual variance of the product. First, we present simulations of experiments and field studies that demonstrate how reduced variances for the components X and Z exacerbate the difficulty of detecting an interaction. Second, we present an intuitive, geometric interpretation of the higher order covariances and their effects on the detection of moderator effects. Finally, we compare the residual variances for various bivariate distributions to the maximum possible residual variance for the product to explicate and to quantify the greater relative statistical efficiency of experiments versus field studies for detecting interactions.

Simulated Experiments and Field Studies

We demonstrate the importance of the product of the component variances (the first term in Equation 2) in determining the residual variance of the product by simulating both experiments and field studies in which all other covariances in Equation 2 are zero.[7] We also use the simulated experiments and field studies to show that the list of factors distinguishing the two types of studies must be incomplete. In these simulations, the overall model error was the same for both types of studies, measurement error did not exist, the underlying models and hence the sizes of β_{XZ} were the same, and the number of observations was held constant. Even after eliminating these possible reasons for differences in statistical power between experiments and field studies, the experiments still had a considerable advantage in their ability to detect interactions and moderator effects.

For the simulations of both the field studies and the experiments, $\beta_0 = 0$, $\beta_X = \beta_Z = \beta_{XZ} = 1$, there were 100 observations, and errors for the model were sampled from the same normal distribution with a mean of 0 and a standard deviation of 4. Hence, both types of simulated studies had (a) the same model, in particular, the same value for the moderator coefficient; (b) the same number of observations; and (c) the same model error, within sampling variation. The distributions of X and Z values were the only way in which the two types of simulated studies differed. For the experiment simulations, we used 2×2 factorial designs, with values of X and Z equal to $+1$ and -1 and an equal number of observations at each of the four combinations of X and Z values. For the field study simulations, we used values of X and Z that varied between the extreme values of $+1$ and -1. More specifically, in the field study simulations, values of X and Z were each sampled independently from a normal distribution with a mean of 0 and a standard deviation of 0.5. Values of X and Z were rounded to create equally spaced 9-point scales ranging from -1 to $+1$ because ranges in field studies are always finite and because ratings are often on scales with discrete intervals.

From 100 simulations each, estimates of the model parameter β_{XZ} for the moderator or interaction effect equaled 0.977 and 0.979 for the field studies and experiments, respectively; both types of studies produced unbiased estimates of β_{XZ}. Also, as intended, the root-mean-square errors were comparable in the two types of studies: 3.99 and 4.02, respectively. However, the standard errors of the estimate of the coefficient for the interaction were much different: 1.72 for the field studies but only 0.41 for the experiments. This difference in the standard errors of the estimate resulted in dramatic differences in the values of the Student's t statistic, median[8] $ts(96) = 0.66$ and 2.41, respectively. The null hypothesis of no moderator effect was rejected in a small minority (only 9%) of the simulated field studies even though it was rejected in a clear majority (74%) of the experiments, $t(96) = 1.98$, $p = .05$ (two-tailed). In other words, 91% of the simulated field studies made Type II errors by failing to reject a false null hypothesis. This difference was also reflected in the PREs or squared partial correlations for the moderator variable: Median PREs were .009 and .057 for the field studies and experiments, respectively, $PRE(96) = 0.039$, $p = .05$.

Figure 1 shows the distributions of the residuals $XZ.X, Z$ (i.e., the values of XZ after the effects of X and Z have been removed) for the simulated experiments and field studies. Note that the range for the field studies was greatly restricted relative to the experiments and that the distribution of the products was much more peaked than the distributions of the components. The variances of the residual products were 1.0 and .05 for the experiments and the field studies, respectively.

In these simulations, the difference in the residual variances of the products, which resulted from the difference in the joint distributions, was solely responsible for the dramatic superiority of the experiments over the field studies in the detection of the moderator effect. Thus, the simulations demonstrate that even if field studies and experiments are identical except for their joint distributions of X and Z, experiments still have dramatically more statistical power for detecting moderator or interaction effects. Clearly, properties of the joint distribution of the predictors must be added to the list of factors contributing to the difficulty of detecting moderator effects in field studies.

Interpretation of Higher Order Covariances

We have just demonstrated that the multiplication of the variances in Equation 2 seriously reduces the residual variance of the product for field studies relative to experiments. Next, we consider the effects of the various covariances. There are three major terms in Equation 2 that adjust the product of the component variances to yield the residual variance of the product. We consider the effects of each term on the residual variance.

The residual variance of the product will be augmented (relative to the product of the component variances) when $C(X^2, Z^2)$, the first adjustment term, is positive. $C(X^2, Z^2)$ is positive whenever extreme values of X, either positive or negative, cooccur with extreme values of Z, either positive or negative (re-

[7] Computer code for these simulations is available from the authors.

[8] We use medians instead of means because of the nonlinear nature of t. Our purpose here is not a definitive estimate of the test statistics but simply a demonstration that the type of design can have a dramatic impact on test statistics even when the underlying model is identical.

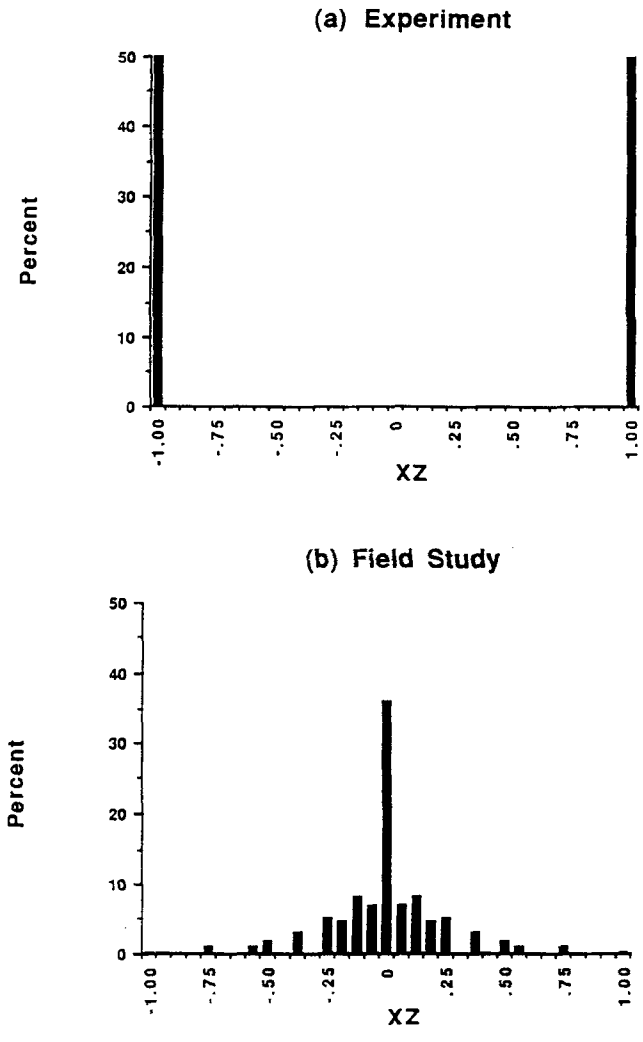

Figure 1. Residual variance of the product from simulations of experiments (a) and field studies (b).

member that X and Z are centered).[9] Such jointly extreme cases are obviously useful for identifying moderator effects. For example, suppose that Z moderates the linear effect of X on Y. Then, if there is no moderator effect, an extreme value of X will have a large impact on Y. However, a simultaneously large value of Z allows a large moderation in the large effect of X on Y. Large moderations of large effects are clearly easier to detect statistically than small moderations of small effects.

In essence, $C(X^2, Z^2)$ measures the extent to which extreme values of X co-occur with similarly extreme values of Z. Consider a view from above of the bivariate frequency distribution with X on one axis and Z on another axis. If X^2 always equaled Z^2, then all of the observations are on the diagonals so that the bivariate frequency distribution of X and Z forms a perfect "X pattern."[10] Thus, $C(X^2, Z^2)$ can be considered a measure of the "X-ness" in the joint distribution. $C(X^2, Z^2)$ has a minimum negative value when the joint distribution has a perfect "+ pattern" in which extreme values of one variable always co-occur with middle (or mean) values of the other variable. It is obvious why such observations are not useful for identifying a moderator effect. If Z is large but $X = 0$, then there is no effect for Z to moderate, and if X is large but $Z = 0$, then there is no moderation of the effect of X. In either case, there can be no moderator effect to detect. If there is a perfect + pattern, then one component of the product is always zero, implying that all values of $XZ = 0$. There is clearly no residual variance of the product in this case; this implies that for a + pattern, $C(X^2, Z^2) = -V(X)V(Z)$ so that $V(XZ.X, Z) = 0$ in Equation 2.

In summary, the first adjustment term in Equation 2 simply implies that the detection of moderators is facilitated to the extent that the joint distribution of X and Z has an X pattern and hindered to the extent that the joint distribution has a + pattern.

$C(X^2, Z^2)$ can be misleadingly large if observations tend to cluster on one of the two diagonals of the joint bivariate distribution. For example, in the extreme case when there is a perfect correlation, either positive or negative, between X and Z, then $X^2 = Z^2$ and maximizes $C(X^2, Z^2)$. To ensure that it is X-ness and not just "diagonal-ness" that is assessed, the second adjustment term in Equation 2 reduces the calculated value of the residual variance by subtracting $C^2(X, Z)$. As we show with a pictorial example later, increasing the correlation between X and Z, with all else being equal, improves the chances of detecting moderator effects because the increase in $C(X^2, Z^2)$ is greater than the adjustment for $C^2(X, Z)$.

The third and final adjustment term in Equation 2 is the most difficult to understand. The components of this term are the two covariances $C(X^2, Z)$ and $C(X, Z^2)$. It is easier to describe when these covariances will be zero, implying no adjustment to the residual variance, than it is to describe when they will have large values. Finney et al. (1984) showed that $C(X^2, Z) = 0$ if either (a) X and Z are stochastically independent or (b) if there is bilateral symmetry such that the frequency of the combination (X, Z) equals the frequency of $(-X, -Z)$. Thus, for either of these two covariances to be nonzero requires both bilateral skewness[11] and a complex form of stochastic dependence. In particular, $C(X^2, Z) \neq 0$ implies that the conditional distribution of X is skewed for at least some

[9] The central moment $E(x^2z^2)$ is closely related to $C(X^2, Z^2)$—see the Appendix—and is a key component in Mardia's (1970) definition of bivariate kurtosis. Unfortunately, there are many other terms in that definition of bivariate kurtosis, so we cannot make exact statements; however, all else being equal, it is generally true that increasing bivariate kurtosis implies greater variability in the residual of XZ.

[10] We are grateful to Charles Berry and Ewart Shaw for suggesting the interpretation of the higher order covariances in terms of their patterns in bivariate contour plots.

[11] The central moments $E(xz^2) = C(X, Z^2)$ and $E(x^2z) = C(X^2, Z)$—see the Appendix—along with $E(x^3)$ and $E(z^3)$, are the key components in Mardia's (1970) definition of bivariate skewness. Thus, although, again, the statement cannot be exact, it is generally true that increasing bivariate skewness decreases the residual variance of the product. Note also that for centered variables, $C(X^2, Z) = C(X, XZ)$ and $C(X, Z^2) = C(Z, XZ)$; in other words, these terms represent the remaining covariances between the individual predictors and their products that are *not* eliminated by centering. In a sense, then, multicollinearity between the predictor variables and their product that is not reduced by centering is a problem for detecting moderator effects.

value of Z and that the degree of the skewness of X depends on Z. When looking at a bivariate frequency distribution from above, large positive values of $C(X^2, Z)$ indicate, with all else being equal, a "left bulge," or "C pattern," and large negative values indicate a "right bulge," or "reversed C pattern." That is, extreme values of X tend to occur with middle values of Z. Similarly, large positive values of $C(X, Z^2)$ indicate a "downward bulge," or "U pattern," and large negative values indicate an "upward bulge," or "inverted U pattern"; either bulge implies that extreme values of Z tend to occur with middle values of X. Thus, the third term in Equation 2 essentially adjusts downward[12] for ways other than + -ness in which extreme values of one variable may co-occur with middle values of the other variable.

In summary, Equation 2 indicates that the residual variance of the product is greater and hence that moderator effects are easier to detect to the extent that (a) extreme values occur (i.e., that the component variances are large) and (b) extreme values of each predictor variable co-occur with extreme values of the other predictor variable. In terms of the bivariate distribution of X and Z, the residual variance of the product is greater the more that observations have an X pattern (rather than a + pattern), the less that observations are concentrated on one diagonal of the X pattern, and the less that there are any asymmetric bulges.

Note that Equation 2 makes it possible to calculate $V(XZ.X, Z)$ for any joint distribution of X and Z that may be anticipated in a study. Calculating the residual variance of the product is useful in any power analysis. However, if a sample is already available, it is usually simpler to obtain the residual variance directly from standard regression programs rather than from Equation 2. $V(XZ.X, Z)$ equals, by definition, the mean square error (MSE) that results from regressing XZ on X and Z. Alternatively, in a full moderated regression model (i.e., Equation 1),

$$V(XZ.X, Z) = \frac{MSE}{n(s_{est}^2)}, \tag{3}$$

where s_{est} is the usual standard error of the estimate for the product.[13]

Maximum Values of V(XZ.X,Z)

The importance of the residual variance of the product in determining the statistical power of the moderator test suggests considering the maximum possible value of $V(XZ.X, Z)$. It is well known that for a finite range, the variance of a predictor variable is maximized when exactly half of the observations are at each extreme. It is easy to see that $V(XZ.X, Z)$ is maximized when both X and Z have maximum variances for their ranges. This in turn implies that the residual variance of the product is maximized when one fourth of the observations are at each extreme corner of a 2×2 design. This and similar results for other designs are proved formally in the literature on the optimal design of experiments. (See Mead, 1988, for a useful textbook presentation of optimal design.) The use of the term *optimal* in this literature means that such designs provide maximum statistical power and the smallest confidence intervals. It does not mean that researchers, especially field researchers, necessarily *ought* to strive for optimal designs. We discuss this issue later.

It is difficult to compare values of $V(XZ.X, Z)$ across different studies because of differences in the ranges and scales of predictor variables. However, the existence of a maximum possible residual product variance for fixed ranges suggests using that maximum to create an index of design efficiency. We define the relative efficiency of a design as the ratio of its $V(XZ.X, Z)$ to the maximum possible value of $V(XZ.X, Z)$ for an optimal design with the same ranges of the predictor variables. Another way to make this comparison is to determine, assuming equal mean square errors, the number of observations a design must have to provide the same efficiency (i.e., the same standard error of the estimate for the moderator coefficient) as an optimal design. If the optimal design has n observations, then to have equal efficiency any other design needs to have a number of observations equal to n times the inverse of the relative efficiency.[14]

As an example, reconsider the simulations described earlier. The experiments were optimal, with one fourth of the observations at each extreme combination of -1 and $+1$. The variances of X and Z were then each 1 and, because all of the covariances were 0, the residual variance of the product equaled $1 \times 1 = 1$. By contrast, the variances of X and Z from the simulations of the field study were only about .24 (the rounding in the sampling design required that these variances be estimated empirically). Furthermore, small sampling variability in the values of the covariance and the bivariate skew and kurtosis components reduced $V(XZ.X, Z)$ to an average value of .052, slightly below the product (.058) of the two component variances. Thus, the relative efficiency of the simulated field studies was only 5.2%. To be as efficient as the experiments with 100 observations, the field studies needed approximately $1/.052 = 19.23$ times as many observations, for a total of $19.23(100) = 1{,}923$ observations. Note that this conclusion did *not* depend on the effect size for the moderator term; no matter what the magnitude of β_{XZ}, the field studies required 19.23 times as many observations as the experiments to produce the same standard error of the estimate of the moderator effect.

---

[12] We conjecture, but cannot prove, that this adjustment must always be downward. The pattern of signs in this last term makes it appear that an upward adjustment might be possible, but we cannot construct a bivariate distribution where this is the case. The last term in this adjustment appears to correct for double counting of bulges along diagonals.

[13] Note that the actual residual variance of the product in the sample is the determinant of power, so n is the appropriate divisor when computing the variances and covariances in Equation 2. Hence, it may be necessary to correct the estimates from regression programs in which other divisors are used; however, such precision is usually not necessary because the goal is to obtain only a general idea of the statistical power in a particular study.

[14] Increasing the number of observations to achieve equivalent efficiency is not the same as obtaining equivalent statistical power because critical values of the test statistics change with n and because of nonlinearities in power functions. However, the n necessary to achieve equivalent efficiency is a good indicator of the n needed to obtain equivalent statistical power.

Examples of Design Relative Efficiency

It would be useful to examine the impact on relative efficiency of all of the different ways in which bivariate distributions arising in field studies might differ from those in optimally designed experiments. This is not feasible because of the infinite number of possible bivariate distributions. Instead, we consider some examples that illustrate various effects of the properties of the bivariate distribution on relative efficiency. Figure 2 displays a variety of possible bivariate distributions for 5×5 designs. The bold number above each distribution in Figure 2 is the efficiency of that design relative to the optimal "four-corners" design in the upper left-hand corner of Figure 2. These distributions are scaled so that the modal category is equally high in each plot in order to emphasize the basic shape of each distribution. The variances and covariances reported shortly were based on assigning the values $-1, -.5, 0, .5$, and 1 to the five levels of each variable.

The first distribution in the first row depicts the optimal design with equal numbers of observations at each corner. The other two distributions in that row illustrate how efficiency decreases as one of the four corners increasingly dominates. This is the classic "unequal n" problem for a two-way ANOVA.

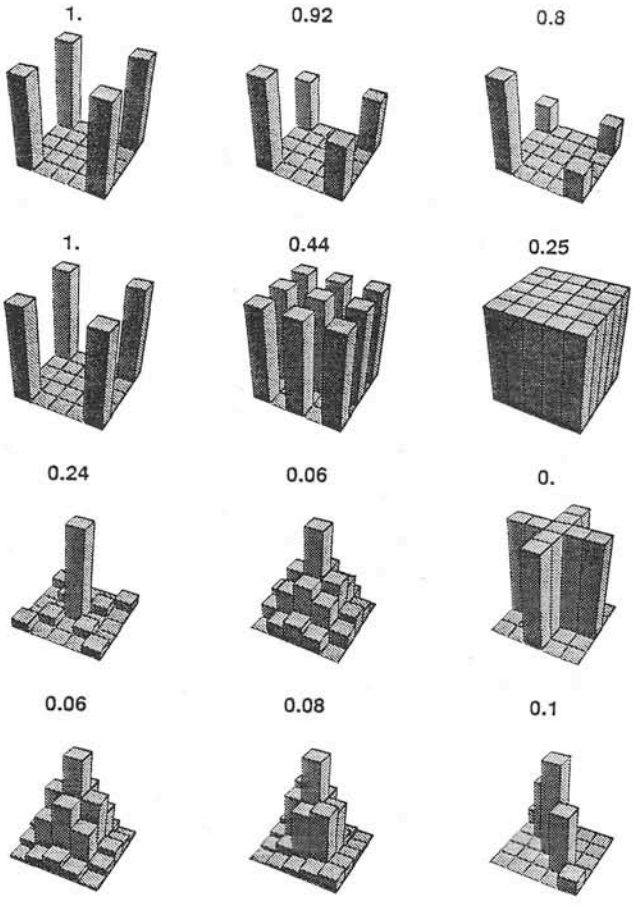

Figure 2. Illustrative joint distributions and their relative efficiencies for detecting interactions. (Note that distributions are rescaled so that the modal category has equal height across distributions.)

Note that a small disparity in the number of observations in each corner has little effect and even when 50% of the observations are concentrated in one of the corners, relative efficiency for the interaction is still 80%.

The second row of Figure 2 illustrates the effect of using intermediate categories. Clearly, adding intermediate categories adversely affects relative efficiency. The relative efficiency of the equal distribution across all combinations of the 5×5 design is only 25% of the maximum possible; in other words, to have equal efficiency for assessing an interaction, the equal distribution design needs four times as many observations as the optimal four-corner design. Of course, using intermediate levels allows the testing of higher order interactions and univariate polynomial effects, tests that are not possible with the optimal design. For example, the middle design in the second row can test for univariate quadratic effects as well as Linear $\times$ Quadratic and Quadratic $\times$ Quadratic interactions, which the first design cannot.

The designs in the third row are calibrated so that the univariate variances of X and Z are equal and then rescaled in the display so that the frequency of the modal category is constant across the row. The left-most design has maximal X-ness (given the constraint of equal univariate variances), and the right-most design has maximal +-ness. The middle design in the third row is as close to being exactly between these two extremes as is possible given the coarseness of the design and the equal variance constraint; that is, $C(X^2, Z^2)$ is approximately zero for the middle design. The relative efficiency of the X design is only 24%, primarily because of the need to stack so many observations in the center to equate the variances. The relative efficiency of the + design is, of course, zero; it cannot detect an interaction. Interestingly, the intermediate design has a relative efficiency of only 6%, much less than half of the relative efficiency of the X design. Apparently, having even a few observations in the most extreme categories is crucial for assessing an interaction. Note that because all of the designs in the third row have equal variances, and indeed equal marginal distributions, they demonstrate that much more than the product of the variances can be important in determining relative efficiency for assessing moderator effects.

The first design in the last row of Figure 2 is as close to the bivariate normal distribution as is possible with a 5×5 design. Its relative efficiency is approximately 6%. The middle design in the last row has the same marginal distributions but a negative correlation between X and Z of approximately $-.50$. The last design in that row also has the same marginal distributions but a *perfect* negative correlation between X and Z. Note that increasing the correlation between X and Z increases the relative efficiency for assessing a moderator effect. For example, for the last design the perfect correlation between X and Z increases $C(X^2, Z^2)$—to .10—more than it increases $C^2(X, Z)$—to .06—so the net effect is an increase in relative efficiency. The middle design illustrates the increase in relative efficiency for a moderate correlation. Of course, the last design in the last row is useless for distinguishing the individual effects of X and Z but, surprisingly, it has more power for detecting the interaction.

In summary, the distributions in Figure 2 illustrate that jointly extreme observations are crucial for detecting interac-

tions. The joint distributions of X and Z can be highly unusual and still provide adequate power for detecting the interaction as long as there are jointly extreme observations. These illustrations also demonstrate that a given design can have much different relative efficiencies for assessing interactions as opposed to first-order linear effects.

A Case Study

It would be useful to calculate the relative efficiency for studies that have failed to show theorized interactions or that indicated statistically significant interactions that reduced the residual error from the additive model by only 1% or 2%. However, few such studies have published the entire bivariate distribution of X and Z and none, as far as we know, published $V(XZ.X, Z)$ or the important covariances $C(X^2, Z^2)$, $C(X, Z^2)$, and $C(X^2, Z)$ that are required to calculate it.

In lieu of a reconsideration of prior studies, we present as a case study an examination of a data set from a field study conducted by Richard Jessor and his colleagues.[15] We focus here on only three of the many variables used in that study. The dependent variable Y was an index of the number and severity of adolescent problem behaviors; X was a 7-point index (0–6) of risk factors that are presumed to predispose an adolescent to problem behaviors; and Z was an 8-point index (0–7) of factors that might protect an adolescent against the risks to which he or she is exposed and thereby moderate the relationship between risks X and problem behaviors Y.

Observations from 1,646 adolescents produced the following parameter estimates:

$$Y = \text{controls} + 8.7X - 1.49Z - 1.23XZ \qquad (4)$$

(overall $R^2 = .24$, $MSE = 600.54$). The intercept and several control variables (such as sex) were grouped into a single term (controls) that was not relevant for the interaction issues discussed here. The test of the moderator effect (i.e., whether -1.23 was reliably different from 0) yielded an F ratio of 17.92 (dfs = 1 and 1638), $p < .0001$. Although the interaction was statistically reliable, it accounted for little additional variation in Y; the PRE, or squared partial correlation, for the interaction was only .01.

Rewriting the estimated model as

$$Y = \text{controls} - 1.49Z + (8.7 - 1.23Z)X \qquad (5)$$

showed that the interaction was consistent with the theory. That is, when there was no protection ($Z = 0$), the slope between risk (X) and problem behavior involvement (Y) was positive (slope = 8.7); however, for each unit that protection increased, the slope between risks and problem behaviors decreased by 1.23, until at $Z = 7$, representing the maximum level of protection, the slope between risks X and problem behaviors Y reached a minimum of 0.09. That is, a high level of protection essentially eliminated the relationship between risks and problem behaviors. A disordinal interaction, which would imply that risk exposure combined with a high level of protection could reduce problem behaviors, seemed theoretically inappropriate in this case. If a disordinal interaction is not appropriate, then given the slope of 8.7 for unprotected adolescents (i.e., when $Z = 0$), the greatest possible magnitude for the moderator coefficient occurs when the moderated slope equals zero; that is, when $8.7 - \beta_{XZ}(7) = 0$, which implies $\beta_{XZ} = -1.24$. Thus, the estimated coefficient of -1.23 is essentially as extreme as it can theoretically be. If the theoretically expected interaction is reliable and is as large as it can possibly be, then why does it account for so little variance? We answer this question by examining the joint distribution of X and Z shown in Figure 3.

Clearly, the joint distribution in Figure 3 is skewed and peaked so that relatively few combinations of X and Z account for most of the observations. This distribution is obviously much different from the optimal four-corner design. The means for X and Z are 1.51 and 2.16, respectively, and the relevant variances and covariances for calculating $V(XZ.X, Z)$ are $V(X) = 1.956$, $V(Z) = 3.033$, $C(X^2, Z^2) = 0.1749$, $C(X, Z) = -1.0088$ (which implies a correlation between X and Z of -0.414), $C(X, Z^2) = -0.6618$, and $C(X^2, Z) = -0.7414$. Using these variances and covariances in Equation 2 to calculate the residual variance of the product yields

$$\begin{aligned} V(XZ.X, Z) &= V(X)V(Z) + C(X^2, Z^2) - C^2(X, Z) \\ &\quad - \frac{C^2(X^2, Z)V(Z) + C^2(X, Z^2)V(X) - 2C(X, Z)C(X^2, Z)C(X, Z^2)}{V(X)V(Z) - C^2(X, Z)} \\ &= 5.9325 + 0.1749 - 1.0177 - 0.7149 \\ &= 4.37. \end{aligned} \qquad (6)$$

The maximum possible value of $V(XZ.X, Z)$ for a design with the same ranges for X and Z would equal the product of the two variances with one fourth of the observations at each extreme combination. In this case, the variances are $3^2 = 9$ and $(3.5)^2 = 12.25$, respectively, and their product equals 110.25. Thus, the relative efficiency is only $4.37/110.25 = 4.0\%$ of the maximum possible for an optimal design. In other words, an optimal design with only 65 observations has the same efficiency as the present joint distribution of X and Z with 1,646 observations. Note that these values are highly similar to those in the simulated studies described earlier and that they do not depend on the magnitude of the moderator effect.

The calculations in Equation 6 also show the primary reason that the residual variance of the product is so small in this case: The two variances are only about one fifth and one fourth, respectively, of the maximum possible values for their ranges. Compared with the maximum possible residual variance, the other adjustments (upward by 0.17 for slight X-ness, downward by -1.01 for covariance between X and Z, and downward by -0.7149 for various asymmetries) are not large. Thus, this example is consistent with our conjecture that the most important components in determining $V(XZ.X, Z)$ are the individual variances and that the effects of the adjustments for the various covariances are relatively small compared with the effect of moving observations from the corners to the center of the joint distribution. Even so, however, the effects of the covariance and

[15] The illustrative analysis presented here should not be considered a definitive representation of these data. We consider only 1 year of a longitudinal data set and only one way of operationalizing the constructs. The research described here was conducted as part of a larger project supported by the William T. Grant Foundation (Grant 88-1194-88 to Richard Jessor, principal investigator).

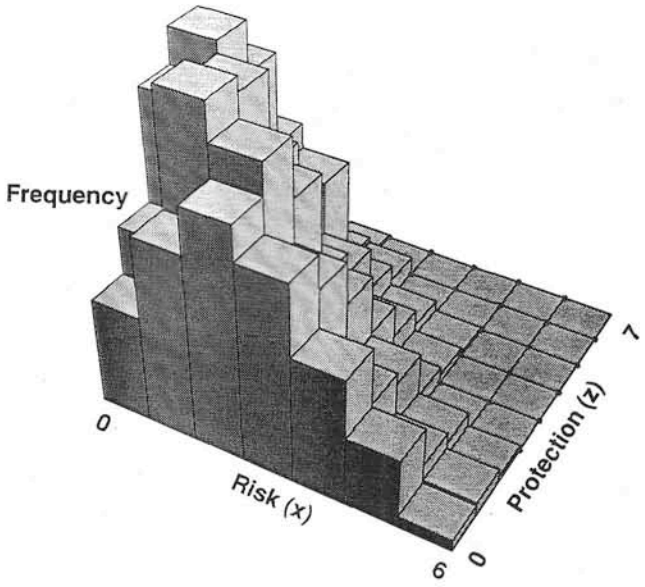

Figure 3. Joint distribution of X (risks) and Z (protective factors) for the case study.

the bivariate skewness components are not trivial. In the present case, if we suppose that the correlation between X and Z and the asymmetry covariances were zero, then the residual variance of the product would be 6.11, almost 40% larger. In other words, a joint distribution with the same central peak but with all those covariance components equal to zero would need only about 1,177 observations to produce the same statistical efficiency for estimating the moderator coefficient as this study with 1,646 observations.

Clearly, the distribution in Figure 3 produces significantly less statistical power than does an optimal distribution. However, the large sample size in this example provides adequate power: A reliable moderator effect is obtained. The question remains as to why the moderator term accounts for so little of the residual variance after removing the effects of X and Z. An expression for PRE from Judd and McClelland (1989) is as follows:

$$\text{PRE} = \frac{1}{1 + \dfrac{\text{MSE}}{b_{XZ}^2 V(XZ.X, Z)}}. \quad (7)$$

Note that Equation 7 implies that for a constant moderator coefficient b_{XZ}, the effect size will vary as a function of the distribution of the predictor variables. If we assume, consistent with the standard assumptions of least squares regression, that MSE and b_{XZ} are the same for an optimal design, then the PRE for an optimal design would equal a respectable

$$\text{PRE} = \frac{1}{1 + \dfrac{600.54}{-1.23^2(110.15)}} = .217. \quad (8)$$

This suggests that it is the nonoptimal distribution of X and Z and not the magnitude of the moderator coefficient that causes the additional variance reduction to be so small. Of course, the joint distribution of X and Z is in fact the one observed in the field study, so that one expects the interaction to produce only a 1% reduction in error in the field. However, it is not appropriate to dismiss a 1% interaction found in the field because, as this example shows, it is equivalent to an interaction reducing error 21.7% in an optimal design. Obviously, there are serious dangers in comparing variance explained by interactions across studies that differ substantially in the optimality of their designs.

Generalizations and Special Cases

Our focus has been on the difficulty of detecting a single moderator effect or two-way interaction. It is now useful to apply our results for the residual variance of the product both to several special, often simpler, cases such as linear effects and quadratic effects and to more complicated cases such as multiple two-way and higher order interactions.

Linear Effects

Field researchers using designs with less statistical power than the maximum possible may believe that because they can detect linear or first-order effects, that they also have adequate power for detecting higher order effects. However, the power issues are much different for the special case of linear effects. Consider the simulation example in which the individual variances for the field studies are approximately one fourth of those for the experiments. To have comparable statistical efficiency for estimating linear effects, the field studies need four times as many observations (400 vs. 100 in the simulation) as the experiments. For detecting linear effects of X and Z in a simple additive model in a field study, one fourth of the variance might be sufficient or might reasonably be compensated by four times as many observations.

The situation is much different for detecting an interaction or moderator effect. Even assuming no complications from the higher order covariances, the residual variance of the product for the field study is only about $(1/4)(1/4) = 1/16$ of that of the experiment. Thus, 16 times as many observations (1,600 vs. 100) are needed in the field study for comparable efficiency. Thus, even though nonoptimal designs or particular sample sizes may be adequate for detecting linear effects, they may simultaneously be grossly inadequate, relative to optimal designs, for detecting moderator effects of comparable magnitude.

Quadratic Effects

The detection of quadratic effects is a special case of detecting moderator effects. Including X^2 in a regression model, equivalent to letting $Z = X$ in Equation 1, asks whether X itself moderates the relationship between X and Y. There are three important reasons for considering the special case of the quadratic effect. First, researchers with good theoretical reasons for expecting to find quadratic effects have similarly lamented the difficulty in finding these effects. Our results regarding moderator effects also illuminate this difficulty. Second, to con-

sider the special case of quadratic effects we need only consider the implications of the univariate distribution X for the residual variance of X^2. The univariate distribution is inherently simpler than the bivariate distributions we considered earlier; this makes it easier to develop intuitions and insights about the practical implications of our mathematical results. Third, Lubinski and Humphreys (1990) demonstrated that it is important to control for the quadratic effects of X and Z before testing for an interaction between them; otherwise, spurious moderator effects may result. Judd and McClelland (1989, p. 274) similarly demonstrated that controlling first for quadratic effects can eliminate interactions that are statistically significant if only the linear effects of X and Z are controlled. However, researchers following the advice of Lubinski and Humphreys need to be aware of the way in which the univariate distributions of X and Z might compromise the statistical power of their quadratic tests.

Using the customary definitions of skewness and excess kurtosis,[16] we show in the Appendix that the residual variance of X^2, controlling for X, is given by the following:

$$V(X^2.X) = V^2(X)[\text{kurtosis}(X) + 2 - \text{skew}^2(X)]. \quad (9)$$

The effects of the univariate distribution of X on the residual variance of X^2, and hence the effects on statistical power, with all else being equal, are clear. $V(X^2.X)$ increases with increasing $V^2(X)$, which is the variability of X itself, with increasing kurtosis, and with decreasing absolute values of skewness.

The maximum possible residual variance for the quadratic effect results when one fourth of the observations are at either extreme of X and the remaining half of the observations are exactly halfway between those two extremes.[17] For example, for five possible levels of X, the optimal distribution is ($^1/_4$, 0, $^1/_2$, 0, $^1/_4$). That is, the relative efficiency of this distribution for assessing the quadratic coefficient is 1.0. This design is not often used by experimentalists to test for quadratic effects. Instead, equal numbers of observations are used at three levels of X giving the distribution ($^1/_3$, 0, $^1/_3$, 0, $^1/_3$), which has a relative efficiency of $8/9 = 88.9\%$. In other words, the usual experimental design needs $9/8 = 1.125$ times as many observations as the optimal design to have the same efficiency.

In most field studies, there are more than three levels of X. If observations are distributed across more categories, then relative efficiency decreases. For example, the relative efficiency of the distribution ($^1/_5$, $^1/_5$, $^1/_5$, $^1/_5$, $^1/_5$) is only 70%, needing 1.43 times as many observations as the optimal design for comparable efficiency. In the extreme, an infinite number of categories is equivalent to a continuous uniform distribution that has a relative efficiency of only 36%, requiring 2.78 times the n for the optimal design. Clearly, spreading observations over many categories has a serious deleterious effect on the ability to detect quadratic effects.

Uniform distributions are, of course, uncommon in field studies. A more typical distribution might be ($^1/_{15}$, $^3/_{15}$, $^7/_{15}$, $^3/_{15}$, $^1/_{15}$), which has a relative efficiency of only 21.8% and requires approximately 4.6 times as many observations to have the same efficiency as an optimal design. These particular numerical examples are not as important as the realization that having many categories or having peaked distributions substantially reduces relative efficiency for assessing the quadratic effects. Note also that Equation 9 implies that any asymmetry in the distribution of X reduces further the variance of the quadratic term and hence reduces statistical power.

Multiple Two-Way Interactions

In many studies, researchers evaluate two or more moderators in a single analysis. For example, a regression model might include terms for XZ, WX, and WZ. Researchers sometimes make inferences about relative importance when, say, one of the three interaction terms is statistically significant and the others are not. However, such inferences require equivalent statistical power for each test. It might well be the case that the interaction terms are equivalent in terms of the sizes of their partial regression coefficients but that there are differences in statistical reliability due entirely to differences in the residual variances of the interaction terms. Thus, when examining multiple two-way interactions, one ought to compare the residual variances of those interactions before making any inferences about their relative importances.

Higher Order Interactions

Our results for two-way interactions are easily generalized to higher order interactions. For example, consider the simulated field studies again and suppose that there is a third variable W with a distribution similar to those of X and Z. In the simulations, the variances of X and Z and, now, W are each a little less than one fourth of the variances in the optimal experimental design. The residual variances of the two-way interactions are then less than $(^1/_4)^2 = {}^1/_{16}$ of the residual variances of the optimal design for detecting two-way interactions. For the three-way interaction WXZ, the residual variance is less than $(^1/_4)^3 = {}^1/_{64}$ of the residual variance in the optimal design. In other words, for the field studies to have efficiency for assessing the coefficient for the three-way interaction comparable to an optimally designed experiment with 100 observations requires $100(64) = 6{,}400$ observations. Thus, field studies are at an even greater disadvantage, relative to experiments, in terms of their ability to detect higher order interactions. The problem is that the interaction is best identified by observations jointly extreme on W, X, and Z, but such observations are extremely rare, more rare than observations that are jointly extreme on any two of the three predictor variables. Similar arguments apply to three-way interactions such as X^3, X^2Z, and XZ^2.

---

[16] *Kurtosis* is defined here as the excess kurtosis relative to the normal distribution; hence, as used here, the kurtosis of the normal distribution is zero.

[17] Note that this is the optimal design for detecting a quadratic effect and that it is not necessarily the best distribution of observations for detecting other effects. For example, it is not the best design for detecting a linear effect and it, of course, has no possibility of detecting a cubic effect with just three distinct values of X. Studden (1982) considered designs optimal for detecting a quadratic effect while still allowing the possibility of detecting higher order effects. However, our focus here is on only the quadratic effect.

Discussion

Summary

We have a clear answer to the question of why field studies have more difficulty detecting moderator effects or interactions than do experiments; field studies, relative to experiments, have nonoptimal distributions of X and Z, and this means the residual variance of the product, $V(XZ.X, Z)$, is relatively lower, which in turn means that the efficiency of the moderator parameter estimate and statistical power is much lower. Simulations, illustrative examples, and a case study demonstrate that the difference in efficiency between field studies and optimally designed experiments is dramatic. Moreover, even when nonoptimal designs successfully detect an interaction, the reduction in variation attributable to adding the interaction term to an additive model is likely to be small because of the important role of $V(XZ.X, Z)$ in the calculation of effect size. These disadvantages of field studies relative to experiments are in addition to problems such as measurement error that have previously been identified. Furthermore, the relative disadvantage of field studies becomes worse when considering higher order interactions.

When will the residual variance of a product be large and when will it be small? Equation 2 for $V(XZ.X, Z)$ and the examples from Figure 2 suggest that the most important determinants of the residual variance of the product are the variances of X and Z. Whatever reduces the individual variances will have an even greater impact on the reduction of the product variance because of the multiplication. Thus, problems caused by reductions in the ranges or variances of the predictor variables are compounded when testing for interactions.

Factors that can reduce the variance of an individual predictor include restrictions on its range, the clustering of observations in the center of its range rather than at the extremes, and the distribution of observations over many categories within its range. Field researchers have little control over such factors and so must contend with whatever range occurs, a unimodal and often skewed distribution with few observations at the extremes, and, if good measures are used, multiple categories between the extremes. The experimentalist, on the other hand, can often use extreme ranges, can concentrate observations at the extremes, and can avoid having observations in middle categories. All of these factors increase the variability of X and Z and therefore increase the variability of XZ. It is important to note that our comparisons between optimal and nonoptimal designs presume a common range. These comparisons show the overwhelming superiority of optimal designs in detecting interactions. However, the ability of experimentalists to use even more extreme ranges *and* optimal designs further enhances their ability to detect interactions. In other words, a bleak assessment of the relative ability of field studies to detect interactions is really a best-case scenario; estimating the relative efficiency using Equation 2 actually understates the true relative superiority of experiments in detecting interactions.

It is a mistake to presume that only the individual variances are important in determining the residual variance of the product. Although the individual variances probably have the greatest effect in reducing the residual variance to a small fraction of its maximum possible value, the adjustments for covariance and bivariate asymmetric patterns can reduce small residual variances even further. This was demonstrated in the case study considered earlier. When statistical power is already low, further reductions caused by covariance and asymmetry are costly. Covariances and asymmetries that reduce the likelihood that extreme values jointly occur on X and Z necessarily reduce relative efficiency for detecting moderator effects.

Implications for Field Research Design

What are the lessons for field researchers? Unfortunately, it is easier to list unwarranted inferences that someone might be tempted to draw from these results than it is to provide helpful advice. For example, an uncritical consideration of the effects of multiple categories and clustering of observations away from the extremes might suggest that field researchers should either use measures of X and Z with fewer categories or use procedures such as median splits to recode observations into fewer categories. However, doing so is a *serious* mistake. Using imprecise measures does not mean that observations are really in few categories and simply recoding observations into extreme categories does not make those observations truly extreme. Using imprecise measures (i.e., those with few categories) implies that X and Z are measured with more error; Busemeyer and Jones (1983) demonstrated the serious deleterious effects of increased measurement error on the detection and interpretation of interactions. Maxwell and Delaney (1993) showed that in some circumstances median splits reduce statistical power and in other circumstances produce spurious interaction effects. The experimentalist's advantage is not due only to having fewer categories but also to being able to ensure that observations in those categories are truly extreme.

Another example of an unwise strategy is collecting data from a random sample and then applying moderated multiple regression on only a subsample that is as close to an optimal design as possible. Analyzing data only from this subsample actually decreases statistical power for detecting the interaction. To see why this is so, suppose we have an optimal subsample. What are the costs and benefits of including another observation from the larger sample in the analysis? The least useful observation that we can add, the one that decreases the residual variance $V(XZ.X, Z)$ the most, is an observation exactly equal to the mean of the residual product. This observation reduces the residual variance to $(n-1)/n$ of what it is for the optimal design. However, to calculate the variance of the error of the coefficient estimate, the crucial term for statistical power, the residual variance is multiplied by the sample size, which in this case is $n/(n-1)$ larger than it is for the optimal design. Hence, in the worst case, the reduction in the residual variance of the product is exactly balanced by the increase in the sample size. For any additional observation that is not exactly equal to the mean of the residual product, the effect of the increase in sample size is greater than the effect of the decrease in the residual variance for a net gain in efficiency and statistical power. Therefore, adding observations to an optimal design never does harm and instead will almost always be beneficial. Thus, field researchers should always use their full samples in moderated multiple regression.

Knowing that adding observations to an optimal design can do no harm suggests that the fundamental problem in field studies is that a subsample satisfying an optimal design simply does not exist. If a large field study does contain a subsample of, say, 60 observations satisfying an optimal design, then that field study has at least as much statistical power, all else being equal, as an optimally designed experiment with 60 observations. For example, note that in Figure 3 an optimal subsample is not even possible—there are simply no observations with maximum values on both X and Z. Moderated regression models used in field studies may make their most interesting predictions about observations that are rare or that may not exist at all.

What, then, can field researchers do to improve their chances of detecting interactions or quadratic effects? Obvious methods for increasing statistical power are to accept higher rates of Type I errors (a strategy not usually acceptable to journal editors) or to increase the number of observations. However, our results suggest that increasing observations may be impractical because typical field study joint distributions of X and Z are so inefficient that enormous samples are required to have the statistical power of optimally designed experiments for detecting interactions.

One interesting alternative strategy, suggested by considering the residual variance of the product, is not to sample randomly but instead to oversample extreme observations so as to ensure that at least a subsample is close to an optimal design. This strategy is controversial. On the one hand, parameter estimates in the moderated regression model will still be unbiased, even with the oversampling, and those estimates will be more efficient in terms of having smaller standard errors and hence narrower confidence intervals. On the other hand, the overall R^2 for the total model and the PRE for the product term will both be higher, perhaps considerably higher, than they are with a random sample. In other words, if one's goal is to estimate the total variation explained by the moderated regression model, then oversampling extreme observations produces a seriously inflated estimate of R^2. However, this seems no greater sin than that committed by experimentalists using optimal but unrepresentative designs.

The bias in estimating PRE for the product term itself is less problematic because the interpretation of the PRE is ambiguous for the case of interactions. Although it is appropriate to view the moderator PRE in the hierarchical test as the proportion of the residual variation in Y attributable to the product after the effects of the "additive" components are removed, it is not appropriate to interpret the moderator PRE as the proportion of the variation uniquely attributable to the multiplicative part of the model.[18] To know the true multiplicative proportion requires knowing the true origins of the scales; in other words, X and Z must be ratio rather than interval scales. If one were allowed to change the origins of X and Z, as is the case in moderated multiple regression, then a change of origin always exists that makes $\beta_X = \beta_Z = 0$; this assigns all of R^2, the variation explained by the entire model, to the multiplicative component, leaving none for the additive components. Birnbaum (1973) and Anderson and Shanteau (1977) demonstrated that even when the true model is entirely multiplicative (i.e., $\beta_X = \beta_Z = 0$), an additive model can "account" for a high proportion of the total variance. Thus, a biased estimate of PRE does not seem to be a serious problem. The more important question is whether any multiplicative effect exists; that is, it is more important to know that PRE does not equal zero than to know exactly what it equals.

Oversampling extreme observations is controversial. However, if it is theoretically important to demonstrate a moderator effect, then oversampling observations extreme on both X and Z may be the best approach. Field researchers might use stratified sampling to obtain the extreme observations in much the same way stratified sampling is used to ensure adequate subgroup sample sizes. The unweighted sample could then be used to test the moderator effect and a weighted sample (on the basis of the stratification scheme) could be used to estimate the population effect sizes. At the least, field researchers should be aware of the consequences of the nonoptimality of their designs and perhaps should publish $V(XZ.X, Z)$ and the relative efficiency of their designs so that readers can better assess the likelihood that an interaction could be detected. Also, field researchers may want to report the comparison of the obtained moderator coefficient to the theoretical maximum value as we did in the case study earlier.

Implications for Experimentalists

What are the implications for our results on residual variances of products for experimentalists? First, experimentalists need to be aware that they often will have enormous statistical power for detecting interactions compared with their colleagues who conduct field research. They need to be aware that the failure of those colleagues to find corresponding interactions may not be due to sloppier procedures but to the field studies' much lower relative statistical power for detecting interactions and moderator effects.

In addition, there is at least one context—the analysis of covariance—in which even the experimentalist is plagued by the relative lack of power in testing interactions that results from less than optimal designs. The distribution of scores on the covariate is likely to be unimodal with few extreme cases so that the residual variance for the product of the covariate and any categorical predictor is likely to be relatively small compared with the maximum possible. Hence, with the sample sizes used in most experiments, it is difficult to detect interactions between the covariate and a categorical predictor. Tests of the important assumption of homogeneity of regression in the analysis of covariance are based on tests of interactions between covariates and categorical predictor variables; our results suggest that such tests are likely to have relatively low statistical power. However, in the same study there might be enormous statistical power for detecting an interaction between two dichotomous categorical predictors.

Conclusion

Our analysis of the relative superiority of experimental designs for detecting interactions implies that unless researchers can select, oversample, or control the levels of the predictor

[18] Jerry Busemeyer drew our attention to this subtle but important issue concerning multiplicative models.

variables, detection of statistically reliable interactions or quadratic effects explaining an appreciable proportion of the variation of the dependent variable will be difficult. This does not mean that researchers should not seek interactions in such conditions; however, they should be aware that the odds are against them.

References

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Newbury Park, CA: Sage.

Allison, P. D. (1977). Testing for interaction in multiple regression. *American Journal of Sociology, 83,* 144–153.

Althauser, R. P. (1971). Multicollinearity and non-additive regression models. In H. M. Blalock, Jr. (Ed.), *Causal models in the social sciences* (pp. 453–472). Chicago: Aldine Atherton.

Anderson, N. H., & Shanteau, J. (1977). Weak inference with linear models. *Psychological Bulletin, 84,* 1155–1170.

Arnold, H. J., & Evans, M. G. (1979). Testing multiplicative models does not require ratio scales. *Organizational Behavior and Human Performance, 24,* 41–59.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Birnbaum, M. H. (1973). The devil rides again: Correlation as an index of fit. *Psychological Bulletin, 79,* 239–242.

Bohrnstedt, G. W., & Goldberger, A. S. (1969). On the exact covariance of products of random variables. *Journal of the American Statistical Association, 64,* 1439–1442.

Busemeyer, J. R., & Jones, L. (1983). Analysis of multiplicative combination rules when the causal variables are measured with error. *Psychological Bulletin, 93,* 549–562.

Champoux, J. E., & Peters, W. S. (1987). Form, effect size, and power in moderated regression analysis. *Journal of Occupational Psychology, 60,* 243–255.

Chaplin, W. F. (1991). The next generation of moderator research in personality psychology. *Journal of Personality, 59,* 143–178.

Cohen, J. (1978). Partialed products *are* interactions; partialed powers *are* curve components. *Psychological Bulletin, 85,* 858–866.

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

Cronbach, L. J. (1987). Statistical tests for moderator variables: Flaws in analysis recently proposed. *Psychological Bulletin, 102,* 414–417.

Dunlap, W. P., & Kemery, E. (1987). Failure to detect moderating effects: Is multicollinearity the problem? *Psychological Bulletin, 102,* 418–420.

Evans, M. G. (1985). A Monte Carlo study of the effects of correlated method variance in moderated multiple regression analysis. *Organizational Behavior and Human Decision Processes, 36,* 305–323.

Evans, M. G. (1991). Problems of analyzing multiplicative composites. *American Psychologist, 46,* 6–15.

Finney, J. W., Mitchell, R. E., Cronkite, R. C., & Moos, R. H. (1984). Methodological issues in estimating main and interactive effects: Examples from coping/social support and stress field. *Journal of Health and Social Behavior, 25,* 85–98.

Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science, 26,* 797–833.

Jaccard, J., Helbig, D. W., Wan, C. K., Gutman, M. A., & Kritz-Silverstein, D. C. (1990). Individual differences in attitude-behavior consistency: The prediction of contraceptive behavior. *Journal of Applied Social Psychology, 20,* 575–617.

Jaccard, J., Turrisi, R., & Wan, C. K. (1990). *Interaction effects in multiple regression.* Newbury Park, CA: Sage.

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

Kenny, D. A., & Judd, C. M. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin, 90,* 201–210.

Lubinski, D., & Humphreys, L. G. (1990). Assessing spurious "moderator effects": Illustrated substantively with the hypothesized ("synergistic") relation between spatial and mathematical ability. *Psychological Bulletin, 107,* 385–393.

Mardia, K. V. (1970). Measures of skewness and kurtosis with applications. *Biometrika, 57,* 519–530.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113,* 181–190.

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

Morris, J. H., Sherman, J., & Mansfield, E. R. (1986). Failures to detect moderating effects with ordinary least squares moderated-regression: Some reasons and a remedy. *Psychological Bulletin, 99,* 282–288.

Pearlin, L. I., Menaghan, E. G., Lieberman, M. A., & Mullan, J. T. (1981). The stress process. *Journal of Health and Social Behavior, 22,* 337–356.

Rohatgi, V. K., & Szekely, G. J. (1989). Sharp inequalities between skewness and kurtosis. *Statistics and Probability Letters, 8,* 297–299.

Saunders, D. R. (1955). The "moderator variable" as a useful tool in prediction. *Proceedings of the 1954 Invitational Conference on Testing Problems* (pp. 54–58). Princeton, NJ: Educational Testing Service.

Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement, 16,* 209–222.

Smith, K. W., & Sasaki, M. S. (1979). Decreasing multicollinearity: A method for models with multiplicative functions. *Sociological Methods and Research, 8,* 35–36.

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

Wise, S. L., Peters, L. H., & O'Connor, E. J. (1984). Identifying moderator variables using multiple regression: A reply to Darrow and Kahl. *Journal of Management, 10,* 227–233.

Zedeck, S. (1971). Problems with the use of "moderator" variables. *Psychological Bulletin, 76,* 295–310.

(Appendix follows on next page)

Appendix

In this appendix we derive an explicit formula for $V(XZ.X, Z)$, the residual variance of the product, expressed entirely in terms of the joint central moments of the bivariate distribution of X and Z. Let $R_{XZ.X,Z}{}^2$ represent the proportion of the variation in XZ that can be predicted by a linear combination of X and Z. Then, the residual variation in XZ not related to either X or Z is given by

$$V(XZ.X, Z) = V(XZ)(1 - R_{XZ.X,Z}{}^2). \quad (10)$$

It is easy to show (e.g., see Cohen & Cohen, 1983) that

$$R_{XZ.X,Z}{}^2 = \frac{\beta_{X.Z}C(X, XZ)}{V(XZ)} + \frac{\beta_{Z.X}C(Z, XZ)}{V(XZ)}. \quad (11)$$

Substituting Equation 11 into Equation 10 yields the following:

$$V(XZ.X, Z) = V(XZ) - \beta_{X.Z}C(X, XZ) - \beta_{Z.X}C(Z, XZ). \quad (12)$$

Note that if neither X nor Z is related to XZ (i.e., the two covariances equal zero), then there is no adjustment; in that case, the residual variance of the product equals the simple variance of the product.

We now use expressions[19] for the variance and covariance of product terms from Aiken and West (1991). In particular, their Equation A.7 (p. 179) was as follows:

$$V(XZ) = V(Z)E^2(X) + V(X)E^2(Z) + E(x^2z^2) + 2E(X)E(xz^2) + 2E(Z)E(x^2z) + 2C(X, Z)E(X)E(Z) - C^2(X, Z). \quad (13)$$

Their Equation A.14 (p. 180) was as follows:

$$C(X, XZ) = E(x^2z) + V(X)E(Z) + C(X, Z)E(X), \quad (14)$$

where $x = X - E(X)$ and $z = Z - E(Z)$ are centered values. Changing the roles of X and Z, we obtain the corresponding equation:

$$C(Z, XZ) = E(xz^2) + V(Z)E(X) + C(X, Z)E(Z). \quad (15)$$

Note that Equations 13–15 simplify considerably if X and Z have been centered so that $E(X) = E(Z) = 0$. However, we continue the proof with no assumptions about the expected values of X and Z. Substituting these expressions for the variances and covariances in Equation 12 yields the following:

$$\begin{aligned}V(XZ.X, Z) &= 2C(X, Z)E(X)E(Z) - C^2(X, Z) \\ &+ 2E(Z)E(x^2z) + 2E(X)E(xz^2) + E(x^2z^2) \\ &+ E^2(Z)V(X) + E^2(X)V(Z) - \{[C(X, Z)E(X) \\ &+ E(x^2z) + E(Z)V(X)][C(X, Z)E(X)V(Z) \\ &+ E(x^2z)V(Z) + E(Z)V(X)V(Z) - C^2(X, Z)E(Z) \\ &- C(X, Z)E(xz^2) - C(X, Z)E(X)V(Z)]\}/[V(X)V(Z) \\ &- C^2(X, Z)] - \{[C(X, Z)E(Z) + E(xz^2) \\ &+ E(X)V(Z)][-C^2(X, Z)E(X) - C(X, Z)E(x^2z) \\ &- C(X, Z)E(Z)V(X)] + C(X, Z)E(Z)V(X) \\ &+ E(xz^2)V(X) + E(X)V(X)V(Z)]\}/[V(X)V(Z) \\ &- C^2(X, Z)]. \quad (16)\end{aligned}$$

This reduces to the following:

$$\begin{aligned}V(XZ.X, Z) &= [C^4(X, Z) + 2C(X, Z)E(x^2z)E(xz^2) \\ &- C^2(X, Z)E(x^2z^2) - E^2(xz^2)V(X) - E(x^2z^2)V(Z) \\ &- C^2(X, Z)V(X)V(Z) + E(x^2z^2)V(X)V(Z)]/ \\ &[V(X)V(Z) - C^2(X, Z)]. \quad (17)\end{aligned}$$

Note that although $E(X)$ and $E(Z)$ appear in the full expression for the residual product in Equation 16, they cancel and do not appear in Equation 17; this proves that $V(XZ.X, Z)$, and hence the statistical test of the moderator effect, is independent of changes in the origin of either the X or the Z scales. Noting that

$$\begin{aligned}&[E(x^2z^2) - C^2(X, Z)][V(X)V(Z) - C^2(X, Z)] \\ &= C^4(X, Z) - C^2(X, Z)E(x^2z^2) - C^2(X, Z)V(X)V(Z) \\ &+ E(x^2z^2)V(X)V(Z), \quad (18)\end{aligned}$$

the expression for the residual variance further simplifies to

$$V(XZ.X, Z) = E(x^2z^2) - C^2(X, Z) - \frac{E^2(xz^2)V(X) + E^2(x^2z)V(Z) - 2C(X, Z)E(x^2z)E(xz^2)}{V(X)V(Z) - C^2(X, Z)}. \quad (19)$$

To express Equation 19 in terms of variances and covariances of centered predictors, we use the following equivalences, which assume that $E(X) = E(Z) = 0$:

$$C(X^2, Z^2) = E(x^2z^2) - V(X)V(Z) \quad (20)$$

$$C(X^2, Z) = E(x^2z) \quad (21)$$

$$C(X, Z^2) = E(xz^2). \quad (22)$$

These equivalences are easily deduced from the general formula for product covariances given by Bohrnstedt and Goldberger (1969) or from the methods used by Aiken and West (1991) in their Appendix. Making appropriate substitutions for the joint central moments in Equation 19 yields the following:

$$V(XZ.X, Z) = V(X)V(Z) + C(X^2, Z^2) - C^2(X, Z) - \frac{C^2(X^2, Z)V(Z) + C^2(X, Z^2)V(X) - 2C(X, Z)C(X^2, Z)C(X, Z^2)}{V(X)V(Z) - C^2(X, Z)}, \quad (23)$$

which is the expression given in the text (Equation 2).

[19] Alternatively, these expressions can be derived easily from the general formula for covariances of arbitrary products provided by Bohrnstedt and Goldberger (1969).

(*Appendix continues on next page*)

The derivation of $V(X^2.X)$ for the special case of the quadratic effect follows directly. Replacing Z with X and z with x in Equation 19 yields

$$V(X^2.X) = E(x^4) - C^2(X, X) - \frac{2E^2(x^3)V(X) - 2C(X, X)E^2(x^3)}{V^2(X) - C^2(X, X)}. \quad (24)$$

Factoring the numerator and denominator of the last term of Equation 24 and canceling yields

$$V(X^2.X) = E(x^4) - C^2(X, X) - \frac{2E^2(x^3)}{V(X) + C(X, X)}. \quad (25)$$

Remembering that $C(X, X) = V(X)$, this becomes

$$V(X^2.X) = E(x^4) - V^2(X) - \frac{E^2(x^3)}{V(X)}. \quad (26)$$

Making the substitutions to common names of $E(x^4) = V^2(X)[\text{kurtosis}(X) + 3]$ and $E(x^3) = V^{3/2}(X)\text{skew}(X)$ yields the final expression:[20]

$$V(X^2.X) = V^2(X)[\text{kurtosis}(X) + 2 - \text{skew}^2(X)], \quad (27)$$

which is the expression given in the text (Equation 9).

[20] Rohatgi and Szekely (1989) showed that $\text{skew}^2(X) \leq \text{kurtosis}(X) + 2$ for all distributions, so values of $V(X^2.X)$ are guaranteed to be nonnegative. Our derivation could be construed as an alternative proof of the Rohatgi and Szekely inequality. *Kurtosis* is defined here to be the excess relative to the normal distribution.

Received April 28, 1992
Revision received March 10, 1993
Accepted March 25, 1993 ■

Call for Nominations

The Publications and Communications Board has opened nominations for the editorships of *Behavioral Neuroscience*, the *Journal of Experimental Psychology: General*, and the *Journal of Experimental Psychology: Learning, Memory, and Cognition* for the years 1996–2001. Larry R. Squire, PhD, Earl Hunt, PhD, and Keith Rayner, PhD, respectively, are the incumbent editors. Candidates must be members of APA and should be available to start receiving manuscripts in early 1995 to prepare for issues published in 1996. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. To nominate candidates, prepare a statement of one page or less in support of each candidate.

- For *Behavioral Neuroscience*, submit nominations to J. Bruce Overmier, PhD, Elliott Hall—Psychology, University of Minnesota, 75 East River Road, Minneapolis, MN 55455 or to psyjbo@vx.cis.umn.edu. Other members of the search committee are Norman Adler, PhD, Evelyn Satinoff, PhD, and Richard F. Thompson, PhD.

- For the *Journal of Experimental Psychology: General*, submit nominations to Howard E. Egeth, PhD, Chair, *JEP: General* Search, Department of Psychology, Johns Hopkins University, Charles & 34th Streets, Baltimore, MD 21218, to egeth@jhuvm.bitnet, or to fax number 410-516-4478. Other members of the search committee are Donald S. Blough, PhD, Martha Farah, PhD, and Edward E. Smith, PhD.

- For the *Journal of Experimental Psychology: Learning, Memory, and Cognition*, submit nominations to Donna M. Gelfand, PhD, Dean, Social and Behavioral Science, 205 Osh, University of Utah, Salt Lake City, UT 84112-1102 or to fax number 801-585-5081. Other members of the search committee are Marcia Johnson, PhD, Michael Posner, PhD, Henry L. Roediger III, PhD, and Richard M. Shiffrin, PhD.

First review of nominations will begin December 15, 1993.