Psychological Methods
1997, Vol. 2, No. 4, 315–328

# Contrast Tests of Interaction Hypotheses

Robert P. Abelson
Yale University

Deborah A. Prentice
Princeton University

This article argues for the use of contrasts to test a priori interaction hypotheses in 2-way analysis of variance designs. It focuses on 3 underused types of interaction contrast tests: a "matching" pattern for cognate levels of row and column factors; the "qualitative quadratic," for monotonic profiles of means in the same direction but with opposed concavities; and a "differential curvilinearity" test for differences in the curvature of two profiles with the same direction of concavity. The circumstances that best capitalize on the potential advantages of a priori contrast tests are indicated, and an effect size measure for contrasts is presented. Investigators are urged to examine residuals after accounting for the variation attributable to the chosen contrast for patterns that may provide hints for more textured hypotheses in further research. If a posteriori contrasts are used, their effect sizes should be noted.

Contrasts are widely recognized as a useful tool for analyzing patterns of systematic variation between means within an analysis of variance (ANOVA; Harris, 1994; Judd & McClelland, 1989; Judd, McClelland, & Culhane, 1995; Keppel & Zedeck, 1989; Myers & Well, 1995; West, Aiken, & Krull, 1996). Compared with the cautious procedure of first testing omnibus effects as a prerequisite to further testing, planned contrasts have the potential advantages of greater power (Myers & Well, 1995), and better focus (Rosenthal & Rosnow, 1985) or "articulation" (Abelson, 1995).

When an investigator has hypothesized a pattern of means believed sufficient to represent the systematic variation exhaustively, an ideal outcome is for the $F$ test of this pattern contrast to yield a large effect size with a small $p$ value, and for the $F$ test of the residual sum of squares to be clearly nonsignificant and small. (As a rule of thumb, $F$s less than 2 may be regarded as small.) Such a parsimonious outcome indicates that the data are well described by an a priori hypothesis for the pattern of systematic variation, and that the deviations from the fitted pattern might be attributable to an orderly random process. We need a word or phrase that refers to this outcome; to emphasize its terseness and economy, we call it a *canonical* outcome.[1]

There are two distinct types of noncanonical outcomes for the test of an a priori contrast: (a) the contrast itself may be clearly weak and nonsignificant, and (b) both the contrast and the residual might be strong and significant. The first outcome suggests that the data do not support the pattern hypothesized by the investigator, either because there is very little systematic variation between means, or because the pattern of the systematic variation is badly fitted by the experimenter's hypothesis. The second outcome indicates that the investigator's hypothesis has received less than total support—the data pattern was somewhat similar to theoretical expectation, but there was additional systematic deviation beyond the expected pattern. Although this outcome is not as parsimonious as a canonical outcome, it can also be satisfying, es-

Robert P. Abelson, Department of Psychology, Yale University; Deborah A. Prentice, Department of Psychology, Princeton University.

We are grateful to Rich Gonzalez, Bob Rosenthal, and Steve West for comments on previous versions of this manuscript and to Allison Baer for assistance with its preparation.

Correspondence concerning this article should be addressed to Deborah A. Prentice, Department of Psychology, Princeton University, Green Hall, Princeton, New Jersey 08544-1010. Electronic mail may be sent via Internet to predebb@pucc.princeton.edu.

[1] This terminology is borrowed from mathematical usage, where its meaning is something like idealized, optimal, and cherished; the term occurs in the phrase, "canonical correlation."

pecially when detective work on residuals yields interesting patterns of systematic variation. We call this type of outcome *ecumenical:* The results are more varied than was initially conceptualized by the investigators.

## Basic Formulas

### *Revealing Interaction*

Interaction patterns are often not obvious from inspection of a two-way table (usually a table of group means), because main effects can mask them. Thus, it is useful to look at the table produced by subtracting the two main effects and the grand mean from the original table of means. Call the original table of means *X;* the equation for making these three adjustments is[2]

$$I_{jk} = (X_{jk} - X_{..}) - (X_{.k} - X_{..}) - (X_{j.} - X_{..}). \quad (1)$$

This simplifies to

$$I_{jk} = X_{jk} - X_{.k} - X_{j.} + X_{..}, \quad (2)$$

where $I_{jk}$ = the "interaction score" or "residual" in cell (*j, k*); $X_{jk}$ = the original entry in cell (*j, K*) of *X;* $X_{.k}$ = the mean, over all rows, of entries in column *k* of *X;* $X_{j.}$ = the mean, over all columns, of entries in row *j* of *X;* $X_{..}$ = the grand mean of all entries in the data matrix. To illustrate the application of Equation 2, suppose *X* is the 3 x 3 array of data shown in the top half of Table 1. Each cell entry gives a mean of *n* cases from a 3 × 3 factorial experiment to be described later. Equation 2 yields the matrix *I* of interaction scores, given in the bottom half of Table 1.

These entries represent the residual (interaction) scores for each cell, after removal from the original data of the observed grand mean and row and column main effects. (The nature and statistical significance of the main effects is not at issue for the present discussion.) In our example, the interaction scores on the main diagonal—cells (1, 1), (2, 2), and (3, 3)—are positive, indicating that the original means in these cells of *X* are greater than can be attributed to the relative effects of their respective rows and columns. Note cell (1, 1) in particular. The original cell mean of 50.50 does not strike the eye as a relatively high value, as it is smaller than the other two entries in its row of *X*. On the other hand, it is quite a bit higher than the other two entries in its column. Equation 1 provides a way to sort this out. We see from the low mean of column 1 that scores in that column are generally depressed. The interpretation of these entries will become evident once the actual study has been described.

Table 1
*Means and Interaction Scores in a 3 × 3 Design*

| | | | |
|---|---|---|---|
| | 50.50 | 67.25 | 59.50 |
| Means = | 34.26 | 67.06 | 54.28 |
| | 25.22 | 57.15 | 68.17 |
| | 8.47 | −1.95 | −6.53 |
| Interactions = | −0.56 | 5.09 | −4.52 |
| | −7.92 | −3.14 | 11.05 |

Two algebraic properties of interaction scores are noteworthy. First, every row and column sums to zero. Within rounding error, this is true of the array *I* in Table 1, for example. Two-way arrays that sum to zero in every row and column are called *doubly centered* (Aiken & West, 1991). Second, the sum of squares for the Row × Column interaction in a two-way ANOVA can be calculated directly by summing the squares of the entries in *I*, and multiplying by the cell *n*. Thus the contribution of cell (*j, k*) to the overall interaction sum of squares is proportional to $I^2_{jk}$.

If the *n*s are unequal, and if the "unweighted means" model (Winer, Brown, & Michels, 1991), also called the "cell means" model (Boik, 1993; Kirk, 1995) is used, then the sum of squared $I_{jk}$s is multiplied by a virtual value: the harmonic mean (ñ) of the cell *n*s, which is calculated by taking the reciprocal of the average over all cells of their reciprocal *n*s. The harmonic mean of any set of numbers is always less than their arithmetic mean. For example, the arithmetic mean of 10, 20, 30, 40, 50, 60, 70, 80, and 90 is 50, but the harmonic mean is 31.81. This difference reflects the fact that the relatively large sampling variance associated with means based on smaller than average *n*s inflates the error term more than the relatively small sampling variance associated with means based on larger than average *n*s dampens the error term. Extremely small *n*s sharply degrade the power of *F* tests. To continue with this example, suppose the first *n* in the set of nine *n*s above was 2 instead of 10. Then the harmonic mean would drop from 31.81 to 13.80. The *F* tests would behave as though there were only 14 cases in each cell.

[2] A good way to understand Equation 1 is through the "sweeping out" method (see Schmid, 1991) that successively corrects the data table for the grand mean, then row, and then column effects (see also, Judd, McClelland, & Culhane, 1995).

### *Choosing Contrast Weights*

In general, contrast weights are chosen to mimic the theoretically predicted pattern of the interaction scores (just as, in the one-way case, a linear trend hypothesis is tested with contrast weights that are themselves linear). Denote a two-dimensional matrix of interaction contrast weights by $W$, with entries $w_{jk}$. Just like the matrix $I$ of interaction numbers, the matrix $w$ must be doubly centered; that is, the $w$s must sum to zero for every row and column if the contrast is to be independent of main effects. The sum of all entries is also zero, as in the one-way case.

### *Calculating a Contrast Sum of Squares*

The equation[3] for a contrast sum of squares, $SS_{\text{con}}$, for an interaction hypothesis is

$$SS_{\text{con}} = n \cdot C^2/\Sigma_j \Sigma_k w_{jk}{}^2, \qquad (3)$$

with the contrast

$$C = \Sigma_j \Sigma_k (w_{jk} I_{jk}). \qquad (4)$$

This equation is the same as the formula for a one-way contrast, except that double summations replace a single summation. Because the $w$s sum to zero, both positive and negative values of $w$ occur. The value of $C$ will be large to the extent that positive interaction scores in $I_{jk}$ fall where the contrast weights $w_{jk}$ are positive, and negative interaction scores fall where the contrast weights are negative, so that the products in Equation 4 accumulate rather than cancel out. When $C$ is then squared, it becomes sizable relative to the sum of squared contrast weights.

A useful alternative equation for $C$ can be derived by entering the algebraic expression in Equation 2 for $I_{jk}$ into the summation in Equation 4 and simplifying. The double centering of $w$ causes the last three terms to drop out, leaving only

$$C = \Sigma_j \Sigma_k (w_{jk} X_{jk}) \qquad (5)$$

In other words, the interaction contrast weights applied to the original data array yield the same result as when applied to the interaction effects.

## The Importance of Residuals

In textbook and journal accounts of applications of contrasts, the display and inspection of residuals is seldom discussed enough. (For one exception, at least, see Anderson, 1982.) In the description-oriented approach to statistical analysis advanced by Tukey (1977), however, detailed treatment of residuals is fundamental, as the goal of statistical analysis is to produce a formulation with the structure: *DATA = FIT + RESIDUAL*, or *MESSAGE = SIGNAL + NOISE*. On this point, as on most others, we hold with Tukey. Without a significance test and careful inspection of the residual variation after extraction of a significant contrast, it is not possible to know whether the outcome is canonical or ecumenical. One cannot argue that the significant a priori contrast gives an adequately complete account of the systematic interaction, because some other systematic pattern or patterns may have been left out of the summary.

There are at least two things that an investigator can do with residuals beyond testing the omnibus null hypothesis that they contain no systematic variance: (a) seek other systematic patterns that would suggest further single *df* contrasts, or (b) plot the distribution of numerical values and look for interesting features. Tukey (1977) gave a very user-friendly procedure—the "stem-and-leaf"—for plotting distributions and identifying outliers; Emerson (1991) explained a graphical procedure for detecting departures of the distribution of residuals from a normal distribution; Abelson (1995) presented a catalogue of underlying process diagnostics associated with various odd features of distributions; and Madansky (1988) provided a nice formal treatment of several precise distributional fits with meaningful interpretations. Discussion of the details of these procedures is beyond our present intent; suffice to say that the tools are available, and investigators should not hesitate to explore patterns for interesting ideas, regardless of the outcome of an omnibus significance test of the original interaction.

## Effect Size Measures for Contrasts

In recent years, commentators (e.g., Cohen, 1992) have emphasized the importance of measures of effect size for giving information beyond or instead of that provided by $p$ values. Effect sizes for mean differences have been treated extensively in the literature (e.g., Rosenthal, 1991), usually in connection with meta-analysis, but psychological researchers seem to lack familiarity with effect size measures for contrasts. We present one such measure here.

[3] Throughout this article, our calculations will presume that there are either equal $n$s in all cells, or, if there are unequal $n$s, that the unweighted means model is used, and the $n$ in the formulas is ñ, the harmonic mean.

The first question to address is the nature of the "effect" whose size is being estimated. The answer is as follows: Contrast tests have the same mathematical structure as $t$ (or $F$) tests of mean difference, with the contrast $C$ replacing the mean difference. The value $C$ from Equation 4 can be conceptualized either as the magnitude of an effect (main effect, interaction effect, etc.), or as the strength of a pattern (linear trend, matching interaction, etc.). (For the distinction between these two types of conceptions, see Abelson, 1995.) The magnitude of C is the item of interest. An index of contrast effect size must remove the arbitrary inflationary influence that the absolute sizes of the $w$s have on the value of $C$, by including the factor $\sqrt{\Sigma w_{jk}{}^2}$ in its denominator. Also, because a simple mean difference can be represented as a contrast with the coefficients $w_1 = +1$ and $w_2 = -1$ for the respective means and zeroes for any other means, the effect size index must reduce to Cohen's $d$ for the case of a simple mean difference.

In light of these considerations, the appropriate contrast effect size measure $d_c$, is given by

$$d_c = (\sqrt{2})\ C/S \cdot (\sqrt{\Sigma w^2}), \qquad (6)$$

where $S$ is the estimated standard deviation within groups. In the simple mean difference case, $\sqrt{\Sigma w_{jk}{}^2}$ equals $\sqrt{2}$, and the formula becomes $d_c = (X_1 - X_2)/S$, which is Cohen's $d$.

A more convenient form of Equation 6 can be derived by replacing $S$ with its estimate as the square root of the mean square error associates with the contrast test, and by expressing $C$ as the right-hand side of Equation 5. This yields

$$d_c = (\sqrt{2})\ \sqrt{F_{con}}/\sqrt{n}, \qquad (7)$$

which is a generalization of the equation $d_c = 2t/(\sqrt{df})$ given by Rosenthal (1991, p. 17), with his $t$ equal to our $\sqrt{F}$, and his $df$ equal to our $2n$.

## Types of Interaction Hypotheses

It would be impossible to provide a general classification of types of interaction contrasts. In principle, there are infinitely many. An elegant interaction-contrast generating device involves "crossing" single degree of freedom polynomial row contrasts with single $df$ polynomial column contrasts. The arithmetic procedure for creating these contrasts of contrasts (Harris, 1994, pp. 179–182) involves calculating the element-by-element products of the individual contrast weights (as in elementary school multiplication tables, or in the more "highfalutin" terminology of vector algebra, taking the "outer product" of the two weight vectors). If the polynomial contrasts are orthogonal (as they typically are in the standard tables; e.g., Winer, Brown, & Michels, 1991, p. 982, or Harris, 1994, p. 467), the scheme produces (R − 1) × (C − 1) orthogonal interaction contrasts that between them exhaust the interaction degrees of freedom. This scheme is sometimes useful (see Anderson, 1982), and has received good textbook coverage. In our view, it is not a good idea to place exclusive reliance on this set of contrasts, however. Higher order polynomial contrasts (cubics, quartics, quintics, etc.) are very difficult to interpret, and do not often yield canonical outcomes. For interactions, the inscrutability of particular contrasts based on products of higher order components is especially daunting. Summary descriptions requiring interpretations of many crossed polynomial contrasts are even more taxing.

Our aim, instead, is to present contrast patterns that often arise from standard sorts of simple theoretical arguments, and are thus appropriate for a priori tests. In principle, there are many of these as well. We focus here on three types that are individually interesting and collectively useful in raising important practical and statistical issues.

### *The Matching Hypothesis*

Sometimes the $m$ rows and $m$ columns of an $m \times m$ experimental design represent respective levels of the two variables that are conceptually similar to each other so that the row and column levels for the cells on the diagonal of the data matrix correspond in some way. For example, rows might comprise a set of social class levels for fathers, and columns might comprise the social class levels of their children, with some measure of closeness of the father–child relationship as the dependent variable. One might hypothesize that having equal social status would make for closer father–child relationships, so that the means in the diagonal cells should be relatively higher than in the other cells.

In this type of example, where some kind of correspondence principle is operative, the expectation of a "matching" pattern can be translated into the interaction hypothesis that the interaction scores for the cells along the main diagonal of the data matrix will all be positive, whereas the scores for the off-diagonal cells will be negative.

An a priori matching hypothesis arose in an experiment by Rabbie (1964), who tested the idea from social comparison theory (Festinger, 1954), that people in an unusual emotional state tend to seek affiliation with people in a similar state. He induced participants to feel either a high, medium, or low level of fear in anticipation of receiving electrical shocks. While they were waiting for the fearful experience, he had them rate, on a 100-point scale, the desirability of having the company of a high-, medium-, or low-fear companion. Mean ratings (with $n = 10$ per cell) are given by the entries shown in the means array in Table 1. Successive rows represent the participant's own fear level—high, medium, or low—and the columns represent the companion's fear level—high, medium, or low.

*Weights for the matching hypothesis.* According to Rabbie's (1964) hypothesis, the 3 x 3 table ought to show a matching pattern—there should be tendency for high-fear participants to prefer high-fear companions, mediums to prefer mediums, and lows to prefer lows. He made no predictions about the main effects of own fear or companion fear; only the interaction is of interest here. Of the 4 *df* available for that interaction, only the 1 *df* contrast between diagonal and off-diagonal cells is pertinent for the predicted matching pattern. In comparing the three interaction scores on the diagonal with the six interaction scores off the diagonal, it assesses the effect of matched fear levels vs. nonmatched levels. If the matching tendency is uniform over rows and columns, the appropriate matrix of weights for a contrast is readily seen to be:

$$W = \begin{matrix} +2 & -1 & -1 \\ -1 & +2 & -1 \\ -1 & -1 & +2 \end{matrix} \quad (8)$$

Like the *I* matrix, this array is doubly centered. As a consequence, it is orthogonal to all possible main effects. (This can be verified by noting that the contrast *C* in Equation 5 would be unaltered by any change in main effects—i.e., *C* is invariant over the set of operations of adding any constant to every entry in any row or column of *X*.)

*The ANOVA framework for the contrast.* The nonbold entries in Table 2 show the standard ANOVA summary, with the background information that there were nine independent groups of 10 participants each,[4] and that the $\mathrm{MS}_{\mathrm{within}}$ was 314. The analysis yielded a significant main effect of companion fear (reflecting general aversion for the high-fear companion), and a significant overall interaction between own fear and companion fear. The unarticulated omnibus interaction conveys no useful information and should be replaced by the subdivision of the interaction into the matching contrast and a residual.

Table 2
*Analysis of Variance of the Rabbie (1964) Data*

| Source | *SS* | *df* | *MS* | *F* | *p* |
|---|---|---|---|---|---|
| Own fear (O) | 1,342 | 2 | 671 | 2.14 | *ns* |
| Companion fear (C) | 13,232 | 2 | 6,616 | 21.07 | <.01 |
| O × C | 3,592 | 4 | 898 | 2.86 | <.05 |
| **Matching contrast** | **3,026** | **1** | **3,026** | **9.64** | **<.01** |
| **Residual** | **566** | **3** | **189** | **0.60** | **>.50** |
| Error | 25,434 | 81 | 314 | | |

*Note.* The *F* ratios for the decomposed interaction are shown in bold.

The value of *C* for the matching contrast in the present example, using Equation 3 with 4 (or 5), is 73.80. The *n* per cell is 10, and the sum of squared *w*s is 18. Thus,

$$SS_{\mathrm{con}} = 10\,(73.80)^2/18 = 3025.80. \quad (9)$$

This contrast has 1 *df*, and captures a huge portion of the 4 *df* sum of squares (3,592) for the interaction. Subtracting out the contrast sum of squares of 3,026 leaves a residual interaction sum of squares of 566 with 3 *df*.

Two *F* ratios for the decomposed interaction are shown in bold in Table 2: an *F* of 9.64, with $p < .01$ for the contrast test of the matching hypothesis, and a nonsignificant *F* of 0.60 for the test of the residual interaction. The first *F* says that the matching pattern observed in the interaction is not at all likely to have been a fluke: A systematic claim of a strong matching is warranted. Indeed the effect size for this matching pattern, using Equation 7, is $d_c = 1.39$; it is a large effect. The second *F* says that one can plausibly argue that no further systematic patterning is needed in accounting for the observed interactions; the matching hypothesis appears to be quite sufficient. This is a canonical outcome.

*Treatment of possible residual variation in the interaction.* Despite the very small *F* value for the residual after accounting for the matching contrast, it

[4] In the actual experimental design, companion fear was a repeated measure, and therefore the calculation of error terms was somewhat more complicated than the presentation here suggests. We defer comment on error terms until our final example, as we want to emphasize the calculation of contrasts, which remain the same regardless of the error terms.

might still be worthwhile to examine the residual between cell variation, even if the 3 *df* test in Table 2 has indicated that sheer noise cannot be ruled out as a reasonable account. Systematic camels can be hidden in tents of apparent randomness, so to speak.

Here it is helpful to consider the theoretical reasoning behind this experiment—the social comparison concept that led to the matching hypothesis. The evidence in Table 2 supported the hypothesis of a relative preference for the companion whose fear level matched the participant's. What if the levels were different? Would it not be reasonable to expect the least preference when the fear levels were most different—in other words, a falling off of preference away from the main diagonal? In fact, the interaction matrix in Table 1 displays exactly such a tendency: The two most negative interaction scores are in the lower-left and upper-right corners. This fall-off effect can be tested with a second set of contrast weights. This set of contrast weights, call it *V*, must be doubly centered, so that it is orthogonal to both main effects (to avoid a confound that would allow a trivializing reinterpretation of the test outcome; see Abelson, 1995). In addition, it must be orthogonal to contrast *W*, else this second contrast might merely mimic the matching contrast. It is important to treat the "falling away" idea as conceptually independent. The steps for creating the matrix *V* are as follows: (a) write a matrix pattern *T* that expresses the qualitative character of the appropriate pattern; (b) double-center it, by Equation 2, and call the result U; (c) orthogonalize to *W* by regressing *U* on *W*, and taking residuals.

A reasonable pattern with high diagonals and a fall-off toward the corners is:

$$T = \begin{matrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1. \end{matrix} \qquad (10)$$

Note that the middle row and column have a positive sum; this is because the biggest drop-off (to −1) does not occur in the middle. After double-centering, standard linear regression formulas can be applied to orthogonalize the matrix to *w:*

$$V = U - b_{U \text{ on } W} \cdot W, \qquad (11)$$

where

$$b_{U \text{ on } W} = \Sigma_j\Sigma_k\,(u_{jk} \cdot w_{jk})/\Sigma_j\Sigma_k\,(w_{jk}{}^2). \qquad (12)$$

This procedure yields a somewhat surprising matrix of weights for the new fall-off contrast:

$$V = \begin{matrix} 1 & 1 & -2 \\ 1 & -2 & 1 \\ -2 & 1 & 1. \end{matrix} \qquad (13)$$

The nonintuitive pattern of weights for a fall-off contrast orthogonal to the matching contrast is a consequence of the double-centering and orthogonalizing. The value of *C* needed to calculate a sum of squares for this extra contrast can be found from the following:

$$C = \Sigma_j\Sigma_k\,(v_{jk}\,I_{jk}), \qquad (14)$$

or from

$$C = \Sigma_j\Sigma_k\,(v_{jk}\,X_{jk}), \qquad (15)$$

which are analogous to the two equivalent forms (Equations 4 and 5) for the initial contrast.

In the numerical example, the value of the *C* turns out to be 28.08. The sum of squares (and thus the mean square) for this "fall-off" contrast is $SS_{\text{con}} = n \cdot C^2/\Sigma_j\Sigma_k\, v_{jk}{}^2 = 10\,(28.08)^2/(18) = 438$.

Referring to the ANOVA in Table 2, we see that this *SS* of 438 exhausts most (77%) of the residual interaction of 566. This is a modest hint that the observed fall-off in Table 1 is systematic. A better hint is that the effect size for this fall-off contrast is an impressive $d = .52$.

These hints, coupled with the theoretical coherence of a fall-off social comparison tendency with decreasing similarity of the companion to the self, create an interesting dilemma. It is difficult to make a formal claim of a fall-off effect that would prove convincing to a skeptic, in as much as the *F* for the fall-off contrast is only $F(1, 81) = 438/314 = 1.39, p > .20$. The appropriate attitude for the investigator in these circumstances might very well be to adopt the private belief that the effect is real, but to refrain from pressing the claim publicly. The significance test essentially serves as a device giving a rhetorical advantage either to the investigator or to critics of the research claim, depending on whether the null hypothesis is rejected or accepted (Abelson, 1995, 1996). Here, the investigator would be in an extremely weak argumentative position. Not only is the raw *p* value greater than .20, but extra conservatism (e.g., in the form of Bonferroni adjustment; see Harris, 1994, pp. 98 ff.) is usually recommended when more than one significance test has been carried out, and even more conservatism (e.g., the Scheffé procedure; see Harris, 1994, pp. 112 ff.) is recommended for a posteriori tests. Therefore, no journal editor is going to allow a confident claim of a fall-off effect to go unchallenged, even if he or she is newly inclined to play down significance tests and to play up effect sizes (Abelson, in press; Hunter, 1997; Schmidt, 1996).

The remaining 2 *df* for the residual interaction (which represent asymmetries between cells above and below the diagonal) account for an *SS* of only 128. It is extremely doubtful that anything else is going on in this 3 × 3 array.

*Is orthogonality really necessary?* The strange appearance of contrast *V*, supposedly designed to capture a tendency for the scores to fall off toward the corners, is unappealing. It doesn't look right. A more direct contrast strategy may seem preferable. Why not write a single contrast that mimics the whole pattern—matching plus fall-off—and test that? This direct strategy, indeed, is what Rosnow and Rosenthal (1995), among others, recommend in general for cases possibly combining two or more effects. If we were not concerned with confounding by main effects, we would pick the weight matrix *T* in Equation 10 (and subtract 1/9 from each entry so that the sum of all the weights is zero; this is a must for a contrast, not an option). This array (call it $T^*$) contains both matching and fall-off effects in a straightforward representation. It yields an $SS_{\text{con}}$ of 4,581, higher than the total interaction of 3,592! The reason, of course, is that portions of the main effects are being picked up, and the analysis threatens to become incoherent. If the $SS_{\text{con}}$ of 4,581 had been the lion's share of the total sum of squares 18,166, then matching + fall-off combined in a single contrast would neatly capture the whole story. However, because this combined pattern contrast only explains about 25% of the total sum of squares, we are left without a principled way to characterize the other 75%, which comprise not only the main effects but also a bit of the residual interaction. Moreover, even if there were no main effects, a strong matching effect with no fall-off would substantially satisfy the contrast $T^*$. Whenever predicted patterns partially overlap and therefore can alias for one another, it will be difficult to interpret the results of contrast tests that do not preserve orthogonality.

*A cautionary note about power.* With a fixed number of means, degrees of freedom for error, and true ratio of the relevant systematic variation to the chance variation (which is unknown but can be estimated), the power of a contrast test varies directly with the degree that the pattern of contrast weights mimics the pattern of true group means. The appropriate measure of the goodness of mimicry is the squared correlation ($\rho^2$) between the contrast weights and the true means. Perfect mimicry would yield $\rho^2 = 1$, and the power of the test would be maximal for the given set of parameters.

In the present case, the matching contrast tested for a pattern of interaction scores with all diagonal entries positive and equal, and all off-diagonal entries negative and equal. The *I* matrix of our example manifested a global pattern approximating this quantitative ideal. One might wonder, however, about the consequences of a partial fit. Imagine an investigator who wanted to predict that even if all three diagonals did not yield matching, at least two of them would—but who was unable to say which two. For the data arriving from such a case, would it be a good idea to plan to test the contrast given by the *W* matrix (Equation 8) with three positive diagonal entries?

The answer is no, not a very good idea. Suppose that the first two rows and columns yielded matching, but the third row and column did not. An idealized *I* matrix of this sort might have $I(1, 1) = I(2, 2) = 10.0$; $I(1, 2) = I(2, 1) = -10.0$; and all elements in the third row and column of *I* equal to zero. For simplicity, let us set aside consideration of error variance and take these idealized values to have been calculated from the data of a 3 × 3 factorial with $n = 10$, as in the Rabbie (1964) experiment. The omnibus interaction sum of squares would be ($n$) times the sum of squared *I* entries, or 4,000. The sum of squares for the matching contrast W would be calculated from Equations 3 and 4, and come out to be 2,000—only half of the overall interaction.

The matching contrast is thus quite mediocre at picking up partial matching. In fact, it can be shown (Abelson, 1962) that with $\rho^2 = .50$ for the test of a contrast in a 3 × 3 design, the power of the contrast test would be less than the power of the omnibus test.

General coverage of issues of degree of correspondence between contrast patterns and underlying true patterns is beyond the scope of this article; the moral of the tale is as follows: When predictions are too loosely qualitative, allowing a broad range of possible quantitative confirmations, contrast tests (whether on main effects or on interactions) cannot be guaranteed to maintain a power advantage over the omnibus test.[5]

[5] It can be shown (Abelson, 1962) that for a contrast test to have as much power as a 4 *df* omnibus test, $\rho^2$ should reach roughly .65. The general problem of designing contrasts to avoid low values of $\rho^2$ was analyzed by Abelson and Tukey (1963).

Nor do they have much articulatory advantage in these cases, because there is a disparity between the apparent precision of the contrast (e.g., the weights in Equation 8), and the vague cloud of confirmatory outcomes. In practice, this means that experimental hypotheses expressed as contrasts should be serious, defensible commitments to particular qualitative and quantitative constraints, with low tolerance for accepting partial pattern fits as confirmatory. The theoretical reasoning behind the contrast should be transparent.

### *The Hypothesis of Opposed Concavities of Profile*

Sometimes two or more levels of the row factor are expected (from past research or simple argument) to show monotonically increasing profiles over the column factor, and the point of real interest lies in the hypothesis that some designated row level or levels will have profiles that are concave upward, and others, concave downward. Such predictions, although qualitative—that is, lacking exact quantitative definition—are nevertheless detailed enough to place a good deal of constraint on the set of outcomes consistent with prediction. We call the contrast testing this prediction the *qualitative quadratic.*

The method of interaction contrast construction for this situation is most simple in the 2 × 3 case. (The 2 × 2 case is trivial, because there is only 1 *df* for interaction and therefore, only one interaction contrast exists—the usual "differential difference.") To take a specific example—modeled very loosely after a study by Pilkington and Lydon (1997)—consider two groups of White participants who differ in some background classification, say, whether the participant's first-grade class did or did not include at least one child from a racial minority. Each group is divided into three random subsamples that are run in a different experimental condition. The three conditions might represent different relationships between the attitudes of a minority group member whom the participant expects to meet and the participant's own attitudes. At Level 1, these attitudes are similar (SIM); at Level 3, they are dissimilar (DIS); at Level 2, the attitudes of the other person are unknown (UNK). The dependent variable is a rating of how much the participant anticipates liking the other person when they meet.

A large mean difference can be predicted between the SIM and DIS conditions. This prediction is not very interesting, however, because a strong, highly replicable similarity–attraction effect has been well established by a great deal of prior research (see Byrne, 1971). The main effect of the group variable—childhood exposure to diversity—might be of more interest, but if that were all the investigators wanted to know, they would not have gone to all the trouble of manipulating the attitude similarity variable. The point of the study revolves around the interaction of the early experience with the attitude similarity factor, and, in particular, the effect of childhood experience for participants in the UNK condition. Denote the means of the diversity group for the three conditions as A, B, and C, and for the homogeneity group as a, b, and c. The hypothesized qualitative pattern might well be A > B >> C for the first group, but a >> b > c for the second. (The double inequality sign indicates larger gaps between means than the single inequality sign.) Expressed verbally, the idea is that participants who had had some degree of childhood exposure to racial diversity would not jump to the conclusion that the attitudes of a minority person about whom they had no information would be different from their own. In other words, the unknown other would be given the benefit of the doubt, and thus be rated almost as high as the SIM other. The participants with no childhood exposure to diversity, on the other hand, might be disposed, in the absence of information, to imagine the worst about the minority other, and thus to rate the UNK other almost as low as the DIS other. In sum, the prediction is that the UNK condition behaves more like the SIM condition for one group (B near A), but more like the DIS condition for the other (b near c).

*Weights for opposed concavities.* Let us develop the 2 × 3 matrix of interaction contrast weights row by row, starting with the predicted relationship between the three means for the diversity group. The hypothesis is SIM > UNK >> DIS, that is, the means drop off from left to right, more sharply between UNK and DIS than between SIM and UNK. Because contrasts should mimic the expected pattern, the goal is to write a prototypic expected profile of the three means. Remember that contrast weights should sum to zero, and that what matters is the relative rather than absolute spacing. Therefore, a reasonable set of weights might be {2 1 −3}. The drop-off is 1 between the first two values, and 4 between the last two, satisfying the ordinal hypothesis on the relative spacing. Of course there are other possible choices that satisfy the ordinal hypothesis equally well, for example, {5 2 −7}. But these variations are inconsequential.

The prediction for the homogeneity group's profile is concave upward instead of downward—that is, a bigger drop-off between the first and second means. A reasonable triplet of weights is a negative reversal of the weights for the first row: {3 −1 −2}.

Assembling the 2 × 3 array yields:

$$W(\text{proposed}) = \begin{matrix} 2 & 1 & -3 \\ 3 & -1 & -2 \end{matrix} \tag{16}$$

This, however, is not an interaction contrast, because the matrix of weights is not doubly centered. In particular, the columns do not sum to zero; this set of weights is highly sensitive to the familiar mean difference between the first and third columns (SIM vs. DIS), known from past research to be a big effect. Again, as we noted earlier, the confounding of main effects and interactions in the same contrast makes the explanation of the test outcome ambiguous. (For an exchange of differing views on this point, see Abelson, 1996; Petty, Fabringer, Wegener, & Priester, 1996; Rosnow & Rosenthal, 1995, 1996.)

The above array of weights can be converted into a set of contrast weights with the same procedure used to convert means into interaction scores. Using Equation 1 or Equation 2 to operate on $W$s instead of $X$s, the main effects contaminating the contrast are removed revealing the pure interaction pattern for the differential concavity.

The result of this operation is:

$$\begin{matrix} -1/2 & 1 & -1/2 \\ 1/2 & -1 & 1/2 \end{matrix}$$

When doubled to yeild all integers, this gives:

$$W = \begin{matrix} -1 & 2 & -1 \\ 1 & -2 & 1 \end{matrix} \tag{17}$$

Each row of three entries is a "quadratic" contrast,[6] with the upper row concave down, the lower row concave up. Remarkably, the same result is obtained after double-centering an initial array with any other numerical choice for the predicted unequal spacings (e.g., {5 2 −7}), paired with its negative reversal. Thus, even though the constraints on the weights are qualitative, the appropriate interaction contrast is numerically exact.

*Applying the weights to the example.* Table 3 shows the means for the six conditions of the Pilkington and Lydon (1997) study. Row and column means are indicated in the margins. The column means drop off from SIM to UNK to DIS, and the row means drop from diversity group to homogeneity group, consistent with prior expectation. The profile of means in the first row is concave upward, and in the second row concave downward, as specified by the experimental hypothesis. Table 4 gives the results of the ANOVA.

Table 3
*Outgroup Member Attractiveness by Attitude Similarity and Participant's Background Experience*

| | Outgroup member's attitudes | | | |
|---|---|---|---|---|
| Background | Similar | Unknown | Dissimilar | *M* |
| Diversity | 5.64 | 5.25 | 4.52 | 5.14 |
| Homogeneity | 4.73 | 3.80 | 4.03 | 4.19 |
| *M* | 5.18 | 4.52 | 4.28 | 4.66 |

*Note.* Entries are means on a 7-point rating scale. Cell $n = 25$.

There are three significant effects. The two main effects have large $F$s with $p$ values less than .001. (The effect size for the background variable is $d = 1.51$; for similarity, if we imagine the 2 $df$ split into two contrasts with equal effect sizes, that value would be $d = .89$.) The interaction, the focus of interest, yields a canonical outcome in support of the hypothesis: The contrast is significant at $p < .05$, and the residual is nowhere near significant ($F < 1.00$). The effect size for the opposing concavity of profiles is only modest ($d = .33$). Note, however, that had the omnibus interaction test been used instead of the contrast test, no significant interaction effect would have been claimed.[7]

[6] If the reader thinks this contrast is obvious, let it be said that graduate students given this example almost invariably create instructively incorrect 2 × 3 contrast matrices. They mimic the pattern too literally, proposing weights like the uncentered array (Equation 16). If these weights are used as contrast, the sum of squares comes 25.08, much bigger than the sum of squares for the whole interaction. What is wrong is that the proposed weights are centered by rows but not by columns, and therefore the contrast is picking up the main effect of columns. The +2 and +3 in the first column and the −3 and −2 in the third column guarantee that the large main effect of similarity will contribute to the proposed contrast. This would happen whether or not there were any interaction; thus this contrast could come out strong because of an obvious main effect rather than the predicted interaction.

[7] As with the last example, a sizable loss of contrast power would occur if only one of the two row profiles had the predicted concavity, and the other had no concavity. However, the precision of the contrast test is actually desirable here, as the spirit of the hypothesis would be violated if only half of it was obtained.

Table 4
*Analysis of Variance of the Similarity–Attraction Data*

| Source | *SS* | *df* | *MS* | *F* | *p* |
|---|---|---|---|---|---|
| Background (B) | 33.84 | 1 | 33.84 | 28.68 | <.001 |
| Similarity (S) | 23.50 | 2 | 11.75 | 9.96 | <.001 |
| B × S | 5.42 | 2 | 2.71 | 2.30 | *ns* |
| Qualitative quadratic | 4.69 | 1 | 4.69 | 3.97 | <.05 |
| Residual | 0.73 | 1 | 0.73 | 0.62 | *ns* |
| Error (within) | 170.30 | 144 | 1.18 | | |

Table 5
*Mean Running Speeds of Rats for Reward*

| | Trial | | | | | |
|---|---|---|---|---|---|---|
| Reward size | 1 | 2 | 3 | 4 | 5 | 6 |
| Large reward | .163 | .306 | .387 | .453 | .475 | .491 |
| Small reward | .156 | .248 | .328 | .369 | .369 | .416 |
| *M* | .160 | .277 | .358 | .411 | .422 | .453 |

*Note.* Running speeds were calculated by taking the reciprocal of the time (in seconds) it took each rat to run the maze.

## *Differential Trends Hypothesis*

Suppose two groups of participants are assessed on a number of trials (or at a number of levels of a quantitative treatment factor). Denote the groups factor by *G*, and the trials (or treatment) factor by *T*. The two trends go in the same direction (signifying learning, development, forgetting, or fatigue) but are hypothesized to differ quantitatively in their rates of change. That is, the shape of the curve is the same for all groups, but they differ in their amplitude; each group's curve is some multiple of a common trend (which is unknown but can be estimated). Most statistics texts restrict their coverage of trend differences to interactions between polynomials (see, e.g., Myers & Well, 1995, pp. 218–225; Winer et al., 1991, pp. 348–351), or contrasts of contrasts (Harris, 1994). The standard textbook advice is to use polynomial trend components (linear, quadratic, cubic) in an exploratory fashion when the column factor is continuous. We view this approach as too wooden. Hypothesis-oriented procedures for developing interaction contrasts are preferable to mechanical search methods.

In the simple two-group case, the method we advocate is to estimate a set of weights for the general trend and apply these weights separately to the profiles of means over *T* of each group, yielding two simple contrasts. The difference between the two simple contrasts gives the interaction contrast. This method is familiar for differences in linear trend (see Boik, 1993). We first review this simple case, and then will outline the method in the much more vexing circumstance of two nonlinear trends, each of which decelerates over trials (as in prototypical learning curves).

*Differential linear trends.* The procedure used to generate the appropriate interaction contrast weights for a differential linear trend is to calculate simple linear column contrasts for Groups 1 and 2, respectively, and subtract the latter from the former. Thus, for a 2 × 6 design

$$C_1 = -5X_{11} - 3X_{12} - X_{13} + X_{14} + 3X_{15} + 5X_{16}; \tag{18}$$

$$C_2 = -5X_{21} - 3X_{22} - X_{23} + X_{24} + 3X_{25} + 5X_{26}; \tag{19}$$

$$C = C_1 - C_2. \tag{20}$$

Removing the $X_{jk}$s from the 2 × 6 array of coefficients, the weights $w_{jk}$ used to calculate *C* are seen to be:

$$W = \begin{array}{cccccc} -5 & -3 & -1 & +1 & +3 & +5 \\ +5 & +3 & +1 & -1 & -3 & -5. \end{array} \tag{21}$$

This set of contrast weights is sensitive to differential rates of linear increase (or decrease) over trials for the two groups. It is doubly centered, as required of interaction contrasts.

As a transparent example, consider an experiment on the effects of size of food reward on instrumental learning in rats.[8] The obvious hypothesis is that learning should occur more rapidly the larger the reward.

Each of 30 rats ran down an alley for several blocks of trials. For 15 of the rats (Group 1), the food reward at the end of the alley was large, and for the other 15 (Group 2), the reward was small. The average running speeds on the first six trial blocks for rats receiving large and small rewards are shown in Table 5.

A standard ANOVA, shown in Table 6, yields significant main effects of reward, $F(5, 140) = 155.40$, $p < .01$, and of trial, $F(5, 140) = 155.40$, $p < .001$, plus a significant Reward × Trial interaction, $F(5, 40) = 3.60$, $p < .01$. This latter outcome is an omnibus indication that the pattern of performance over trials

[8] The data set for this example, which is a piece of a larger study carried out many years ago, was kindly supplied by Allan Wagner.

Table 6
*Conventional Analysis of Variance of the Running Speed Data*

| Source | *SS* | *df* | *MS* | *F* | *p* |
|---|---|---|---|---|---|
| Reward groups | .189 | 1 | .189 | 7.10 | <.01 |
| Rats within groups | .745 | 28 | .0266 | | |
| Trials | 1.8426 | 5 | .3685 | 155.40 | <.001 |
| Reward × Trials | .0430 | 5 | .0086 | 3.60 | <.01 |
| Reward × Linear Trials | .0284 | 1 | .0284 | 5.94 | <.05 |
| Residual | .0146 | 4 | .0037 | 2.06 | <.10 |
| Rats × Trials | .332 | 140 | .0024 | | |
| Rats × Linear Trials | .1322 | 28 | .0047 | | |
| Residual error term | .1998 | 112 | .0018 | | |

differs in some unspecified way for rats in the two reward conditions.[9]

A focused statement of the interaction hypothesis is that although running speed will increase over trials for both groups, it will increase more rapidly for rats in the large reward compared with the small reward condition. If this increase were assumed to be linear over trials, the appropriate weight matrix $w$ would be given by Equation 21.

Applying these weights to the raw data using Equation 5 yields an $SS_{\mathrm{con}}$ of .0284, and by subtraction from the overall Reward × Trials sum of squares, a residual interaction *SS* of .0146. In this repeated measures design, the error term for the Rewards × Trials interaction is a within-subjects error. When the Group × Trial interaction is partitioned into a differential trend contrast and a residual, the possibility arises of also partitioning the error term into a portion for the between-rats variability of trends within groups, and a portion for the residual rat variability over trials. (See Keppel & Zedeck, 1989, pp. 282–289 for computational details.)

When the contrast and residual tests are carried out, we find a significant $F$ for the contrast, $F(1, 28) = 5.94$, $p < .05$; $d = .88$, and a marginally significant $F$ for the residual, $F(4, 112) = 2.04$, $p < .10$. This outcome is encouraging but not absolutely ideal; it suggests that the differential linear pattern provides an approximate description of the data, but only approximate. The learning trends, as expected, are not linear, but curvilinear, and a more subtle method is necessary to do adequate justice to this nonlinearity.

*Differential curvilinear trends.* A very common type of nonlinear trend hypothesized in this type of experiment is a relatively smooth decelerating curve. Running speeds generally increase more rapidly across earlier trials than later trials, as is displayed visually in Figure 1 for the present data.

It would be desirable to incorporate this feature into a set of contrast weights. Suppose the general trend across trials was known to be some multiple of, say, the profile $\{0\ 6\ 10\ 13\ 15\ 16\}$. Here, the successive increases in the response measure are 6 between Trials 1 and 2, 4 between Trials 2 and 3, and then successively smaller increments of 3, 2, and 1. This profile pattern could be made into a set of contrast weights by centering them to add to zero, preserving the relative spacings. Subtracting the mean value of 10 from each yields the weights $\{-10 \quad -4 \quad 0 \quad 3 \quad 5 \quad 6\}$.

Because the hypothesis of this experiment is a differential trend across groups, the appropriate interaction contrast would be formed by using the above weights for the first group, and their negatives for the second group, as in the linear case. The differential trend contrast would then be

$$W = \begin{matrix} -10 & -4 & 0 & 3 & 5 & 6 \\ 10 & 4 & 0 & -3 & -5 & -6, \end{matrix} \qquad (22)$$

but a problem remains. We almost never, in fact, know beforehand the exact shape of the curvilinear profile. Curves can differ in their relative curvature at

[9] Caution is warranted in the interpretation of learning curves. A trial in the midst of a learning sequence both tests the effects of previous training, and gives a new training experience. By the last trial, what is measured is any effect of pretraining, plus the lag $(n - 1)$ effect of first trial training, the lag $(n - 2)$ effect of second trial training, and so on. Rescorla (1988) faulted such a measure as theoretically opaque. Further difficulty arises if there is a nonlinear relation between a theoretical variable, such as response strength, and its overt measurement. However, these objections do not spoil the descriptive use of trend tests; they affect the theoretical interpretation of the outcomes.

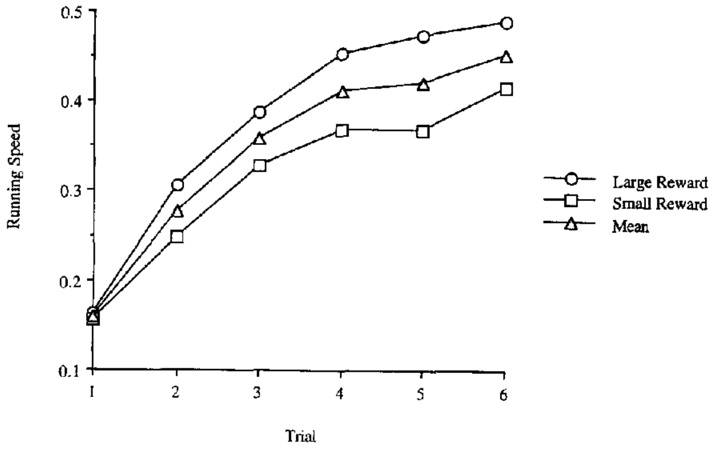

*Figure 1.* Differential trends in rat running speeds under conditions of low and high reward.

different points, and the above $W$ is mathematically arbitrary.

A simple insight saves the day.[10] Because it is impossible to know beforehand the form of the expected curvilinearity, why not use the data themselves to estimate it? Differential trend tests, like all interaction contrasts, have doubly centered weights and are orthogonal to all main effects. In particular, the differential trend test using as a profile of weights the main effect of trials in the actual data would be independent of that effect. Using observed means to form contrast weights for interaction tests is an idea that goes back to Tukey (1949).

It is a counterintuitive idea. Obviously a vector of weights derived from the observed trial means could not be used to test the main effect of trials; it would capture 100% of the variance by definition. But the interaction is not constrained to follow the same profile as the main effect. The pattern of one does not imply anything about the pattern of the other. The prediction of this experiment—that the same (main effect) profile will obtain in both conditions, amplified in one compared with the other—is a specific, a priori hypothesis about the form of the interaction.

The procedure for deriving contrast weights from the profile of trial means is a straightforward extension of the linear case. In the rat example, the mean running speeds, averaged across both reward conditions, (shown numerically at the bottom of Table 5, and graphically in Figure 1) are as follows: {.160 .277 .358 .411 .422 .453}. Centering yields a vector of contrast weights: {−.187 −.070 .011 .064 .075 .106}. We give the name *graded weights* to such a vector. The term *grade*, in the meaning familiar to truckers and trail hikers, denotes relative steepness of climb between adjacent locations. The profile is a succession of grades.

Reversing signs to obtain the second row of the 2 × 6 weight matrix yields the contrast testing differential graded trend:

$$\begin{aligned} W = &-.187 \; -.070 \; +.011 \; +.064 \; +.075 \; +.106 \\ &+.187 \; +.070 \; -.011 \; -0.64 \; -.075 \; -.106. \end{aligned} \tag{23}$$

As shown in Table 7, the application of this contrast to the data gives a highly significant contrast, $F(1, 28) = 8.45$, $p < .01$, $d = 1.06$, and a small and nonsignificant $F$ for the residual, $F(4, 28) = 1.03$, *ns*. This canonical outcome indicates that the experimenter's differential learning hypothesis provides a parsimonious account for the data. The effect size of 1.06 is an improvement over the value of 0.88 for the linear interaction, although it is not an overwhelming increase in effect size. However, the key fact is that the differential graded test leaves a residual of the same magnitude as the random error.

We call the interaction contrast using graded weights the *differential curvilinear trend test.* It is similar in intent to the differential linear trend test, but in coping flexibly with curvilinearity, its statistical power exceeds that of its linear cousin. It is a superior alternative for testing the hypothesis of differing trends in two groups, given two expectations: (a) successive increments in one group's profile of trial means are a constant multiple of the successive incre-

[10] We are indebted to Harry Gollob, who many years ago realized the applicability of this procedure.

Table 7
*Alternative Analysis of Variance of the Running Speed Data*

| Source | *SS* | *df* | *MS* | *F* | *p* |
|---|---|---|---|---|---|
| Reward groups | .189 | 1 | .189 | 7.10 | <.01 |
| Rats within groups | .745 | 28 | .0266 | | |
| Trials | 1.8426 | 5 | .3685 | 155.40 | <.001 |
| Reward × Trials | .0430 | 5 | .0086 | 3.60 | <.01 |
| Reward × Graded Trials | .0350 | 1 | .0350 | 8.45 | <.01 |
| Residual | .0080 | 4 | .0020 | 1.03 | ns |
| Rats × Trials | .332 | 140 | .0024 | | |
| Rats × Graded Trials | .1150 | 28 | .0041 | | |
| Residual error term | .2170 | 112 | .0019 | | |

ments in the other group's profile; and (b) the degree of error variance in the profiles of group means across trials is relatively small.

The first condition implies that the difference between the two groups increases or decreases as the average level of the means increases or decreases. In other words, the test for graded trend is appropriate when the magnitude of response (averaged over both groups) is expected to covary with the distance between the two groups. This relationship entails and is entailed by a multiplicative relation between the profiles of relative change for the two groups.

Observed means differ from true means, of course, so that even if condition a was to hold for errorless data, it might not for actual data. Because the trials main effect is used to construct the graded weights, those weights might show random fluctuations if the data contain a large error component. The learning curve example above suggests a mild influence of error. The graded weights were based on the mean speeds over trials: {.160 .277 .358 .411 .422 .453}. The successive *increases* in mean speeds are .177, .081, .054, .011, and .031; note an inversion in monotonicity by the last two figures. This uncharacteristic feature for a learning profile could be smoothed out with further refinements to the graded weight procedure, though the numerical consequences would be relatively tiny in the present example. When the degree of error is moderately large, the graded weights procedure loses its advantage over the straightforward (albeit rough) use of linear weights, because $\rho^2$ drops too much.

## Summary

The goal of this exposition has been to provide some guidelines for the creative, yet judicious, use of contrasts to test interaction predictions in the two-way ANOVA. The specific examples have been chosen to illustrate interaction patterns that are often hypothesized but rarely tested directly in psychological research. We have offered more conceptual advice than computational detail; the reader is advised to consult his or her favorite statistics text for additional guidance.

All of the procedures described in this article are well within the capabilities of most popular computer statistical packages. SAS (PROC GLM), for example, can handle all of the analyses reported in this article; SPSS, BMDP, and SYSTAT can also be programmed (with appropriate ingenuity on the part of the user) to perform contrasts on interaction effects. Moreover, the use of interaction contrasts should be facilitated by one final advantage of the technique—its computational simplicity. In the occasional case in which statistical packages give satisfactory error terms but do not calculate interaction contrasts themselves, the required calculations can be carried out on the back of an envelope.

## References

Abelson, R. P. (1962). *A priori* contrasts in the analysis of variance. Unpublished manuscript, Yale University, New Haven.

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Erlbaum.

Abelson, R. P. (1996). Vulnerability of contrasts to simpler interpretations: An addendum to Rosnow and Rosenthal. *Psychological Science, 7,* 242–246.

Abelson, R. P. (in press). A retrospective on the significance test ban of 1999. In L. Winslow & S. A. Mulaik (Eds.), *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Abelson, R. P., & Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Annals of Mathematical Statistics, 34,* 1347–1369.

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

Anderson, N. H. (1982). *Methods of information integration theory.* New York: Academic Press.

Boik, R. J. (1993). The analysis of two-factor interactions in fixed effects linear models. *Journal of Educational Statistics, 18,* 1–40.

Byrne, D. (1971). *The attraction paradigm.* New York: Academic Press.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1,* 98–101.

Emerson, J. D. (1991). Graphical display as an aid to analysis. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Fundamentals of exploratory analysis of variance.* New York: Wiley.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140.

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

Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8,* 3–7.

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

Judd, C. M., McClelland, G. H., & Culhane, S. C. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. In J. T. Spence, J. M. Darley, & D. J. Foss (Eds.), *Annual review of psychology* (Vol. 46, pp. 433–465). Palo Alto, CA: Annual Reviews.

Keppel, G., & Zedeck, S. (1989). *Data analysis for research designs.* New York: W.H. Freeman.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Madansky, A. (1988). *Prescriptions for working statisticians.* New York: Springer-Verlag.

Myers, J. L., & Well, A. D. (1995). *Research design and statistical analysis.* Hillsdale, NJ: Erlbaum.

Petty, R. E., Fabringer, L. R., Wegener, D. T., & Priester, J. R. (1996). Understanding data when interactions are present or hypothesized. *Psychological Science, 7,* 247–252.

Pilkington, N. W., & Lydon, J. E. (1997). The relative effect of attitude similarity and attitude dissimilarity on interpersonal attraction: Investigating the moderating roles of prejudice and group membership. *Personality and Social Psychology Bulletin, 23,* 107–122.

Rabbie, J. (1964). Differential preference for companionship under threat. *Journal of Abnormal and Social Psychology, 67,* 643–648.

Rescorla, R. (1988). Behavioral studies of Pavlovian conditioning. *Annual Review of Neuroscience, 11,* 329–352.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: Sage.

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

Rosnow, R. L., & Rosenthal, R. (1995). "Some things you learn aren't so": Cohen's paradox, Asch's paradigm, and the interpretation of interaction. *Psychological Science, 6,* 3–9.

Rosnow, R. L., & Rosenthal, R. (1996). Contrasts and interactions redux: Five easy pieces. *Psychological Science, 7,* 253–257.

Schmid, C. H., (1991). Value splitting: Taking the data apart. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Fundamentals of exploratory analysis of variance* (pp. 72–113). New York: Wiley.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for the training of researchers. *Psychological Methods, 1,* 115–129.

Tukey, J. W. (1949). One degree of freedom for nonadditivity. *Biometrics, 5,* 232–242.

Tukey, J. W. (1977). *Exploratory data analysis.* Reading, MA: Addison Wesley.

West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality, 64,* 1–48.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

Received May 8, 1996
Revision received December 20, 1996
Accepted December 26, 1996 ■