Technical Commentary

UNDERSTANDING DATA WHEN INTERACTIONS ARE PRESENT OR HYPOTHESIONED

Richard E. Petty, Leandre R. Fabrigar, Duane T. Wegener, and Joseph R. Priester

Ohio State University, Queen's University, Yale University, and University of Michigan

Abstract—The traditional approach to interpreting data when an interaction is present is to interpret cell means as reflecting a difference between differences. An alternative is to interpret any main effects and separately interpret interaction residuals. Interpretation of interaction residuals can sometimes lead to nonsensical conclusions. Thus, in deciding between the approaches, researchers should consider (a) the conceptual nature of the variables involved, (b) relevant theories, and (c) the additional data that might be necessary to distinguish among competing plausible representations. Issues have also arisen regarding how to test a hypothesis that involves interaction variance. Some researchers have argued that the use of a focused contrast designed to test a specific ordering of cell means (e.g., a linear contrast) is often the best strategy. We note potential problems with the use of such contrasts and discourage researchers from blanket use of contrasts that combine main effect and interaction variance.

Statistical interactions often play a crucial role in testing psychological theory. Yet recent articles have raised important issues regarding how researchers should interpret their data when a statistical interaction is present, and how they should best test an interaction hypothesis. In this article, we discuss conceptual issues regarding understanding data from factorial designs (especially the common 2 × 2) when an interaction is present or hypothesized. We focus on the 2 × 2 in part because of the ubiquity of this design, and because appropriate procedures for testing and interpreting the data when interactions are obtained in this design have been the subject of considerable recent attention and controversy (e.g., see Bobko, 1986; Meyer, 1991; Rosnow & Rosenthal, 1989a, 1989b, 1991, 1995; Ross & Creyer, 1993; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993).

UNDERSTANDING THE DATA WHEN AN INTERACTION IS PRESENT

In order to make psychological sense of data, one must have some agreed-upon method of representing the pattern of results obtained. This is particularly important when an interaction is present because there are multiple ways to represent or depict interactions (cf. Judd, McClelland, & Culhane, 1995). For example, Rosnow and Rosenthal (1989a, 1989b, 1991, 1995) have suggested that understanding the data is fostered by representing and interpreting the interaction as the residual remaining after main effects have been removed. We argue that a blanket use of this approach rather than the more traditional approach of representing and interpreting an interaction as a difference between differences in cell means (e.g., Keppel, 1991) can lead to misleading conclusions about the psychological processes underlying the results. It is important to note that representation of interactions is not a statistical issue per se. The two ways of representing interactions are mathematically equivalent (i.e., interaction variance is the same, whether calculated through residuals or differences between differences; see Guilford & Fruchter, 1978; Keppel, 1991; Rosnow & Rosenthal, 1989b). However, the different representations can suggest different (and conflicting) substantive conclusions.

Traditional Versus Residual Approach to Interactions

Following the finding of a significant interaction in an analysis of variance (ANOVA), researchers have traditionally described the interaction effect through comparisons of original cell means (e.g., through simple effects tests; Keppel, 1991). Take, for instance, a hypothetical example presented by Rosenthal and Rosnow (1985). In a 2 × 2 design, one of two versions of a political advertisement was shown to either liberals or conservatives. A traditional ANOVA indicates a significant main effect of advertisement, no main effect of political ideology, and a significant interaction between the two independent variables (see Fig. 1a). According to the difference-between-differences approach (e.g., see Keppel, 1991), one acceptable interpretation of these data is that there is a greater effect of the differing ad types for liberals than for conservatives. Keppel (1991) recommended analyses of the two simple effects (of ad type, in this example) as one method for examining the meaning of the interaction (pp. 236–245; see also Myers, 1979; Pedhazur, 1982). This approach to interpretation resonates with the views espoused by Fisher (1947, pp. 92–93) and other researchers (e.g., Hoaglin, Mosteller, & Tukey, 1991).

An alternative view has been espoused by proponents of the residual approach. Rosenthal, Rosnow, and their colleagues (e.g., Rosenthal & Rosnow, 1985, 1991; Zuckerman et al., 1993) have noted correctly that cell means reflect not only the influence of the interaction term in an ANOVA, but also the influence of all lower order effects. Therefore, according to Rosenthal and Rosnow, if one is interested in understanding the psychological meaning of an interaction in a 2 × 2 design, one

Address correspondence to Richard E. Petty, Department of Psychology, Ohio State University, 1885 Neil Avenue Mall, Columbus, OH 43210; e-mail: petty.1@osu.edu.

VOL. 7, NO. 4, JULY 1996

Copyright © 1996 American Psychological Society 247
Interactions

Fig. 1. Traditional representation of the difference between differences (a) and residual representation of the interaction (b) in a 2 (political ideology) x 2 (ad version) factorial.

should not do so by examining a representation that also reflects main effects. Rather, one must remove the influence of the main effects from the cell means prior to imparting meaning to the interaction.

Consider the Ad Version x Political Ideology data in Figure 1. According to Rosenthal and Rosnow (1985), a researcher following the traditional approach might be tempted to interpret the presence of an interaction in the data as indicating that liberals are more strongly influenced by version A than version B, whereas conservatives are equally influenced by both advertisements. Rosenthal and Rosnow (1985) asserted, however, that such a conclusion "would be wrong!" (p. 5). Instead, they prefer to interpret the representation of the interaction as depicted in Figure 1b. Consistent with this representation, Rosenthal and Rosnow stated that "the interaction actually shows that conservatives and liberals reacted in exactly opposite ways to the two types of propaganda" (p. 8; see Rosenthal & Rosnow, 1985, for calculation of residuals).

Obviously, this residual representation suggests something very different about the behavior of conservatives than does the representation following from the traditional approach. That is, a difference-between-differences approach encourages one to develop theory that explains why liberals are affected differently by the two ads but conservatives are not, whereas the residual approach encourages one to explain why liberals and conservatives are influenced in opposite ways by the two ads. The nature of the variables in this example does not preclude generating a meaningful psychological interpretation of the interaction residuals. Therefore, the representations encouraged by the traditional and the residual views are both potentially accurate.

Comparing the Approaches

Rosenthal and Rosnow (1991) asserted that "before an interaction effect can be understood, . . . the residuals defining the interaction must be displayed" (p. 367), and Zuckerman et al. (1993) stated that "interaction effects cannot be interpreted on the basis of comparisons between cell means" (p. 53). If these scholars meant to suggest only that researchers who describe cell means are not describing solely interaction variance, this is correct. However, if they meant that researchers should necessarily impart psychological meaning to interaction residuals (as suggested by their assertion about the "reactions" of liberals and conservatives depicted in Fig. 1), then we disagree. In particular, it is not clear that authors who wish to understand the data containing an interaction should invariably devote journal space to displaying and interpreting interaction residuals instead of interpreting cell means. In fact, interpretation of residuals can sometimes lead to nonsensical conclusions because the interaction per se is a statistical entity that might or might not directly correspond to any meaningful underlying psychological process.

Consider a pattern of data similar to that in Figure 1a, but with different variables. Specifically, imagine an experiment investigating the influence of exposure duration (subliminal [5 ms] vs. supraliminal [500 ms]) and actual letter height (⅛ in. vs. 9 in.) on subjects' verbal estimates of the height of letters. See Figure 2a for the expected cell means.

A 2 x 2 ANOVA shows a main effect of letter size and an Exposure Duration x Letter Size interaction. The difference-between-differences approach suggests that the interaction exists because there is a greater impact of letter size on perceptions of height when the letters are presented supraliminally than when the letters are presented subliminally. Advocates of the residual approach would presumably suggest that this characterization is wrong. That is, they would assert that the meaning of the interaction can be understood only by interpreting the residuals represented in Figure 2b. According to the residual approach, then, one would interpret the presence of the interaction as indicating that supraliminal- and subliminal-exposure subjects reacted to the presented letters in exactly opposite ways. That is, supraliminal-exposure subjects reacted to the letters by perceiving 9-in. letters as taller than ¼-in. letters, but 1. In the traditional approach, the description and interpretation of the interaction (i.e., difference between differences) often encompasses a description and interpretation of any main effects. Thus, main effects often receive little attention when an interaction is present because the interaction shows that these effects are not uniform across levels of at least one other factor (e.g., Kirk, 1968). Representing interaction variance separately from main effects in the residual approach encourages people to interpret main effect and interaction variance separately, however.
Computing residuals for a significant interaction in the $2 \times 2$ design will always produce a crossover pattern. This does not imply that a crossover process is responsible for the effect or that a crossover interpretation should be offered.² Just because an interaction is obtained does not mean that one should necessarily impart psychological meaning to the interaction residuals. If the residual approach leads to nonsensical conclusions given the variables under study (as in the letter-size example), the residual approach can be discarded. Nevertheless, the residual approach can be useful if one has a theory that accounts separately for the main effect and interaction variance. As noted, however, if a researcher prefers a crossover (i.e., residual-based) interpretation of the interaction, he or she must also develop an explanation of the removed main effect (or effects). If two or more plausible interpretations of the interaction still exist, it might then be possible to distinguish among the alternatives using additional data (e.g., measures of the proposed underlying mechanisms, or a new study in which the "main effect" process is removed or inhibited so a proposed crossover process can emerge in the cell means).

**USING CONTRASTS IN PLACE OF INTERACTION TESTS**

Given the controversy over how to interpret data when an interaction is present, it is not surprising that some controversy has also arisen over how to provide the best statistical test when a pattern that includes interaction variance is hypothesized. In particular, some scholars have recommended that because interaction predictions often suggest a specific pattern of cell means, such predictions should be tested using focused contrasts that conform to the hypothesized pattern (e.g., see Rosenthal & Rosnow, 1985, 1991; Rosnow & Rosenthal, 1995). In this section, we examine the utility of using focused contrasts to test hypothesized patterns that include interactions.

**Hypotheses That Imply an Ordering of Cell Means**

Contrasts are comparisons of two or more cell means. Contrasts are often used to test the predicted pattern of means by

---

² Of course, neither the residual nor the traditional approach necessitates a particular interpretation of the psychological processes underlying the interaction. A researcher using the residual approach could determine that one process accounts for both main effect and interaction variance. If one were to regard both main effect and interaction variance as accounted for by one process, however, there would seem to be little reason to represent and interpret the interaction variance separately from the main effect (or effects). Maintaining that different sources of variance must be separated from each other to be interpreted suggests that the sources of variance have some psychological meaning separate from one another. Although one can imagine some cases in which separable processes are responsible for main effects versus interactions, there are many cases in which moderation of the same process creates both main effect and interaction variance. Therefore, use of residuals as the default method for understanding interactions appears unwise.
assigning to each cell a weight that corresponds to the predicted pattern. For instance, if a researcher expects that the means of four cells in a study will be ordered in a decreasing linear fashion with equal spacing between cells, then contrast weights assigned to test that prediction would be +3, +1, −1, and −3, respectively. Any number of possible combinations of weights is possible as long as the weights sum to zero (see Rosenthal & Rosnow, 1985, 1991). The primary purported advantage of contrast analyses is the focused nature of the research questions that can be addressed. For instance, in a one-way design with four cells, investigating a linear trend of cell means by using the contrast weights just listed is more focused on the hypothesis than an omnibus F test in the traditional one-way ANOVA. In fact, ordering contrasts such as the linear contrast described have even been advocated to replace the single-df main effect and interaction tests in a 2 × 2 design when hypotheses imply an ordering of cell means (see Rosnow & Rosenthal, 1991, 1995).

For instance, consider a 2 × 2 design discussed by Rosnow and Rosenthal (1989a, 1991) and Meyer (1991). The investigator hypothesizes that the health status of one’s child (healthy vs. unhealthy) has a larger impact on parental grief when the child dies than when the child is male rather than female (see Fig. 3). Note that this hypothesis implies the following ordering of cell means on the measure of grief: B (healthy male) > A (healthy female) > C (unhealthy female) > D (unhealthy male). Because contrasts specify particular orders of cell means, Rosnow and Rosenthal (1991) stated that “contrasts sharpen our conception of the hypothesized trends” and “our preference to test [the predicted pattern] would be a contrast with weights +3, +1, −1, −3” (p. 574; see Rosnow & Rosenthal, 1995, for a similar recommendation).

One reason for favoring the contrast approach is that a difference-between-differences interaction test (df = 1) could show significant results in a 2 × 2 design even if the ordering of means were quite different from the hypothesized ordering. Therefore, the linear contrast is viewed as answering a more focused question than the interaction test. A perusal of the journals on our shelves turned up many instances of the use of contrasts to examine hypothesized patterns that included interactions.3

Potential Problems With the Contrast Approach

In order to demonstrate some potential problems with using certain contrasts to replace the traditional interaction test when a difference between differences is hypothesized, we reanalyzed the data from one published study that used such a contrast. The researchers used a Variable X (high vs. low) × Variable Y (high vs. low) design, and hypothesized a “specific interaction” such that Y would have a greater impact on the dependent measure when X was high than when X was low.4

The results obtained in their study are graphed in Figure 4a.

Because the specific interaction implied that the Y-high/X-high and Y-low/X-high cells would be most extreme (with the Y-high/X-low and Y-low/X-low cells between the extremes), the authors decided to test the ordering of means using the linear contrast recommended by Rosnow and Rosenthal (1991). That is, the authors weighted the Y-high/X-high cells as +3, Y-high/ X-low cell as +1, Y-low/X-low cell as −1, and Y-low/X-high cell as −3. The linear contrast on the data yielded a highly significant result (p < .001), and the authors concluded that their hypothesis was supported. Yet when the data are examined using a traditional ANOVA, there is a strong main effect for variable Y (p < .001), but no significant X-by-Y interaction (p > .25). That is, Y does not differentially influence the dependent measure across levels of X. The discrepancy between these tests lies in the fact that the linear contrast combines both main effect and interaction variance, whereas the ANOVA interaction reflects the unique variance accounted for by the difference between differences. It would thus appear that the linear contrast used to analyze a 2 × 2 design can be highly significant when only the main effect portion of the contrast is truly present in the data. This property of linear contrasts implies that a variety of unanticipated patterns of cell means could produce a significant contrast effect.

For example, imagine that the effect of Y was exactly the same across levels of X in the data (see Fig. 4b). The same linear contrast supposedly testing a hypothesized specific interaction would produce an even more significant result than for the original data (p < .00006), despite the fact that the interaction sum of squares for these data would be zero! Finally, the linear contrast could be significant even if there were a significant difference between differences in the opposite direction of the hypothesized pattern (p < .023; see Fig. 4c)! Given that the chosen contrast is designed to test a specific linear ordering of means, it might surprise some readers to learn that cell means that deviate dramatically from the hypothesized ordering can produce even larger test statistics than cell means that conform to the hypothesis.

3. Rosnow and Rosenthal might not have intended the focused contrast to replace the difference-between-differences approach, but, nonetheless, their comments have apparently been interpreted by some researchers as suggesting this.

4. We do not identify the study because we are simply using these data for convenience. That is, we could have made up some hypothetical data to make the same points.
does not preclude the possibility that other contrasts account for the data as well as or better than the specified contrast. This difficulty is illustrated in Figure 4b: The linear contrast is significant (and the residual left after the linear contrast is nonsignificant) even though there is no interaction variance in the data. Even though a researcher using such a linear contrast might profess that he or she is not interested in main effects or interactions, it is impossible for the data to entirely match the proposed linear ordering (i.e., B > A > C > D, as in Fig. 3) unless both main effect and interaction variance are present in the data (i.e., if one or the other is absent, then at least some of the proposed ordering will fail to occur). Thus, if the data can be completely characterized by a main effect, then a main effect interpretation is a more accurate (and more conceptually parsimonious) characterization of the data than one that implies both main effect and interaction variance should be present.

**Two Problems With “Focused” Contrasts**

This example illustrates two potential problems with the use of contrasts to test a specific ordering of cell means implied by a hypothesis that includes both main effect and interaction variance. The first problem concerns the match between the research hypotheses and the statistical tests. When researchers propose “differential impact” hypotheses, using statistical techniques that confound “differential impact” variance (i.e., interaction variance) with “overall impact” variance (i.e., main effect variance) is inappropriate. Using a test that combines interaction variance with main effect variance provides no diagnostic information about the differential impact portion of the hypothesis. If a researcher hypothesizes a pattern of results that should produce both main effect and interaction variance, the hypothesis is fully supported only if both sources of variance are present in the data. Unfortunately, the use of ordering contrasts in such cases provides information only that the combination of both sources of variance is significant.

The second problem is that a significant focused contrast

**What Should Researchers Do?**

There are at least two possible approaches to address the problem of confounding sources of variance in a contrast: creating contrast weights that orthogonize the relevant sources of variance and simultaneously testing multiple nonorthogonal contrasts in regression analyses. In the present example, the authors’ predictions included a main effect of Y and the differential impact of Y at different levels of X (which is orthogonal to the main effect). Thus, the statistical tests should have examined the significance of each source of variance, not some combination of the two. Thus, we believe that authors with such hypotheses would be better served by using a traditional ANOVA (which tests orthogonal main effects and interactions) than by using the confounded linear contrast.

An alternative strategy would be to specify several contrasts, each of which might theoretically account for the pattern of data observed. These contrasts could then be compared through the use of simultaneous regression analyses in order to determine which contrast accounts for the most unique variance (i.e., partialing for the effects of the other contrasts being examined; cf. Rosenthal & Rosnow, 1991, pp. 477-478). If the authors of our example study had tested the linear contrast controlling for the main effect of Y, they would have found that the linear contrast becomes nonsignificant ($p > .25$). Of course, computing a contrast that combines main effect and interaction variance while controlling for another contrast that includes only the main effect leaves one with a test that is basically the ANOVA test of the interaction.

**GENERAL CONCLUSIONS**

When researchers obtain a statistical interaction, there are different approaches to understanding its psychological mean-
Although some scholars have suggested that the psychological meaning of interactions in data can be understood only by removing variance due to lower order effects, we believe that there is no canned way to interpret data when an interaction is present. Researchers should not necessarily assume that the significant sources of variance in an ANOVA separately represent some meaningful psychological process (or processes). We believe that in representing and interpreting data containing interactions, researchers should consider the conceptual nature of the variables under study (which can often rule out certain representations of the interaction as implausible), the relevant psychological theories (which can direct researchers toward an appropriate understanding of the interaction), and any additional data necessary to distinguish among plausible interaction interpretations. It is only after these considerations that researchers can make an informed decision regarding how an interaction can best be understood. When researchers use focused contrasts as tests of trends in data, they should be aware that a single contrast sometimes confounds multiple sources of meaningful variance (e.g., the linear contrast we described earlier included both main effect and interaction variance). In such cases, the contrast is not diagnostic of the existence of either effect included (i.e., confounded) in the test. Thus, if one's hypothesis depends on a specific source of variance being present, one cannot adequately test for the existence of that effect using a contrast that combines that variance with other sources of variance.

Acknowledgments—We are grateful to Robert Abelson, Anthony Greenwald, Geoffrey Keppel, Robert C. MacCallum, Gifford Weary, and Timothy Wilson for comments on the July 1993 version of this article. Preparation of this article was supported by National Science Foundation Grants BNS-9021647 and SBR-9320854 to R.E.P., a National Science Foundation predoctoral fellowship to L.R.F., and National Institute of Mental Health traineeships (T32 MH19728) to L.R.F., D.T.W., and J.R.P.
Technical Commentary

UNDERSTANDING DATA WHEN INTERACTIONS ARE PRESENT OR HYPOTHESIZED

Richard E. Petty, Leandre R. Fabrigar, Duane T. Wegener, and Joseph R. Priester

1Ohio State University, 2Queen's University, 3Yale University, and 4University of Michigan

Abstract—The traditional approach to interpreting data when an interaction is present is to interpret cell means as reflecting a difference between differences. An alternative is to interpret any main effects and separately interpret interaction residuals. Interpretation of interaction residuals can sometimes lead to nonsensical conclusions. Thus, in deciding between the approaches, researchers should consider (a) the conceptual nature of the variables involved, (b) relevant theories, and (c) the additional data that might be necessary to distinguish among competing plausible representations. Issues have also arisen regarding how best to test a hypothesis that involves interaction variance. Some researchers have argued that the use of a focused contrast designed to test a specific ordering of cell means (e.g., a linear contrast) is often the best strategy. We note potential problems with the use of such contrasts and discourage researchers from blanket use of contrasts that combine main effect and interaction variance.

Statistical interactions often play a crucial role in testing psychological theory. Yet recent articles have raised important issues regarding how researchers should interpret their data when a statistical interaction is present, and how they should best test an interaction hypothesis. In this article, we discuss conceptual issues regarding understanding data from factorial designs (especially the common 2 × 2) when an interaction is present or hypothesized. We focus on the 2 × 2 in part because of the ubiquity of this design, and because appropriate procedures for testing and interpreting the data when interactions are obtained in this design have been the subject of considerable recent attention and controversy (e.g., see Bobko, 1986; Meyer, 1991; Rosnow & Rosenthal, 1989a, 1989b, 1991, 1995; Ross & Creyer, 1993; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993).

UNDERSTANDING THE DATA WHEN AN INTERACTION IS PRESENT

In order to make psychological sense of data, one must have some agreed-upon method of representing the pattern of results obtained. This is particularly important when an interaction is present because there are multiple ways to represent or depict interactions (cf. Judd, McClelland, & Culhane, 1995). For example, Rosnow and Rosenthal (1989a, 1989b, 1991, 1995) have suggested that understanding the data is fostered by representing and interpreting the interaction as the residual remaining after main effects have been removed. We argue that a blanket use of this approach rather than the more traditional approach of representing and interpreting an interaction as a difference between differences in cell means (e.g., Keppel, 1991) can lead to misleading conclusions about the psychological processes underlying the results. It is important to note that representation of interactions is not a statistical issue per se. The two ways of representing interactions are mathematically equivalent (i.e., interaction variance is the same, whether calculated through residuals or differences between differences; see Guilford & Fruchter, 1978; Keppel, 1991; Rosnow & Rosenthal, 1989b). However, the different representations can suggest different (and conflicting) substantive conclusions.

Traditional Versus Residual Approach to Interactions

Following the finding of a significant interaction in an analysis of variance (ANOVA), researchers have traditionally described the interaction effect through comparisons of original cell means (e.g., through simple effects tests; Keppel, 1991). Take, for instance, a hypothetical example presented by Rosenthal and Rosnow (1985). In a 2 × 2 design, one of two versions of a political advertisement was shown to either liberals or conservatives. A traditional ANOVA indicates a significant main effect of advertisement, no main effect of political ideology, and a significant interaction between the two independent variables (see Fig. 1a). According to the difference-between-differences approach (e.g., see Keppel, 1991), one acceptable interpretation of these data is that there is a greater effect of the differing ad types for liberals than for conservatives. Keppel (1991) recommended analyses of the two simple effects (of ad type, in this example) as one method for examining the meaning of the interaction (pp. 236–245; see also Myers, 1979; Pedhazur, 1982). This approach to interpretation resonates with the views espoused by Fisher (1947, pp. 92–93) and other researchers (e.g., Hoaglin, Mosteller, & Tukey, 1991).

An alternative view has been espoused by proponents of the residual approach. Rosenthal, Rosnow, and their colleagues (e.g., Rosenthal & Rosnow, 1985, 1991; Zuckerman et al., 1993) have noted correctly that cell means reflect not only the influence of the interaction term in an ANOVA, but also the influence of all lower order effects. Therefore, according to Rosenthal and Rosnow, if one is interested in understanding the psychological meaning of an interaction in a 2 × 2 design, one
Interactions

![Diagram](attachment:image.png)

Fig. 1. Traditional representation of the difference between differences (a) and residual representation of the interaction (b) in a 2 (political ideology) × 2 (ad version) factorial.

should not do so by examining a representation that also reflects main effects. Rather, one must remove the influence of the main effects from the cell means prior to imparting meaning to the interaction.

Consider the Ad Version × Political Ideology data in Figure 1. According to Rosenthal and Rosnow (1985), a researcher following the traditional approach might be tempted to interpret the presence of an interaction in the data as indicating that liberals are more strongly influenced by version A than version B, whereas conservatives are equally influenced by both advertisements. Rosenthal and Rosnow (1985) asserted, however, that such a conclusion "would be wrong!" (p. 5). Instead, they prefer to interpret the representation of the interaction as depicted in Figure 1b. Consistent with this representation, Rosenthal and Rosnow stated that "the interaction actually shows that conservatives and liberals reacted in exactly opposite ways to the two types of propaganda" (p. 8; see Rosenthal & Rosnow, 1985, for calculation of residuals).

Obviously, this residual representation suggests something very different about the behavior of conservatives than does the representation following from the traditional approach. That is, a difference-between-differences approach encourages one to develop theory that explains why liberals are affected differently by the two ads but conservatives are not, whereas the residual approach encourages one to explain why liberals and conservatives are influenced in opposite ways by the two ads.

The nature of the variables in this example does not preclude generating a meaningful psychological interpretation of the interaction residuals. Therefore, the representations encouraged by the traditional and the residual views are both potentially accurate.

Comparing the Approaches

Rosenthal and Rosnow (1991) asserted that "before an interaction effect can be understood, . . . the residuals defining the interaction must be displayed" (p. 367), and Zuckerman et al. (1993) stated that "interaction effects cannot be interpreted on the basis of comparisons between cell means" (p. 53). If these scholars meant to suggest only that researchers who describe cell means are not describing solely interaction variance, this is correct. However, if they meant that researchers should necessarily impart psychological meaning to interaction residuals (as suggested by their assertion about the "reactions" of liberals and conservatives depicted in Fig. 1), then we disagree. In particular, it is not clear that authors who wish to understand data containing an interaction should invariably devote journal space to displaying and interpreting interaction residuals instead of interpreting cell means. In fact, interpretation of residuals can sometimes lead to nonsensical conclusions because the interaction per se is a statistical entity that might or might not directly correspond to any meaningful underlying psychological process.

Consider a pattern of data similar to that in Figure 1a, but with different variables. Specifically, imagine an experiment investigating the influence of exposure duration (subliminal [5 ms] vs. supraliminal [500 ms]) and actual letter height (¼ in. vs. 9 in.) on subjects' verbal estimates of the height of letters. See Figure 2a for the expected cell means.

A 2 × 2 ANOVA shows a main effect of letter size and an Exposure Duration × Letter Size interaction. The difference-between-differences approach suggests that the interaction exists because there is a greater impact of letter size on perceptions of height when the letters are presented supraliminally than when the letters are presented subliminally. Advocates of the residual approach would presumably suggest that this characterization is wrong. That is, they would assert that the meaning of the interaction can be understood only by interpreting the residuals represented in Figure 2b. According to the residual approach, then, one would interpret the presence of the interaction as indicating that supraliminal- and subliminal-exposure subjects reacted to the presented letters in exactly opposite ways. That is, supraliminal-exposure subjects reacted to the letters by perceiving 9-in. letters as taller than ¼-in. letters, but

1. In the traditional approach, the description and interpretation of the interaction (i.e., difference between differences) often encompasses a description and interpretation of any main effects. Thus, main effects often receive little attention when an interaction is present because the interaction shows that these effects are not uniform across levels of at least one other factor (e.g., Kirk, 1968). Representing interaction variance separately from main effects in the residual approach encourages people to interpret main effect and interaction variance separately, however.
Fig. 2. Traditional representation of the difference between differences (a) and residual representation of the interaction (b) in a 2 (exposure) \( \times \) 2 (letter height) factorial.

subliminal-exposure subjects reacted to the letters by perceiving 1/4-in. letters as taller than 9-in. letters! Because people who participated in the subliminal conditions could not even see the letters, such a representation of how the results were altered seems absurd. In fact, in psychological research, there are many variables that do not lend themselves well to the removal of main effects when interpreting the meaning of interactions.

What Should Researchers Do?

As a matter of practice, we believe that researchers should not confuse statistical sources of variance with underlying psychological processes when interpreting their data. For example, three sources of variance in a 2 \( \times \) 2 ANOVA (two main effects and one interaction) could result from three psychological processes, or two, or one, or even four or more! Thus, the residual approach—which encourages interpretation of interaction variance separate from main effect variance—should not be viewed as the default way to impart psychological meaning to data when an interaction is present. For example, it would be inappropriate to take results from a study that has produced an interaction in which the cell means do not show a crossover pattern, remove the main effects from the original cell means, and then argue that the pattern of residuals necessarily provides evidence for a theory that predicts a crossover pattern. Computing residuals for a significant interaction in the 2 \( \times \) 2 design will always produce a crossover pattern. This does not imply that a crossover process is responsible for the effect or that a crossover interpretation should be offered.

Just because an interaction is obtained does not mean that one should necessarily impart psychological meaning to the interaction residuals. If the residual approach leads to nonsensical conclusions given the variables under study (as in the letter-size example), the residual approach can be discarded. Nevertheless, the residual approach can be useful if one has a theory that accounts separately for the main effect and interaction variance. As noted, however, if a researcher prefers a crossover (i.e., residual-based) interpretation of the interaction, he or she must also develop an explanation of the removed main effect (or effects). If there are more plausible interpretations of the interaction or in which the interaction still exist, it might then be possible to distinguish among the alternatives using additional data (e.g., measures of the proposed underlying mechanisms, or a new study in which the "main effect" process is removed or inhibited so a proposed crossover process can emerge in the cell means).

USING CONTRASTS IN PLACE OF INTERACTION TESTS

Given the controversy over how to interpret data when an interaction is present, it is not surprising that some controversy has also arisen over how to provide the best statistical test when a pattern that includes interaction variance is hypothesized. In particular, some scholars have recommended that because interaction predictions often suggest a specific pattern of cell means, such predictions should be tested using focused contrasts that conform to the hypothesized pattern (e.g., see Rosenthal & Rosnow, 1985, 1991; Rosnow & Rosenthal, 1995).

In this section, we examine the utility of using focused contrasts to test hypothesized patterns that include interactions.

Hypotheses That Imply an Ordering of Cell Means

Contrasts are comparisons of two or more cell means. Contrasts are often used to test the predicted pattern of means by

2. Of course, neither the residual nor the traditional approach necessitates a particular interpretation of the psychological processes underlying the interaction. A researcher using the residual approach could determine that one process accounts for both main effect and interaction variance. If one were to regard both main effect and interaction variance as accounted for by one process, however, there would seem to be little reason to represent and interpret the interaction variance separately from the main effect (or effects). Maintaining that different sources of variance must be separated from each other to be interpreted suggests that the sources of variance have some psychological meaning separate from one another. Although one can imagine some cases in which separable processes are responsible for main effects versus interactions, there are many cases in which moderation of the same process creates both main effect and interaction variance. Therefore, use of residuals as the default method for understanding interactions appears unwise.
assigning to each cell a weight that corresponds to the predicted pattern. For instance, if a researcher expects that the means of four cells in a study will be ordered in a decreasing linear fashion with equal spacing between cells, then contrast weights assigned to test that prediction would be $+3$, $+1$, $-1$, and $-3$, respectively. Any number of possible combinations of weights is possible as long as the weights sum to zero (see Rosenthal & Rosnow, 1985, 1991). The primary purported advantage of contrast analyses is the focused nature of the research questions that can be addressed. For instance, in a one-way design with four cells, investigating a linear trend of cell means by using the contrast weights just listed is more focused on the hypothesis than an omnibus $F$ test in the traditional one-way ANOVA. In fact, ordering contrasts such as the linear contrast described have even been advocated to replace the single-df main effect and interaction tests in a $2 \times 2$ design when hypotheses imply an ordering of cell means (see Rosnow & Rosenthal, 1991, 1995).

For instance, consider a $2 \times 2$ design discussed by Rosnow and Rosenthal (1989a, 1991) and Meyer (1991). The investigator hypothesizes that the health status of one’s child (healthy vs. unhealthy) has a larger impact on parental grief when the child dies when the child is male rather than female (see Fig. 3). Note that this hypothesis implies the following ordering of cell means on the measure of grief: B (healthy male) > A (healthy female) > C (unhealthy female) > D (unhealthy male). Because contrasts specify particular orders of cell means, Rosnow and Rosenthal (1991) stated that “contrasts sharpen our conception of the hypothesized trends” and “our preference to test [the predicted pattern] would be a contrast with weights $+3$, $+1$, $-1$, $-3$” (p. 574; see Rosnow & Rosenthal, 1995, for a similar recommendation).

One reason for favoring the contrast approach is that a difference-between-differences interaction test (df = 1) could show significant results in a $2 \times 2$ design even if the ordering of means were quite different from the hypothesized ordering. Therefore, the linear contrast is viewed as answering a more focused question than the interaction test. A perusal of the journals on our shelves turned up many instances of the use of contrasts to examine hypothesized patterns that included interactions.3

Potential Problems With the Contrast Approach

In order to demonstrate some potential problems with using certain contrasts to replace the traditional interaction test when a difference between differences is hypothesized, we reanalyzed the data from one published study that used such a contrast. The researchers used a Variable X (high vs. low) x Variable Y (high vs. low) design, and hypothesized a “specific interaction” such that Y would have a greater impact on the dependent measure when X was high than when X was low.4 The results obtained in their study are graphed in Figure 4a.

Because the specific interaction implied that the Y-high/X-high and Y-low/X-high cells would be most extreme (with the Y-high/X-low and Y-low/X-low cells between the extremes), the authors decided to test the ordering of means using the linear contrast recommended by Rosnow and Rosenthal (1991). That is, the authors weighted the Y-high/X-high cell as $+3$, Y-high/X-low cell as $+1$, Y-low/X-low cell as $-1$, and Y-low/X-high cell as $-3$. The linear contrast on the data yielded a highly significant result ($p < .001$), and the authors concluded that their hypothesis was supported. Yet when the data are examined using a traditional ANOVA, there is a strong main effect for variable Y ($p < .001$), but no significant X-by-Y interaction ($p > .25$). That is, Y does not differentially influence the dependent measure across levels of X. The discrepancy between these two tests lies in the fact that the linear contrast combines both main effect and interaction variance, whereas the ANOVA interaction reflects the unique variance accounted for by the difference between differences. It would thus appear that the linear contrast used to analyze a $2 \times 2$ design can be highly significant when only the main effect portion of the contrast is truly present in the data. This property of linear contrasts implies that a variety of unanticipated patterns of cell means could produce a significant contrast effect.

For example, imagine that the effect of Y was exactly the same across levels of X in the data (see Fig. 4b). The same linear contrast supposedly testing a hypothesized specific interaction would produce an even more significant result than for the original data ($p < .000006$), despite the fact that the interaction sum of squares for these data would be zero! Finally, the linear contrast could be significant even if there were a significant difference between differences in the opposite direction of the hypothesized pattern ($p < .023$; see Fig. 4c)! Given that the chosen contrast is designed to test a specific linear ordering of means, it might surprise some readers to learn that cell means that deviate dramatically from the hypothesized ordering can produce even larger test statistics than cell means that conform to the hypothesis.

3. Rosnow and Rosenthal might not have intended the focused contrast to replace the difference-between-differences approach, but, nonetheless, their comments have apparently been interpreted by some researchers as suggesting this.

4. We do not identify the study because we are simply using these data for convenience. That is, we could have made up some hypothetical data to make the same points.

---

Fig. 3. Hypothesized ordering and corresponding contrast weights in a $2 \times 2$ (health of child) x (gender of child) factorial design.

---

VOL. 7, NO. 4, JULY 1996
does not preclude the possibility that other contrasts account for the data as well as or better than the specified contrast. This difficulty is illustrated in Figure 4b: The linear contrast is significant (and the residual left after the linear contrast is nonsignificant) even though there is no interaction variance in the data. Even though a researcher using such a linear contrast might profess that he or she is not interested in main effects or interactions, it is impossible for the data to entirely match the proposed linear ordering (i.e., B > A > C > D), as in Fig. 3) unless both main effect and interaction variance are present in the data (i.e., if one or the other is absent, then at least some of the proposed ordering will fail to occur). Thus, if the data can be completely characterized by a main effect, then a main effect interpretation is more accurate (and more conceptually parsimonious) characterization of the data than one that implies both main effect and interaction variance should be present.

What Should Researchers Do?

There are at least two possible approaches to address the problem of confounding sources of variance in a contrast: creating contrast weights that orthogonalize the relevant sources of variance and simultaneously testing multiple nonorthogonal contrasts in regression analyses. In the presented example, the authors' predictions included a main effect of Y and the differential impact of X at different levels of Y (which is orthogonal to the main effect). Thus, the statistical tests should have examined the significance of each source of variance, not some combination of the two. Thus, we believe that authors with such hypotheses would be better served by using a traditional ANOVA (which tests orthogonal main effects and interactions) than by using the confounded linear contrast.

An alternative strategy would be to specify several contrasts, each of which might theoretically account for the pattern of data observed. These contrasts could then be compared through the use of simultaneous regression analyses in order to determine which contrast accounts for the most unique variance (i.e., partialing for the effects of the other contrasts being examined; cf. Rosenthal & Rosnow, 1991, pp. 477–478). If the authors of our example study had tested the linear contrast controlling for the main effect of Y, they would have found that the linear contrast becomes nonsignificant (p > .25). Of course, computing a contrast that combines main effect and interaction variance while controlling for another contrast that includes only the main effect leaves one with a test that is basically the ANOVA test for the interaction.5

GENERAL CONCLUSIONS

When researchers obtain a statistical interaction, there are different approaches to understanding its psychological mean-

---

5. In the preceding sections, we have outlined potential pitfalls in the use of contrasts to test specific orderings of cell means, especially within factorial designs. We do not wish to imply by this critique that contrasts have no place in testing or articulating psychological theory. Contrasts can be especially useful in designs with more than two levels of one or more factors (see Abelson & Prentice, 1995) or as a technique for articulating results (see Abelson, 1995).
Interactions

Although some scholars have suggested that the psychological meaning of interactions in data can be understood only by removing variance due to lower order effects, we believe that there is no canned way to interpret data when an interaction is present. Researchers should not necessarily assume that the significant sources of variance in an ANOVA separately represent some meaningful psychological process (or processes). We believe that in representing and interpreting data containing interactions, researchers should consider the conceptual nature of the variables under study (which can often rule out certain representations of the interaction as implausible), the relevant psychological theories (which can direct researchers toward an appropriate understanding of the interaction), and any additional data necessary to distinguish among plausible interaction interpretations. It is only after these considerations that researchers can make an informed decision regarding how an interaction can best be understood. When researchers use focused contrasts as tests of trends in data, they should be aware that a single contrast sometimes confounds multiple sources of meaningful variance (e.g., the linear contrast we described earlier included both main effect and interaction variance). In such cases, the contrast is not diagnostic of the existence of either effect included (i.e., confounded) in the test. Thus, if one’s hypothesis depends on a specific source of variance being present, one cannot adequately test for the existence of that effect using a contrast that combines that variance with other sources of variance.

Acknowledgments—We are grateful to Robert Abelson, Anthony Greenwald, Geoffrey Keppel, Robert C. MacCallum, Gifford Weary, and Timothy Wilson for comments on the July 1993 version of this article. Preparation of this article was supported by National Science Foundation Grants BNS-9021647 and SBR-9520854 to R.F.P., a National Science Foundation predoctoral fellowship to L.R.F., and National Institute of Mental Health traineeships (T32 MH19728) to L.R.F., D.T.W., and J.R.P.

REFERENCES


*(Received 3/23/95; Accepted 7/13/95)*
CONTRASTS AND INTERACTIONS REDUX: Five Easy Pieces

Ralph L. Rosnow\(^1\) and Robert Rosenthal\(^2\)
\(^1\)Temple University and \(^2\)Harvard University

Abstract—This reply to Abelson (this issue) and Petty, Fabrigar, Wegener, and Priester (this issue) is couched within the framework of five basic principles advising that we (1) hang on to what we predicted long enough to test it (Tarzan's leap), (2) be wary of beguiling statistical designs that may not address the question of interest (the Sirens' song), (3) not allow well-worn habits of thinking to ensnare our perceptions (Lavoisier's crease), (4) weigh the possibility of more than one correct hypothesis (the dayyan's decree), and (5) not confuse unplanned with planned contrasts (the archer's aim).

We welcome this opportunity to comment on articles by Abelson and by Petty, Fabrigar, Wegener, and Priester in this issue of Psychological Science. We greatly admire Abelson's thoughtful tone and spirit, though still there may be some differences in our relative emphases. We emphasize that the precision of the contrast answer is no greater than the precision of the contrast question, and we think "overall pattern" contrasts should be used only if they embody the theory, hypothesis, or question of interest. Contrast weights never know whether (in Abelson's terminology) they were local effects or global patterns, and one person's effect contrast is another's pattern contrast. Thus, Abelson's distinction is not intrinsic to any lambda weights, but is entirely in the mind or theory of the data analyst. We see no intrinsic virtue in setting up a \(2 \times 2\) design in preference to a \(1 \times 4\) design in which a set of wired-in contrasts will give row, column, and interaction effects that may or may not be of substantive interest. Testing hypotheses calls for the right questions, because asking the wrong questions leads to the wrong answers.

Nonetheless, we certainly agree with Abelson that "investigators often do not think clearly about what point they should try to make, nor even whether their results support any worthwhile point at all" and that "there often may be more than one 'right answer'" (p. 245). Abelson says many things we find interesting and use as a springboard for proposing some further ideas. We also agree with Petty et al. when they note that there are "different approaches to understanding . . . psychological meaning" and advise that "researchers should consider the conceptual nature of the variables under study" (pp. 251–252). We resonate to these statements as quite consistent with our own understanding of the nature of knowledge and emphasis on methodological pluralism and theoretical ecumenism (e.g., Rosenthal & Rosnow, 1991; Rosnow, 1981), a view that has been eloquently expressed by Campbell (1988); Fiske and Shweder (1986); Houts, Cook, and Shadish (1986); McGuire (1986); and other researchers.

Address correspondence to Ralph L. Rosnow, Department of Psychology, Temple University, Philadelphia, PA 19122.

However, Petty et al. make many other statements that leave us puzzled and concerned, because these statements imply confusion in thinking about contrasts and the meaning and interpretation of interaction in analysis of variance (ANOVA). At various points, for example, they note or imply four ways of interpreting any interaction in ANOVA: examining the residuals, examining the differences between orthogonal simple effects, examining the simple effects, and examining the differences between nonorthogonal simple effects. The first and second approaches are identical and accurate definitions of interaction in \(2 \times 2\) designs; the third and fourth approaches are inaccurate definitions of interaction under most circumstances. Petty et al. also do not seem to realize that row, column, and interaction effects in a \(2 \times 2\) are in fact three orthogonal contrasts. They say that "the linear contrast is viewed as answering a more focused question than the interaction test" (p. 250). But no contrast is more focused than any other contrast; all have a single degree of freedom. Space limitations prevent us from parsing their commentary item by item, and it would be wasteful simply to rehash what is readily available elsewhere (e.g., Rosenthal & Rosnow, 1985, 1991; Rosnow & Rosenthal, 1989a, 1995b, 1995). It is hard to avoid some redundancy, but we focus our discussion within the framework of five anecdotal principles.

PRINCIPLE 1. TARZAN'S LEAP

Petty et al. claim we "have suggested that understanding the data is fostered by representing and interpreting the interaction as the residual remaining after main effects have been removed" (p. 247). This is simply not true. We do not even care if the interaction is computed. All we say is that if someone cites the presence of an interaction, only the residuals (or leftover effects) tell what the claimed interaction is. Perhaps the reader will recall the dinner scene in the movie Five Easy Pieces. In this scene, Jack Nicholson asked for a side order of toast and the waitress told him, "We don't serve side orders of toast." He replied, "You make sandwiches, don't you?" and ordered a sandwich, telling her to remove all the ingredients until all that was left was what he wanted—toast. It is the same with interaction in ANOVA. Interaction is always and exclusively defined by the leftover effects (residuals); anyone claiming an obtained interaction should be prepared to display the residuals that define that interaction.

Petty et al. express dismay that a researcher who predicted an interaction might be misled by examining the residuals. They make puzzling citations to a section in Fisher's 1947 The Design of Experiments and work coauthored by Hoaglin, Mosteller, and Tukey to support their claim that the "residual approach" [sic] is an "alternative" or unconventional way of analyzing
even he suffers a momentary lapse: "a test of the residual mean square produces \( F = 2.70, p = .08 \), suggesting (weakly) . . ." (p. 244). The \( p \) value, of course, tells us something very different from whether there was a "weak" or "strong" effect.

There is no heavenly sanctuary for people with \( ps < .05 \), but there may be a dreaded warmer place for people who say that failure to reject the null hypothesis implies an effect size of "zero." Petty et al. and Abelson, of course, do not make this mistake, but claims of "no effect" when "obtained \( p > .05 \)" abound in the psychological literature. Cohen (1994) has urged researchers to examine the confidence intervals around their obtained effects, and Rosenthal and Rubin (1994) have proposed a new statistic, called the counternull value, that gives the nonnull magnitude of effect size that is supported by the same amount of evidence as is the null value of the effect size. When the effect-size estimate is based on a symmetric reference distribution (e.g., Hedges's \( g \) or Cohen's \( d \)), the obtained effect falls halfway between the null value of the effect size and the counternull value.

We prefer the \( r \) effect-size index because of its computational convenience and ease of interpretation (e.g., Rosenthal & Rubin, 1982; Rosnow & Rosenthal, 1988). Because \( r \) is based on an asymmetric reference distribution, one can use the Fisher \( Z \) transformation to approximate the counternull value of \( r \) (see Rosenthal & Rubin, 1994). That method gives a slightly inflated estimate; an accurate estimate of the counternull value (Rosenthal, Rosnow, & Rubin, 1996) can be obtained by

\[
r_{\text{counternull}} = \frac{4r^2}{1 + 3r^2},
\]

where \( r \) is the effect size, and the \( r_{\text{counternull}} \) is interpreted as described previously. When researchers imply that failure to reject the null hypothesis means "no effect" or that finding a statistically significant \( p \) value means an effect of important magnitude, research consumers can often use contrasts, null to counternull intervals, and other simple procedures to reach into the results and decide for themselves what the reported data reveal (Rosnow & Rosenthal, 1996).

**PRINCIPLE 4. THE DAYYAN'S DECREE**

Abelson and Petty et al. are concerned with, as Abelson puts it, how "one large effect within the pattern can inappropriately make the whole pattern seem impressive" (p. 244). Abelson goes on to say that "wariness of pattern contrasts that subsume local effects, and careful consideration of which factors are ground and which are figural, should lower the risk of misleading claims" (p. 245). Although we do not agree entirely with Abelson's conceptualization, we think he is correct that there are times when researchers may want to adjust one contrast for another. The procedure of de-meaning should help unwary investigators, who can remove what is not of interest and operate on the residuals. Table 1 shows the results that Abelson gives for a 2 \( \times \) 2 design. Suppose we wanted to remove the old (or background) effect to get a view of what is going on over and above that effect. Part A of Table 1 shows the results before and after removing the row effects. Part B shows the ANOVA on the de-meaned results and the \( r \) effect sizes for 1-\( df \) tests. Researchers have the option of testing and examining whatever pattern is of interest.

We also advise researchers to consider the plausibility of more than one right hypothesis in a social world of infinite causal possibilities. Suppose a researcher designed a study with four ordered groups (A, B, C, and D) and hypothesized a cyclic pattern in the specified sequence. The lambda weights chosen to represent A, B, C, and D are \(-1, +3, -3, +1\), respectively. A well-grounded older hypothesis (of no interest in this situation) implies \( \lambda < 0 \) and no prediction about B or C. This old hypothesis is denoted by weights of \(-1, 0, 0, +1\) (i.e., no prediction for B or C). The two hypotheses are nonorthogonal. Following Abelson's argument, the contrast to test the new hypothesis is vulnerable to the simpler interpretation of the old hypothesis, which is only a partial explanation and far less risky to boot (i.e., it predicts no reversal of direction). The vulnerability argument seems to foreclose on two or more correlated

<table>
<thead>
<tr>
<th>Table 1. De-meaning the row means in Abelson's 2 ( \times ) 2 example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Basic data</strong> **</td>
</tr>
<tr>
<td>**</td>
</tr>
<tr>
<td>High **</td>
</tr>
<tr>
<td>Low **</td>
</tr>
<tr>
<td>Mean **</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>B. Analysis of variance</strong></th>
<th>**</th>
<th>SS **</th>
<th>** df **</th>
<th>** MS **</th>
<th>** F **</th>
<th>** p **</th>
<th>** r **</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between groups **</td>
<td>** 28.34 **</td>
<td>** 3 **</td>
<td>** 9.45 **</td>
<td>** 3.36 **</td>
<td>** .03 **</td>
<td>** — **</td>
<td></td>
</tr>
<tr>
<td>Rows **</td>
<td>** 0.00 **</td>
<td>** 1 **</td>
<td>** 0.00 **</td>
<td>** 0.00 **</td>
<td>** .00 **</td>
<td>** .00 **</td>
<td></td>
</tr>
<tr>
<td>Columns **</td>
<td>** 14.88 **</td>
<td>** 1 **</td>
<td>** 14.88 **</td>
<td>** 5.30 **</td>
<td>** .03 **</td>
<td>** .36 **</td>
<td></td>
</tr>
</tbody>
</table>


Contrasts and Interactions

processes operating simultaneously (i.e., more than one right answer). Because it is always possible for someone with a fertile imagination to spy some partial effect in the background, the unfortunate implication of the vulnerability argument seems to be that psychological science may as well close up shop.

There is an old Yiddish anecdote that helps to underscore the idea of, as Abelson puts it, "more than one 'right answer.' " A dayyan, or rabbinical judge, was asked by a married couple to mediate a disagreement. The wife told her side and the dayyan said, "You are right." The husband then told his side and the dayyan again said, "You are right." An inculcating student who overheard the conversation addressed the dayyan: "Rebbe, you told them they were both right, but surely they both can't be a hundred percent right." The dayyan replied, "You are right, too." Many years ago, Campbell and Stanley (1963) cautioned psychological researchers that when two well-grounded hypotheses disagree, they may both be right to some extent. The lesson of our fourth principle is not to foreclose prematurely on the possibility of more than one right hypothesis.

In the spirit of Abelson's advice that contrasts should be evaluated for how well they do relative to the sum of squares (SS) for the noncontrast between-groups effect, we can suggest some procedures. One intuitive method is to compute the proportion of the between-groups SS that is accounted for by particular contrasts. We compute the correlation between the lambda weights and group means (which we call "alerting r") and then square it to answer our question. Table 2 shows Abelson's E and P contrasts (from his Table 1) and, in the last column, the proportion of the between-groups SS accounted for by each contrast (i.e., r^2_{alerting}). We get an intuitive idea of how well each contrast performed by comparing r^2_{alerting} with the proportion theoretically associated with any randomly chosen contrast. Any randomly chosen contrast should account for the proportion defined by the reciprocal of the degrees of freedom of the between-groups SS (Rosenthal & Rosnow, 1985). In the case of three groups, as in Table 2, any randomly chosen contrast should be associated with \( r^2_{df} = .50 \). In the case of four groups, as in Table 1, the expected proportion is \( r^2_{df} = .33 \).

We can also compare two contrasts by creating a new contrast out of the differences between the contrast weights (Rosenthal et al., 1996). Suppose we wanted to evaluate the accuracy or predictive power of Contrast 1 relative to Contrast 3 in Table 2. When contrast weights are added or subtracted, their sums or differences are influenced more by the contrast weights with larger variance than by the weights with smaller variance. Therefore, in order to be sure that the comparison is fair (i.e., not simply reflecting the contrast with greater variance), the weights of each contrast should be divided by the standard deviation of the weights as defined by

\[
\sigma_k = \sqrt{\frac{\sum \lambda_k^2}{k}},
\]

where the numerator is the sum of the squared lambda weights, and the denominator is the number of groups or conditions (k).

The contrast SS for the difference between the two contrasts is

\[
SS_{contrast} = \frac{nL^2}{\sum \lambda^2},
\]

where \( L \) is the weighted sum of all condition means (M), and the weights are the corresponding lambda weights (\( \lambda \)) called for by the contrast, or

\[
L = \sum M \lambda = M_1 \lambda_1 + M_2 \lambda_2 + \ldots + M_k \lambda_k,
\]

and n is the number of observations in each condition (given equal n per condition).

To illustrate with our continuing example, substitution in Equation 1 gives \( \sigma_k = 0.816 \) for Contrast 1 and 1.414 for Contrast 3. Dividing the weights listed for Contrast 1 (+1, 0, -1) by 0.816 yields standardized weights of +1.225, 0, and -1.225; dividing the weights listed for Contrast 3 (+1, +1, -2) by 1.414 yields standardized weights of +0.707, +0.707, and -1.414. The differences between the corresponding weights are +0.518, -0.707, and +0.189. With n = 12, substitution in Equation 2 gives

\[
SS_{contrast} = \frac{12(5.5)(+0.518) + (4.7)(-0.707) + (3.5)(+0.189))^2}{(+0.518)^2 + (-0.707)^2 + (+0.189)^2}=rac{12(0.1876)^2}{0.8039} = 0.4223 = 0.53
\]

The table below shows Abelson's different ways of looking at his 1 × 3 example.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Weights</th>
<th>( F(df) )</th>
<th>( p )</th>
<th>Cohen's ( d )</th>
<th>( r^2_{alerting} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategy E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>+1</td>
<td>0</td>
<td>-1</td>
<td>15.01(1,33)</td>
<td>.00048</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>+2</td>
<td>-1</td>
<td>0.21(1,33)</td>
<td>.66</td>
</tr>
<tr>
<td>Strategy P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>+1</td>
<td>+1</td>
<td>-2</td>
<td>12.8(1,33)</td>
<td>.002</td>
</tr>
<tr>
<td>4</td>
<td>+1</td>
<td>-1</td>
<td>0</td>
<td>2.4(1,33)</td>
<td>.13</td>
</tr>
</tbody>
</table>

Note: The \( r^2_{alerting} \) is the squared correlation between lambda weights and group means; it gives the proportion of between-groups SS that is accounted for by the particular contrast. In Abelson's data, the three group means are 5.5, 4.7, and 3.5.
Because $S_{contrast} = M_{contrast}$, we divide this $S_{contrast}$ by Abelson's $M$ error, which equals 1.60, to find $F(1, 33) = 0.32$, $r = .10$, showing a modest difference between Contrast 1 and Contrast 3.

PRINCIPLE 5. THE ARCHER'S AIM

Abelson admonishes us about failing to warn against the "too flexible" nature of contrasts. The flexibility of contrasts is a virtue, not an inherent problem; the problem, as Abelson recognizes, lies with researchers who do not "know what they really want to test with their contrasts" (p. 245). Nonetheless, it is prudent to caution researchers not to confuse planned and unplanned contrasts. We are reminded of the Bohemian legend about a fabled archer who was offered an empire if he could teach the king how to become a great marksman. The king came upon the archer standing next to a grove of trees. Each tree had a chalked circle and an arrow in the exact center of the circle. One arrow quivered in its circle even as the king approached. "Keep your empire," the archer told the king, "for the secret of my skill is that I shoot first and draw the circle afterward."

It is no great trick to create contrast weights that will reflect any pattern of interest. One simply calculates the mean of the condition means and subtracts this grand mean from each condition mean, which gives a perfect set of lambdas. Previously, we described how to create a table of predicted values (Rosnow & Rosenthal, 1995), which we hoped would encourage precision in specification. Petty et al. note that researchers are often encouraged to develop theory to explain ad hoc results, and Abelson tells how he devised contrast weights for a predicted effect from the obtained means. As a precaution against people reading too much into these statements, our final principle is that in describing and interpreting contrasts, it is important to indicate which came first—the "circle" or the "arrow."

Acknowledgments—We thank Abelson for his generous acknowledgments and for the many contributions he has made to the improvement of data analysis in psychology. Indeed, had our discipline listened to him in 1962 when he wrote about contrast analysis, our field would be further advanced in data analytic sophistication than it is today. We wish to acknowledge the funding received by Rosnow from Temple University's Bolton Endowment and that received by Rosenthal from the Spencer Foundation and a James McKee Catell Sabbatical Award. The content of this article is solely the responsibility of the authors.

REFERENCES