Definition and Interpretation of Interaction Effects

Ralph L. Rosnow
Temple University

Robert Rosenthal
Harvard University

When interaction is claimed in a factorial arrangement, the results almost always require more detailed analysis than is typically reported in our primary journals. In reporting interactions, research psychologists have gotten into the habit of examining only the differences between the original cell means (the simple effects) instead of more properly examining the residuals, or leftover effects, after the lower order effects have been removed. The logic of decomposing the original cell or condition means into the main effects and the effects of the interaction is described and illustrated by an algorithm for 2 × 2 interactions.

The mathematical meaning of interaction effects is unambiguous, and textbooks of mathematical and psychological statistics routinely include proper definitions of interaction effects. Despite this, a substantial number of research psychologists reporting results in our primary journals interpret interactions incorrectly. This was revealed by a systematic review of 320 articles published in 1985 in the first and last issues of the following American Psychological Association journals: *Developmental Psychology* (Vol. 21, Nos. 1 & 6), *Journal of Abnormal Psychology* (Vol. 94, Nos. 1 & 4), *Journal of Comparative Psychology* (Vol. 99, Nos. 1 & 4), *Journal of Consulting and Clinical Psychology* (Vol. 53, Nos. 1 & 6), *Journal of Counseling Psychology* (Vol. 32, Nos. 1 & 4), *Journal of Educational Psychology* (Vol. 77, Nos. 1 & 6), *Journal of Experimental Psychology: General* (Vol. 114, Nos. 1 & 4), *Journal of Experimental Psychology: Learning, Memory, and Cognition* (Vol. 11, Nos. 1 & 4), and *Journal of Personality and Social Psychology* (Vol. 48, No. 1; Vol. 49, No. 6).

There were 191 research articles in which reports of analysis of variance (ANOVA) results alluded to some factorial arrangement. Of these, 34% reported nonsignificance or reported significance but did not proceed further because the observed interaction was stated to be irrelevant. Another 28% interpreted observed interactions but did not specify whether the interpretation was based on inspecting the residuals or on simple effects comparisons; in only 1% of these cases did the authors clearly specify that their interpretation was based on evaluating the residuals for the interaction that was present. In the remaining 37% of cases, there were clear-cut indications that demonstrated confusion in thinking about interactions; that is, observed interactions were followed only by simple effects comparisons (e.g., contrasts computed on original cell or condition means), or they were followed by referring the reader to a table of original cell or condition means. None of the journals surveyed was immune from this kind of error in reporting the pattern of an observed interaction.

Thus, the nature of the error is quite consistent. Once investigators find significant interactions, they attempt to interpret them by examining differences between original cell (or condition) means. We do not, of course, claim that there is anything wrong about comparing means, and elsewhere, in fact, we have illustrated the advantages of employing contrasts when a comparison of means is the objective (Rosenthal & Rosnow, 1985; Rosnow & Rosenthal, 1988). The problem we discuss in this article, however, is that these means or *simple effects* are made up only partially of interaction effects; main effects may contribute to simple effects even more than interactions do. In order to correct what seems to be a widespread misunderstanding surrounding the concept of interaction, we will first illustrate the nature of the problem with a specific example and will then describe the logic of decomposing cell (condition) means into interaction effects and lower order effects by an algorithm for 2 × 2 interactions.[1]

Nature of the Problem

An investigator is interested in studying the despair of bereavement of family members when a child dies. He predicts (with grief intensity as the dependent variable) that health of child and sex of child interact, but instead of stating the form and degree of relationship of the expected interaction, he specifies only the rank ordering of the "uncorrected" cell means: healthy male child > healthy female child > unhealthy female

We thank Nanette Nelson for gathering the information concerning the reporting of research results in APA journals. Preparation of this article was supported in part by the National Science Foundation and by the Bolton Fund of Temple University.

Correspondence concerning this article should be addressed to Ralph L. Rosnow, Department of Psychology, Temple University, Philadelphia, Pennsylvania 19122.

[1] The situation may be more complicated in the case of unbalanced designs (unequal *N*s per cell). However, when the unweighted means approach is employed, all of our discussion applies (Armor & Couch, 1972; Rosenthal & Rosnow, 1984; Walker & Lev, 1953; Winer, 1971).

Psychological Bulletin, 1989, Vol. 105, No. 1, 143–146

Table 1
Predicted and Obtained Effects

Health status	Overall effects: Girl	Overall effects: Boy	Row effects	Residual effects: Girl	Residual effects: Boy
		Predicted[a]			
Healthy	+1	+3	+2	−1	+1
Unhealthy	−1	−3	−2	+1	−1
Column effects	0	0			
		Obtained[b]			
Healthy	−1	+3	+1	−1	+1
Unhealthy	−1	−1	−1	+1	−1
Column effects	−1	+1			

[a] Analysis of variance (ANOVA): $SS_{\text{total}} = 20$; $SS_{\text{health}} = 16$; $SS_{\text{sex}} = 0$; $SS_{\text{h×s}} = 4$.
[b] ANOVA: $SS_{\text{total}} = 12$; $SS_{\text{health}} = 4$; $SS_{\text{sex}} = 4$; $SS_{\text{h×s}} = 4$.

child > unhealthy male child. In the top half of Table 1, the positive and negative values of the four conditions under Overall effects indicate the overall or simple effects predicted, keeping to single-digit integers for simplicity. The row and column effects denote the size of the two main effects implicit in the investigator's designated rank ordering. Row effects are defined for each row as the mean of that row minus the grand mean; column effects are defined for each column as the mean of that column minus the grand mean. The values under Residual effects show the twice-corrected cell means; these are the interaction effects implicit in the designated rank ordering of overall effects. For example, to calculate the residual effect corresponding to the first column, first row of overall effects, we subtracted +2 (row effect) and 0 (column effect) from +1 (overall effect), which gives us −1 (residual effect). Were the residual effects all equal to the grand mean (0), that would tell us there was no interaction implicit in the designated rank ordering of overall effects. The sum of squares values underscore the assumption implicit in the investigator's designated rank ordering that the variance due to the main effect of health is expected to be four times greater than the variance due to the interaction effect. What the investigator calls an interaction should be called an interaction plus a (dominating) main effect.

In the bottom half of Table 1, we see the overall effects and ANOVA as published by the investigator. In his report the investigator accurately states that there was a significant interaction between sex of child and health of child. The pattern of interaction was not as expected, he reports, but was instead healthy male child > healthy female child = unhealthy female child = unhealthy male child. The positive and negative values of the overall or simple effects indicate the reported pattern of results; the twice-corrected means or residual effects are given in the last two columns. We see that the row effects, column effects, and residual effects have identical values (+1, −1), indicating that what was called an interaction should more properly be called an interaction plus equally large main effects. Only ⅓ of the variance attributed to the interaction effect was actually due to the interaction. Also, seemingly unbeknown to the investigator, we see that he found exactly the pattern of interaction that was predicted.

In summary, the interaction effect is defined basically in terms of the cell means corrected for the main effects. This is true even though the mean square for interaction can be viewed as variability of the differences between the (uncorrected) cell means for the various rows of the table of overall effects; the mean square for interaction will have a nonzero value if the difference between any two cell means in any row differs from the corresponding difference in any other row. Suppose we have a 3 × 2 table of cell means in which the rows are three grade levels and the columns are control versus experimental treatment. Cell values for rows 1, 2, and 3 are 7, 8, and 9, respectively, in column 1, and they are 9, 10, and 11 in column 2. Because observed differences between the control and experimental are identical in all three rows, we need go no further to conclude that the mean square for interaction is zero. Even if only two such differences differed from one another, there would be a nonzero mean square for interaction. If in column 2, for example, the cell values for rows 1 and 2 had been 10 and 9, respectively, the mean square for interaction would have a value other than zero, because the observed difference in row 1 now differs from that in row 2.

Nonetheless, in focusing attention only on the original cell means (as in this and the preceding example), one is essentially ignoring the form and degree of relationship of the interaction itself. Again, the reason is that the original cell means are the combined effects of the interaction, the row effects, the column effects, and the grand mean. It is a mistake to think that the uncorrected cell or condition means automatically reveal the conditional relationships of the residual effects. Like peeling away the skins of an onion, we need to peel away the lower order effects in order to separate the effects of the interaction from the main effects. Table 1 telescopes this procedure, and we now describe in more detail the logic of decomposing cell or condition means by an algorithm for 2 × 2 interactions generally.

Decomposing Cell or Condition Means

In the 2^2 approach the investigator compares the four treatments that can be formed by combining two levels each of two factors or substances. For example, the owners of a baseball team want to evaluate a method called "Ralphing" for improving a player's ability to withstand the pressures of competition. They also want to evaluate two different levels of experience at

Table 2
2^2 Factorial

Factor	b_0	b_1	M
a_0	$\overline{AB}_{00}$	$\overline{AB}_{01}$	$\bar{A}_0$
a_1	$\overline{AB}_{10}$	$\overline{AB}_{11}$	$\bar{A}_1$
M	$\bar{B}_0$	$\bar{B}_1$	$\bar{G}$

Table 3
ANOVA Summary

Source	SS	df	MS	F	p	eta
Method (A)	81	1	81	43.2	<.001	.758
Experience (B)	9	1	9	4.8	.036	.361
$A \times B$	9	1	9	4.8	.036	.361
Error	60	32	1.875			

which Ralphing might be effective for this team, namely, whether a player has never previously participated (inexperienced player) or has previously participated in championship competition (experienced player). These two factors, method and experience, are to be evaluated simultaneously using the 2^2 design shown in Table 2, where $\overline{AB}_{ij}$ symbolizes the dependent variable means for n elements in condition ab_{ij}, $\bar{A}_i$ and $\bar{B}_j$ are the average observations at levels a_i and b_j, respectively, and $\bar{G}$ is the mean of all the means. Inexperienced and experienced ball players will be randomly assigned to a zero control group or to Ralphing, and the performance of each player will be measured by the total number of "hits" in an experimentally devised game in which players are kept independent of each other for purposes of analysis.

In reporting the results, the investigator uses F to test the hypothesis of zero effects and to reject at the 5% level. The performance scores (raw scores) of 18 inexperienced and 18 experienced ball players who were assigned in equal numbers to the control group or to Ralphing (n = 9) are given as follows: ab_{00} = 1, 1, 1, 3, 3, 3, 5, 5, 5; ab_{01} = 1, 1, 1, 3, 3, 3, 5, 5, 5; ab_{10} = 4, 4, 4, 5, 5, 5, 6, 6, 6; ab_{11} = 6, 6, 6, 7, 7, 7, 8, 8, 8. Table 3 shows the ANOVA (computed in the usual manner) and Table 4 shows a table of means. Both main effects and interaction are significant beyond the 5% critical level (cf. Nelson, Rosenthal, & Rosnow, 1986) and are of moderate to substantial effect size as conventionally defined (cf. Cohen, 1977). To evaluate method (Factor A), the investigator compares the average performance of players assigned to the zero control ($\bar{A}_0$) with the average performance of players assigned to Ralphing ($\bar{A}_1$). To evaluate experience (Factor B), he compares the average performance of inexperienced ball players ($\bar{B}_0$) with that of experienced ball players ($\bar{B}_1$). At this point, he would be making a logical mistake were he to identify the rank ordering of the cell means ($\overline{AB}_{11} > \overline{AB}_{10} > \overline{AB}_{00} = \overline{AB}_{01}$) as the interaction that is present. Referring back to Table 2, we can say that the error would be in mistaking the values of $\overline{AB}_{ij}$ for the residual effects remaining once the row and column effects (the effects of $\bar{A}_i$ and $\bar{B}_j$) were peeled away from $\overline{AB}_{ij}$. It is not absolutely necessary to remove $\bar{G}$ from $\overline{AB}_{ij}$ in order to interpret an interaction properly, but (as is now shown) doing so further focuses the interaction by freeing it of all encumbrances.

Table 4
Table of Means

Method	b_0	b_1	M
a_0	3	3	3
a_1	5	7	6
M	4	5	4.5

The F test of the AB interaction (Table 3) directs the investigator's attention to calculating the residual effects, because factorial ANOVA routines of many computer packages fail to provide the residual variability in the cell means after corrections have been applied. Table 5 provides an algorithm for obtaining two-way interaction effects, in which the independent effects of the rows, columns, and grand mean are also given for more detailed study. Row 1 shows the means of the four conditions; comparisons among group (or condition) means are called tests of simple effects when both means are at the same level of Factor A or Factor B. Thus, an analysis of the comparison $\overline{AB}_{11} - \overline{AB}_{01}$ would estimate the effect of Factor A if Factor B were held constant at level 1; the comparison $\overline{AB}_{10} - \overline{AB}_{00}$ would estimate the effect of Factor A if Factor B were held constant at level 0. Row 2 shows the row effects, which are defined for each of the four treatment conditions as the mean of the relevant row minus the grand mean. Row 3 shows the column effects, which are defined for each of the four treatment conditions as the mean of the relevant column minus the grand mean. Row 4 shows the grand mean. Finally, row 5 shows how to obtain the two-way interaction effect for each cell, that is, the residual remaining

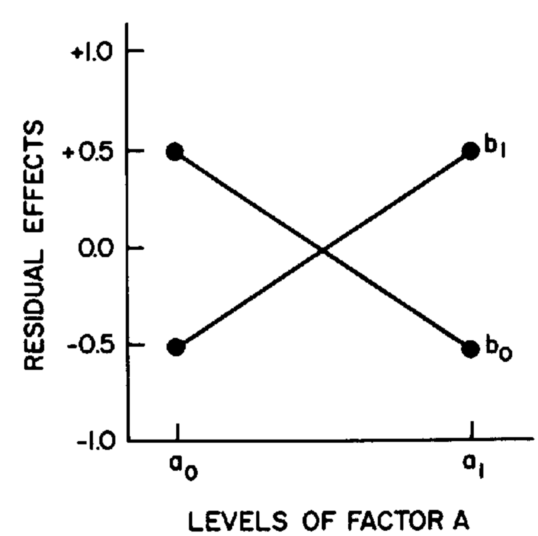

Figure 1. Graphical representation of two-factor interaction for 2 × 2 factorial.

Table 5
Obtaining Interaction Effects in the 2^2 Factorial

	a_0		a_1	
Effect	b_0	b_1	b_0	b_1
(1) Group mean	$\overline{AB}_{00}$	$\overline{AB}_{01}$	$\overline{AB}_{10}$	$\overline{AB}_{11}$
(2) Row effect	$\bar{A}_0 - \bar{G}$	$\bar{A}_0 - \bar{G}$	$\bar{A}_1 - \bar{G}$	$\bar{A}_1 - \bar{G}$
(3) Column effect	$\bar{B}_0 - \bar{G}$	$\bar{B}_1 - \bar{G}$	$\bar{B}_0 - \bar{G}$	$\bar{B}_1 - \bar{G}$
(4) Grand mean	$\bar{G}$	$\bar{G}$	$\bar{G}$	$\bar{G}$
(5) Interaction	$1 - (2 + 3 + 4)$	$1 - (2 + 3 + 4)$	$1 - (2 + 3 + 4)$	$1 - (2 + 3 + 4)$

after subtracting rows 2, 3, and 4 from row 1. In other words, Interaction = Group Mean − (Row Effect + Column Effect + Grand Mean).

A similar procedure would be used to obtain higher order interactions in more complex arrangements. In general, a higher order interaction is defined basically as the residual set of effects after the main effects and all lower order interactions relevant to the higher order interaction have been removed. Were this a 2^3 factorial, the two-way interactions of A × B, A × C, and B × C would be the effects remaining after removal of the two main effects designated by the letters naming the interaction. The three-way interaction of A × B × C would be the set of residuals remaining after removal of the three main effects and the three two-way interactions. Similarly, in a 2^4 factorial the four-way interaction would be the set of residuals remaining after the four main effects, six two-way interactions, and four three-way interactions were subtracted from the total of all between-condition effects.

Continuing with our example, Table 6 shows the separation of the two-way interaction from the main effects of Factors *A* and *B*. Row 5 gives the residuals after subtracting the row effects (row 2), column effects (row 3), and grand mean (row 4) from the group means (row 1). Had both main effects been large, the small interaction shown might merely have implied a minor variation in one factor as the other factor was at its higher or lower level. In the present instance, inspecting the residuals may be of some interest, as Figure 1 shows that (in relation to the control) the experienced ball players benefited moderately from Ralphing to the same degree that the inexperienced ball players were harmed by it. Note that the interaction is X-shaped, which will be the case for any 2^2 factorial when interaction is present.

Table 6
Obtaining the A × B Interaction

	a_0		a_1	
Effect	b_0	b_1	b_0	b_1
Group mean	3	3	5	7
Row effect	−1.5	−1.5	+1.5	+1.5
Column effect	−0.5	+0.5	−0.5	+0.5
Grand mean	4.5	4.5	4.5	4.5
Interaction	+0.5	−0.5	−0.5	+0.5

In other 2^k and more complex factorials, other types of interaction are possible and would be described accordingly (cf. Rosenthal & Rosnow, 1984).

A Final Note

We have stressed the description and interpretation of interaction effects in this article. In a given study, however, the scientific goals may require the investigator to focus attention on a particular ordering of the cell means (contrasts) rather than on the row, column, and interaction effects. This discussion should not be viewed as an argument against comparing group means, because it often makes good sense to focus on a comparison of means using planned contrasts and to deemphasize the traditional row, column, and interaction effects when they are based only on omnibus (or diffuse) tests of statistical significance (cf. Rosenthal & Rosnow, 1985; Rosnow & Rosenthal, 1988). The point of this article is to emphasize that if investigators are claiming to speak of an interaction, the exercise of looking at the corrected cell (or condition) means is absolutely necessary.
will be the case for any 2^2 factorial when interaction is present. In other 2^k and more complex factorials, other types of interac-

References

Armor, D. J., & Couch, A. S. (1972). *Data-text primer: An introduction to computerized social data analysis.* New York: Free Press.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Nelson, N., Rosenthal, R., & Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 41,* 1299–1301.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis.* New York: McGraw-Hill.

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

Rosnow, R. L., & Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. *Journal of Counseling Psychology, 35,* 203–208.

Walker, H. M., & Lev, J. (1953). *Statistical inference.* New York: Holt.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Received July 3, 1986
Accepted February 5, 1988 ■