AD-A076 775 NAVAL HEALTH RESEARCH CENTER SAN DIEGO CA F/G 12/1
SPECIFIC INTERACTION CONTRASTS: A STATISTICAL TOOL FOR REPEATED--ETC(U)
DEC 76 V K THARP, A LUBIN
UNCLASSIFIED NAVHLTHRSCHC-76-78 NL

1 OF 1
AD-
A076775

END
DATE
FILMED
12 79
DDC

SPECIFIC INTERACTION CONTRASTS: A STATISTICAL TOOL FOR REPEATED-MEASURES DESIGNS

V. K. THARP, JR.

ARDIE LUBIN

REPORT NO. 76-78

NAVAL HEALTH RESEARCH CENTER

P. O. BOX 85122

SAN DIEGO, CALIFORNIA 92138

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND

BETHESDA, MARYLAND

2

Specific Interaction Contrasts:

A Statistical Tool for Repeated-Measures Designs

Van K. Tharp, Jr. and Ardie Lubin

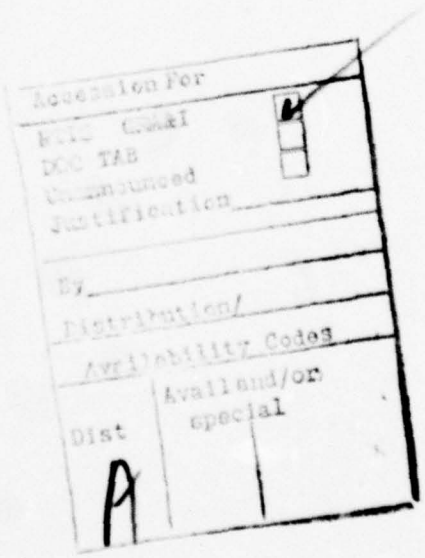
Naval Health Research Center

San Diego, California 92152

Running head: Specific Interaction Contrasts

Abstract

Experimental psychologists often use multifactor repeated-measure designs in which interactions are the most important effects to be assessed. An experimenter has at least five ways to evaluate such interactions: (1) a univariate repeated-measures analysis of variance (ANOVA), with (probably) inflated estimates of the degrees of freedom; (2) a univariate repeated-measures ANOVA with the Greenhouse-Geisser conservative estimate of the degrees of freedom; (3) the Greenhouse-Geisser stepwise analysis; (4) a multivariate ANOVA; and (5) specific interaction contrasts. We show that no matter which of the above paths is chosen, the careful experimenter must compute specific interaction contrasts (i.e., t-tests). A worked example is given.



Specific Interaction Contrasts:
A Statistical Tool for Repeated-Measures Designs

R. A. Fisher (Fisher & MacKenzie, 1923) used the concepts of interaction and additivity when he invented the Analysis of Variance (ANOVA) method for dealing with multifactor experiments. Until recently, these concepts did not play a primary role in any theoretical models commonly used in experimental psychology. However, in 1969, Sternberg incorporated both concepts directly into his additive-factor method for decomposing reaction time (RT) into processing stages.

Assuming that RT is composed of a number of additive stages in a known order, Sternberg proposed that each stage be studied by influencing its duration with various treatments. Two treatments which influence one or more stages in common should have an interaction effect on RT. But if the two treatments influence different stages, then each should have an additive effect on RT.

Taylor (1976) recently extended this methodology to conditions in which some dependence may occur between processing stages. His primary restrictive assumption is that stage dependence must be expressable as a linear function of the stage times involved. Under these conditions, Sternberg's hypothesis concerning additivity does not hold. That is, the absence of a significant interaction between the effects of two variables does not imply that these variables influence different processing stages. However, two treatments influencing one or more processing stages in common still should have an interaction effect on RT. Thus, the important question becomes: How do we test for interactions?

The traditional significance test of interactions is the F-ratio in the analysis of variance (ANOVA). We find two major difficulties in using this test with the form of experimentation advocated by Sternberg (1969) and Taylor (1976). First, multifactor repeated-measure designs, such as those required by Sternberg (1969) and Taylor (1976), do not meet a critical assumption of the ANOVA--independent scores. Second, the F-ratio is a vague test, telling the experimenter almost nothing.

R. A. Fisher's first use (1923) of the ANOVA was to study the effect of treatments on plots of ground. His most important assumption was that the criterion score (yield) for any particular plot was independent of the score for any other plot. The F-ratio, consequently, has some problems in its application to a repeated-measures experiment. For example, Sternberg (1969) and Taylor (1976) advocate exposing a subject to all possible conditions in a multifactor experiment. Even if we assume that the carryover effects of every treatment on all subsequent treatments are negligible, we must deal with the fact that any two scores measured on the same organism, can have (and usually do have) non-zero correlations. This makes the usual univariate repeated-measure F-ratio misleading or uninterpretable (e.g., see Lana & Lubin, 1963 and the justification section of the present paper).

An even greater difficulty is that the calculation of F-ratios in conjunction with a completely general (i.e., unspecified) ANOVA model tells the experimenter almost nothing about a hypothesized interaction effect. A significant F-ratio tells us nothing about the direction, amount, or location of the underlying interaction; while a nonsignificant F-test may simply lack the power to detect a single large inter-

action effect. J. A. Nelder (see Plackett, 1960, p. 213) has criticized such general ANOVA models on the grounds of vagueness. He suggested, instead, that the model be specialized to fit the particular application, thus gaining power and supplying more information to the experimenter.

Professor G. A. Barnard, also commenting on the paper by Plackett (1960), noted that the ANOVA essentially reduces to a set of independent contrasts[1] and that we are free to select groups of contrasts in any manner we choose (see Plackett, 1960, p. 215). Since efficient use of the ANOVA involves selecting a specific model to fit the chosen application, why not devise specific contrasts to test one's hypotheses? As Geisser has said: "When there are contrasts of scientific importance the omnibus F is irrelevant" (personal communication). Contrast weights specify the size, sign, and location of the putative effects. A contrast can be tested by a t-ratio. In this way we avoid the vagueness and lack of power of the F-ratio. In particular, contrasts do not require the many assumptions necessary for a repeated-measure ANOVA.

Sternberg (1969), in fact, recommended that interactions be evaluated with specific interaction contrasts. Unfortunately, he provided no detailed worked examples of his procedure. Furthermore, the interaction in his one example, in which computation is briefly discussed, has only one possible degree of freedom, so it does not reveal the true potential of the procedure. We have seen no other use of this technique, so we can only assume that Sternberg's description was insufficient for most of his readers. The purpose of this paper, therefore, is to detail the computational procedures involved and give some theoretical justification for the use of contrasts.

Specific Interaction Contrasts: Computation

A specific interaction contrast is an estimate of the variation due to interaction based on an explicit set of contrast weights. In a multi-factor, repeated-measures design, the contrast procedure requires that one first obtain a set of contrast weights. These may be obtained by prior reasoning or from prior empirical knowledge. Each of the p conditions in the interaction must have a contrast weight, and the sum of these weights will be zero, by definition. The next step involves calculating the residuals from additivity (i.e., the interaction effect) for each subject, obtained by subtracting the overall mean and the main effects from a subject's score for each of the p conditions. Again, by definition, the residual from additivity should be a set of numbers which sum to zero. The contrast weights are then applied to the interaction effect of each subject to obtain a contrast score. Given n subjects, the set of n contrast scores can be used in a routine t-test of the null hypothesis, H_o, that the interaction effect is zero. If H_o is accepted, then additivity of the main effects is implied only for this particular weighted combination of treatment levels. Generally, the maximum number of independent, specific interaction effects is limited by the degrees of freedom for the interaction term in the ANOVA model.

Let Y_k be the contrast score for the kth subject. Then

$$Y_k = \sum_i \sum_j \omega_{ij} q_{ijk} \qquad \begin{matrix} i = 1 \text{ to } r \\ j = 1 \text{ to } c \end{matrix} \qquad (1)$$

where ω_{ij} is the contrast weight for the ijth cell, and q_{ijk} is the residual from additivity for the ijth cell within the kth subject. First, we will give the standard method of computing q_{ij}, the residual

from additivity. Then we will discuss the more difficult problem of estimating the contrast weights, ω_{ij}.

Let us look at an example (Tharp, 1975) in which the time required to name pictures is measured under alcohol and baseline conditions. The designated correct names to the stimuli came from five different word-frequency (WF) categories. The deleterious effect of alcohol on verbal reaction time is expected to increase as word-frequency decreases--an interaction effect.

In this example[2] there are five levels of WF and two levels of drug, giving us (5-1) (2-1) = 4 degrees of freedom for the interaction. Thus, there are four possible independent specific interaction effects. Table 1 illustrates summary data for one subject. There are ten scores per subject.

- - - - - - - - - - - - - -

Table 1 about here

- - - - - - - - - - - - - -

Residuals from Additivity

Let $\bar{X}_{ij}$ be the average score in row i and column j. First, the effect of the ith row, R_i, is equal to:

$$R_i = \frac{\sum_j \bar{X}_{ij}}{2} - \bar{X}.. \qquad j = 1,2 \qquad (2)$$

where $\bar{X}..$ is the grand mean. Second, the effect of the jth column, C_j, is equal to:

$$C_j = \frac{\sum_i \bar{X}_{ij}}{5} - \bar{X}.. \qquad i = 1 \text{ to } 5 \qquad (3)$$

Finally, the residual from additivity q_{ij}, in row i and column j is equal to:

$$q_{ij} = \bar{X}_{ij} - R_i - C_j - \bar{X}.. \qquad (4)$$

Note that R_i and C_j are row and column effects, respectively, not means. For example, for the score in row 1 and column 1 of Table 1, the row effect (R_1) is equal to the row mean minus the grand mean (i.e., 859.5 - 1035.3 = -175.8). The column effect (C_1) is equal to the column mean minus the grand mean (i.e., 1124.8 - 1035.3 = 89.5). Finally, to obtain the residual from additivity (q_{11}), the row effect, the column effect, and the grand mean are all subtracted from the cell score (i.e., 901 + 175.8 - 89.5 - 1035.3 = 48.0). All the residuals from additivity, computed from Table 1, are shown in Table 2.

- - - - - - - - - - - - - - -

Table 2 about here

- - - - - - - - - - - - - - -

Each residual is equal to a random error term (e_{ijk}), plus a putative interaction effect, τ_{ij}.

Determination of Contrast Weights

Assume that the true interaction effect in the ijth cell is τ_{ij}. Then the best estimate of the contrast weight, ω_{ij}, is τ_{ij}. If the set of ten contrast weights is equal to (or proportional to) the set of ten interaction effects, then the contrast score of equation 1 is maximized.

Theoretical Weights. To the extent that an experiment is based on experience, sound judgment and prior scientific knowledge, one should have little difficulty predicting the amount and direction of any

hypothesized interactions. In our example, the effect of alcohol was expected to remain relatively constant from WF_1 to WF_3 and to increase dramatically at WF_4 and WF_5. These results were expected because (1) an exponential increase in *a priori* stimulus-response uncertainty occurs from WF_1 to WF_5 and (2) the effect of alcohol increases as a function of experimentally defined stimulus-response uncertainty (Tharp, Rundell, Lester, & Williams, 1974). Thus, based on prior knowledge, we might postulate the set of weights, ω_{ij}, shown in Table 3.

- - - - - - - - - - - - - - -

Table 3 about here

- - - - - - - - - - - - - - -

There are two restrictions on these interaction contrast weights. First,

$$\sum_j \omega_{ij} = 0 \text{ for every row} \qquad j = 1,2 \qquad (5)$$

and second

$$\sum_j \omega_{ij} = 0 \text{ for every column} \qquad i = 1 \text{ to } 5 \qquad (6)$$

One can adjust any set of weights to fit these rules by subtracting from each weight the appropriate row mean, finding column means of the row-adjusted scores, and then subtracting the appropriate column mean from each cell.[3]

Contrast weights are usually given as single digits lying between -9 and +9. Single digit weights might be obtained by smoothing and dividing all scores by their lowest common denominator. Two-digit accuracy might be justified with 50 or more subjects.

Cross-Validation Procedure. If one cannot ask questions about interaction effects in terms of prior contrast weights, then empirical

post-hoc contrast weights can be estimated from the cell means for half the experimental subjects (analysis group) by setting $\omega_{ij} = a_{ij}$. The remaining subjects (cross-validation group) then can be used to get an unbiased estimate of each putative interaction effect and to test the obtained effect for significance. This cross-validation method is standard operating procedure for psychometricians in multiple regression, test construction, etc., (Mosier, 1951). Cross-validation is rarely used as such by statisticians. However, the "jack-knife method", popularized by Tukey, as well as the Geisser "predictive sample reuse method", can be viewed as a generalization of the usual two-group cross-validation procedure (see Mosteller & Tukey, 1968; Geisser, 1975).

Our example consisted of 24 subjects, so 12 of them (i.e., the analysis or training group) are used to estimate the weights while the remaining 12 are the cross-validation (testing) group. Table 4 gives the set of weights derived from the analysis group.

- - - - - - - - - - - - - - -

Table 4 about here

- - - - - - - - - - - - - - -

Notice that these values are quite similar to our theoretical values in that most of the alcohol effect occurs at WF_5.

Significance Tests

The significance test will have maximum power when the interaction effect, τ_{ij}, for each cell is equal to, or proportional to, the contrast weight for that cell. For simplicity in what follows, we omit the case of proportionality. Thus, the experimenter's hypothesis is:

$$\underline{H}_1: \quad \tau_{ij} = \omega_{ij}, \tag{7}$$

and the null hypothesis is:

$$H_o: \quad \tau_{ij} = 0 \tag{8}$$

Both hypotheses assume that τ_{ij} equals the expected value of q_{ij}, so H_o is equivalent to $\bar{q}_{ij} = 0$, where

$$\bar{q}_{ij} = \frac{\sum_{k=1}^{n} q_{ijk}}{n} \tag{9}$$

and n refers to the number of subjects.[4] The contrast score for the kth subject is defined by equation 1. The expected value of $\bar{Y}$, computed over the n subjects, will equal zero if H_o is true. If H_1 is true, then the expected value is:

$$E(\bar{Y}) = \sum_i \sum_j \omega_{ij}^2 = \sum_i \sum_j \tau_{ij}^2 \tag{10}$$

We now have a between-subjects t-test, eliminating the repeated-measures problems.

Theoretically, under H_1 the Y_k scores for the cross-validation group will range from zero to $\sum_i \sum_j \omega_{ij}^2$.[5] Under H_o, the Y_k scores can be positive or negative, centering on zero. The conventional t-test is simple to obtain.

$$t = \frac{\bar{Y}\sqrt{n}}{s_y} \tag{11}$$

where H_o is $E(\bar{Y}) = 0$

H_1 is $E(Y) > 0$

and the t-ratio has $(n-1)$ degrees of freedom.

To complete our example, Table 5 illustrates the Y_k scores for the cross-validation group using the weights obtained from the analysis group (right) or the theoretical weights (left).

- - - - - - - - - - - - - -

Table 5 about here

- - - - - - - - - - - - - -

For Sternberg (1969) additive effects between two treatments are as meaningful as interactive effects. Thus, he suggests that

> ... one might present findings in terms of mean interaction contrasts of theoretically interesting magnitudes, and adjust tests so that errors of Types 1 and 2 have equal probabilities with respect to such alternatives. (p. 310)

One could, for example, use the appropriate t-ratio at the 50% level as a rejection point. This adjustment cannot be used to infer additivity, however, within the context of the test we recommend for interactions. That is, rejection of a specific interaction hypothesis does not imply additivity. In view of Taylor's (1976) cautions against interpreting additivity, this inability to infer an additive relationship from a specific test does not appear to be a serious drawback.

Several Interaction Hypotheses

The major interest of the study may not be the comparison of H_0 with H_1. In some cases, divergent theories might lead to divergent hypotheses as to the nature and direction of the interaction. For example, suppose that some evidence predicted a "golden-mean" theory

in which the optimal effect of alcohol occurs at the median word frequency level, WF_3, while decrement occurs at the extreme word frequency levels, WF_1 and WF_5. Thus, the theoretical weights might be those shown in Table 6. We now have two experimental hypotheses,

- - - - - - - - - - - - - - -

Table 6 about here

- - - - - - - - - - - - - - -

$\underline{H}_1$ and $\underline{H}_2$. Let the vector of weights given in Table 4, $\underline{V}_1$, represent $\underline{H}_1$; and the vector of weights given in Table 6, $\underline{V}_2$, represent $\underline{H}_2$. Each vector will yield a $\underline{t}$-ratio--the higher the $\underline{t}$-ratio, the better the hypothesis fits the data. To make an accurate estimate of the significance of the $\underline{t}$-ratios, the rejection levels must be adjusted to take account of the fact that two similar significance tests have been obtained from the same data.[6]

Although there are many solutions to this multiple-comparison problem (e.g., Miller, 1966), we prefer the Dunn-Bonferroni method (Dunn, 1959). Assume that you want to hold your experimentwise Type 1 error at .05 (i.e., when $\underline{H}_o$ is true then either $\underline{t}$-ratio or both $\underline{t}$-ratios will be significant in five percent of all comparisons). In the simplest version of the Dunn-Bonferroni, the chosen alpha level is divided by the number of comparisons to obtain a critical level of significance for each comparison. Thus, in this example we would divide .05 by two to obtain .025--the level at which each $\underline{t}$-test would be evaluated. This is a very conservative test which guarantees that the Type 1 error will be .05 or greater, so it lacks some of the power of other tests. In general, it will be more conservative (i.e., the Type 1 error will

be larger) as the correlation between $\underset{\sim}{V}_1$ and $\underset{\sim}{V}_2$ increases. Thus, the experimenter should be careful to avoid testing redundant hypotheses.[7] When the weights of $\underset{\sim}{V}_1$ correspond exactly to the true interaction effects, ω_{ij}, then the contrast scores for $\underset{\sim}{V}_1$ will account for all interaction variance (i.e., any vector which is orthogonal to $\underset{\sim}{V}_1$ will have a $\overline{\underset{\sim}{Y}}_k$ of zero).

Most experiments contain several families of statistical hypotheses. The Dunn-Bonferroni adjustment may apply separately to each family (Miller, 1966). For example, one might be interested in the effect of a new treatment on various information processing stages, and thus introduce that treatment into a multifactor design with several treatments whose locus of effect has already been "established" by means of the Sternberg-Taylor procedure. Such an experiment would involve a two-step analysis. Step one, constituting one family of statistical tests, would involve confirming the interactions between the established treatments in data which did not include the new treatment. The second step, constituting the second family, would be a search for interactions between the effects of the new treatment and the established ones.

Justification

Let us review the remaining four ways to evaluate interactions in a repeated measures design: (1) the univariate repeated-measures ANOVA using the (probably) inflated estimate of degrees of freedom; (2) multivariate ANOVA: (3) conservative degrees of freedom with a univariate ANOVA: and (4) the Greenhouse-Geisser stepwise analysis. We will demonstrate that no matter what the outcome of the given procedures, the careful experimenter must use specific interaction contrasts.

Univariate Repeated-Measures ANOVA

For the conscientious experimenter few problems are so exasperating as the analysis of a repeated-measures design. In 1954, Box showed that the exact analysis of a repeated-measures design is a multivariate analysis and that the basic answers must be given in terms of multivariate F-ratios (see, also, Winer, 1971, sections 4.4 and 4.9). The usual univariate F-ratio approach, set forth in almost every psychological statistics text, is strictly valid only under a set of necessary and sufficient conditions known variously as (a) the "circularity property"[8] (Rouanet and Lepine, 1970); (b) "equality of variances of differences" (Huynh and Feldt, 1970); or (c) "homogeneity of interaction variances" (McNemar, 1962, pp. 315-316).

All F-ratios with only one degree of freedom for the numerator are valid under the standard multivariate normal assumptions (see Appendix). With three or more measures per subject, one could test for the "circularity property" and drop the univariate approach when it does not hold. When the circularity property does hold, one need only worry about interpreting the meaning of a significant F-ratio.

Multivariate Analysis of Variance

One possible solution for analyzing repeated-measures data is to avoid the univariate approach--do a multivariate ANOVA as soon as we have three or more measures per subject. Unfortunately, the multivariate ANOVA presents many problems unless one has a large number of subjects.

Given rc scores per subject, the number of subjects must be equal to or greater than (rc-1) in order to compute a multivariate ANOVA. This absolute bar is present because every multivariate ANOVA demands the computation of the inverse of a within-group covariance matrix. When

n+1 is less than the number of measures per subject, the determinant of the covariance matrix must be zero, and the inverse does not exist.

Most experimenters probably can slip under the (rc-1) barrier with one or two degrees of freedom to spare. If so, then we usually run into the problems of the ill-conditioned covariance matrix and the inherently large sampling variance of correlation coefficients and variances. The idea of estimating the inverse of a 4x4 or 5x5 covariance matrix, which is based on less than 20 degrees of freedom, only appeals to those who have an overwhelming faith in small samples.

The inherent sampling instability of second-order statistics (e.g., correlations, variances, etc.) will interact with small sample size to emphasize the problems of the multivariate-normal model. The unbiased estimate of a covariance matrix demands many more assumptions than the unbiased estimate of a difference between two independent means. Some of these assumptions are given in the Appendix. Ordinarily, even when we do not have exact normality or homogeneous variances, the Central Limit Theorem guarantees that the difference between a pair of independent means goes very quickly towards a normal distribution with homogeneous variance, provided that scores are independent and that there are enough of them. But second-order statistics are very sensitive to even slight deviations from normality (e.g., the fourth moment), and the Central Limit Theorem has very little effect with small samples. Furthermore, if some fundamental assumption such as linearity or independence has been violated, then the Central Limit Theorem simply does not apply. For example, if one measure has a non-monotonic relation to another, then no increase in the sample size will linearize that relation.

In summary, then, multivariate analysis of variance is an exact way to evaluate repeated-measure interactions. Nevertheless, a multivariate ANOVA is practical only (a) if one has fifty cases or more (given that rc is more than 5) and (b) if the important assumptions of linearity, a well-conditioned covariance matrix, etc., hold (see Appendix).

Conservative Estimate of the Degrees of Freedom

Some investigators have attempted to avoid the repeated-measures problems of the univariate ANOVA by using the conservative test advocated by Greenhouse and Geisser (1959). This approach is based on a statistic, epsilon, developed by Box (1954). When epsilon equals unity,[9] then the usual univariate ANOVA of repeated-measures is valid. When epsilon is less than unity, multiplying the degrees of freedom for the numerator and the denominator of the univariate F-ratio by epsilon, gives the approximate degrees of freedom for a valid evaluation of the univariate F-ratio.

The Greenhouse and Geisser (1959) conservative test avoids the estimation of epsilon. It uses the fact that epsilon cannot go any lower than $1/(p-1)$ where p is the number of measures per subject. Consequently, by using this minimum value of epsilon, one usually underestimates the degrees of freedom and the significance level of the F-ratio. If the conservative test is significant, then the multivariate analysis would also be significant. But, the Greenhouse-Geisser conservative test will only reject the null hypothesis for rather large F-ratios. When the conservative test is not significant, the experimenter is given very little information.

Greenhouse-Geisser Stepwise Analysis

Greenhouse and Geisser in describing their stepwise analysis (1959, p. 110) advised that one should test the univariate F-ratio first, using the nominal degrees of freedom without the epsilon correction. When this F-ratio is significant, one must then perform the conservative test. The epsilon correction is needed only when the F-ratio with inflated degrees of freedom is significant and the conservative test is not. This stepwise approach, also recommended by Lana and Lubin (1963), assumes that no further analysis is needed when the univariate F-ratio with possibly inflated degrees of freedom is not significant. This assumption is wrong. Davidson (1972) showed that one could easily obtain a significant Hotelling T^2 (i.e., a significant multivariate F-ratio) on the same data.[10] Thus, a non-significant F, with possibly inflated degrees of freedom, also gives the experimenter very little information.

In summary, if a significant F is obtained with the conservative test or by using epsilon to approximate the valid degrees of freedom, then the only problem is to interpret the meaning of the F-ratio. Otherwise, one can turn to a multivariate ANOVA if enough subjects are available (e.g., 50 or so, given that p is about 5) and the important multivariate assumptions are met.

Specific Interaction Contrasts Versus the Omnibus F-Ratio

Justification. When a non-significant F is obtained either with the Greenhouse-Geisser conservative test or by using epsilon to approximate the valid degrees of freedom and the conditions are not appropriate for a multivariate ANOVA, only one of our suggested solutions remains--specific interaction contrasts. Such a situation always justifies the use of specific interaction contrasts.

However, the most powerful argument for the use of interaction contrasts is a pragmatic one. Even when F-ratio tests (univariate or multivariate) are appropriate, only two outcomes are possible--a significant F-ratio or a nonsignificant F-ratio. Both outcomes should lead the careful experimenter to test for specific interactions.

A nonsignificant F-ratio may be due to a lack of power, the result of considering all possible interaction contrasts simultaneously. Therefore, a careful experimenter should always apply any prior interaction contrast (derived from hypothesis or an analysis group) to the data to see if the t-test is significant even though the overall F-ratio was not.

If the F-ratio is significant, then the experimenter still has to determine which cells yielded the significant interaction effect, as well as the direction and amount of each cell effect. A significant F-ratio does not guarantee that the experimenter's hypothesis about the interaction is correct. For example, the cell sizes may be as hypothesized, but with opposite signs; all signs may be as hypothesized, but the amounts may be wrong; or possibly the experimental hypothesis may be wrong about both the direction and magnitude of the interaction effects.

In summary, a nonsignificant F-ratio tells one almost nothing. A significant F-ratio is merely a hunting permit, with the interaction contrast and its associated t-test as weapon.

Advantages. A specific interaction contrast is pragmatic, powerful, robust and easy to compute. Moreover, as we have shown, it is unavoidable when evaluating interactions in a repeated-measures design.

Specific interaction contrasts allow the experimenter to test his interaction hypothesis exactly. Since a specific hypothesis is tested,

the *t*-test performed is one-tailed and always more powerful than the comparable *F*-ratio.

Only three basic assumptions are required of the data in order to test an interaction contrast: (1) independence of the *n* contrast scores, (2) a normal distribution, and (3) homogeneous variance. Independence is guaranteed by independent selection and scoring of the *n* subjects. The latter two assumptions--normality and homogeneous variance--are not guaranteed, but can be tested by using the Wilcoxon signed rank test in tandem with the *t*-test. As *n* increases, normality and homogeneous variance become irrelevant. Thus, specific interaction contrasts, in conjunction with appropriate rank-order tests, are robust.

References

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effects of inequality of variance and of correlation between errors in the two-way classification. Annals of Mathematical Statistics, 1954, 25, 484-498.

Davidson, M. L. Univariate versus multivariate tests in repeated-measures experiments. Psychological Bulletin, 1972, 77, 446-452.

Dixon, W. J. (Ed.). Biomedical computer programs. Berkeley: University of California Press, 1975.

Dunn, O. J. Confidence intervals for the means of dependent normally distributed variables. Journal of the American Statistical Association, 1959, 54, 613-621.

Fisher, R. A., & MacKenzie, W. A. Studies in crop variation. II. The manurial response of different potato varieties. Journal of Agricultural Science, 1923, 13, 311-320.

Gaito, J. Repeated measurements, designs and counterbalancing. Psychological Bulletin, 1961, 58, 46-54.

Geisser, S. The predictive sample reuse method with applications. Journal of the American Statistical Association, 1975, 70, 320-328.

Geisser, S., & Greenhouse, S. W. An extension of Box's results on the use of the F distribution in multivariate analysis. Annals of Mathematical Statistics, 1958, 29, 885-891.

Greenhouse, S. W., & Geisser, S. On the methods in the analysis of profile data. Psychometrika, 1959, 24, 95-112.

Huynh, H., & Feldt, L. S. Conditions under which mean square ratios in repeated measurements designs have exact F-distribution. Journal of the American Statistical Association, 1970, 65, 1582-1589.

Lana, R. E., & Lubin, A. The effect of correlation on the repeated measures design. Educational and Psychological Measurement, 1963, 23, 729-739.

McNemar, Q. Psychological statistics. New York: Wiley, 1962.

Miller, R. G., Jr. Simultaneous statistical inference. New York: McGraw-Hill, 1966.

Mosier, C. I. Problems and designs of cross-validation. Educational and Psychological Measurement, 1951, 11, 5-11.

Mostellier, F., & Tukey, J. W. Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), Handbook of social psychology (Vol. 2). Reading, Mass: Addison-Wesley, 1968.

Plackett, R. L. Models in the analysis of variance. Journal of the Royal Statistical Society, 1960, 22, 195-217.

Rouanet, H., & Lepine, D. Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. British Journal of Mathematical and Statistical Psychology, 1970, 23, 147-163.

Sternberg, S. The discovery of processing stages: Extension of Donder's method. In W. G. Koster (Ed.), Attention and performance II. Amsterdam: North Holland, 1969. (Reprinted from Acta Psychologia, 1969, 30, 276-315.

Taylor, D. A. Stage analysis of reaction time. Psychological Bulletin, 1976, 83, 161-191.

Tharp, V. K. Alcohol and response selection: An information processing analysis. (Doctoral dissertation, University of Oklahoma, 1975). Dissertation Abstracts International, 1975, 36, 6438B. (University Microfilms No. 76-14,096).

Tharp, V. K., Rundell, O. H., Lester, B. K., & Williams, H. L. Alcohol and information processing. Psychopharmacologia, 1974, 40, 33-52.

Votaw, D. F., Jr. Testing compound symmetry in a normal multivariate distribution. Annals of Mathematical Statistics, 1947, 19, 447-473.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill; 1st ed., 1962, 2nd ed., 1971.

Footnotes

This research was supported by the Department of the Navy, Bureau of Medicine and Surgery, under Work Unit Numbers MR000.01.01.6011 and MR041.01.03-0152. The views presented in this paper are those of the authors. No endorsement by the Department of the Navy has been given or should be inferred.

We thank S. Geisser, H. Williams, J. Callan, and the staff of the Psychophysiology Division at the Naval Health Research Center for their helpful comments during the preparation of this manuscript.

Address requests for reprints to: Van K. Tharp, Jr., Ph.D., Naval Health Research Center, San Diego, California 92152.

[1]A contrast is a weighted combination of scores where the sum of the weights equals zero.

[2]Since there are only two levels of drug (i.e., 1 df), one can subtract baseline from alcohol scores to simplify the computational procedures for computing residuals from additivity. We have not simplified in the above example in order to show how such residuals are computed with both rows and columns. Finding residuals for the simplified scores is a straightforward generalization of the example given.

[3]This procedure is equivalent to the method detailed for finding residuals from additivity.

[4]We assume that the interaction effect for the ijth cell, τ_{ij}, is a constant for all n subjects. Any interaction with the subjects is thrown into the error deviance.

[5]When the specific interaction hypothesis is correct, under most circumstances the contrast weights will be proportional to the actual

interaction effect. For example, we recommend smoothing the contrast weights, but not the residuals from additivity for each subject, to single digit numbers. Thus, the theoretical maximum value can be stated more accurately as being $k \sum_i \sum_j \omega_{ij}^2$, where k is the constant of proportionality.

[6]Each t-ratio may have a different between-subjects variance in this example. The completely general ANOVA demands that every F-ratio and t-ratio for interaction must have the same error variance to comply with the assumption of homogeneous variance.

[7]The routine application of stepwise multiple regression (e.g., Dixon, 1975) to the matrix of vectors representing the hypotheses will guarantee linear independence and thus eliminate redundancy. If there are h degrees of freedom for the interaction deviance in the ANOVA, then one can construct h vectors that are mutually orthogonal to one another.

[8]These "circularity properties" are less restrictive than the "compound symmetry" property (Votaw, 1948). Compound symmetry holds when, given p repeated-measures, the p variances are equal and the $p(p-1)/2$ correlations are identical. Rouanet and Lepine (1970) have shown that compound symmetry is sufficient but not necessary to guarantee the validity of all F-ratios in a univariate ANOVA of a repeated-measures design. To confuse the issue further, the Greenhouse and Geisser 1959 article has a slip (p. 95): the word necessary was applied to the compound symmetry model rather than sufficient, as was implied by their first paper (Geisser and Greenhouse, 1958). This slip was copied by Lana and Lubin (1963), Gaito (1961), Winer (1962) and others. Winer corrected this slip in his second edition (1971, pp. 282-283).

[9]Rouanet and Lepine (1970), as well as Huynh and Feldt (1970), verified that when the circularity property holds, the Box epsilon criterion equals unity.

[10]If some of the measures with small variances have very high correlations with one another, and if the remaining measures have high variance and low inter-correlations with all other measures, then the multivariate ANOVA will unerringly pick out the linear compound that gives maximum weight to the differences between the means of the highly correlated measures. The univariate ANOVA gives equal weight to all differences, and so may end up with a nonsignificant result.

Table 1

Verbal RT in Milliseconds, N=1

	Alcohol	Baseline	
WF_1	901	818	$\bar{X}_1$ = 859.5
WF_2	932	854	$\bar{X}_2$ = 893
WF_3	949	870	$\bar{X}_3$ = 909.5
WF_4	1109	982	$\bar{X}_4$ = 1045.5
WF_5	1733	1205	$\bar{X}_5$ = 1469
	$\bar{X}_{.1}$ = 1124.8	$\bar{X}_{.2}$ = 945.8	$\bar{X}_{..}$ = 1035.3

Table 2

Residuals From Additivity For One Subject

	Alcohol	Baseline	Row Effects
WF_1	-48.0	48.0	-175.8
WF_2	-50.5	50.5	-142.3
WF_3	-50.0	50.0	-125.8
WF_4	-26.0	26.0	10.2
WF_5	174.5	-174.5	433.7
Column Effects	89.5	-89.5	$\bar{X}_{..}$ = 1035.3

Table 3

A Set of Theoretical Contrast Weights

	Alcohol	Baseline	Sum
WF_1	-1	1	0
WF_2	-1	1	0
WF_3	-1	1	0
WF_4	0	0	0
WF_5	3	- 3	0
	0	0	

Table 4

Contrast Weights Obtained Empirically

from the Analysis Group

	Condition	
	Baseline	Alcohol
WF_1	-1	1
WF_2	-5	5
WF_3	1	-1
WF_4	-4	4
WF_5	9	-9

Table 5

Contrast Scores

Subjects	Using Hypothesized Weights	Using Analysis Group Weights
Y_1	648	2,458
Y_2	3,101	8,293
Y_3	2,055	6,547
Y_4	4,446	13,388
Y_5	438	1,096
Y_6	3,616	9,800
Y_7	271	2,249
Y_8	580	1,888
Y_9	-524	-1,594
Y_{10}	155	-15
Y_{11}	530	1,528
Y_{12}	818	618
$\bar{Y}$	1,344.500	3,854.667
s_y	1,574.214	4,570.038
t	2.959	2.922
p	.006	.007
r^2	.443	.437

Table 6

A Golden-Mean Set of Contrast Weights

	Alcohol	Baseline
WF_1	-1	1
WF_2	0	0
WF_3	2	-2
WF_4	0	0
WF_5	-1	1

Appendix

Some Assumptions of Multivariate Analysis of Variance

(1) Independence of Subjects. Each subject was selected independently of any other subject.

(2) Normality. Each of the p measures has a marginal normal distribution.

(3) Homogeneity of Residual Variance (Homoscedasticity). When we predict one of the p measures from any linear component of the remaining (p-1) measures, all errors of prediction must have the same variance.

(4) Linearity. Each of the p measures has a linear regression on any weighted combination of the remaining (p-1) measures.

(5) Homogeneous Universe. Each subset of subjects in the sample has the same covariance matrix as any other subset.

(6) Well-conditioned Covariance Matrix. The determinant of the p by p covariance matrix is clearly greater than zero.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 76-78
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): Specific Interaction Contrasts: A Statistical Tool for Repeated-Measures Designs.

5. TYPE OF REPORT & PERIOD COVERED: Interim rept.

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): Van K. Tharp, Jr. & Ardie Lubin

8. CONTRACT OR GRANT NUMBER(s): 61153N

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92138

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: MR000 01 01, 6011 and MR041 01 03, 0152

11. CONTROLLING OFFICE NAME AND ADDRESS:
Naval Medical Research & Development Command
Bethesda, MD 20014

12. REPORT DATE: December 1976

13. NUMBER OF PAGES: 30

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
Department of the Navy
Bureau of Medicine and Surgery
Washington, D. C. 20372

15. SECURITY CLASS. (of this report): UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report):
Approved for public release; distribution unlimited.

61153N

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES: NAVHLTHRSCHC-76-78

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Repeated measures-designs
Interactions
Contrasts
MR000.01
MR041.01

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(U)....Experimental psychologists often use multifactor repeated-measure designs in which interactions are the most important effects to be assessed. An experimenter has at least five ways to evaluate such interactions: (1) a univariate repeated-measures analysis of variance (ANOVA), with (probably) inflated estimates of the degrees of freedom; (2) a univariate repeated-measures ANOVA with the Greenhouse-Geisser conservative estimate of the degrees of freedom; (3) the Greenhouse-Geisser stepwise analysis; (4) a multivariate

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Copy

391 642

20. (Continued)

ANOVA; and (5) specific interaction contrasts. We show that no matter which of the above paths is chosen, the careful experimenter must compute specific interaction contrasts (i.e., t-tests). A worked example is given.