Psychological Bulletin
1980, Vol. 87, No. 1, 51-57

Nonorthogonal Analysis of Variance—Once Again

Elliot M. Cramer and Mark I. Appelbaum
L. L. Thurstone Psychometric Laboratory
University of North Carolina at Chapel Hill

In a previous article, Appelbaum and Cramer discussed the relationship between orthogonal and nonorthogonal analysis of variance (ANOVA), noting that the problems traditionally associated with the nonorthogonal ANOVA are easily resolved by viewing ANOVA as a series of model comparisons. Overall, Spiegel, and Cohen described another method that they argued estimates the same parameters and tests the same hypotheses in both the orthogonal and nonorthogonal cases. In the current article it is noted that since this is true of all the methods under consideration, in the context of comparing linear models, a preference must be based on other grounds. In particular, it is shown that our method is more powerful when there is no interaction or when there are small interaction effects. For the interactive case it is noted that their test is a test of equally weighted marginal means and that although marginal means may be of interest in some circumstances, the weights may be freely chosen, depending on the substantive problem. Any tests of significance in the interactive case are dependent on the weights used.

Considering the bulk of the literature on the subject, nonorthogonal analysis of variance (ANOVA) must seem a complex and paradoxical area of statistics. This is not so, provided the problem is seen as one of comparing competing models of nature and estimating the parameters of nature within the context of a model.

We first presented (Appelbaum & Cramer, 1974) a general critique of the substantial literature that appeared prior to 1972. Our stern words were directed to the literature as a whole and not to any one article. We argued that there is a single method of least squares but that there are many different tests of significance, each of which corresponds to a comparison of two different linear models. Overall and Spiegel (1969) discussed three "methods," each involving comparisons of specific linear models. They seemed to favor their Method 2, which involved what can be called the tests of A eliminating B, of B eliminating A, and of AB eliminating A and B. These tests imply the comparison of the following:

A eliminating B:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon,$$

and

$$Y_{ij} = \mu + \beta_j + \epsilon;$$

B eliminating A:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon,$$

and

$$Y_{ij} = \mu + \alpha_i + \epsilon;$$

AB eliminating A and B:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon,$$

and

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon.$$

In our article we emphasized that in a non-

This work was supported in part by Grant NE-G-00-3-0111 from the National Institute of Education, U.S. Department of Health, Education and Welfare. The opinions expressed do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred. The work was also supported in part by Grant MH-10007 from the National Institute of Mental Health, U.S. Public Health Service.

Requests for reprints should be sent to Elliot M. Cramer or Mark I. Appelbaum, L. L. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill, North Carolina 27514.

orthogonal design one must also consider the tests of A ignoring B and B ignoring A, that is, the comparison of the models,

A ignoring B:

$$Y_{ij} = \mu + \alpha_i + \epsilon,$$

and

$$Y_{ij} = \mu + \epsilon;$$

B ignoring A:

$$Y_{ij} = \mu + \beta_j + \epsilon,$$

and

$$Y_{ij} = \mu + \epsilon.$$

We specifically showed how the failure to consider these tests could lead to incorrect conclusions. None of the Overall and Spiegel (1969) "methods" consider examining all of these tests (i.e., making all of the model comparisons), and hence, in our view, none can be considered as complete.

Analysis of the Recommendation Presented by Overall, Spiegel, and Cohen

More recently, Overall, Spiegel, and Cohen (1975) have advocated Method 1 over Method 2. Method 1 can be described as testing main effects eliminating interactions or as comparing a full interactive model with one in which only a single main effect term is dropped, but the interactive term is retained. Describing how they have arrived at their conclusions, Overall et al. note that

> the strategy that appeared most often to be recommended in applied statistics texts involves basically a "main effects" model with tests for interaction effects included as a safeguard against departures from additivity (Rao, 1965; Snedecor & Cochran, 1967; Winer, 1971). The analysis proposed by Appelbaum and Cramer (1974) follows this logic. (p. 184)

The argument against this approach, as developed by Overall et al., rests on a single principle that we believe is correct and proper and on a single procedure that we believe is incorrect.

The principle is

> that the method for the analysis of variance of data from nonorthogonal designs should estimate the same parameters and test the same hypotheses as can otherwise be estimated and tested in a balanced analysis of variance experimental design involving the same factors. (Overall et al., 1975, p. 184)

This principle is consistent with our views, since in our (Appelbaum & Cramer, 1974) article we said, "Having decided to employ the method of least squares . . . one is left only with the selection of possible models and model comparisons. The models selected are logically independent of the observed numbers of observations per cell" (p. 336). Since the models do not involve the numbers of observations per cell, it follows that our methods test the same hypotheses and estimate the same parameters whether there are equal numbers of observations or not.

By the same logic it follows that all of Overall and Spiegel's (1969) methods test the same hypotheses and estimate the same parameters in both the orthogonal and nonorthogonal cases, once one has specified the competing models. The choice among different methods must be based on other criteria. The procedure that Overall et al. used to demonstrate the superiority of Method 1 is faulty, since it involves duplication of observations to make the design orthogonal. This, of course, violates the assumption of independence in ANOVA.

We now examine our reasons for rejecting the routine use of Overall and Spiegel's (1969) Method 1, as it has been advocated. First, we consider some problems of estimation that were not treated in our earlier article.

Estimation and the Overall et al. Criterion

It is not true, as suggested, that if two methods estimate the same parameters, they must yield the same estimates. To estimate a population mean, we could use the sample mean or simply use the first observation, discarding the others. Both the sample mean and the first observation are unbiased estimators of the population mean, but they will, in general, yield different estimates. The sample mean is better, since it will be closer to the population mean on the average. This precision of estimates is the crucial distinction between the methods that Overall et al. (1975) advocate and the methods that we advocate.

We did not deal with estimation in the nonorthogonal ANOVA in our earlier article (Appelbaum & Cramer, 1974), since we did not believe that there was any disagreement as to what was appropriate. We now feel that this topic does require some attention.

The estimation problem is easily and completely solved, once one decides on the model that one believes applies to the real world. The usual role of significance testing is to determine, based on the data of the real world, which model is the most reasonable one from among a set of competing models. Having made a decision as to which model obtains, one may then estimate the parameters of the model, but estimation may occur only in the context of a particular model.

One possible model—the two-factor interactive model—is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon. \qquad (1)$$

It is a trivial matter to obtain a set of least squares estimates of the parameters of this model. We say "a set" because there are infinitely many sets that are equivalent, in the sense that they will yield identical predicted values of $\hat{Y}_{ij}$. It is, however, a standard practice to impose additional constraints on the model, to obtain a unique set of estimates. The purpose of the constraining system, however, is solely computational convenience. It is obvious that the very best we can do in this model is to predict the cell means exactly, since there are no parameters that are unique to any single observation. Any two models that predict the cell means exactly must be equivalent. It is also a consequence of the mathematics of the system that any model that has as many independent parameters as cells must predict the cell means exactly.

There exist infinitely many constraining systems that may be applied to the full interactive model, to produce the computational determinacy desired. The simplest of these is

$$\mu = \alpha_i = \beta_j = 0,$$

for all i and j, leaving the model,

$$Y_{ij} = \gamma_{ij} + \epsilon. \qquad (2)$$

In this case the least squares estimates of γ_{ij} are simply the observed cell mean, $\bar{Y}_{ij}$.

The more conventional constraining system, however, is

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0. \quad (3)$$

If the design has a levels of Factor A and b levels of Factor B, there are then (after the imposition of the constraints) $1 + (a - 1) + (b - 1) + (a - 1)(b - 1) = ab$ independent parameters, which is also the number of independent parameters in the model specified by Equation 2. This equals the number of cells in the design, and it then follows that the model constrained in this way must be equivalent to that in the model specified by Equation 2. For those familiar with the matrix approach to the ANOVA, this result is readily apparent, since the model matrix for this constrained design must have $a \times b$ columns.

It is also a trivial matter to directly write the least squares estimates of the parameters of the interactive model constrained by Equation 3. They are

$$\hat{\mu} = \bar{Y},$$

$$\hat{\alpha}_i = \bar{Y}_{i.} - \bar{Y},$$

$$\hat{\beta}_j = \bar{Y}_{.j} - \bar{Y}_{..},$$

and

$$\hat{\gamma}_{ij} = \bar{Y}_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}, \qquad (4)$$

where $\bar{Y}_{..}$ is the unweighted average of the cell means, and $\bar{Y}_{i.}$ and $\bar{Y}_{.j}$ are the unweighted averages of the cell means for row i and column j, respectively. Substituting these estimates into the model specified by Equation 1 gives

$$\hat{Y}_{ij} = \mu + \hat{\alpha}_i + \hat{\beta}_j + \hat{\gamma}_{ij} = \bar{Y}_{ij},$$

again showing the equivalence of the models specified by Equations 1 and 2.

We thus see that estimation in the interactive model is an easy matter, with the estimates of the parameters being simple linear functions of the observed cell means and being free of the n_{ij}.

In fact there is no gain in talking of estimating parameters in the model specified by Equation 1, since it is equivalent to the model specified by Equation 2, which is a cell means model requiring no constraints. We have $a \times b$ populations (one per cell), and the only parameters of interest are their means and their common variance.

The situation, in general, is not nearly so simple when there is no interaction, that is, when estimation proceeds within the main effects model,

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon. \qquad (5)$$

In general, we would have to solve a set of simultaneous least squares equations to obtain estimates of parameters in the model specified by Equation 5. An exception, however, is in the orthogonal case, in which the estimates of μ, α_i, and β_j have the same form as in the model specified by Equation 4. In the more general nonorthogonal case, there will again be infinitely many solutions to the unconstrained least squares equations, although estimates of $\mu + \alpha_i$ and $\mu + \beta_j$ will be unique.

An interesting consequence of least squares estimation is that if the main effects model is the correct model, but if one uses the estimates of μ, σ_i, and β_j obtained from the interaction model, one still obtains unbiased estimates of Y_{ij}. These estimates of Y_{ij} are, however, less precise; that is, they have larger variances than do the estimates obtained from the correct main effects model. For this reason is it desirable to estimate the parameters of the main effects model when we have accepted that model as the true one rather than to use the estimates from the interaction model.

Tests of Significance Following a Nonsignificant Test of Interaction

We can, of course, never know except in the probabilistic sense of a statistical test whether there is truly an interaction present in nature. We must in the final analysis rely on the results of statistical tests to direct us to reasonable models on which to base our estimation procedure. We must then consider which behavior is appropriate when the data dictate an interaction-free model and what the consequences of such behavior are. There are, in this respect, only three cases that need concern us; in all three we assume that the statistical test of interaction is nonsignificant, whereas the state of nature is as indicated.

Case 1: No interaction in the population and nonsignificant test. The first case we consider is the case when there is indeed no interaction present in nature. No empirical demonstration is needed to verify that if one has the form of the true linear model, the least squares estimates of the parameters in that model will be the best unbiased linear estimates. Furthermore, it is obvious that if one fits an interactive model, when there is in fact no interaction, one will obtain unbiased estimates that will not be minimum variance. For this reason, it is a mistake to include worthless effects in an ANOVA model, just as it would be to include worthless variables (variables that do not improve prediction) in a regression problem. The additional sampling error causes the main effect parameters and the estimated cell means to have larger standard errors than would the estimates from a main effects model. This point has been noted in a regression context by Walls and Weeks (1969) and is exactly what would occur if Overall and Spiegel's (1969) Method 1 were applied in this case. The increase in sampling error may be substantial and will result in less powerful tests of main effects.

Case 2: Small interaction in the population and nonsignificant test. The second case is the situation in which there is a true interaction, but its magnitude is too small to be detected by a conventional test of interaction. We have previously argued that the main effect parameters are not meaningful for the interactive model but that the predicted cell means are. The predicted cell means will have a smaller variance when estimated in the main effects model than when estimated in the interactive model, since the variance depends only on the variance of the dependent variable and on the design matrix. (**X** in the usual matrix approach to ANOVA.) The predicted cell means will, however, be biased in this case. Since minimum variance unbiased estimators do not exist, the mean square error becomes relevant for comparison. We must add the mean squared bias (which will be a function of the magnitude of the small but nonzero interaction terms) to the variance to obtain the mean square error. This term will be small, if the interactive effects are small, as would be the situation under Case 2. Operating under Case 2, we still estimate the same parameters and test the same effects in both the orthogonal and nonorthogonal cases, but we simply estimate and test with a small amount of bias. We gain substantially, in that the estimates will be more precise and the tests will be more powerful than if we followed Overall and Spiegel's (1969) Method 1, which is based on the interactive model.

To see the difference, let us compare the variances of the estimated parameters and estimated cell means for the data used by Overall et al. (1975). In Table 1 we have computed the variances of the estimated main effect parameters and the predicted cell means for both the main effects model (our procedure) and the interactive model (Overall & Spiegel's, 1969, Method 1). If $\mathbf{X}$ is the matrix of independent variables, the variance–covariance matrix of the estimated parameters is $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$, whereas the variance–covariance matrix of the predicted cell means is $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2$. The variances are on the diagonals of these matrices and do not depend on which model is correct in nature. Since σ^2 serves only as a scale factor, we have assumed in Table 1 that it is equal to one. We see that the estimated parameters of the main effects model have slightly smaller variances than those of the interactive model, whereas the corresponding predicted cell means have substantially smaller variances, when estimated from the main effects model. The effect on the predicted cell means is particularly marked for the cells with a small number of observations, since the variance of a sample mean (the predicted value for an interactive model) is simply σ^2/n.

Case 3: Large interaction in the population and nonsignificant test. The third case covers the situation in which a large interaction is somehow not detected by the interaction test. In this situation the reverse of Case 2 will occur, and the mean square errors of the estimated parameters and cell means will be smaller for the interactive model. The probability of this third case occurring is remote, however, for if the magnitude of the interaction effects is large and if the sample size is reasonable, the power of the interaction test is large.

In the case of a nonsignificant interaction with a high p value (e.g., .2), many investigators would pool the interaction with error. The gain from this will be slight unless the error degrees of freedom are small, and only in this case would we choose to pool.

We conclude, then, that if the true interaction is null or small, one should not use the Method 1 tests, since they will be less powerful than the ones we have recommended. The examples we gave in our earlier article also apply to Method 1 tests. We showed there how one would be led to falsely conclude that there was no main effect, when the proper conclusion was that there was one main effect of indeterminate origin. In the case of a large interaction, we expect a test of reasonable power to detect the interaction, and we consider that situation below.

Table 1

Variances of Parameter Estimates and Predicted Cell Means for 2 × 5 Factorial Design With Unequal Numbers of Observations, Assuming $\sigma^2 = 1$

	Type of model	
Parameters	Main effects	Interactive
	Estimates of parameters	
μ	.161	.169
α_1	.241	.244
α_2	.235	.244
β_1	.241	.244
β_2	.224	.231
	Estimated cell means	
$\hat{Y}_{11}$	.116	.167
$\hat{Y}_{12}$	.151	.333
$\hat{Y}_{13}$	.158	.333
$\hat{Y}_{21}$	.158	.333
$\hat{Y}_{22}$	.112	.167
$\hat{Y}_{23}$	.154	.333
$\hat{Y}_{31}$	.151	.333
$\hat{Y}_{32}$	.109	.167
$\hat{Y}_{33}$	.112	.167

On Main Effects, Marginal Effects, and Interaction

Overall and Spiegel (1969) have suggested specific tests of main effects regardless of whether there is a significant interaction. We believe that this is based on the mistaken notion that orthogonality rather than nonorthogonality is the norm. We believe that a good deal of the confusion in nonorthogonal ANOVA is due to the fact that in the orthogonal case, tests of different model comparisons yield numerically identical results. We argue that even in the orthogonal case, the standard tests of main effects should not routinely be used in the presence of an interaction. Indeed, we have made in our earlier article specific recommendations as follows:

1. Begin with the full model including main effects and interaction effects.
2. Test for a significant interaction; if this

Table 2
Illustration of Marginal Means

	Factor		Weights		
Factor	B_1	B_2	$w = .5$	$w = 1$	$w = 0$
	Interactive model				
A_1	10	20	15	10	20
A_2	20	10	15	20	10
	Noninteractive model				
A_1	10	20	15	10	20
A_2	30	40	35	30	40

test is significant, no tests of main effects are appropriate; however, one may wish to test certain contrasts in the cell means to aid in interpretation of the results. In making this second recommendation, we had in mind a distinction between main effects and marginal effects that we discuss later. We have never disputed the claim that the so-called tests of main effects in the presence of interaction are valid tests of something. We question, however, whether they test hypotheses that are typically of interest, when there is an interaction.

Additional insight into the nature of this problem can be gained through a more careful consideration of the problems of testing and estimating "main effects" in the presence of interaction. At this point it is necessary to introduce a basic logical distinction between two concepts that have, unfortunately, come to be held as virtually synonymous—a main effect and a marginal effect. By a main effect, we mean the effect of a particular experimental treatment or state of nature that is the common and consistent effect of that treatment or state of nature, irrespective of which other treatments or states of nature it is combined with. By a marginal effect, we mean simply the effect of the experimental treatment (state of nature), averaged, in some sense, over all occurrences of that treatment. These two concepts are equivalent only in the noninteractive model. In the case of a model in which there is an interaction, the two concepts are distinct; in fact, under the interactive model, the concept of a main effect does not apply, for an interaction implies that there is no consistent effect of the treatment but rather that one must consider that treatment in combination with some other treatment(s) to assess its effect. This distinction can also be demonstrated by the concept of a simple row (or column) effect that is commonly defined as the difference between a cell mean and its corresponding row (or column) mean. If, for a given factor, the simple effect of the several treatments should be identical for all levels of other factors, this constant simple effect is the main effect.

Given then that one is operating with a model that contains an interaction term, it is, at best, misleading to speak of main effects, for one is considering marginal effects. These marginal effects will be averages of cell means across rows or columns of the design. There is no particular reason for using a simple average rather than a weighted average. If the model is truly interactive, the weights used will have a substantial effect on the marginal effects. Suppose, for example, that the cell means are like those shown in Table 2 for a 2×2 ANOVA. If we define a marginal A mean as

$$\hat{Y}_{i.} = w\hat{Y}_{i1} + (1 - w)\hat{Y}_{i2},$$

we find that the difference in marginal means for A will be 0, -10, or 10, depending on whether w is .5, 1, or 0. For the data from a noninteractive model shown in Table 2, we find that the difference in marginal means is 20, regardless of what the weights are.

The tests of main effects proposed by Overall et al. (1975) in Method 1 are in fact tests of equally weighted marginal means for an interactive model. It can also be shown that these tests are equivalent to the method of unweighted squares of means proposed by Yates (1934) and discussed by Bancroft (1968). These are tests of the equivalence of row or column marginal means

$$\mu_{i.} = \sum_{j} \mu_{ij}/b,$$

and

$$\mu_{.j} = \sum_{i} \mu_{ij}/a.$$

These particular marginal means are but one of many possible sets of marginal means that can be constructed, and it is by no means clear that this is the most desirable set to test in any particular situation (see Appelbaum & Cramer, 1976).

We agree that the procedures advocated by Overall et al. (1975) test the same hypotheses in both the orthogonal and nonorthogonal case; further, we agree that they are valid tests of certain hypotheses, but we doubt that they are hypotheses of particular interest in either the orthogonal or nonorthogonal case. We believe that informed statistical analysts would not routinely perform a test of main effects in the presence of a significant interaction in the orthogonal case; why then in the nonorthogonal case? They might consider tests of weighted marginal means, as we have discussed here and elsewhere (Appelbaum & Cramer, 1976), in which the weights would depend on the substantive problem. More typically we expect that they would be concerned with explaining the interaction through tests of simple effects within levels of the interacting effect or through subdividing the design into interactive and noninteractive pieces.

Conclusion

In this article we have shown that the significance testing procedures that we have previously recommended for the nonorthogonal ANOVA are consistent with the basic principle advocated by Overall et al. (1975), namely, that in the nonorthogonal case one should estimate the same parameters and test the same hypotheses that one would estimate and test if there were equal numbers of observations in the cells. Indeed, that principle is implicit in our original article (Appelbaum & Cramer, 1974). We have further shown that their method for achieving the stated goal, if routinely applied, will not lead to optimal tests or estimates. In our discussion of the relationship between estimates and hypothesis testing, we have made clear the reasons for preferring our procedure, since it leads to more powerful tests and more precise estimates.

It must be recalled that the issues of how to estimate effects and how to test hypotheses are distinct. The methods discussed by Overall et al. and by us are methods for testing hypotheses and not for the estimation of effects. Once one has decided on a model, one uses least squares to obtain estimates of the parameters in that model. Given a tentative model, one tests hypotheses to determine if a simpler model is plausible. Of course, given a simpler model, one may wish to estimate the parameters of that model in preference to the more complex one. This distinction is not one that we uniquely make. Bock (1975), for instance, regards these as distinct processes. He begins with some initial model, performs tests of significance to determine if a simpler model is appropriate, and then estimates the parameters in the simplest reasonable model. We have seen no evidence that suggests that the methods advocated by Overall et al. are preferable. We continue to maintain, along with Rao (1965), Snedecor and Cochran (1967), and Winer (1971), that one should test main effects, assuming that no interaction is present, when this is what is suggested by the data at hand.

References

Appelbaum, M. I., & Cramer, E. M. Some problems in the nonorthogonal analysis of variance. *Psychological Bulletin*, 1974, *81*, 335–343.

Appelbaum, M. I., & Cramer, E. M. Balancing—Analysis of variance by another name. *Journal of Educational Statistics*, 1976, *1*, 233–252.

Bancroft, T. A. *Topics in intermediate statistical methods* (Vol. 1). Ames: Iowa State University Press, 1968.

Bock, R. D. *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill, 1975.

Overall, J. E., & Spiegel, D. K. Concerning least squares analysis of experimental data. *Psychological Bulletin*, 1969, *72*, 311–322.

Overall, J. E., Spiegel, D. K., & Cohen, J. Equivalence of orthogonal and nonorthogonal analysis of variance. *Psychological Bulletin*, 1975, *82*, 182–186.

Rao, C. R. *Linear statistical inference and its applications*. New York: Wiley, 1965.

Snedecor, G. W., & Cochran, W. G. *Statistical methods*. Ames: University of Iowa Press, 1967.

Walls, R. C., & Weeks, D. L. A note on the variance of a predicted response in regression. *American Statistician*, 1969, *23*, 24–26.

Winer, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

Yates, F. The analysis of multiple classifications with unequal numbers in the different classes. *Journal of the American Statistical Association*, 1934, *29*, 51–66.

Received August 1, 1978 ■