Iacobucci, Dawn (ed.) (2001), *Journal of Consumer Psychology's Special Issue on Methodological and Statistical Concerns of the Experimental Behavioral Researcher*, 10 (1&2), Mahwah, NJ: Lawrence Erlbaum Associates, 5-35.

# Analysis of Variance

## I.A. CAN I TEST FOR SIMPLE EFFECTS IN THE PRESENCE OF AN INSIGNIFICANT INTERACTION?

On more than a few occasions I have encountered the situation in which a hypothesized two-way interaction was not significant according to the analysis of variance (ANOVA), and yet simple effects tests yielded results consistent with the expected interaction (e.g., a significant simple effect in one condition but an insignificant simple effect in the other condition). What can account for this apparent inconsistency? Moreover, if an interaction is hypothesized, is it appropriate to proceed with the simple effects tests even though the interaction term fails to achieve significance in the ANOVA? I have seen this done in published articles and believe that it is appropriate when a priori expectations exist, but I have also heard others (including reviewers) argue that simple effects tests are appropriate only if the ANOVA interaction term is significant.

---

Editor: I see two elements in this question—namely, (a) Why might a simple effect be significant when the overall interaction was not?, and (b) Must the overall interaction be significant to conduct tests of the simple effects?

To answer to the first question, consider the apparent interaction depicted in Figure 1. One's $MS_{\text{error}}$ term may be large enough that the interaction would not be significant. The simple effect of Factor A at Level $b = 1$ may also be insignificant. However, a test of the simple effect of Factor A at Level $b = 2$ could be significant. In the assessment of the overall interaction, which is, after all, an aggregate of these components, the significant simple effect may be washed out by the insignificant one.

It would be a superiorly clean result if both simple effects were significant—for example, either a qualitative difference in that the slopes were going in opposite directions, or a quantitative difference with both slopes going in the same direction, but one being steeper than the other. Either of these situations capture what is meant intuitively by an "interaction" (that the effect of one factor is contingent on the other). When one slope, or difference between means, is significant, but the other is not, the researcher is left to conclude that one half of the data are interesting, and one half yield the philosophically and statistically troublesome null result.

Why is such a result, like that depicted in the figure, suboptimal? The analysis of "simple effects" breaks down sums of squares, not just those attributable to the $A \times B$ interaction, but also the $A$ main effect for the simple effects of $A$ at each level of $B$ (or the $B$ main effect for the simple effects of $B$ at each level of $A$). Most researchers will test all four simple effect combinations—$SS_{A@b1}$, $SS_{A@b2}$, $SS_{B@a1}$, $SS_{B@a2}$—and then write up the two that illuminate the theorizing most clearly. Let us say we focus on the simple effects of Factor A at each level of Factor B. The sum of those simple effects' sums of squares, $SS_{A@b1} + SS_{A@b2}$, will equal not $SS_{A \times B}$, but $SS_A + SS_{A \times B}$ (cf. Keppel, 1991, pp. 241–242). The interaction in the typical 2 × 2 design has a single degree of freedom. Each simple effect also has one numerator degree of freedom. Testing two simple effects then uses two degrees of freedom, one more than may be derived from the interaction. The second degree of freedom is borrowed from the main effect of Factor A—that is, the simple effects will in part reflect the main effects; the simple effects do not reflect only the interaction. Therefore, for example, an $SS_{A@b1}$ effect is partly determined by the $A \times B$ interaction effect (which we want) and partly capitalizes on the $A$ main effect in the presence or even absence of an interaction (which is not good, and mathematically the only thing we can do about that is to verify that the interaction itself is significant).

An intuitive way to understand these relations is again by examining the plot. One pattern of data that is consistent with

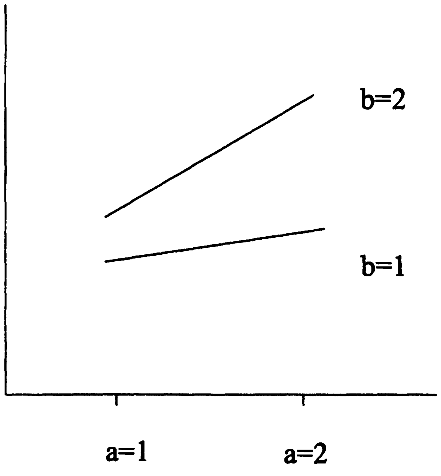


**FIGURE 1** Hypothetical two-way interaction plot of means.

the scenario under discussion would be if there was a strong main effect for Factor A. If there were, even in the absence of a significant interaction, the simple effect of $A$ at $b = 2$ could be significant. As depicted, the $a = 1$ mean is lower than the $a = 2$ mean, both for the main effect (i.e., collapsing over $b = 1$ and $b = 2$), and specifically also for the $b = 2$ case, the focal simple effect.

If one's data looked like those in the figure, and one concluded that there was a significant simple effect of Factor A at Level $b = 2$, but not for $b = 1$, it is somewhat misleading to suggest that there is an element of interaction in one's data. The significant simple effect states that the top line's slope is significantly different from (greater than) zero. However, the insignificant interaction indicates that it is nevertheless not significantly different from the slope of the line beneath it.

Abstractly, the second part of the question addresses the appropriate relation between an omnibus, overall $F$ test of a hypothesis and tests for follow-up comparison of means that comprise a part of that global hypothesis. Overall $F$ tests in our typical experimental world consist of two kinds—those for main effects and those for interactions. The statistics texts that address this issue tend to present the argument in the simpler, main effects context. We begin with main effects, but subsequently will address the case of interactions. One may expect that the argument for the interaction is a simple extension of that for the main effect, but stay tuned, because it is not.

Within the main effects context, we can further distinguish two classes of questions: What to do when the factor has two levels, and what to do when the factor has three or more levels. Even after exposure to the most elemental ANOVA material, researchers know that factors with two levels are mathematically simple; if a 2 × 2 factorial yields significant main effects for Factors A or B, there is no further analytical work to be done to understand the nature of those main effects. The researcher knows the null hypothesis, $H_0$: $\mu_1 = \mu_2$, is to be rejected, and taken together with the means on the two levels of the factor, the interpretation of such a finding is unambiguous. (The big mean is significantly larger than the small one.) Furthermore, if one were to conduct a "contrast" between the only two means available, the mean square for the contrast would equal exactly the mean square for the omnibus test of that main effect—the tests are identical and therefore redundant. Given that the extraordinarily vast majority of the experiments we conduct and report have factors with only two levels (usually 2 × 2s), this scenario is predominant, so the question of contrasts on main effects is usually moot. The situation is more complicated for a significant interaction in a 2 × 2 factorial because there are four means; the comparison of several pairs of which may be theoretically interesting. We turn to interactions shortly.

Consider now the situation of a factor with three or more levels. The null hypothesis tested is of the form, $H_0$: $\mu_1 = \mu_2 = \ldots = \mu_k$. Rejecting this null indicates that at least one mean, or a combination of means, is different from others, but follow-up contrasts are necessary to determine the nature of the mean differences (Scheffé, 1959, pp. 55, 66; Winer, Brown, & Michels, 1991, p. 140; cf. "A significant $F$ allows, if not demands, a further analysis of the data," Keppel, 1991, p. 111; "An overall $F$ test is often merely the first step in analyzing a set of data," Kirk, 1982, p. 90). Furthermore, we know that our follow-up analysis will be rewarding: "If an overall $F$ test is significant, an experimenter can be certain that some set of orthogonal[1] contrasts contains at least one significant contrast among the means" (Kirk, 1982, p. 95; Scheffé, 1959, p. 71). Accordingly, we operate by the decision rule: If my $F$ is significant, I must conduct contrasts to compare the means more precisely.

We also operate by the obverse rule: If my $F$ is not significant, there is no point in conducting contrasts because I will not find any significant differences. For example, Keppel (1991) stated,

> With a nonsignificant omnibus $F$, we are prepared to assert that there are no real differences among the treatment means and that the particular sample means we have observed show differences that are reasonably accounted for by experimental error. We will stop the analysis there. Why analyze any further when the differences can be presumed to be chance differences? (p. 111)

If there would seem to be a one-to-one relation between a significant overall $F$ and the presence of at least one significant contrast (albeit sometimes not the one we want), how do we reconcile this intuition with statements that say we can do "planned" comparisons whether the overall $F$ is significant? Consider these three exemplars: (a) "In some circumstances ... comparisons are undertaken following a significant $F$ test conducted on the entire experiment. In other circumstances, they are conducted instead of the overall $F$ test, in which case they are often referred to as planned comparisons," or planned "comparisons can be conducted directly on a set of data without reference to the significance or nonsignificance of the omnibus $F$ test" (Keppel, 1991, pp. 110, 112); (b) "It is not necessary to perform an overall test of significance prior to testing planned orthogonal contrasts" (Kirk, 1982, p. 96); (c) "Tests may ... arise because of the particular hypotheses the experimenter has which (s)he wants to evaluate. Those hypotheses can be evaluated with or without the overall analysis of variance" (Winer et al., 1991, p. 141).

We know that a main effect for Factor A has $(a - 1)$ degrees of freedom, and so, if we tested a set of $(a - 1)$ orthogonal contrasts, $\psi_1, \psi_2, \ldots, \psi_{a-1}$ (and there would be a multitude of possibilities, some sets being more interesting theoretically than others), the sum of these contrasts' sums of squares, $SS\,\psi_1 + SS\,\psi_2 + \ldots + SS\,\psi_{a-1}$, would equal $SS_A$. (Of course, not all

[1] Do not be overly concerned with the adjective *orthogonal* here, because, as Kirk (1982) acknowledged, any contrast "can be expressed as linear combinations of those in an orthogonal set" (p. 93).

the interesting research questions involve orthogonal contrasts, cf. Kirk, 1982, p. 95, but this assumption helps to simplify this discussion.) For some sets of $(a - 1)$ $\psi_i$s, we could order them from largest to smallest and know that if $MS_A$ was significant, so too will at least the largest $MS\ \psi$ be, as well as perhaps others. On the other hand, it could be that we have a (barely) significant $MS_A$, but that among the $\psi_i$s we have our heart set on, even the maximum $SS\ \psi$ may not exceed the critical value. This relation may be seen by considering the inherent power of the main effect test, wherein all the data from all the levels of Factor A are lent to the test, compared with that of the one degree of freedom contrast. For example, imagine our design is 5 × 2, with the typical 10 participants per cell. The degrees of freedom to test $MS_A$ are $(5 - 1)$ and $5*2*(10 - 1)$, for a critical $F_{\alpha=.05} = 2.53$. The degrees of freedom to test $MS\ \psi$ are 1 and 90; $F_{\alpha=.05} = 4.00$. That is, the main effect test may detect some variation from the null, but we may have to search a while before finding it in one contrast or two. This result suggests to me that even though we are given license to test a contrast without testing the overall $F$, I would always be interested in the diagnostic information of the overall $F$, particularly in case my prior guesses or hypotheses had been wrong (perhaps this has never happened to you?)—the data in this case would be indicating that although my pet theory was not supported, something else interesting may be found. (And, in the "what is the big deal" category, no one computes the overall or contrast $F$s by hand, and even if one was only intent on studying and reporting the contrast $F$s, the statistical computing packages all produce the overall $F$s in the ANOVAs, so why would one ignore complete information?)

We have seen that if an overall $F$ is significant, then at least the maximum contrast will achieve significance. What we have not seen yet is a scenario for which a contrast may be significant and the overall $F$ not significant. That is because this scenario is logically and statistically impossible, unless one begins to compare apples to oranges. Two examples cf where the overall $F$ test is an apple and the follow-up contrasts are oranges are these: First, Keppel (1991, p. 124) discussed planned comparisons for which only a subset of the data are used to compute an error term for each contrast—namely, only those cells being compared contribute to the temporary error term (also, see his response to the question on identifying appropriate mean square error [$MSE$] terms in this special issue). The argument is that the within-cell variability may be smaller for the focal cells, so that the fuller, aggregate $MS_{error}$ term provides insufficient sensitivity to detect the contrast. Thus, the overall $F$ test would not be significant, whereas the test of the contrasts against a presumably smaller error term would be more sensitive and could yield significance.

However, note that the variances would need to be fairly heterogeneous to benefit from this adjustment (not to mention reviewers going bonkers); for example, in our 5 × 2 example with 10 participants per cell, we saw the $MS_A$ would be tested on 4 and 90 degrees of freedom, for an $F_{\alpha=.05}$=2.53. If we conducted a contrast on $a = 1$ versus $a = 2$, and created a temporary $MS_{error}$ term involving only the data in those cells, the degrees of freedom would be 1 and $2*2*9$, for an $F_{\alpha=.05} = 4.17$. In any event, this sort of test is changing slightly the setting in which the contrast is tested from that in which the overall $F$ was tested.

A second example of apples to oranges is due to the numerous comparison test procedures, each of which is designed to detect somewhat different alternative hypotheses (i.e., noncentrality parameters; e.g., Hochberg & Tamhane, 1987; Klockars & Hancock, 1992; Seaman, Levin, & Serlin, 1991; also see Wahlsten, 1991). For example, Scheffé's follow-up test uses the $F$ distribution, whereas Tukey's test (for comparing pairs of means) uses a different distribution to obtain critical values (Scheffé, 1959, p. 73). Scheffé (p. 67) said that if cell sample sizes are equal, and you really only want to compare means in pairs (e.g., $\mu_i - \mu_j = 0$, not, e.g., $\mu_i + m_j - 2\ \mu_k = 0$), then Tukey's test provides greater precision (narrower confidence intervals). Thus, it may be possible to have a significant overall $F$; a significant Scheffé test of some complicated combination like $\mu_i + \mu_j - 2\ \mu_k$ (i.e., the average in conditions $i$ and $j$ differs from k); no significantly different pairs of means using the Scheffé test, but significantly different pairs of means using the Tukey test; or, an insignificant overall $F$; no significant Scheffé follow-up tests; and a significant Tukey test.

(It is even conceivable that the computation of contrasts on portions of one's data rather than on the fuller data set had a simple sociological explanation, e.g., the computation of such statistics by hand in the days before computers. Even so, this situation no longer applies, so lingering practices should also be updated.)

There remains to be discussed the issue regarding the alpha level of contrast tests that are a priori or planned versus comparisons that are *a posteriori* or post hoc (Hays, 1988, pp. 384–385; Kirk, 1982). Most recommendations allow each planned comparison to be tested at the .05 error rate, relying on the logic that the hypothesized contrast was constructed before the researcher saw the data, so there would have been no possibility for capitalization on chance. Note the maximum numbers of contrasts are still constrained by their degrees of freedom (e.g., $a - 1$ for $A$, etc.), because otherwise, the researcher could delineate a very long list of contrasts and claim them all as planned, for who among us is not clever enough to create some rationalization if need be.

Recommendations usually urge a correction on the .05 error rate for post hoc contrasts, usually of a Bonferroni form in which .05 is divided by the number of tests run for Factor A (and then .05 is adjusted separated for the family of tests run on B, etc.; cf. Holland & Copenhauer, 1988). Alternatively, one could use the macho (statistically conservative) Scheffé (1959) test, which permits testing of any sort of post hoc comparison (e.g., pairs of means or more complicated combinations) and "is applicable only in a situation where a preliminary overall $F$ test for the treatment has shown significance. If such a test has shown significance, then there exists at least one comparison for which the null hypothesis will be rejected at the same level of significance" (Minium, 1978, p. 420; Scheffé, 1959, p. 70). Given the conservative nature of the Scheffé test, statisticians classify it as an alternative to the Bonferroni correction on alpha (i.e., if one is using a Scheffé test, one need not also correct for the number of tests on the alpha rate; Hays, 1988, pp. 414–415).

My theoretician colleagues see statistics as a tool, not a master, and hold the data to the standard that a working hypothesis is to be judged superior, at least for the moment, if it is more parsimonious and explains more of the data patterns than any competitive theory. Certainly, and one key modifier, as I think they would agree, is the phrase "for the moment." Significance tests are indexes to aid the researcher in understanding what effects are likely replicable. If a post hoc analysis temporarily yields a better theoretical explanation, yet is not replicated with repeated, concerted efforts, then one must deduce that the result had been a sampling blip and that the formerly superior theory may need to be replaced with its apparently less elegant predecessors. It would be interesting to explore in greater detail the intersections of philosophy and statistics, given their common heritage in logic and mathematics. Just as we should pay attention to substantive theories, we must also play by the rules of statistical theories: Many statisticians are unforgiving of post hoc tests unless their foundational assumptions are met, like the corresponding overall omnibus tests being significant—for example, "What one cannot do is to attach an unequivocal probability statement to such*post hoc* comparisons, unless the conditions underlying the method have been met" (Hays, 1988, p. 418). Accordingly, I do not think it hurts to be a little cautious—namely, statistically conservative—in post hoc tests by correcting alpha or using the Scheffé (1959) test (cf. Cliff, 1983, pp. 123–124).

Finally, let us return to the question of the application of this logic to the heretofore seemingly analogous overall *F* test for an interaction and the follow-up simple effects. Given the answer previously to (a), wherein it was described that simple effects break down not just the interaction $\mathrm{SS}_{\mathrm{A}\times\mathrm{B}}$, but also the sums of squares for one of the main effects, it would be inappropriate to test and report simple effects in the absence of a significant overall *F* test for that interaction. We have seen that a simple effect can be significant driven by an interaction or a main effect. The autonomy in the main effect case between the overall *F* and the follow-up contrasts is due to the fact that all the contrast sums of squares are component portions of the sums of squares for the main effect—that is, the components sum up to the whole. By comparison, the tests of an overall interaction and its simple effects cannot be viewed as similarly autonomously; if a simple effect is significant, one hypothesis is that the interaction is significant, but for many data patterns, an alternative explanation—a completely reasonable and likely plausible competing theory—is that there is no interaction, only a main effect. Consider Keppel's (1991) words on the issue:

> Occasionally you will see simple effects tested with no mention of the statistical test of interaction. The typical pattern of results will consist of two simple effects—one that is significant and one that is not—and the researcher concludes that interaction is present. The trouble with this argument is that the formal test of interaction assesses the differences in simple effects. An inference of interaction that is based on the test of simple effects alone does not provide this necessary information ... A simple effect of Factor A does not reflect "pure" interaction but is also influenced by the average or main effect of the fact. ... Thus, testing the simple effects alone does not provide an unambiguous picture of interaction. (pp. 244–245)

However, let us look at a few important exceptions, occasions for which one may proceed to interpret simple effects unambiguously. Recall the basic equations that testing a simple effect of Factor A at $b_1$ or $b_2$ breaks down the $SS_{A\times B}$ and $SS_A$, and that testing a simple effect of Factor B at $a_1$ or $a_2$ breaks down the $SS_{A\times B}$ and $SS_B$. In Figure 2, we see a pattern of means that is consistent with there being no significant main effect for *A* but a significant main effect for *B*. To say $SS_A$ is not significant is to say it is statistically not greater than zero. If so, then the simple effects of Factor A at $b_1$ or $b_2$ may be attributed more certainly to the interaction, given that the $SS_A$ term is statistically zero. Note, though, that the interpretation of the simple effects of Factor B at $a_1$ or $a_2$ is still ambiguous due to the significant main effect for *B*. Also, note that the $SS_A$ will not be precisely zero; if the main effect for *A* is approaching significance, it would be conservative to not interpret the simple effects of *A* at $b_1$ or $b_2$ either.

Figure 3 presents the converse situation: the main effect fo·· *A* is significant and for *B* it is not. Given that $SS_B$ is statistically negligible, the interpretation of the simple effects of

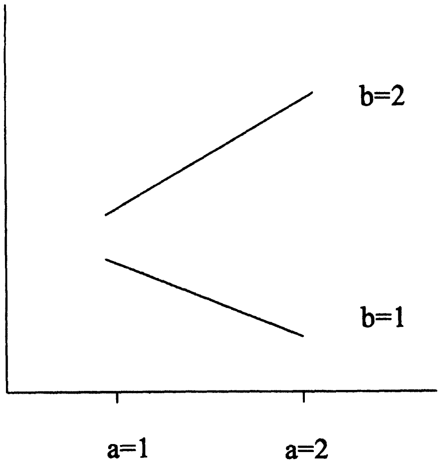


FIGURE 2 Now, imagine $SS_A$ was not significant and $SS_B$ was as in the figure.

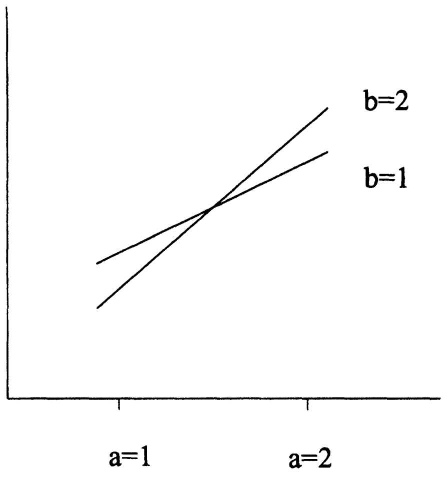


FIGURE 3 Now, imagine $SS_B$ was not significant and $SS_A$ was as in the figure.

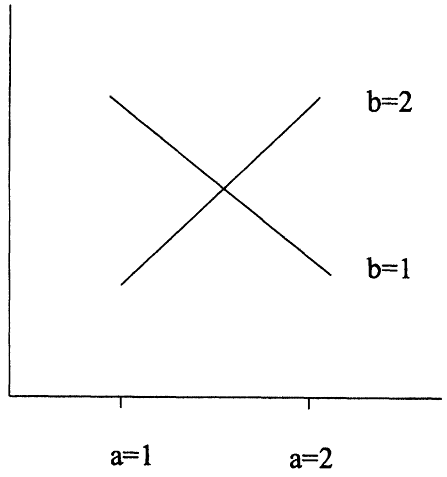


FIGURE 4 Finally, the result if neither $SS_A$ nor $SS_B$ were significant.

Factor B at $a_1$ or $a_2$ are interpretable unambiguously as attributable to the interaction. The simple effects of Factor A at $b_1$ or $b_2$ are, of course, still problematic due to the significant main effect for *A*.

Finally, Figure 4 depicts a scenario for which neither main effect is significant. In cases like these, $SS_A$ and $SS_B$ are both statistically zero, so "have at it." Either set of simple effects, those that examine the effect of Factor A at $b_1$ or $b_2$ or those that examine the effect of Factor B at $a_1$ or $a_2$, are interpretable unambiguously.

In conclusion, with regard to testing contrasts on main effects without testing the omnibus main effect hypothesis, there is some difference of opinion among experts. For the vast majority of the behavioral research in our field, the debate is moot because most of our factors have only two levels. (My personal approach, had I a factor with three or more levels, would be to examine full information—the omnibus main effect, as well as a priori and post hoc contrasts; why would we neglect certain aspects of understanding our data?)

Regarding the testing of simple effects, clearly the preference is to do so after having demonstrated a significant interaction. If the interaction itself is not significant, one must proceed carefully. If the main effect of *A* is significant and *B* is not, one may conduct the tests of $SS_{B@a1}$ and $SS_{B@a2}$, but not $SS_{A@b1}$ or $SS_{A@b2}$. If the main effect of *B* is significant and *A* is not, one may conduct the tests of $SS_{A@b1}$ and $SS_{A@b2}$, but not $SS_{B@a1}$ or $SS_{B@a2}$. If neither main effect is significant, any of the four simple effects is defensible. Finally, there is no debate among the experts that proceeding to test simple effects without demonstrating a significant interaction in the presence of both main effects being significant is indefensible.

## REFERENCES

Cliff, Norman. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research, 18,* 115–126.

Hays, William L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart, & Winston.

Hochberg, Yosef, & Tamhane, Ajit C. (1987). *Multiple comparison procedures.* New York: Wiley.

Holland, Burt S., & Copenhauer, Margaret DiPonzio. (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin, 104,* 145–149.

Keppel, Geoffrey. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Kirk, Roger E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Klockars, Alan J., & Hancock, Gregory R. (1992). Power of recent multiple comparison procedures as applied to a complete set of planned orthogonal contrasts. *Psychological Bulletin, 111,* 505–510.

Minium, Edward W. (1978). *Statistical reasoning in psychology and education* (2nd ed.). New York: Wiley.

Seaman, Michael A., Levin, Joel R., & Serlin, Ronald C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin, 110,* 577–586.

Scheffé, Henry. (1959). *The analysis of variance.* New York: Wiley.

Wahlsten, Douglas. (1991). Sample size to detect a planned contrast and a one degree-of-freedom interaction effect. *Psychological Bulletin, 110,* 587–595.

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

The issues of a priori versus post hoc tests and logics are intriguing, so I invited two of my philosophically-inclined colleagues to speak to the issue, and their response follows.

Professors Alice Tybout and
Brian Sternthal
Northwestern University

To address the a priori versus post hoc explanation issue, consider the following situation: A study is conducted in which the outcomes predicted from theory are not confirmed by the data. The investigator develops an alternative explanation for the data that appears to provide a compelling account of the findings. No rival explanations can be thought of that account for the data as well as the investigator's theorizing. Does this represent a rigorous test of theory?

Intuitively, it would seem that a test of theory is more rigorous if the phenomena observed are predicted prior to conducting the research rather than if an explanation is fashioned after the outcomes are known. Accurate prediction is viewed as a formidable task because the researcher must anticipate outcomes that are unseen and that may be influenced by poor measurement and extraneous events. In contrast, post hoc analysis is often pejoratively referred to as a fishing expedition that is simplified by the availability of data at the time the explanation is being constructed. The concern is that the explanation offered post hoc is likely to be particular to the data at hand and unlikely to be confirmed in subsequent tests. The conventional wisdom is that when explanation is post hoc, another test in which the theory's predictions are confirmed is required for the test to be rigorous.

The view that prediction offers a more compelling test of theory than post hoc explanation thus rests on the notion that prediction is a more difficult task. Even if this were the case, it is inappropriate to equate task difficulty with the rigor of a theory test. The rigor of a theory test is determined by whether one theoretical view offers a more parsimonious explanation for the findings than do its rivals (Sternthal, Tybout, & Calder, 1994). An explanation is preferred if it accounts for the observed outcomes with fewer concepts than other explanations, or if it can explain more phenomena with the same number of concepts as its rivals.

Defining rigor in terms of parsimony has an important implication with regard to prediction and post hoc explanation. It implies that when an explanation is conceived in relation to data, collection is immaterial. For the moment, the best available account for this and other relevant data has been identified. Another study may be conducted to test the predictions a priori, but if the theory's predictions are simply confirmed, the study would not represent further theoretical progress. Given the limitation of induction, another study would not increase confidence about the adequacy of the theory or the rigor of the test.

The view that post hoc explanation advantages the researcher is fallacious on other grounds. First, if the explanation is derived after the data are collected, the inclusion of experimental controls that would be useful in distinguishing a preferred explanation from its rivals are less likely to have been anticipated and administered than if the theory test were based on prediction. Thus, finding a unique explanation for the data despite this impediment would seem impressive.

Furthermore, it should also be noted that the seeming advantage of developing an explanation post hoc is offset by the fact that for an explanation to be parsimonious and for a test to be rigorous, the explanation must account for relevant data beyond that reported by the investigator. As Brush (1989) noted, in this case postdiction is even more impressive than prediction because it offers a better account of data in a situation in which rivals have had an opportunity to be formulated and tested. In contrast, prediction makes it far more likely that later views will be found to offer more parsimonious explanations than the currently preferred theory.

The focus on parsimony as the criterion for the rigor of a theory test circumvents the problem attendant to when an explanation is conceived. Post hoc explanation is typically suspected when predictions are counterintuitive or complex. In this situation, the investigator is often challenged to document that the outcomes obtained were predicted and, regardless of the reality, may have difficulty doing so. By focusing on parsimony, the thorny issue of when explanation was conceived becomes immaterial.

In developing a post hoc explanation, how well the data are explained by alternative theories is examined. If two or more views offer equally parsimonious explanations for the data, additional research is needed if the test of theory is to be rigorous. If the data in this and related studies are explained more parsimoniously by one theory than another, the test is rigorous and the research is in some sense complete.

## REFERENCES

Brush, Stephen G. (1989). Prediction and theory evaluation: The case of light bending. *Science, 246,* 1124–1129.

Sternthal, Brian, Tybout, Alice M., & Calder, Bobby J. (1994). Experimental design: Generalization and theoretical explanation. In Richard P. Bagozzi (Ed.), *Principles of marketing research* (pp. 195–223). Cambridge, MA: Blackwell.

## I.B. CAN I MAKE CONCLUSIONS BASED ON AN INTERACTION WITHOUT TESTING THE SIMPLE EFFECTS?

A researcher predicts a crossover interaction between two variables (one continuous, the other dichotomous) that is tested via a linear contrast regression model containing the appropriate main effect and interaction terms. The interaction term receives a significant coefficient. The researcher then concludes that the data support the predicted interaction. Is this an appropriate conclusion? More specifically, given that a significant interaction could occur for data patterns that differ from the one predicted (e.g., a noncrossover pattern, a crossover pattern in the opposite direction), is it not necessary to undertake the appropriate simple main effects tests to establish whether the data actually support the differences predicted by the crossover interaction?

---

Editor: Yes, it is imperative to defend a statement purported to describe data with a statistic. The scenario you have described for regression would be like obtaining a significant $F$ test for the $MS_{A \times B}$ interaction in ANOVA and then doing no further investigations—that is, both the plot of the cell means and the tests of simple effects to substantiate claims about precisely which means are significantly different from each other. You are absolutely right that a mere significant regression coefficient associated with an interaction term yields no detailed information about the nature of that interaction. If you see someone trying to do this—make a claim about their data without a statistic to support that claim (regardless of whether in regression or ANOVA or anything else for that matter), nail them. (You may find Questions I.A. and II.H. in this special issue helpful also.)

## I.C. IDENTIFYING THE APPROPRIATE *MSE* TERM FOR AN *F* TEST

When you have a design involving multiple factors, what is the appropriate *MSE* to employ in estimating follow-up contrasts involving a subset of the overall design? For example, if I have two factors (*A* and *B*) with two levels each, and I want to run a contrast analyzing the simple effect of *A*, what *MSE* should be used? When all factors are between subjects in nature, a general recommendation I have seen is that the correct error term is that associated with the overall design. However, in some recent e-mails I have exchanged with Geoff Keppel (personal communication, 1997), he indicated that in the new edition of his book he will advocate using the *MSE* only from the data relevant to the contrast—that is, in the earlier example, using a one-way ANOVA based on data from participants exposed to the first level of *B* for one contrast, then a separate one-way using data from those exposed to the second level of *B* only.

A more complicated, but related, issue is when a mixed design is involved—for example, where *A* and *B* are between-subjects factors and *C* and *D* are within-subjects factors. In analyzing the simple effect of *A*, the problem here is that there is no *MSE* from the overall design. Rather, there are *MSE* terms for relevant subsets of the design (e.g., one for the subset of the design including only *A* and *B*, one that includes *A*, *B*, and *C*, etc.). I have seen an article recently (I cannot recall which one in particular) in which the *MSE* from the most relevant subset of the design is used. However, given that this may include data not entirely relevant to the contrast of interest, a more precise way to estimate the simple main effect of the variable of interest would be, again, to use a one-way ANOVA using an error term based only on data contributing to that effect.

Although more precise (although that may be debatable), a problem with running the contrasts using separate error terms is the lower power associated with this alternative relative to that for an error term based on a larger part of the design. Therefore, this is an interesting dilemma: more precision or more power?

Professor Geoffrey Keppel
University of California, Berkeley

Although this question focuses on the nature of the error terms for evaluating follow-up contrasts involving a subset extracted from a larger design, I would prefer to broaden this discussion to include the planned contrasts that usually are conducted in lieu of an overall or omnibus test.

*Between-subjects designs.* When all factors are between subjects (a between-subjects design), the frequent choice for the error term is the one based on the variances of the independent groups, pooled or averaged over all treatment conditions. This choice is appropriate provided that the group variances are homogeneous and, less important, that the scores are normally distributed for all groups. Under these circumstances, the overall error term provides a more precise estimate of the population variances for the treatment groups than error terms based on subsets of the experiment and with greater power due to the increased number of degrees of freedom associated with this estimate. When the assumption of homogeneous variances cannot be reasonably met, however, the average error term either overestimates or underestimates the error variability appropriate for any given subset of the data, the direction depending on the particular variances included in the subset of groups (e.g., see Keppel, 1991, pp. 123–124.) Under these circumstances, then, I suggest that the correct choice of error term is dictated by concerns for *precision,* which would be an error term based only on those groups included in the subset under consideration. On the issue of power, a researcher should include sufficient numbers of observations to provide acceptable power for these critical comparisons—comparisons that in effect should be viewed as separate experiments whose power is determined in part by the researcher's choice of the number of participants.

*Single-factor designs.* In a single-factor design, most planned and follow-up analyses would consist of comparisons between one treatment condition and another. With heterogeneity present, one might use the Welch test, which calculates the error term as a weighted average (with two groups and equal numbers of participants, the error term is simply an average of the two group variances), with its degrees of freedom adjusted somewhat, depending on the degree of heterogeneity present (see Keppel, 1991, pp. 124–128).

*Two-factor designs.* In a two-factor factorial design, the possibilities multiply. Analyses focusing on subsets of treatment groups may consist of comparisons between two means contributing to either of the two main effects, interaction contrasts that focus on different processes contributing to the overall interaction, simple effects that focus on the effects of one of the independent variables (IVs) at a particular level of the other, and so-called simple comparisons that usually contrast two of the treatment means making up any given simple effect. If variance homogeneity is a safe assumption, the error term for all of these analyses is the average group variance for the entire factorial design. On the other hand, if this assumption cannot be reasonably met, a researcher should consider calculating individual error terms for each analysis. One way to visualize these analyses is to block out or identify the groups contributing to a given analysis and use as the error term an average of those particular group variances. Again, the logic behind this approach is based on using a more precise estimate of error variance for the particular analysis involved. Concerns for power should be addressed in the design of the experiment and not be dependent on the small gain in apparent power afforded by the degrees of freedom associated with the error term from the overall analysis.

*Within-subjects designs.* Let us turn now to the specification of error terms in within-subjects designs—designs in which participants serve in all possible treatment conditions or treatment combinations. The assumptions associated with these designs are more complicated than those associated with between-subjects designs. Moreover, violations of these assumptions appear to be more common and more serious. Fortunately, many of these complications are safely solved by following an analogous procedure to the one I suggested for between-subjects designs when heterogeneity is present.

*Single-factor designs.* For the single-factor design, which I refer to as an $(A \times S)$ design, comparisons between two treatment means are evaluated with an error term based only on the data contributing to that particular analysis. To illustrate, a comparison between two means becomes in effect a miniature single-factor design—a (Acomp. × $S$) design in which "Acomparison" designates the factor represented by the two conditions being compared (see Keppel, 1991, pp. 356–361). The resulting $F$ is associated with one degree of freedom for the numerator (Acomp.) and $(a - 1)(n - 1) = (2 - 1)(n - 1) = n - 1$ degrees of freedom for the numerator (Acomp. × $S$).

*Two-factor designs.* The detailed analysis of the two-factor within-subjects design follows a similar procedure, covered in detail in Keppel (1991, pp. 468–478). Any analysis of the two main effects starts by combining or collapsing the data for each participant over the variable not involved in the analysis (over the levels of Factor B in the analysis of the $A$ main effect and over the levels of Factor A in the analysis of the $B$ main effect). For the purposes of this analysis, then, we can treat these combined data as if they were derived from an actual single-factor within-subjects design, using the miniature single-factor arrangements of the data described in the preceding paragraph—an (Acomp. × $S$) design for a comparison based on the $A$ main effect and a (Bcomp. × $S$) design for a comparison based on the $B$ main effect. The analysis of simple effects divides the factorial design into a number of single-factor within-subjects designs. For example, an analysis of the simple effects of Factor A produces a set of single-factor within-subjects designs, one $(A \times S)$ design for each level of Factor B. The error term for each analysis is based only on the data relevant for the analysis. For the analysis of a simple effect of Factor A, the error term is the interaction of Factor A with participants at a particular level of Factor B, whereas the error term for a simple comparison between two means contributing to a simple effect is an interaction based on this particular comparison and participants (Acomp. × $S$). The analysis of interaction contrasts is also based on miniature experiments—namely, a minimal two-factor withins-subjects design in which each factor represents a comparison between two means; the error term is the three-way interaction based on this miniature experiment (Acomp. × Bcomp. × $S$).

In summary, we can see an overall analysis strategy emerging from this discussion. If the homogeneity assumption for the between-subjects designs and the additional assumptions for the within-subjects designs are in question, the safest course of action is to avoid error terms that include estimates of error variance that are not directly relevant to the analysis under consideration. Specialized error terms for either type of design are derived from miniature designs formed by extracting the relevant data from the overall design and basing the analysis on this restricted set of data.

*Mixed factorial designs.* The analysis of mixed factorial designs in which both types of IVs, between-subjects and within-subjects, are present further complicates the specification of error terms for analytical comparisons and analyses, when there is reason to believe that the homogeneity assumptions are violated, which, in my opinion, is a highly likely possibility. Under these circumstances, you might prefer to seek out detailed expositions available in books, such as the two chapters I devoted to the discussion of these sorts of analyses within the context of the mixed two-factor design (Keppel, 1991, chaps. 17 and 18), which provide extensive elaboration and worked numerical examples. For this relatively short comment, however, I simply note that the justification for the error terms I recommend in those chapters are extensions of the principles discussed here for the analysis of pure between-subjects and within-subjects designs. To be more specific, you first isolate that portion of the data matrix that contributes directly to the specific analysis you are contemplating. Doing so guarantees that you will base your estimates of error variability on the relevant portion of your data matrix. Second, you should determine the nature of the type of design that is reflected after you have isolated the relevant data matrix. Let us see how this works out for the analysis of the two sets of simple effects and of an interaction contrast, using the two-factor mixed design as an example.

### *Simple effects involving the between-subjects factor.*

Suppose you are considering the analysis of the simple effects of the between-subjects factor (Factor A) at Level $b1$. Once you have identified the relevant portion of the data matrix, you should be able to see that the resulting design is a single-factor between-subjects design and analyze the data accordingly—that is, use for the error term the average within-cell variance for the different $A$ conditions at $b1$. This is a between-subjects design because each participant supplies only one observation for this particular analysis. If the simple effect is significant, you would probably consider conducting meaningful comparisons between means that are contributing to this significant effect. Because this subset of data is in effect a single-factor between-subjects design, you would proceed as I suggested in the first section of this note if the cell variances are not reasonably homogeneous—namely, by calculating separate error terms and adjusting the degrees of freedom for the error term. You would repeat this entire process with the subsets of data representing the remaining levels of Factor B (for a numerical example, see Keppel, 1991, Section 17.5).

### *Simple effects involving the within-subjects factor.*

Let us consider next the analysis of the simple effects of the repeated- or within-subjects factor at the different levels of Factor A. Starting with Level $a1$, you would identify the relevant portion of the data matrix, which you will find is a single-factor within-subjects design—a $(B \times S)$ design at Level $a1$. From this point on, you would follow the procedures I described for analyzing an actual single-factor within-subjects design. In this case, the error term for the simple effect of Factor B at Level a1 is the $B \times S$ interaction obtained from this data subset. The error term for each comparison between means you conduct would be based on that portion of this data subset that is relevant to this particular comparison. You would repeat this entire process with the subsets of data representing the remaining levels of Factor A (for a numerical example, see Keppel, 1991, Section 17.6).

*Mixed interaction contrasts.* Finally, I briefly describe the analysis of interaction contrasts. Assuming that the two interacting contrasts (Acomp. and Bcomp.) consist of comparisons between two means, the resulting design is a mixed 2 × 2 design extracted from the larger, original mixed factorial. Once you have isolated the relevant portion of the data for this analysis, you should realize that the error term for the interaction contrast is the interaction of the contrast with participants (Acomp. × Bcomp. × $S$; for a numerical example, see Keppel, 1991, Section 18.3).

*Higher order factorial designs.* A discussion of the detailed analysis of higher order factorial designs, such as the four-factor design described by the poser of this question, would require a detailed and lengthy answer and, perhaps, a numerical example to render the answer comprehensible. On the other hand, it should be possible to divine a strategy by considering what we do with simpler designs and then extending that strategy to the more complex ones. The strategy I have been advocating is one of isolating the portion of the data directly relevant to the question being asked and then analyzing this subset as if it were a separate experiment—for example, an analysis of the simple effects in a two-factor design as separate one-way designs. Offhand, I do not see why this strategy could not be applied to complex designs.

## REFERENCE

Keppel, Geoffrey. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

---

Editor: If you obtain a significant interaction between Factors A and B, having tested an $F$ statistic of the form, $F = MS_{A \times B}/MS_{error}$, where $MS_{error}$ is typically the $MS_{S/AB}$ (in Keppel's, 1991 notation), where participants are said to be nested in $A$ and $B$ because each participant is exposed to one and only one combination of levels of those factors—it is the "error term that is associated with the overall design," as phrased in the question—and you then wish to follow up that omnibus interaction $F$ test with simple effects of $A$ at each level of $B$ (or $B$ at each level of $A$), the mean squares for those simple effects must also be divided by the same $MS_{error}$ term if you want the statistical testing to be coherent; that is, if you do not find a significant interaction, you will not find any significant simple effect.

Bird and Hadzi-Pavlovic (1983) defined coherence in the context of a multivariate analysis of variance (MANOVA). If you do not reject an omnibus null in the MANOVA using a particular test statistic (e.g., Wilks's likelihood ratio test $\Lambda$, Roy's $R$, Pillai–Bartlett trace $V$, Hotelling–Lawley trace $T$), there will not be any significant multivariate contrasts to be found as long as you are consistent in examining the follow-up contrasts using the same choice of statistic (the $F$ based on $\Lambda$, $R$, $V$, or $T$). Thus, Bird and Hadzi-Pavlovic stated,

> In a coherent two-stage analysis, the overall test can be regarded as a theoretically unnecessary but practically useful

> starting point in the exploration of data: theoretically unnecessary because the overall test could be omitted without affecting the outcomes of tests on contrasts; useful in practice because a nonsignificant overall test indicates that there would be no point in carrying out follow-up tests. (p. 168)

Analogously, in the context of an ANOVA, Scheffé (1959, p. 67) described the relation between the overall *F* test and the follow-up contrasts among subsets of means using his multiple comparison procedure. These analytical situations are close analogs to the issue raised in the question.

The logic of testing contrasts (for main effects) or simple effects (the term for a contrast in an interaction) is that of further breaking down variance terms to try to understand what occurred in the data. Think of a Venn diagram with the largest circle representing $SS_{total}$; within that circle are the components due to the variability of $SS_A$, $SS_B$, $SS_{A \times B}$, and the remainder is $SS_{S/AB}$. A significant interaction indicates that the variability due to the $A \times B$ term is large relative to the within-cell variability, hence we call it significant. Simple effects are attempts to break down the interesting *A*, *B*, and $A \times B$ variabilities into even smaller units that make better sense theoretically—that is, the main effects and interaction sums of squares portions of the larger Venn diagram can be seen as comprised of units representing $SS_{A@b1}$ and $SS_{A@b2}$ (or $SS_{B@a1}$ and $SS_{B@a2}$; however you wish to slice the data further).[2] If you test the overall interaction against $MS_{S/AB}$, but test the simple effects against some other error term (a variant or portion of $MS_{S/AB}$), your overall test of the interaction and your simple effects will represent results that compare apples to oranges. You can almost always do something mathematically or statistically, but that does not mean it is particularly good logic.

It is not even persuasive to argue that computing an overall interaction *F* test ($MS_{A \times B}/MS_{S/AB}$) and then running a one-factor ANOVA on the $b = 1$ data to simulate the simple effect of *A@b1* (and then running another ANOVA on the $b = 2$ data to simulate *A@b2*) is more precise because only the sums of squares due to within-cell variability for the cells of focus in each test are contributing (i.e., $a_1b_1$ and $a_2b_1$ for the first test, and $a_1b_2$ and $a_2b_2$ for the second test). The only condition under which this could be more precise is if you believed the assumption of homogeneity did not hold for your data. For example, if the within-cell variabilities are tiny in the $a_1b_1$ and $a_2b_1$ cells, but large in the $a_1b_2$ and $a_2b_2$ cells, then the pooled overall error term $MS_{S/AB}$ would seem too big for the *A@b1* test and too small for the *A@b2* test.

Even then, two criticisms prevail:

1. We know that the univariate analysis of variance is rather robust to fairly disparate heterogeneity—we count on this robustness so much that we rarely investigate whether the assumption holds or not, because we know that it largely does not matter (i.e., we will still be led to the proper decision of rejecting the null hypothesis or not). If the variances are so different that the conditions of purported precision as just described held, the analyst has far bigger problems to worry about than whether the simple effects are run properly (with the overall error term) or as simulated within the single-factor ANOVAs.
2. As the questioner implies, an even bigger problem results from the fact that power is lost due to the new degrees of freedom being much smaller than for the overall $MS_{S/AB}$ term (usually $ab(n-1)$, now $a(n-1)$, etc.). It really is optimal to analyze the data properly, in the sense of representing the actual two-factor design and using all the data simultaneously to enhance one's power and accuracy. Sometimes researchers do the pseudo follow-up ANOVAs (e.g., parsing the data set into the $a = 1$ data and separately the $a = 2$ data), simply because they had not figured out how to coax the statistical computing package they used into computing simple effects. This can be a nontrivial pragmatic problem, but it is solvable (e.g., in Statistical Analysis System, or SAS, to test the simple effect of *A* at level $b = 1$ in a 2 × 2 factorial, include the statement, "contrast 'A@b1' a 1 –1 a*b 1 0 –1 0"; to test *A* at level $b = 2$, the statement, "contrast 'A@b2' a 1 –1 a*b 0 1 0 –1"; to test for the simple effect of *B* at $a = 1$, include "contrast 'B@ a1' b 1 –1 a*b 1 –1 0 0"; and to test *B* at $a = 2$, include "contrast 'B@a2' b 1 –1 a*b 0 0 1 –1.")

I had conditioned my first statement on the typical choice of $MS_{error}$ being $MS_{S/AB}$. It is important to remember that sometimes $MS_{\text{"error"}}$ takes on a form different from $MS_{S/AB}$. For example, if Factor A is a fixed-effects factor and Factor B is a random-effects factor, then the tests of the main effect of *B* and the interaction are each tested against $MS_{S/AB}$, but the test of the main effect of *A* is tested against $MS_{A \times B}$ (with corresponding loss of error degrees of freedom, from $ab(n-1)$ to $(a-1)(b-1)$). Most experiments are run with all factors as fixed, and so this issue does not usually arise—both main effects and the interaction are tested against the $MS_{S/AB}$ term; the overall error term as it is phrased in the question. (See Hicks, 1982, Iacobucci, 1995, and Jackson & Brashers, 1994, for the issues of fixed vs. random effects and the computation of expected mean squares [EMS], which will guide you to the correct error term. Other references useful with regard to experimental design include Box, Hunter, & Hunter, 1978; Cochran & Cox, 1957; Maxwell & Delaney, 1990.)

Regarding the issue in the second paragraph of the question, it is true that mixed designs can be more complicated to track which *MS* term serves as an error term for which *MS* effect terms; however, the basic logic of the simpler two-factor,

[2] The simple effects (of *A* at each level of *B*, $SS_{A@b1} + SS_{A@b2}$) break down the $SS_{A \times B} + SS_A$ (and the simple effects of *B* at each level of *A* break down $SS_{A \times B} + SS_B = SS_{B@a1} + SS_{B@a2}$). In essence, the Venn diagram components of $SS_{A \times B}$ and $SS_A$ are remerged and then reapportioned. The terms $SS_{A@b1}$ and $SS_{A@b2}$ should be tested against the same error term as were $SS_{A \times B}$ and $SS_A$. For more detail, see Question I.A. in this special issue.

between-subjects design as discussed thus far still holds. Hence, in the four-factor example given in the question (assuming all factors as fixed for simplicity), the $MS_{A \times B}$ interaction would be tested against $MS_{S/AB}$, and the $MS_C$ main effect would be tested against $MS_{C \times S/AB}$. Simple effects on the $A \times B$ interaction would also be tested against $MS_{S/AB}$, and contrasts on Factor C would be tested against $MS_{C \times S/AB}$. Therefore, although it is true that there is no one overall error term for such a design, what is important is that the error term used in each omnibus test (e.g., the overall $A \times B$ interaction, the overall $C$ main effect) is also consistently used as the error term when doing follow-up tests (e.g., simple effects within $A \times B$, contrasts within $C$, respectively).

In conclusion, determining the appropriate $MS_{error}$ term for an $F$ test depends on the EMS that may be derived as a function of whether the factors are fixed or random and the nature of the experimental design. (The usual case is that all factors are fixed, which leads all $MS_{errors}$ to be of the form $MS_{S/AB}$, the simple typical analysis.) If any factors other than participants are random, or the experimental design is not a straightforward factorial, consult design books like Box et al. (1978), Cochran and Cox (1957), Hicks (1982), Jackson and Brashers (1994), or Maxwell and Delaney (1990). (You may also find Question I.G. in this special issue of interest.) When one determines the error term for the test of the omnibus hypothesis, the same error term should be used to test the follow-up contrasts and comparisons.

## REFERENCES

Bird, Kevin D., & Hadzi-Pavlovic, Dusan. (1983). Simultaneous test procedures and the choice of a test statistic in MANOVA. *Psychological Bulletin, 93,* 167–178.

Box, George E. P., Hunter, William G., & Hunter, J. Stuart. (1978). *Statistics for experimenters.* New York: Wiley.

Cochran, William G., & Cox, Gertrude M. (1957). *Experimental designs* (2nd ed.). New York: Wiley.

Hicks, Charles R. (1982). *Fundamental concepts in the design of experiments.* New York: Holt, Rinehart, & Winston.

Iacobucci, Dawn. (1995). *Review of random factors in ANOVA. Journal of Marketing Research, 32,* 238–239.

Jackson, Sally, & Brashers, Dale E. (1994). *Random effects in ANOVA.* Thousand Oaks, CA: Sage.

Keppel, Geoffrey. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Maxwell, Scott E., & Delaney, Harold D. (1990). *Designing experiments and analyzing data.* Belmont, CA: Wadsworth.

Scheffé, Henry. (1959). *The analysis of variance.* New York: Wiley.

## I.D. POOLING ACROSS FACTORS

What is the appropriate basis for pooling across the levels of a factor in a multifactor design (e.g., pooling across $C$ in a design including Factors A, B, and C.) I have seen or heard of several recommendations, but each independently of the other (of course). One criterion is that pooling may be allowed if there is no theoretical basis for expecting the interaction between $A$ and $B$ to be qualified by $C$ (i.e., no $A \times B \times C$ interaction would be anticipated). A second, related criterion is whether, empirically, $C$ is found to interact with $A$ and $B$. A third criterion is suggested by MacKenzie, Lutz, and Belch (1986) that involves testing the equality of variances for all relevant dependent variables (DVs) across all levels of $C$—pooling is appropriate if no significant variances are detected. Given that one criterion can be met without the others (e.g., significant interactions may be detected even though none were anticipated conceptually, variances may be equal even though interactions involving $C$ are anticipated or detected), some consolidation or hierarchy of these criteria may be useful as a means of reconciling any potential conflicts.

## REFERENCE

MacKenzie, Scott B., Lutz, Richard J., & Belch, George E. (1986). The role of attitude toward the ad as a mediator of advertising effectiveness: A test of competing explanations. *Journal of Marketing Research, 23,* 130–143.

Editor: You are certainly right that the criteria you describe can yield conflicting results, leaving the researcher to ponder whether 'tis nobler to pool or not to pool. The first two of your stated criteria represent the classic interplay between theory and data. We usually put theory on a pedestal, in this application, saying that no matter what the data look like, we intend to collapse over levels of Factor C (although, then, one must wonder why Factor C was included in the experimental design). Doing so would be foolhardy if there is a significant $A^*B^*C$, $A^*C$, or $B^*C$ interaction; even a $C$ main effect suggests that the $A \times B$ design would be muddied a bit without recognizing the underlying more complicated factorial structure. The most practical danger of aggregating in the presence of any of these significant terms would be that the remaining $A \times B$ effects could be washed out or misrepresented (i.e., the bias that results from missing variables in a model).

An advantage of maintaining the $A \times B \times C$ structure in reporting is that it is a more valid representation of the actual study, and the significance tests should be more precise because the $MSE$ term has been parsed more finely (i.e., the effects of $C$, on its own and jointly with the other factors, have been removed). Of course, we might acknowledge that if an $A \times B \times C$ factorial were submitted to a journal, with no significant main effect for $C$ and no interactions involving $C$, surely the reviewers would pounce, wondering what was the point of (the probably exploratory) Factor C. Although collapsing over levels of $C$ may not retain the sensitivity of the $A \times B$ tests, given the $MSE$ issue, the power may still be strong due to the same effective sample size.

Empirics need to dominate theory in this decision. If there is a significant effect involving $C$, the fuller factorial ought to be represented because the data are essentially indicating that different things are happening depending on Factor C. This more complex effect indicates that the theorizing needs to be enriched to account for the empirically demonstrated contingencies. Of course, one of the ways we allow theory to dominate is to question the data—in this case, questioning the importance of the significant effects involving *C*; for example, "well, it is only a small effect"; "the results are qualitatively similar, it is just that *A* × *B* at *C1* does not achieve significance, whereas *A* × *B* at *C2* does"; or even, "there are a couple of significant effects involving *C*, but I do not know what to do with them, and the *A* × *B* effects that I wish to present are still significant even though I have collapsed the design in the analysis" (though, no, you do not get bonus points for the test being more conservative here), and so on. Rationalizations to oneself or others can be persuasive, as long as the researcher is not missing an opportunity to advance the theoretical explanation.

The MacKenzie et al. (1986, p. 135) criterion was a different kind of animal. They sought aggregation to have bigger sample sizes to enable them to execute structural equations models. (They tested for homogeneity of variances on each of the 12 measures they were proposing to model and did not find differences across conditions. Perhaps they could have used simultaneous, i.e., multivariate, tests to examine whether the 12 × 12 variance–covariance matrices were equal across the conditions. It is also interesting that, although we expect our manipulations to affect means, we expect variances, or in this case, covariances, as input to the structural equations models, to remain constant.) The focus of the aforementioned question seems to be on the appropriateness of collapsing across a factor in the context of analysis of variance, not structural equations models, and ANOVA is known to be fairly robust to violations of homogeneity of variance. That is, even if one were to apply the MacKenzie et al. (1986) criterion and find that the variances differed significantly across conditions, the subsequent ANOVA is less likely to be affected. Therefore, I do not really see their criterion as a third viable candidate for the typical data analysis scenario in ANOVA.

## I.E. UNEQUAL CELL SIZES (MISSING DATA)

How critical is the problem of unequal cell sizes (missing data), and what should be done about it? My experimental design professor in our educational psychology department says that ending up with unequal cell sizes gives the appearance that the researcher did not have complete control over the experiment, in addition to creating even more potentially serious concerns about differential dropout rates (experimental mortality) from different conditions and violating equal variance assumptions. However, despite our best efforts (e.g., signing up equal numbers of participants for each condition), not all participants show up for their assigned sessions, complete the task, or supply usable data. I have heard solutions such as randomly dropping one or two participants to even out cell sizes. Is this preferred to collecting more data to add to the size of the smaller cells? And, how should we report or describe this problem (and how we handled it) in writing up the results of the experiment?

---

Editor: It seems a bit harsh to me to make an attribution regarding a researcher's (lack of) control when participants do not show up for their assigned sessions or fail to yield data for other reasons. If it is any consolation, I think most researchers regularly experience the problems you have described.

It is certainly true that perfectly balanced designs are highly desirable; for a given sample size, equal cell sizes contribute to the overall power in detecting significant effects and to the robustness in detecting those effects in the presence of violated assumptions (e.g., nonnormality or heteroscedasticity; cf. Algina & Oshima, 1990; Hakstian, Roed, & Lind, 1979; Sawilowsky & Blair, 1992). Your solutions are commonly offered, with the usual counterobjections and caveats—it seems cost-ineffective and wasteful to drop participants, and in contrast, it may be effortful to gather additional data that now carry a possible confounding of having been collected later in time than the earlier, previously unbalanced data (e.g., Efron, 1994; Little, 1992; Little & Rubin, 1987; Searle, 1987).

The problem of unequal cell sizes is actually severe. In a 2 × 2 design, cell sizes of {10, 10, 10, 11} are clearly better than {10, 10, 10, 20} (a pattern with high variance) or {8, 12, 14, 6} (some cell sizes are quite small), for example, but strictly speaking, all three of these sets are unbalanced.

One of the issues is easy to see. For example, in the tiny data set of $N = 6$ observations depicted in the following, the cell means suggest there will be no significant main effect for Factor A, nor an interaction, but perhaps a main effect for *B*. The marginal means will vary depending on whether one computes them as means of cell means (e.g., $8 + 5/2 = 6.5$ for $a = 1$ and $a = 2$) or means of raw data (e.g., $(7 + 9 + 5)/3 = 7$ for $a = 1$ and $(8 + 4 + 6)/3 = 6$ for $a = 2$).

| data: | b=1 | b=2 |
|---|---|---|
| a=1 | 7<br>9 | 5 |
| a=2 | 8 | 4<br>6 |

| cell means: | b=1 | b=2 |
|---|---|---|
| a=1 | 8.0 | 5.0 |
| a=2 | 8.0 | 5.0 |

The bigger pickle actually comes in that the unbalance creates a problem analogous to multicollinearity in regression, in that the hypotheses to be tested are affected (e.g., Iacobucci, 1995; Milligan, Wong, & Thompson, 1987; Perreault & Darden, 1975).

Imagine writing those six data points in terms of the ANOVA model parameters ($Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$), or at least the first several of the model terms, to keep it simple.

To test for the main effect of Factor A, we would posit $H_0$: $\alpha_1 = \alpha_2 = 0$, or $\alpha_1 - \alpha_2 = 0$.

|  | b=1 | b=2 |
|---|---|---|
| a=1 | $7=\mu+\alpha_1+\beta_1$<br>$9=\mu+\alpha_1+\beta_1$ | $5=\mu+\alpha_1+\beta_2$ |
| a=2 | $8=\mu+\alpha_2+\beta_1$ | $4=\mu+\alpha_2+\beta_2$<br>$6=\mu+\alpha_2+\beta_2$ |

If we compared the mean for $a = 1$ to the mean for $a = 2$, where the means are computed as means of cell means, we would have

$$
\begin{aligned}
&\tfrac{1}{2}\,\{\tfrac{1}{2}[(\mu + \alpha_1 + \beta_1) + (\mu + \alpha_1 + \beta_1)] + (\mu + \alpha_1 + \beta_2)\}\\
&-\tfrac{1}{2}\,\{(\mu + \alpha_2 + \beta_1) + \tfrac{1}{2}\,[(\mu + \alpha_2 + \beta_2) + (\mu + \alpha_2 + \beta_2)]\}\\
&= \tfrac{1}{2}\,[\,(\mu + \alpha_1 + \beta_1) + (\mu + \alpha_1 + \beta_2)] - \tfrac{1}{2}\,[(\mu + \alpha_2 + \beta_1) +\\
&\quad (\mu + \alpha_2 + \beta_2)]\\
&= \tfrac{1}{2}\,(2\mu + 2\alpha_1 + \beta_1 + \beta_2) - \tfrac{1}{2}\,(2\mu + 2\alpha_2 + \beta_1 + \beta_2)\\
&= \mu + \alpha_1 + \tfrac{1}{2}\beta_1 + \tfrac{1}{2}\beta_2 - \mu - \alpha_2 - \tfrac{1}{2}\beta_1 - \tfrac{1}{2}\beta_2\\
&= \alpha_1 - \alpha_2
\end{aligned}
$$

which is precisely the hypothesis we wish to test.

If, instead, we used the means for $a = 1$ and $a = 2$, where they were computed as means of raw data, we would have

$$
\begin{aligned}
&(\tfrac{1}{3})\,[(\mu + \alpha_1 + \beta_1) + (\mu + \alpha_1 + \beta_1) + (\mu + \alpha_1 + \beta_2)]\\
&-(\tfrac{1}{3})\,[(\mu + \alpha_2 + \beta_1) + (\mu + \alpha_2 + \beta_2) + (\mu + \alpha_2 + \beta_2)]\\
&= (\tfrac{1}{3})\,(3\mu + 3\alpha_1 + 2\beta_1 + \beta_2) - (\tfrac{1}{3})\,(3\mu + 3\alpha_2 + \beta_1 + 2\beta_2)\\
&= \mu + \alpha_1 + (\tfrac{2}{3})\beta_1 + (\tfrac{1}{3})\beta_2 - \mu - \alpha_2 - (\tfrac{1}{3})\beta_1 - (\tfrac{2}{3})\beta_2\\
&= \alpha_1 - \alpha_2 + (\tfrac{1}{3})(\beta_1 - \beta_2)
\end{aligned}
$$

which is some interesting bizarro hypothesis, but not one we wish to test. We want to isolate the model terms purely—the $\alpha$s for the main effect of *A*, the $\beta$s for the main effect of *B*, and the $(\alpha\beta)$s for the interaction. The difference in the previous examples is picking up on some effect of *B* in addition to the focal effect on *A*—that is, we would say that the effect of *A* is biased by that of *B*. We need to adjust the effect of *A* to remove this contamination of Factor B. Other ways of stating this goal include the following: We want our effects to be orthogonal (i.e., the tests of one factor should be statistically independent of the other effects), we wish to examine the effects having been adjusted for the other effects, or we want the other effects to have been partialled out.

Thankfully, as nasty as the problem is, the solution is truly simple (and it is endorsed with fairly clear consensus across statisticians). The major statistical computing packages make available more than one type of sums of squares computation, and it is simply a matter of reporting the statistics (*SS*s, *MS*s, *F*s, and *p* values) associated with the proper type of computation that is testing the right set of hypotheses.

In SAS, the analyst should use the Procedure General Linear Model (PROC GLM; PROC ANOVA assumes perfectly balanced data). Within GLM, SAS computes four types of sums of squares and, by default, reports Types 1 and 3. It is the set of results listed under Type 3 that should be reported and interpreted. (It is actually good practice to obtain only the Type 3 sums of squares from SAS, so as not to get confused, by including the option SS3 on the model statement; e.g., "model depvar = *a b a*b/ss*3).

Thanks go to my colleague, Angela Lee, who demonstrated to me that a current version of Windows PC Statistical Package for the Social Sciences (SPSS) has a menu from which to select one of the four types of sums of squares, and the proper Type 3 is the correct default. However, check the documentation or help function, because depending on whether one is using a PC or mainframe–network package version of SPSS, and depending on which version is installed, the defaults vary and the types of sums of squares have different labels. If the type of sums of squares is not labeled "3," use those labeled "unique," "partial," or "regression," but not "sequential." In some versions of SPSS, the default for the MANOVA procedure is correct, but the default for the ANOVA procedure is not. (Be choosy with statistical packages—be wary of purchasing one that is user friendly with neat graphics with documentation that does not even address the issue of unbalanced designs—probably signaling that the programmers were unaware of the statistical issues—how user friendly is a package that gives the wrong results? "Hop." That was the sound of me getting off my soapbox.)

In conclusion, aim for balance—the mechanical effort of obtaining more participants will pay off statistically (in terms of power and robustness), but if the resulting data are not perfectly balanced, use SAS Type 3 sums of squares (or Type 3 in SPSS) to analyze them.

## REFERENCES

Algina, James, & Oshima, Takako C. (1990). Robustness of the independent samples Hotelling's $T^2$ to variance–covariance heteroscedasticity when sample sizes are unequal and in small ratios. *Psychological Bulletin, 108,* 308–313.

→ Efron, Bradley. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association, 89,* 463–479.

Hakstian, A. Ralph, Roed, J. Christian, & Lind, John C. (1979). Two sample T2 procedure and the assumption of homogenous covariance matrices. *Psychological Bulletin, 36,* 1255–1263.

Iacobucci, Dawn. (1995). The analysis of variance for unbalanced data. In David W. Stewart & Naufel J. Vilcassim (Eds.), *Marketing theory and applications* (Vol. 6, pp. 337–343). Chicago: American Marketing Association.

→ Little, Roderick J. A. (1992). Regression with missing x's: A review. *Journal of the American Statistical Association, 87,* 1227–1237.

Little, Roderick J. A., & Rubin, Donald B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Milligan, Glenn W., Wong, Danny S., & Thompson, Paul A. (1987). Robustness properties of nonorthogonal analysis of variance. *Psychological Bulletin, 101,* 464–470.

→ Perreault, William D., & Darden, William R. (1975). Unequal cell sizes in marketing experiments. *Journal of Marketing Research, 12,* 333–342.

Sawilowsky, Shlomo S., & Blair, R. Clifford. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111,* 352–360.

Searle, Shayle. (1987). *Linear models for unbalanced data.* New York: Wiley.

## I.F. DIRECTIONAL PREDICTIONS IN MEAN DIFFERENCES

When a directional prediction exists for an interaction in a 2 × 2 ANOVA, is it appropriate to divide in half the significance of the *F* or is it necessary to run a specific contrast?

Professor Joseph Verducci
Ohio State University

Yes. In the 2 × 2 case, the *F* test for an interaction is just the square of a *t* test, so it is okay to test for a positive interaction at the $\alpha = .05$ level by rejecting only if the estimated interaction effect is positive and the *p* value for the associated *F* test is less than .10.

Editor: Okay, but why? In a 2 × 2, if either main effect is significant, once you have the means, you know immediately which is larger or smaller than the other. Theoretically, it is the rare circumstance in which a researcher could argue persuasively that a hypothesis was truly one-tailed; if one expects $\bar{A}_1 > \bar{A}_2$, is it truly of no interest if instead $\bar{A}_2 > \bar{A}_1$? (It might be of great interest at least because evidently the a priori guess was inaccurate.) Statistically, if an effect is significant at a two-tailed level (.025), the test statistic will exceed the critical value at the one-tailed level (.05), so one-tail tests usually strike me as a move of desperation by the researcher. Finally, in testing simple effects (follow-up testing on interactions), it is not clear what a directional test would look like.

## I.G. SHOULD I TREAT A PARTICIPANT ATTRIBUTE (E.G., GENDER) AS A BLOCKING FACTOR?: PART 1

One question that sorely needs to be addressed is about a common misconception that I have encountered when reviewing for journals and serving on dissertation committees. Authors will often say something like, "A two-factor factorial design was used, with X levels of the first factor and two levels of the gender factor." Then, the authors will proceed to treat the experiment and analysis as a between-subjects, two-factor analysis design, instead of as a one-factor randomized block design. Often, the theory section has a prediction of an interaction, and they perform the test as such, which is clearly incorrect. I would like to see this matter be given attention. (Other things that are commonly done is to treat repeated measures as fixed effects; to discount the relevance of aberrations in structural equation models, such as Heywood cases, and to not use log-linear analysis in an appropriate manner; or to simply apply a probit model without explicitly testing for interactions among the predictors.)

Professor Joan Meyers-Levy
University of Minnesota

Assuming that authors who are making statements about two-factor factorial designs like the one you mentioned are studying theoretical issues like those that may pertain to gender differences, I see no problem in treating the analysis as a between-subjects, two-factor design rather than a one-factor randomized block design. Given that predictions have been made involving gender and the other factor, gender is not a factor that is introducing undesirable error variance that could obscure the observation of some other more focal (main) effect. Rather, any effects involving gender are an object of study and interest in their own right as the effect of the first factor is indeed anticipated to vary as a function of gender. Although Keppel (1973, p. 504; cf. Keppel, 1991) noted that the use of randomized blocks generally "results in an experiment that is more sensitive than the corresponding experiment without blocks" because the error term "reflects the variability of participants from populations in which variation in the blocking factor is greatly restricted," controlling the variance associated with gender does not serve an essential purpose here because the effects associated with all factors (including gender) are equally the focus of interest. In fact, one may argue that if the anticipated effect (e.g., an interaction involving gender) emerges despite the fact that the data have been analyzed as a two-factor design with the variance associated with gender allowed to vary unchecked, the observed effect could be viewed as all the more robust.

### REFERENCES

Keppel, Geoffrey. (1973). *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: Prentice Hall.

Keppel, Geoffrey. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Professor James Barnes
University of Mississippi

When treatment levels are combined such that each level of Factor A appears once with each level of Factor B, the treatments are said to be completely crossed, a characteristic of all factorial designs. Participants are then randomly assigned to one of the treatment combinations (Hicks, 1993; Kirk, 1995).

In the situation that you describe, participants cannot be randomly assigned to the different levels of gender (excluding surgical changes).

A randomized block design allows for differences (gender) before the experiment to be evaluated. A randomized block design enables the researcher to test two null hypotheses—namely, that treatment populations means are equal and that block population means are equal (Kirk, 1995).

The presence of block–treatment interaction effects alters the expected values of the mean squares producing a nonadditive model (see Table 7.3– in Hicks, 1993, p. 267). Furthermore, the research problem that you describe is either a fixed-effects model (Factor A and block fixed) or a mixed-effects model (Factor A random and block fixed). Thus, researchers must insure that the appropriate mean sum of squares is being used for the particular situation that is being examined.

A repeated measures design (each participant is observed under all, or a subset, of the conditions in the experiment) is a special case of the randomized block design. That is, each participant serves as his or her own blocking variable (Hicks, 1993; Kirk, 1995).

The general situations in which models of discrete DVs are relevant are cases in which the researcher is interested in choice behavior or the occurrence or nonoccurrence of an event. One is usually not interested in estimating the value or size of the DV but in analyzing the underlying probability of the given event or choice. Logit and probit have been used extensively in this situation. Both models require different assumptions concerning the underlying probability distribution function of the DV. As the name suggests, logit assumes the logistic function as the underlying distribution. Probit, on the other hand, assumes that the underlying probability distribution function is normal.

## REFERENCES

Hicks, Charles R. (1993). *Fundamental concepts in the design of experiments* (4th ed.). New York: Holt, Rinehart, & Winston.

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

Editor: Hmm. Well, it is true that gender is not typically randomly assigned. However, all the levels of both factors appear in all possible combinations—that is, the factors are crossed. It is also true that blocking is usually done to mechanically control for an extraneous factor, whereas in the question posed, gender is a theoretically focal factor. Therefore, perhaps a different tack is to determine when it actually matters statistically. There are two issues to address: First, are the factors random or fixed?; second, does one fit the full or reduced block ANOVA model?

Factors that characterize participants are usually thought to be random, but as Barnes notes here, gender would be considered fixed, because the entire population of levels of gender is captured with the two—the researcher has access to the entire population of levels of that factor. (If one's subject factor or potential blocking factor were something like citizenship or country of origin, and representatives truly were sampled randomly, one could make a case for the factor being random, with the possibility of generalizing to the participant population.) The *F* statistics for the main effects for *A*, *B*, and the *A* × *B* interaction in a 2 × 2 factorial (following Hicks, 1993, etc.) would be

1. If the subject factor, *B*, is fixed, as here,
   (a) and if Factor A were fixed, the usual case, $F = MS_A/MS_{S/AB}$, $F = MS_B/MS_{S/AB}$, $F = MS_{A \times B}/MS_{S/AB}$
   (b) if Factor A were random, $F = MS_A/\mathrm{MS}_{\mathrm{S/AB}}$, $F = MS_B/MS_{A \times B}$, $F = MS_{A \times B}/MS_{S/AB}$.
2. To be complete, if the subject Factor B were random,
   (a) and Factor A fixed, F = $MS_A/MS_{A \times B}$, $F = MS_B/MS_{S/AB}$, $F = MS_{A \times B}/MS_{S/AB}$
   (b) or Factor A random, F = $MS_A/MS_{A \times B}$, $F = MS_B/MS_{A \times B}$, $F = MS_{A \times B}/MS_{S/AB}$.

It is rare that researchers in our field make claims that they have selected the levels of their factors randomly. Thus, cases 1(b) and 2(b) are less prevalent. If we keep our focus on the question's subject variable of gender, then Factor B will be considered fixed also—that is, scenario 1(a) is relevant, not 2(a). The *F* statistics in circumstance 1(a) are the usual *F*s. The tests for *B* and *A* × *B* are the same. The tests for *A* differ; furthermore, 1(a) will be more powerful than 2(a).

Another issue arises in blocking designs regarding whether one should be fitting a full or reduced ANOVA model. If the blocking factor is a nuisance variable—one that is expected to yield differences on the DV, yet is not of focal interest—one groups respondents into homogenous blocks and then distributes those respondents within each block across the experimental conditions. For example, if the DV of interest is memory (recall or something), a potential blocking factor may be intelligence. (Smarter people are likely to recall more, regardless of the experimental treatment, but this finding would not be a great theoretical insight.) Blocks would be formed in which similarly intelligent people (as measured by some indicator variables) would be distributed across levels of the experimental factor—a portion of bright respondents would go into Levels a1, a2, and so on. A portion of moderately bright respondents would similarly be assigned to Levels a1, a2, and so on. A portion of the low wattage group similarly would be assigned across Levels a1, a2, and so on. At this point, the nuisance variable of intelligence is approximately equally represented across the groups, with the important implication that the A(factor)*B(intelligence) interaction is assumed to be zero. Highly intelligent people would be expected to remember more, regardless of the experimental treatment to which they were exposed; the same applies for the other groups. Therefore, in condition a1, there will be a group of highly intelligent people who recalled the most, a group of less intelligent people who recalled the next most, and so on; and these comparative results would be true for a2, and so on. The question regarding the effectiveness of the experimental intervention is simply whether the overall performance (of all three groups) was better in a1, or a2, and so on, regardless of the individual differences caused by intelligence, which hopefully have been made uniform across the conditions.

This scenario is not that which describes the gender research in our journals. In gender research, the question is in fact whether men and women (or boys and girls) will react differently in different treatment conditions—that is, the assumption is not made that men (or women) will behave the same under all conditions. Perhaps men or women will be more or less sensitive or affected by exposures to stimuli of different sorts. Thus, the theoretical questions revolve around the interaction between the Factor A and the subject property B. Why does this matter?

If one believes that the blocking has created an orthogonal setup for which an interaction has no logical meaning, one would expect that by definition, $SS_{A \times B} = 0$. Accordingly, it is suggested that one fits a reduced model. A full model includes all terms, $A$, $B$, and $A \times B$; a reduced model would include only terms $A$ and $B$ (e.g., in SAS with a statement: Model y = *a* *b*;). Note that doing so means the $A \times B$ interaction cannot be tested because the model will not produce an estimate for its $SS$. The assumed zero or negligible $SS_{A \times B}$ term is aggregated into the $SS_{S/AB}$ term, the old error term, resulting in a new error term, $SS_{\text{“}E\text{”}}$ (= $SS_{S/AB} + SS_{A \times B}$, with degrees of freedom = $df_{S/AB} + df_{A \times B} = ab(n-1) + (a-1)(b-1)$). The $F$ tests would be, $F = MS_A/MS_E$ and $F = MS_B/MS_E$, for all possibilities of $A$ and $B$ being fixed or random.

In conclusion, the first resolution must be a determination of the theoretical intent of the research—if one is trying to simply control for gender, the blocking approach makes sense and a reduced model can be defended (and the interaction not tested). (Alternatively, one could run an analysis of covariance, or ANCOVA, with gender introduced as a dummy variable, discussed elsewhere in this special issue.) If one is trying to document differences between men and women in their reactions to experimental stimuli, a different philosophical approach is embraced, the full model is fit to allow for the testing of the interaction, and then one must simply verify the status (random or fixed) of the experimental factor to choose between the $F$ tests in 1(a) or 1(b). If gender is not the subject variable, one might instead be choosing between 2(a) and 2(b). A related question follows.

## I.H. SHOULD I TREAT A PARTICIPANT ATTRIBUTE (E.G., GENDER) AS A BLOCKING FACTOR?: PART 2

If I measure an IV, such as participant gender, and expect an interaction between this IV and another, manipulated IV, would this still be considered a blocking variable? My understanding is that blocking is intended to remove error, to control for the effects of an extraneous source of variance possibly not handled through randomization. But, if the blocked variable exhibits theoretically interesting differential effects on the other IV, would it not make sense to treat it as another IV? Cook and Campbell (1979) said I can do this, but I have had a reviewer insist that mixing IVs and blocking variables is not "clean," and I should focus on main effects. What is really the difference between blocking variables and measured variables?

### REFERENCE

Cook, Thomas D., & Campbell, Donald T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Professor Joan Meyers-Levy
University of Minnesota

I agree with your view that if one draws on a theory that suggests gender should interact with some manipulated IV, gender need not be treated as a blocking variable. Keppel (1973) noted that, consistent with this view, "blocking may be thought of as a control factor, where its introduction is motivated by a desire to reduce error variance rather than to study the effect of the subject factor per se" (p. 509). In such a situation, the use of randomized blocks generally "results in an experiment that is more sensitive than the corresponding experiment without blocks" because the error term "reflects the variability of participants from populations in which variation in the blocking factor is greatly restricted" (p. 504). Yet, Keppel (1973) also noted that a randomized block design may be used when factors such as gender are "the object of study" and the design is "not merely a means to reduce error variance" (p. 509). Using a randomized block design in this case, however, seems to be appropriate when there is a desire to assess the generalizability of some other effect across certain classifications of situations or individuals (e.g., gender). Nonetheless, if as in the case you outline, a study is intended to test a theory that specifically predicts that a subject factor like gender and some other factor interact, such a study does not constitute simply an assessment of whether some particular effect generalizes across gender. As such, I see no reason why the study must or necessarily should be treated as a randomized block design.

### REFERENCE

Keppel, Geoffrey. (1973). *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: Prentice Hall.

Professor James Barnes
University of Mississippi

A randomized block design allows for differences, such as gender, before the experiment to be evaluated. In the situation that you describe, participants cannot be randomly assigned to the different levels of gender, so blocking is still necessary even with an expected gender–factor interaction. However, the presence of block–treatment interaction effects alters the expected values of the mean squares producing a nonadditive model (see Table 7.3–1 in Hicks, 1993, p. 267). Furthermore, the research problem that you describe is either a fixed-effects model (Factor A and block fixed) or a mixed-effects model (Factor A random and block fixed). Thus, you need to be sure that the appropriate mean sum of squares are being used for statistical testing. The reviewers' comments concerning "clean" results most likely result from the inability to interpret main effects when interactions are present (Hicks, 1993).

When treatment levels can be combined such that each level of Factor A appears once with each level of Factor B and participants can be randomly assigned to one of the treatment combinations, then a factorial design would be appropriate (Hicks, 1993; Kirk, 1995). If randomization can be accomplished, then Cook and Campbell (1979) are correct and either a blocking or a factorial design should produce similar results.

## REFERENCES

Hicks, Charles R. (1993). *Fundamental concepts in the design of experiments* (4th ed.). New York: Holt, Rinehart, & Winston.

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

## I.I. HOW TO ANALYZE DATA FROM A NESTED DESIGN

I often run experiments in which a construct is operationalized by choosing two or more types of advertisements or brands, and a number of real ads or brands are nested within type (T). A typical design would look like this:

| | Type 1 | | | Type 2 | | |
|---|---|---|---|---|---|---|
| | Brand 1 | Brand 2 | Brand 3 | Brand 4 | Brand 5 | Brand 6 |
| Al | S1–S20 | Sl–S20 | Sl–S20 | Sl–S20 | S1–S20 | S1–S20 |
| A2 | S21–S40 | S21–S40 | S21–S40 | S21–S40 | S21–S40 | S21–S40 |

I am interested in the A × T interaction. I always run into problems when analyzing this design. My questions are the following:

1. How do I decide whether it is better to make *type* a within-subjects factor (as shown earlier) or a between-subjects factor?

2. Suppose I have the choice between adding more brands or adding more participants per cell. How do I make that trade-off?

3. I realize that I should have as many brands as possible within each type, but how do I decide how many are enough? Does it matter whether type is within-subjects or between-subjects?

4. Usually, I want to add a Latin square to balance out order effects. How do I do this, and how should I specify all my effects afterwards? Maybe I should not do this at all.

5. Working with real brands or ads, I once had an unequal number of brands or ads in each level of type: for example, three brands of Type 1 and five brands of Type 2. I did not succeed in programming the analysis in SAS. Is this due to SAS or is there a real problem? (Note that SAS problems are real problems for me; I am not a statistician.) Can I only do this kind of analysis if I have equal numbers?

6. What if my DV is categorical?

7. My psychologist friends tell me I should not bother with trying to model the brand effect and just average across brands—that is, compute an average for each cell in the $A \times T$ design and analyze the 2 × 2 interaction. I know that this is wrong, but how wrong is it? What are the consequences for Type 1 and Type 2 error?

8. What about the selection of brands? Ideally, I assume, both samples of brands are random from an infinitely large sample from each type. I treat brands as random factor, because I want to generalize beyond the chosen brands.

(a) However, in reality, I often do this experiment with brands from a single product category (e.g., domestic brands and imported brands of cars). This is not an infinitely large population. What if my sample exhausts the population? What if I know the sample is large relative to the population, but I do not exhaust it (say I include half of the existing brands). What if I know the sample must be large relative to the population, but I do not know exactly how large?

(b) If I am honest, I know that my samples of brands or ads are not really random; I take what I can get. Does this change anything to the inferences I can make? What exactly?

Professor Joan Meyers-Levy
University of Minnesota

1. In most cases, it is likely to be better to make *type* a between-subjects factor, because doing so eliminates artifactual influences (e.g., demand effects, carry over effects, reactivity) that can emerge from treating it as a within-subjects variable. At the same time, exceptions may occur if the anticipated effects are likely to be very subtle and jeopardized by the heightened variance introduced by treating type as a between-subjects factor. In this

case—and it will be a judgment call—treating type as a within-subjects factor may be preferable (see Keppel, 1973).

2. Adding more brands would seem advisable to the extent that you anticipate (for theoretical reasons) that different outcomes may emerge for different sorts of brands. Adding more brands also may be advisable if your research is primarily application versus theory oriented. For more discussion of this issue, see Calder, Phillips, and Tybout (1981). On the other hand, if multiple brands are being included simply to offer some evidence for the generalizability of the effects of type and A, I do not see all that much value in greatly increasing the number of brands—after all, the observation that the anticipated effects were robust for $n$ brands can never rule out the possibility that they will not generalize for an $n + 1$ brand.

3. The decision of how many brands is enough depends largely on whether your research is theory or application focused (cf. Calder et al., 1981). If your research is primarily theoretical, including more brands that are likely only to show more and more evidence of generalizability does not seem to be all that worthwhile. Again, as noted earlier, assembling a huge number of replications of an effect does not add much theoretically, and it can never eliminate the possibility that the focal effect may not replicate in yet some other instance.

## REFERENCES

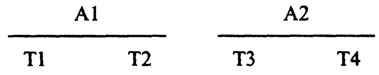
Calder, Bobby J., Phillips, Lynn W., & Tybout, Alice M. (1981). Designing research for application. *Journal of Consumer Research, 8,* 197–207.

Keppel, Geoffrey. (1973). *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: Prentice Hall.

Professor James Barnes
University of Mississippi

The author of this question poses an interesting problem dealing with hierarchical or nested factorial designs and repeated measures. Such a design could result from a study of cooperative advertising where a retailer features several different brands in a single advertisement.

Kirk (1995, p. 476) provided an excellent graphical example of the difference between hierarchical and crossed designs. In a nested or hierarchical design (*T*(*A*)), *T1* and *T2* appear only with *A1*, and as a result, there can be no $A \times T$ interaction (see the following figure).



In a crossed design, each level of *T* appears once and only once with each level of Treatment A, and an $A \times T$ interaction can be estimated (Kirk, 1995). Winer (1971, p. 540) considered two cases of three-factor studies with repeated measures. Case 1 has two of the three factors (fixed) crossing (or repeated on) all subjects. The model for Case 1 would be

$$Y_{ijkm} = \mu + A_i + S_{j(i)} + T_k + AT_{ik} + TS_{kj(i)} + B_m + AB_{im} + BS_{mj(i)} + TB_{km} + ATB_{ikm} + TBS_{mkj(i)}$$

where $A_i$ and $S_{j(i)}$ are the only between-subjects effects: $A_i$ = advertisement, $i = 1, 2$; $S_{j(i)}$ = subject, $j = 1, 2, \ldots 20$ for all $i$; $T_k$ = type, $k = 1, 2$; and $B_m$ = brand, $m = 1, 2, \ldots 6$.

In Case 2, subjects are nested within two (fixed) factors and the third factor crosses (or is repeated on) all subjects. The model for Case 2 is

$$Y_{ijkm} = \mu + A_i + T_k + AT_{ik} + S_{j(ik)} + \quad \text{(between-subjects terms)}$$
$$B_m + AB_{im} + TB_{km} + ATB_{ikm} +$$
$$BS_{mj(ik)} \quad \text{(within-subjects terms)}$$

In Case 1, type is a within-subjects factor, whereas in Case 2, it is a between-subjects factor. To answer your first question, one needs to determine whether type is to be a crossed or a nested factor. Your design appears to have type crossed with subjects and Factor A.

In my experimental design class, I always stress to my students that they develop an EMS table when first formulating an experimental project. An EMS table provides a deeper understanding of the experiment, shows which tests can be conducted, and reveals areas of potential problems before making a heavy investment in time and resources. Detail procedures for creating an EMS table are provided by Hicks (1993, pp. 160–163) and Kirk (1995, pp. 86–95). If I am correctly reading your example, your model is analogous to Case 1 because type and brand are repeated on all subjects, but subjects are nested within Factor A (e.g., advertisement type), and brand is nested in type. The following is an EMS table determined by this design and the sample sizes that you provide. (I am assuming that subjects and brands are random.)

| | A | Ss | Type | Brand | | |
|---|---|---|---|---|---|---|
| | 2 | 20 | 2 | 3 | | |
| | F | R | F | R | (F=Fixed, R=random) | |
| *Source* | *i* | *j* | *k* | *m* | *Expected Mean Square* | *F-test* |
| $A_i$ | 0 | 20 | 2 | 3 | $\sigma^2_{BS} + 20\sigma^2_{AB} + 6\sigma^2_{S} + 120\phi^2_{A}$ | no test |
| $S_{j(i)}$ | 1 | 1 | 2 | 3 | $\sigma^2_{BS} + 6\sigma^2_{S}$ | vs. $MS_{BS}$ |
| $T_k$ | 2 | 20 | 0 | 3 | $\sigma^2_{BS} + 40\sigma^2_{B} + 3\sigma^2_{TS} + 120\phi^2_{T}$ | no test |
| $B_{m(k)}$ | 2 | 20 | 1 | 1 | $\sigma^2_{BS} + 40\sigma^2_{B}$ | vs. $MS_{BS}$ |
| $AT_{ik}$ | 0 | 20 | 0 | 3 | $\sigma^2_{BS} + 20\sigma^2_{AB} + 3\sigma^2_{TS} + 60\phi^2_{AT}$ | no test |
| $AB_{im(k)}$ | 0 | 20 | 1 | 1 | $\sigma^2_{BS} + 20\sigma^2_{B}$ | vs. $MS_{BS}$ |
| $TS_{kj(i)}$ | 1 | 1 | 0 | 3 | $\sigma^2_{BS} + 3\sigma^2_{TS}$ | vs. $MS_{BS}$ |
| $BS_{m(k)j(i)}$ | 1 | 1 | 1 | 1 | $\sigma^2_{BS}$ | |

Direct tests are indicated in the previous EMS table. For example, the test of the null hypothesis concerning brand is against the Brand × Subject interaction. Notice that there are no direct tests for the $A$ or type main effects, nor their $A \times T$ interaction. It is possible to compute a pseudo $F$ test (Hicks, 1993, pp. 167–169) by combining and subtracting other EMS components. I am not aware of any software packages that compute pseudo $F$ tests. However, because most statistical packages output mean square values, manually constructing such a test in not very difficult.

An EMS table is also invaluable when deciding where additional resources should be allocated. Consider a table of critical values for the $F$ distribution. For small degrees of freedom, critical values are quite large. For example, $F_{1,1,.95} = 161$, whereas $F_{2,2,.95} = 19.0$. The implication of degrees of freedom is that for small degrees of freedom, a larger effect size is necessary to be able to reject the null hypothesis. Because much of what one studies in consumer psychology consists of small effect sizes, small degrees of freedom make the research task more difficult. At an alpha level of 0.05, the critical value of $F$ for the test of the $A \times$ Brand interaction null hypothesis (on $k(i-1)(m-1) = 4$ and $(m-1)k(j-1)i = 152$ $df$) is $F_{4,152} = 2.43$. If all other things are equal, increasing the number of types by one ($F_{6,228}$) produces a critical value of 2.14. In comparison, increasing the levels of Factor A by one ($F_{8,228}$) reduces the critical value to 1.98. Even though these differences can be slight, it could mean the difference between rejecting and not rejecting the null hypothesis for a small effect size. (However, practical issues are also a concern. Note that increasing the levels of Factor A to achieve the slightly more powerful, lower critical value would require running more subjects because Factor A is a between-subjects factor; whereas the drop in $F$ values from 2.43 to 2.14 requires an additional type category, but no more subjects.) Thus, the answer to your question of whether to add more brands or more subjects or whatever should be based on an assessment of the effect size you wish to be able to detect relative to the critical $F$ value.

A Graeco-Latin square can be a useful design if one is dealing with three factors and each of those factors has the same number of treatment levels (Hicks, 1993). The smallest such design would be two levels of ad, two levels of type, and two levels. Increasing the number of levels of any factor requires increasing the other. The Graeco-Latin square is like a cube in that it is the same size in all three dimensions. Although this design deals very well with the problem of order effects, there are several constraints that one must be willing to accept. A major restriction is that a factor level can appear once and only once in each row, column, and depth of the cube. Finding naturally occurring ads, brands, and so on may be a very difficult task. Because of the design, it is not possible to recover any interactions between factors—only main effects can be examined. In addition, missing values pose a greater problem because the combination of factor levels occurs only once in the design. Taking additional replications can be used to offset missing cells, and there are some procedures for estimating missing values (Hicks, 1993).

ANOVA is based on three assumptions: (a) Observations are normally distributed on the DV in each group, (b) homogeneity of the population varies for the groups, and (c) observations are independent. Stevens (1992, pp. 240–241) provided a succinct summary in table form describing the consequences of violating each of these assumptions. Unequal $n$s have the greatest impact on the actual alpha level if the assumption violation results from heterogeneous variances. If the $n$s are equal, there is only a slight effect on the actual alpha level. Your psychologist friend's suggestion of computing an average for each cell and then proceeding as if there is a single observation in each cell is often suggested as a method for dealing with unequal $n$s. Provided your effect size is sufficient to be detected by an $F$ test with single degrees of freedom, there should be no statistical differences from this approach.

The SAS program will test all effects against the error term if one exists in the model. Therefore, the program must be instructed as to which tests are appropriate (yet another use for the EMS table). This instruction is given by the "test" command, where hypothesis ($H$) equals the term to be tested, and error ($E$) equals the proper divisor for testing. I do not profess to be an expert SAS programmer, but I think the instructions would look something like this (PROC GLM; Class A T B; Model Y = $A|T(S)|B(S)$; Test H = $T(S)$ $E = T*B$) to test type against the Type × Brand interaction.

A categorical DV turns the analysis into a log-linear model. Although there is insufficient space to discuss details of log-linear analysis, I will highlight some difference between the different approaches. Recall that in ANOVA, one of the basic assumptions is the normal distribution. In log-linear models, the appropriate distribution is the multinomial. In log-linear analysis, one fits a series of models to the data, whereas in ANOVA, we generally think of fitting a model to the data. As a result, we need to reverse our thinking about tests of significance. In ANOVA, we usually want the statistic to be significant, whereas in log-linear analysis, a test that is not significant is good in the sense that the given model fits the data. Stevens (1992) devoted a chapter in his text to log-linear analysis in the multivariate case.

Finally, turning to your last question concerning the brand factor as a fixed variable, it would be useful to return to the EMS table and change the status of brand from random to fixed and recompute the EMS. Changing brand to a fixed factor would create different statistical tests, with different degrees of freedom.

## REFERENCES

Hicks, Charles R. (1993). *Fundamental concepts in the design of experiments* (4th ed.). New York: Holt, Rinehart, & Winston.

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

Stevens, James. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Editor: Special note to the Northwestern alumni of my ANOVA PhD seminar: See? I was not (merely) being a sadistic professor when I tortured you with EMS.

## I.J. HOW TO ANALYZE DATA FROM A "YOKED" DESIGN

I am interested in the effects of attentional self-control—that is, the opportunity of a consumer to control the speed of information processing or the rate at which information items are processed (e.g., Internet vs. television). I want to be able to assess the effect of being in control independently of the actual speed at which information is processed.

A long time ago, in an undergraduate psychology class, we read an old article on the relation between behavioral control and stress (I forgot who the author was) in which monkeys were delivered shocks (this was before animal rights existed). They were tested in pairs. One monkey had the opportunity to control the shocks (they were preceded by an auditory signal, and the monkey could press a button to avoid some of them). The other monkey received exactly the same shocks at the same times but was not able to control them. The difficulty of the control task was manipulated such that the control monkey did not succeed in avoiding all the shocks. If a shock was not avoided, it was delivered to both monkeys. The actual number of shocks varied from pair to pair. The DV was the formation of ulcers; paired monkeys were compared on ulcer formation. Monkeys with control developed many less ulcers than monkeys not in control. This design was called a yoked design. At the time we did not focus on the data analysis, but this problem seems methodologically similar to the research problem I introduced earlier.

Therefore, the question is how to design and analyze experiments in which respondents are randomly assigned to pairs (i.e., in a yoked design). Both members of a pair receive exactly the same treatment, except for the opportunity to exert control. There is variability between pairs, though, in the actual situation to which they are exposed. In the proposed scenario, both members of the pair receive exactly the same information at the same rate about a product, but only one is able to control the rate at which the information is delivered. There may be differences between pairs in the amount of information that is delivered.

How are data like these analyzed? What if the main IV (control) is crossed with another between-subjects variable? What if I want to test the interaction of the control manipulation with an individual difference variable?

Editor: It is very important not to scramble the analysis of yoked experiments, but rather to conduct the analysis "eggxactly" right. Thankfully, the *Journal of Consumer Psychology* (*JCP*) reader is a hard-boiled researcher, so this discussion should go over easy. Okay, enough yokes about yolks.

One can envision your lab setup in which one person would be making mouse clicks to investigate some consumer purchase, while another person would be seated passively watching a networked screen across which the same images flew, with no opportunity to send input to the system. The two conditions would be labeled "control" and "no control" and the dyad is paired due to their shared information that may differ somewhat across yoked pairs.

The data from this yoked design would be analyzed via a straightforward two-group matched *t* test, which is the simplest version of a repeated measures model. Matched *t* tests are easily envisioned when we hear of classic studies on twins (i.e., put one twin in one condition, the other twin in the other condition, and their data are linked given their special relationship). Sometimes the criterion of matching is self-selected, like when husband–wife pairs are distributed across counseling sessions. (And, cohort analysis is essentially the linking of large groups of people who are approximately the same age, who therefore would have experienced similar cultural phenomena; cf. Menard, 1991.) In this arrangement, the matching, or yoking, has been imposed by the experimenter.

For any of these scenarios, the data would look like this:

| yoked pair | score for person in Condition 1 | score for person in Condition 2 |
|---|---|---|
| 1 | $X_{11}$ | $X_{21}$ |
| 2 | $X_{12}$ | $X_{22}$ |
| 3 | $X_{13}$ | $X_{23}$ |
| ... | | |
| *n* | $X_{1n}$ | $X_{2n}$ |

When matched *t* tests are presented in introductory statistics books (e.g., Hays, 1988, pp. 313–315; Mann, 1995, pp. 494, 521–528; Weinberg & Goldberg, 1979; pp. 310–316), the examples illustrate explicitly how the performance of people in the first condition is compared to that of the matched persons in the second, literally by taking difference scores:

| yoked pair | score for Person 1 | score for Person 2 | difference score |
|---|---|---|---|
| 1 | $X_{11}$ | $X_{21}$ | $d_1 = \mathrm{X}_{11} - \mathrm{X}_{21}$ |
| 2 | $X_{12}$ | $X_{22}$ | $d_2 = X_{12} - X_{22}$ |
| 3 | $X_{13}$ | $X_{23}$ | $d_3 = X_{13} - X_{23}$ |
| ... | | | |
| n | $X_{1n}$ | $X_{2n}$ | $d_n = X_{1n} - X_{2n}$ |

A one-sample *t* test (with *n* – 1 *df*) is computed on the difference scores to test H0: $\mu_d = 0$:

$$t = \frac{(\bar{d} - \mu_d)}{\left(\frac{s_d}{\sqrt{n}}\right)},$$

where $\bar{d}$ is the mean of the $d_i$s and $s_d$ is their standard deviation.

As the design becomes more complex (adding conditions or factors, etc.), the matched *t* test becomes a repeated measures design (or mixed design if there are also between-subjects factors), and one proceeds as usual for a within-subjects data analysis. (Those designs are applicable whether we are comparing one person in multiple conditions or structurally linked persons in multiple conditions.) The computer or model uses the instructions regarding which participants are matched with which others and conducts the analysis accordingly—the user need not actually compute difference scores.

## REFERENCES

Hays, William L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart, & Winston.

Mann, Prem S. (1995). *Introductory statistics* (2nd ed.). New York: Wiley.

Menard, Scott. (1991). *Longitudinal research.* Newbury Park, CA: Sage.

Weinberg, Sharon L., & Goldberg, Kenneth P. (1979). *Basic statistics for education and the behavioral sciences.* Boston: Houghton Mifflin.

## I.K. DEALING WITH HIGHER ORDER INTERACTIONS IN REPEATED MEASURES DESIGNS

How should one deal with high-order statistically significant interactions in multifactor repeated measures designs? I often use scenarios as experimental stimuli in the context of my research. These scenarios vary according to several characteristics. Participants generally react to a subset or all of the scenarios, so I use repeated measures ANOVA for statistical analysis. For instance, in a recent study on sponsorship, I varied four factors in a 2 × 2 × 2 × 2 factorial design. It is not rare that I use more complex designs. The problem I often have when I proceed to analyze the data is observing several statistically significant high-order interactions, because, as you know, repeated measures designs are more sensitive. Now, I understand that significant interactions have to be interpreted. However, you must admit that the interpretation of a quadruple interaction is not an easy thing to do. In addition, the variance explained by such interactions is often small.

What should be done? Concentrate on the interpretation of significant main effects only? Concentrate on those interactions that are more important? How do you judge the importance of an effect (apart from looking at the range of mean differences and omega squared statistics—which are problematic in the context of repeated measures ANOVA)?

I find this problem coming back every time I do this type of experimental research, particularly when sample size is large (which ought to be a good thing). I think most of the time the interactions are not worth looking at, but what should I say in reporting the results? There are so many reviewers out there eager to find the little bug in an article who will notice the problem quickly if you try to evade it. What is most frustrating is the fact that increasing sample size is likely to lead to this type of situation. One is tempted to reduce sample size to avoid these problems.

A related question from a different researcher follows.

Although my major research interests are main effects or two-way interaction effects, my results showed a three-way interaction (which is not really important to my research objective). In such a case, is the interpretation of the main effect or the two-way interaction effect still meaningful, and should we make any conclusions based on the main effects or two-way interaction effect?

### Professor Scott Maxwell
### Notre Dame

It is important to realize that a main effect represents differences in marginal means averaging over all other factors in the design. As such, the main effect is an average (which may be either unweighted or weighted) of relevant simple effects (e.g., Maxwell & Delaney, 1990, p. 244; also see Cole, Maxwell, Arvey, & Salas, 1994; Maxwell, 1992; Maxwell & Delaney, 1990). A statistically significant interaction implies that these various simple effects are not all equal to one another. Even so, it is still legitimate to regard the main effect as a correct statement regarding average effects. However, when an interaction is present, the individual components of this average are known (probabilistically) to be different from one another. Thus, the real question here is whether it is meaningful to interpret an average effect when the individual effects making up the average are different from one another. This question cannot be answered simply from a statistical perspective, because meaningfulness derives largely from the purpose of the study. However, one aspect of the data that is often important is whether the interaction that qualifies the interpretation of the main effect is ordinal or disordinal (cf. Maxwell & Delaney, 1990, pp. 244–247, 298). When the interaction is ordinal, the simple effects all have the same sign and differ only in magnitude. In a case such as this, it can make sense to interpret an average effect. However, when the interaction is disordinal, the simple effects have different signs, making an average less meaningful. Finally, it should be noted that even in a simple two-way design, the two-way interaction can be ordinal with respect to one factor but disordinal with respect to the other.

## REFERENCES

Cole, David A., Maxwell, Scott E., Arvey, Richard D., & Salas, Eduardo. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin, 115,* 465–474.

Huberty, Carl J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105,* 302–308.

Maxwell, Scott E. (1992). Recent developments in MANOVA applications. In Bruce Thompson (Ed.), *Advances in social science methodology* (pp. 1–40). Greenwich, CT: JAI.

Maxwell, Scott E., & Delaney, Harold D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Pacific Grove, CA: Brooks/Cole.

Editor: As Maxwell's answer suggests, the response does not depend on whether the experimental design contains within-subjects factors. As the questioner seems to know, we speak of the interpretation of lower order effects (e.g., main effects, two-way interactions) as being moderated by the higher order interactions (three- and four-way interactions, etc.). I have empathy for this researcher who, seeking reliable data, evidently gathers relatively large samples and then "pays the price" by watching all the terms manifest significant. Significance is usually cause for celebration, but it does seem tedious to explain away the significant interactions that do not matter—for example, theoretically they were intended as replications (i.e., the stimuli exposures). I have also seen reviewers criticize the fact that an effect was not perfectly consistent across the stimuli (e.g., demonstrations over purchase categories). To be fair, spotting such patterns of results is surely part of a reviewer's job, because if there could be found consistencies among the inconsistencies, additional contingent factors would yield more complex theorizing. If a researcher's theory largely explains the data, but the inconsistencies are troubling the reviewer, it may be the burden of the reviewer to offer a theory that explains the data better (as parsimoniously and also encompassing the inconsistencies). The questioner's instincts are good, for although it may seem like some index of effect size could prove valuable (because they are independent of sample size), that class of indexes is not without its own problems (cf. Fern & Monroe, 1996). (See also the discussion in this special issue of stimuli as replications in a nested design, Question I.I.)

## REFERENCE

→ Fern, Edward F., & Monroe, Kent B. (1996). Effect-size estimates: Issues and problems in interpretation. *Journal of Consumer Research, 23*(2), 89–105.

## I.L. MULTIPLE STIMULI AND LEARNING STUDIED THROUGH A REPEATED MEASURES DESIGN

We have completed collecting data in a computer interactive experiment, the goal of which was to ascertain how individuals react to outcome feedback in probabilistic environments given externally set expectation levels. To achieve this objective, graduate students made sales estimates given various levels of advertising budgets for seven consecutive periods, receiving outcome feedback after each trial. Experiment participants were randomly assigned to one of four treatment conditions: small–large forecasting errors × attainable–unattainable preset expectation levels. The pattern of errors (over- and underestimating sales) was randomized across participants to mitigate the possibility of cheating, but in all cases, the sum of the errors was zero. After determining the pattern of errors for a participant in the small error condition, the pattern was multiplied by four and presented to an individual in the large error condition. After each decision, outcome feedback was provided, thus, the participants could react to feedback on the second to the seventh trials. The externally set performance expectation levels were embedded, but highlighted, in the task description, which was presented in a case format. Expectations were manipulated so that they could or could not be met.

Numerous hypotheses are being examined. For clarification purposes, we share one. We hypothesize that attainable versus unattainable expectation levels will affect the way managers interpret their success or failure, and that this in turn will affect subsequent decision-making processes. In early iterations solving recurring problems, we expect that when participants cannot meet performance expectations, they will change their behavior. Specifically, we expect them to interpret their failure to meet performance goals as due to their inability to correctly estimate the parameter value in the provided advertising response decision model, rather than the result of inherent unpredictability (noise) in the decision environment. As a result, they will adjust the model's parameter more than will participants who succeeded in meeting performance goals. Thus, if we measure the extent to which participants change the parameter attached to the DV (advertising), in the attainable expectation condition cue weights should be revised more than in the attainable expectation condition.

To analyze the data, we have used two methodological approaches. The first is to use the MANOVA routine in SPSS where the six trials are the repeated measures and the crossed treatment conditions are the IVs. As noted earlier, the pattern of errors received by each subject was shuffled (though in all cases the sum of the errors was zero). Thus, one subject may have received a small negative error on the first trial and another in the same treatment condition a positive error after the first trial. We did not "back shuffle" the responses so that the order of the errors was the same for each subject prior to analysis. We believe we are correct in not reordering the responses, because one of our goals is to tap the speed of learning over trials. As we understand it, this means we cannot examine learning across treatments at each trial; rather, we must examine the averaged *F* test by each of the two main effects as well as the interaction. If we are correct thus far, what appears in the printout is the univariate *F* test by trial, followed by the average $F$ test. Would we just report the latter in a publication? There are no treatment means associated with the latter. How do we know that significant effects are significant in the intended direction?

Given that one of the goals is to examine how quickly people learn to react appropriately in a specific decision-making environment, a more complicated but interesting way to analyze the data is as follows. Assuming we believe one treatment condition will cause participants to learn faster (i.e., the changes to their decision weights will more rapidly decrease toward zero), we fit a regression model to the changes in the cue weights over trials for each individual participant. The DV is the percentage change in cue weight over trials; the IV is the trial number. This is done for each participant. We anticipate that the slopes in the fitted models for those in the "learn faster" condition would be larger, negative numbers. Although the individual regression models are unreliable (there are only six observations), these betas would then be used as DVs in a 2 × 2 full factorial ANOVA model. Does the unreliability at the individual level washout when analyzed with similar data from numerous individuals? The slopes to the fitted line do appear to be a wonderful way of capturing the speed with which people learn. We are not aware of anyone using this technique.

Professor Greg Allenby
Ohio State University

The goal here is to produce a model in which noisy data produces learning and accurate data produces more learning. The researcher knows the extent of noisiness (uncertainty or variance), but the key is to understand the perceived or operational amount of uncertainty. I would study this using a Bayesian model of updating, but instead of knowing the prior and likelihood and then deriving the posterior (as in a standard analysis), I would specify the prior and posterior and then derive the likelihood. The variance parameter in the likelihood reflects the operational level of uncertainty. I would need to know more about the problem to flesh out the exact models that I would use, but the key here is that the measurement must condition on what is observed, and what is observed in this problem is the prior and posterior amount of knowledge.

Editor: Bayesian updating models do sound apropos and worth investigating. Your choice of the repeated measures ANOVA model sounds perfectly well suited for addressing your concerns. I am not sure what you mean by back shuffle (was that a dance in the late 1970s?), but I am guessing that you are trying to make a case that the order of stimuli presented over trials is unimportant. Often researchers who test for "order," hoping it will not have an effect or interaction with any other factor, find their hope fulfilled partly because so few participants were assigned to any given order that there is no power to detect an order effect.

Your situation seems a bit more complicated. I sense some ambiguity in how you characterize the stimuli in your design. On the one hand, you do not wish to examine the $F$ tests per trial, because you exposed different participants to different stimuli as they proceeded along, so an aggregate statistic (across respondents within a trial) may be meaningless. In this case, as you intuit, the best you could do is report the final $F$ test, as capturing the outcome of the process, but not the steps along the process itself (nor, as you recognize, the means at each trial).

On the other hand, you make it sound like you could argue philosophically that the stimuli presented over trials are expected to be theoretical replicates, in which case you could presumably interpret the $F$ tests and means per trial. Another perspective would be to not worry about the stimuli within a trial whatever and rather create an index that flags when (i.e., during which trial) the respondent has proceeded up the learning curve, as determined by some criterion that seems suited to your learning task (correct answers, speed in responding, etc.). This suggestion is related to your investigation into slopes—again, a nice intuition. However, you need not shift paradigms to test the shapes of individuals' learning curve functions or their points of "inflection" (essentially at which trials do you see substantial increments of improvement for each respondent?). Within the ANOVA framework, you would treat trials as quantitative factors, and you could test "trend" contrasts (i.e., linear, quadratic, etc.; cf. Keppel, 1991). Related questions in this special issue may also be useful, including (a) studying effects in regression analogous to simple effects in regression (Section II.H.) and (b) hierarchically nested designs in ANOVA (Section I.I.).

## REFERENCE

Keppel, Geoffrey. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

## I.M. SHOULD WE PLOT MEANS OR MEAN DIFFERENCES WHEN ANALYZING WITHIN-SUBJECTS DESIGNS?

Should we plot means or mean differences when analyzing within-subjects designs?

Professor Jan-Benedict Steenkamp
Catholic University of Leuven, Belgium

I do not feel strongly about this. Mean differences are easily computed from the means. In any case, the data suffer from correlated errors, so I am not sure whether mean differences is a great help.

---

Editor: Steenkamp is right, though the intuitions behind the question are good, given that the underlying philosophy in a within-subjects, or repeated measures, design is one of comparing participants' responses under one condition to their responses under another condition, to see what kinds of changes have been established. McNeil (1992) had some useful suggestions for plotting data observations that are structurally tied to each other. To that same end, you may find Harris's (1985) presentation of "profile analysis" interesting, in plotting and analyzing shapes of curves across trials or conditions.

## REFERENCES

Harris, Richard J. (1985). *A primer of multivariate statistics* (2nd ed.). New York: Academic.

McNeil, Don. (1992). On graphing paired data. *American Statistician, 46,* 307–311.

## I.N. USES OF ANCOVA

I understand the typical use of the ANCOVA. However, is it appropriate to use ANCOVA to measure whether the experiment result was confounded by some uncontrolled variables (e.g., demographic variables)? If yes, how do we interpret the ANCOVA result in such a case? If not, what are other statistical tools to detect such confounding effects?

---

Editor: Yes, the use of the ANCOVA as suggested in this question is legitimate and could be implemented beneficially in this manner more frequently. When conducting experiments, we randomly assign respondents to conditions to enhance internal validity—our hopes at being able to attribute group differences on various DVs to our experimental interventions. If we have randomly assigned persons to conditions, then a priori, before our manipulations, these groups should be, statistically speaking, roughly (i.e., within sampling variability) equal on nearly any measure one could concoct. However, random and statistically speaking always work in the long run (over one's career?), which is to say that any given random assignment might look peculiar indeed to the eye. In particular, we worry about confounding—extant group differences on extraneous (i.e., nonfocal) attributes that could serve as alternative explanations for the group differences observed on the dependent measures of interest. In such situations, we can use the ANCOVA to correct for, or statistically control for, such observed differences. In these circumstances, ANCOVA serves as a post hoc equating of the groups when the random assignment did not appear sufficient.

In paradigms, where we know certain extraneous factors to be persistent, we may opt to mechanically control for differences by "blocking" on that extraneous factor—purposively distributing equal numbers of persons high and low on the blocking attribute across the experimental conditions to assure there will be no confounding. Blocking can be clumsy, however, and it is difficult to know just which nuisance factors will flare up on any given occasion. An advantage of the ANCOVA is that all the potential covariates may be measured (after the DVs so as to minimize any possible contamination), and then these variables may be tested as candidate covariates, one at a time or in groups, to see which corrections should be maintained in the analysis.

As an example of using ANCOVA to correct for initial group differences, I have been playing with data that arose from a classic pretest–posttest design in which we found the predicted posttest results. However, on further inspection, we also found the pretest results to be significant, with a similar pattern of means. Once the pretest attitude scores were included as a covariate, the adjusted effects on the DV were no longer impressive (nor significant); the groups' attitudes simply differed across conditions before we ever intervened.

How does one interpret ANCOVA results when using the model to correct for extant group differences? Hopefully, the effect for the covariate (and perhaps its interaction with one or more of the experimental factors) is significant, but even if it is not, including the covariate in the model usually helps to clean up the analysis by reducing the *MSE* term that enhances the likelihood of finding significance in the DV. In reporting regressions, researchers frequently describe the theoretically important findings, mentioning that they also controlled for certain other variables—for example, by having included variables and dummy variables representing the extraneous factors simultaneously in the model. The logic for ANCOVA is similar, and this sort of reporting would seem to be a reasonable format—for example, "using a covariate to control for the a priori group differences observed on [weight], we nevertheless saw significant differences on [self-reported cravings for chocolate] across the various [Advertisement × Room Scent] conditions."

For additional information on the ANCOVA, consult Edwards (1979), Keppel (1991, pp. 297–236), Kirk (1982, pp. 715–757), and Wildt and Ahtola (1978).

## REFERENCES

Edwards, Allen L. (1979). *Multiple regression and the analysis of variance and covariance.* New York: Freeman.

Keppel, Geoffrey. (1991). *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: Prentice Hall.

Kirk, Roger E. (1982). *Experimental design procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Wildt, Albert R., & Ahtola, Olli. (1978). *Analysis of covariance.* Beverly Hills, CA: Sage.

## I.O. WHEN TO USE MANOVA AND SIGNIFICANT MANOVAS AND INSIGNIFICANT ANOVAS OR VICE VERSA

Consider an experimental study that involves multiple dependent measures hypothesized to respond in the same manner to an experimental manipulation. When the data are analyzed with MANOVA, the experimental manipulation has a nonsignificant effect. Nonetheless, the univariate ANOVAs show a significant effect in the predicted direction for each of the dependent measures. Why does this inconsistency occur, and which set of results should the researcher rely on?

A related question follows: I have run into confusion when determining when to use MANOVA versus separate ANOVAs. Most textbooks are vague on this subject and just mention that MANOVA can be used with correlated DVs. The question is, What is the desired level of correlation to justify use of MANOVA? I asked my resident statistics expert, and he feels that DVs correlated from about .3 to about .7 are eligible, because anything lower indicates variables that really are not very related, and anything higher indicates redundancy. He was also quick to point out that it must make sense conceptually to package DVs together before doing so. What is the conventional wisdom on this matter? I have had reviewers that insist MANOVA must be used if there are multiple DVs, period, without considering the conceptual implications.

Another question is, What do I do if MANOVA is not significant, but ANOVAs on the individual DVs are? Is it possible that the DVs share variance and that prevents the MANOVA from approaching significance? Is there any way to justify using the individual ANOVAs instead?

### Professor Scott Maxwell
### Notre Dame

The MANOVA null hypothesis is that group mean differences are zero on each and every DV. Thus, all it takes is for groups to differ on a single DV for the null hypothesis to be false. Therefore, with a large enough sample, the multivariate null hypothesis will be rejected with a probability approaching 1.0 even if the groups are identical on all but a single variable. Most researchers, however, are not blessed with an indeterminately large sample. Thus, statistical power becomes an important consideration in planning as well as evaluating group comparisons.

In some situations, univariate tests may be more powerful than the multivariate test, leading to one or more statistically significant univariate results but a nonsignificant multivariate test. The opposite is also possible—the multivariate test can be more powerful than any of the univariate tests, in which case the multivariate test may be statistically significant, but there are no statistically significant group differences on any of the individual DVs, according to univariate tests. The reason such discrepancies can occur is because the power of each univariate test depends on mean differences and variances (and obviously sample size), but the power of MANOVA depends not just on group mean differences and variances, but also on correlations among the DVs. Although some researchers have suggested that high correlations lead to greater power, others have suggested just the opposite. Cole, Maxwell, Arvey, and Salas (1994) showed that the power of MANOVA can either increase or decrease as a function of the intercorrelations among the DVs.

In particular, to the extent that the various DVs tend to measure the same underlying construct, the power of MANOVA is likely to be less than if the measures are more disparate. Thus, especially when variables are highly related, approaches other than MANOVA may be preferable. For example, Maxwell (1992) pointed out the potential benefits of forming a composite dependent measure such as a simple unweighted average or first principal component of the original DVs.

Alternatively, a latent variable model could be adopted and latent variable mean differences investigated (cf. Cole, Maxwell, Arvey, & Salas, 1993). More generally, researchers are advised to heed Huberty and Morris's (1989) advice about the extent to which their multiple dependent measures constitute a variable system.

If the researcher's only reason for using MANOVA is to control the family-wise Type 1 error rate, Huberty and Morris (1989) suggested that better methods may be available. On the other hand, when researchers are interested in examining linear combinations of the DVs, MANOVA is likely to be the method of choice.

## REFERENCES

Cole, David A., Maxwell, Scott E., Arvey, Richard D., & Salas, Eduardo. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin, 114,* 174–184.

Cole, David A., Maxwell, Scott E., Arvey, Richard D., & Salas, Eduardo. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin, 115,* 465–474.

Huberty, Carl J., & Morris, John D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105,* 302–308.

Maxwell, Scott E. (1992). Recent developments in MANOVA applications. In Bruce Thompson (Ed.), *Advances in social science methodology* (Vol. 2, pp. 200–222). Greenwich, CT: JAI Press.

Maxwell, Scott E., and Delaney, Harold D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Pacific Grove, CA: Brooks/Cole.

Editor:

MANOVA is a generalization of analysis of variance that allows the researcher to analyze more than one dependent vari-

> able. Many times it is of interest to compare group means on several variables simultaneously, and MANOVA allows the researcher to do this in both experimental and observational research. (Bray & Maxwell, 1985, p. 5)

Bray and Maxwell (p. 31ff) provided a nice demonstration of one of the arguments for MANOVA, that it can be more powerful in detecting group differences—for example, the marginal distributions on either of two DVs may yield insignificant ANOVA results, yet jointly the groups are cleanly separable in the 2-D multivariate space, via MANOVA (cf. Iacobucci, 1994).

Usually the selection of the variables to model simultaneously in a MANOVA is conceptually driven. Multiple indicators of a single construct are likely to be correlated, and they make for a sensible variable system (as stated previously) or conceptual package (as per the second question stated previously). Alternatively, a researcher may be interested in modeling the joint behavior of multiple complementary DVs, whose measurement may not be redundant, but whose information is correlated, as in components of a behavioral process system. In either case, "the dependent variables should be theoretically correlated as well as empirically correlated" (Weinfurt, 1995, p. 251).

If, as is typically argued, an advantage of MANOVA over ANOVA is that the former takes into account the correlations among the dependent measures in these conceptual sets, as the second question poses, just how correlated ought those measures be? I agree whole heartedly with the questioner's resident stats experts—a range of .3 to .7 seems sensible, albeit a subjective rule of thumb. If all the DVs are pairwise correlated greater than .7, it may be simpler to create a composite score—a simple average—and execute a univariate ANOVA.

If all the correlations fall short of .3 in magnitude, the DVs are essentially contributing unique information, and a series of ANOVAs may be conducted with little loss of information. However, if one were to do so, it would be important to consider the Type 1 error rates and compensate in a Bonferroni fashion, correcting alpha by dividing it by the number of DVs. Protection from excessive Type 1 errors has been another argument for MANOVA, though researchers remind us that this protection is true only of the first stage in hypothesis testing (i.e., the MANOVA overall test that determines whether anything is significant). In the follow-up tests to identify just which groups differ on just what variables, Type 1 errors may abound (Bird & Hadzi-Pavlovic, 1983; Ramsey, 1980, 1982; see also Huberty & Morris, 1989).

Having said all that, all the right things, I have to say that I personally cannot remember ever reading a single journal article in which the hypotheses to be tested were actually multivariate. ANOVA and MANOVA are linear models (one term, like a main effect, gets added to or subtracted from another term like an interaction). Linear models are inherently compensatory, meaning that if a respondent scores high on one variable and low on another, their data are treated like a respondent with the reverse pattern, or one who scores in the middle on both variables (within limits, e.g., if the variables are roughly equally weighted). Stated another way, linear models treat data in a disjunctive form (is *X* high or is *Y* high). Most journal articles state hypotheses in a conjunctive form (e.g., H1: "We expect *X* to be high under these conditions," and H2: "We expect *Y* to be high under these conditions"). Researchers with these predictions would not conclude a success if *X* were extra high and *Y* were low—the form of the hypothetical priors are not compensatory. If a univariate hypothesis is stated, it may make most sense to simply test it univariately. If a number of indicator variables have been measured and they are shown (through a factor analysis) to be tapping the same construct, a composite (e.g., average) may be formed and a simple ANOVA conducted. If a number of DVs are to be tested and they cannot be aggregated, for example, they are hypothesized to represent different constructs with different anticipated results, then the researcher should do separate ANOVAs and probably operate conservatively by adjusting alpha.

Nevertheless, a most definitive scenario in which MANOVA should be used rather than ANOVA is when modeling repeated measures (within-subjects) variables or mixed designs (i.e., in conjunction with between-subjects factors). The ANOVA treatment for a repeated measures or mixed design is known to be nonrobust to violations of the homogeneity of treatment difference variances assumption, which is relaxed in the MANOVA modeling approach on such data.

In theory, a MANOVA is a straightforward extension of ANOVA. And, in theory, if several DVs are at least somewhat correlated and their results are qualitatively similar (showing the same patterns of means, etc.), then the conclusions drawn between the two techniques should converge.☺

## REFERENCES

Bird, Kevin D., & Hadzi-Pavlovic, Dusan. (1983). Simultaneous test procedures and the choice of a test statistic in MANOVA. *Psychological Bulletin, 93,* 167–178.

Bray, James H., & Maxwell, Scott E. (1985). *Multivariate analysis of variance.* Newbury Park, CA: Sage.

Iacobucci, Dawn. (1994). Analysis of experimental data. In Richard Bagozzi (Ed.), *Principles of marketing research* (pp. 224–278). Cambridge, MA: Blackwell.

Ramsey, Philip H. (1980). Choosing the most powerful pairwise multiple comparison procedure in multivariate analysis of variance. *Journal of Applied Psychology, 65,* 317–326.

Ramsey, Philip H. (1982). Empirical power of procedures for comparing 2 groups on p variables. *Journal of Educational Statistics, 7,* 139–156.

Weinfurt, Kevin P. (1995). Multivariate analysis of variance. In Laurence G. Grimm & Paul R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 245–276). Washington, DC: American Psychological Association.

## I.P. TESTING FOR SIGNIFICANT DIFFERENCES IN VARIANCES, RATHER THAN MEANS

Is there a way to compare whether variances on some variable in two different groups are different? I have two groups in my sample of survey respondents, and I have noticed that mean ratings are higher in one group compared to the other, but their judgments also seem to have lower variance. This finding makes theoretical sense and helps the research, but I do not know how to demonstrate statistically that the variance is smaller in the one group versus the other.

---

Professor Greg Allenby
Ohio State University

Sure, you bet. You treat the variance as any other parameter in a model and build two models of the data: (a) The variance is constrained to be the same in the two subsamples, and (b) the variance parameter is allowed to differ in the two subsamples. Fit these two models to the data via likelihood methods. Minus two times the difference in the log likelihood is distributed chi-square with 1 *df* in this problem. If –2*log likelihood is greater than the magic cutoff number for a chi-square with 1 *df,* then the variances are significantly different.

---

Professor Sachin Gupta
Northwestern University

Assume these are random samples from two normal populations with means and variances ($\mu_1$, $\sigma_1^2$ ) and ($\mu_2$, $\sigma_2^2$ ). Several hypotheses about the population variances may be tested (see Mood, Graybill, & Boes, 1974):

$$\begin{aligned} &\text{a) } H_0: \sigma_1^2 \le \sigma_2^2; && H_1: \sigma_1^2 > \sigma_2^2 \\ &\text{b) } H_0: \sigma_1^2 \ge \sigma_2^2; && H_1: \sigma_1^2 < \sigma_2^2 \\ &\text{c) } H_0: \sigma_1^2 = \sigma_2^2; && H_1: \sigma_1^2 \ne \sigma_2^2 \end{aligned}$$

The test is derived from the basic idea that if $X_{11}, X_{12}, \ldots, X_{1n1}$ is a random sample from a normal density with mean $\mu_1$, and variance $\sigma_1^2$ and if $X_{21}, X_{22}, \ldots, X_{2n2}$ is a random sample from a normal density with mean $\mu_2$ and variance $\sigma_2^2$, and if the two samples are independent, then we know that the statistic

$$Q = \frac{\sum(X_{1i} - \bar{X}_1)^2 / (n_1 - 1)\sigma_1^2}{\sum(X_{2i} - \bar{X}_2)^2 / (n_2 - 1)\sigma_2^2}$$

has the $F$ distribution with ($n_1$ – 1) and ($n_2$ – 1) *df* when $\sigma_1^2 = \sigma_2^2$. The statistic $Q$ tends to be large when $\sigma_1^2 = \sigma_2^2$ and small when $\sigma_1^2 = \sigma_2^2$. We use this characteristic to formulate tests for the hypotheses given previously. For example, the test of hypothesis (c) is the following: Prefer $H_1$, if $k_1 < Q < k_2$, where $k_1$ and $k_2$ are selected so that the test has size $\alpha$. If we let the two tails have equal areas of $\alpha/2$, then $k_1 = F_{\alpha/2}(n_1 - 1, n_2 - 1)$, and $k_2 = F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$.

## REFERENCE

Mood, Alexander, Graybill, Franklin A., & Boes, Duane C. (1974). *Introduction to the theory of statistics.* New York: McGraw-Hill.

---

Editor: I do not know whether the typical *JCP* reader will be familiar with the likelihood methods suggested in the first scenario, so I am going to focus on the second. Within the second response, let me focus on Hypothesis c. It may help to make this statistic more familiar looking if we restate the hypotheses from

$$H_0: \sigma_1^2 = \sigma_2^2 \quad \text{vs.} \quad H_1: \sigma_1^2 \ne \sigma_2^2$$

to

$$H_0: \frac{\sigma_1^2}{\sigma_2^2} = 1 \quad \text{vs.} \quad H_1: \frac{\sigma_1^2}{\sigma_2^2} \ne 1$$

The $F$ statistic, afterall, is a ratio of two variances—specifically, the null hypothesis in the ANOVA scenario involves $MS_{effect}/MS_{error}$.

In the statistic $Q$ mentioned previously, we do not really do anything with the $\sigma_1^2$ in the numerator or the $\sigma_2^2$ in the denominator, because, according to the null hypothesis, these are theoretically equal, and so they essentially drop out of the equation. If you look at the pieces of the equation that remain, you have the variance of Sample 1 ($S_1^2$) in the numerator and the variance as estimated in Sample 2 ($S_2^2$) in the denominator; for example

$$S_1^2 = \left(\frac{1}{n_1 - 1}\right)\sum_{i=1}^{n1}(X_{1i} - \bar{X}_1)^2.$$

Therefore, you simply compute $Q$ as the ratio of ($S_1^2$) over ($S_2^2$); literally, $Q = S_1^2 / S_2^2$. You would reject the null hypothesis and claim that $S_1^2 > S_2^2$ if $Q$ exceeded the right-hand critical value in the $F$ distribution, and you would support the claim that $S_1^2 < S_2^2$ if $Q$ was less than the left-hand critical value in the $F$ distribution.

This test of variances may look a little odd because our typical experience with $F$ tests is that we reject our null hypotheses only for large values of $F$s—we have put the entire rejection region in the right-hand side of the distribution, caring to demonstrate only whether the numerator variance ($MS_{effect}$) is larger than the denominator ($MS_{error}$). (In fact, this approach so dominates the uses and tests of comparative variances that most of our statistical texts provide only right-hand tail critical values in their tables.) Therefore, if you put the bigger of your two variances in the numerator, call it Sample 1, and then test against the new alternative:

$$H_0: \sigma_1^2 / \sigma_2^2 = 1 \quad \text{vs.} \quad H_2: \sigma_1^2 / \sigma_2^2 > 1,$$

you would use the critical value $F_\alpha(n_1 - 1, n_2 - 1)$, where the entire rejection region is in the right-hand tail—for example, $F_{.05,30,36} = 1.78$; $F_{.01,12,10} = 4.71$, and so on. (Statistical purists would say that by examining the two variances and seeing which is larger and putting it in the numerator as representing Sample 1 is already a heuristic of a significance test, and doing so increases the likelihood of rejecting the null. If this issue is of concern, a more stringent significance level could be implemented. I was simply trying to make salient the similarity between this test and something likely to be familiar to the reader—the *F* test in ANOVA.)

By the way, if you consider this test the variance analog to a two-sample *t* test for testing mean differences, we also know there is a *t* test for testing a single mean against some hypothetical value (i.e., $H_0$: $\mu_1 = \mu_0$ compared to the two-sample version, $\mu_1 = \mu_2$). There is a similar test to see whether a single variance estimate is equal to some given value—that is, $H_0: \sigma_1^2 = \sigma_0^2$, where $\sigma_0^2$ is some constant "c," say, $H_0: \sigma_1^2 = 4.0$). The statistic takes the following form: $s_1^2 / c$. For example, for $\alpha$ = .05, sample size *n* = 30, *df* = 29, two-tailed chi-square cutoff values are 16.05 and 45.72 (i.e., reject and conclude your observed variance estimate is significantly less than c if this statistic falls short of 16.05, reject and conclude your observed variance is significantly greater than c if it exceeds 45.72, and do not reject the null or maintain your variance is plausibly equal to c in the population if the statistic falls within the range). The one-tailed cutoff value is $\chi^2 = 42.56$ (if you care only whether your variance is greater than c; Hogg & Tanis, 1979, pp. 250–251).

## REFERENCE

Hogg, Robert V., & Tanis, Elliot A. (1979). *Probability and statistical inference.* New York: Macmillan.

## I.Q. MANIPULATION CHECK: RELATIVE VERSUS ABSOLUTE MEAN DIFFERENCES

Consider an experimental study containing two levels of a construct that includes a manipulation intended to represent the ends of the continuum embodied by the construct (e.g., high vs. low involvement; near vs. far brand extension). Participants' responses to an appropriate manipulation check are affected significantly by the manipulation. However, inspection of the mean responses reveals that, in one condition, they averaged around the midpoint of the manipulation check scale (i.e., the mean does not differ significantly from the scale midpoint), but were more polarized in the other condition (i.e., the mean does differ significantly from the midpoint). What labels should the researcher use in describing these two conditions (e.g., the original high–low labels vs. higher–lower vs. high–moderate or low–moderate)? And, given that the scale is interval rather than ratio in nature, is it inappropriate to assign meaning to any given point (e.g., the midpoint) on the scale?

Professor Jan-Benedict Steenkamp
Catholic University of Leuven, Belgium

I am bothered by the fact that sometimes reported means on the manipulation check suggests that the one condition was not low at all (or high) but was actually close to the midpoint. The theory may still hold for, say, average versus high, but people still often use the labels low–high. I believe that is not correct, and the correct labels should be used, reflecting the actual scores given. I believe in any case that most phenomena are continuous so that some kind of dichotomization is a simplification of reality. Your question shows that this may not be the ideal strategy.

I see no particular problem to assign meaning to the scale. After all, the scale is verbally anchored, and, hence, we know that on a 5-point scale ranging from 1 (*very low*) to 5 (*very high*), 4 means *high.* Of course, the scale may not be interval scaled at all. In that case, we do not know whether it is rather high, pretty high, or close to very high. As you know, we can get better insight into these issues by applying "Homals" (a program for correspondence analysis in SPSS; see Hair, Anderson, Tatham, & Black, 1998, p. 553) or a related technique to these scores. However, in short, to make the score meaningful, we have to give it verbal meaning, which is also what the respondent does.

## REFERENCE

Hair, Joseph F., Jr., Anderson, Rolph E., Tatham, Ronald L., & Black, William C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Editor: I am not sure just how definitively to take a midpoint, or a participant's interpretation of one. Having been baptized in measurement, I guess I have a healthy (overactive?) skepticism of taking any measured variable literally. Furthermore, I think the important point in the application posed is that researchers are trying to provide demonstrations of phenomena that occur at relatively high and relatively low ends of a scale. Therefore, labeling two means as high and low instead of high and middle would not keep me up at night. The empirical question is whether the high–middle less extreme outcomes provide sufficient variance for interesting effects to result—that is, if one can produce the results of interest under the high–middle conditions, presumably the focal effects would be even more pronounced had the calibration yielded high–low scores (i.e., the former contrast, perhaps produced by using a more subtle manipulation, yields a conservative test).

In addition, even if the researcher thought high and middle were more precise labels for their current data, the next researcher might attempt to replicate the study conceptually (i.e., same construct), but perhaps doing so with a different operationalization and measure (i.e., different variable), thereby likely creating an altogether new calibration, one that was even more exaggerated or truncated, depending on the purpose of the new study. If this issue is really troublesome, the extent of variation on these operationalizations can be captured easily by indexes of effect size (omega-squared, eta-squared, etc.; cf. Cramer & Nicewander, 1979; Dodd & Schultz, 1973; Dwyer, 1974; Keppel, 1991; Maxwell, Camp, & Arvey, 1981; O'Grady, 1982).

## REFERENCES

Cramer, Elliot M., & Nicewander, W. Alan. (1979). Some symmetric, invariant measures of multivariate association. *Psychometrika, 44,* 43–54.

Dodd, David H., & Schultz, Roger F., Jr. (1973). Computational procedures for estimating magnitude of effect for some analysis of variance designs. *Psychological Bulletin, 79,* 391–395.

Dwyer, James H. (1974). Analysis of variance and the magnitude of effects: A general approach. *Psychological Bulletin, 81,* 731–737.

Keppel, Geoffrey. (1991). *Design and analysis: A researcher's handbook* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Maxwell, Scott E., Camp, Cameron J., & Arvey, Richard D. (1981). Measures of strength of association: A comparative examination. *Journal of Applied Psychology, 66,* 525–534.

O'Grady, Kevin C. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin, 92,* 766–777.

## I.R. EFFECTIVE MANIPULATION IN THE RESOURCE MATCHING CONTEXT

One recent stream of empirical efforts in the persuasion literature has focused on the resource matching framework that predicts that persuasion depends on the extent to which the cognitive resources needed for advertising processing match the cognitive resources consumers have available at the time of advertising processing. A typical study would entail different experimental conditions that vary the match of available resources (*AR*) and required resources (*RR*). For example, one condition would be designed to represent a match (*AR* = *RR*), whereas other conditions would be designed to represent a mismatch (*AR* > *RR* or *AR* < *RR*).

To validate assumptions about whether a given condition represents a match or mismatch, a recent *Journal of Consumer Research* (*JCR*) article reports evidence in the form of response times (RTs) to a secondary task. RTs were fastest in the condition designed to represent *AR* > *RP*, slowest in the condition intended to represent *AR* < *RF*, with the RTs in the *AR* = *RR* condition falling between these two extremes. Based on the results, it was concluded that the conditions represented the desired levels of match–mismatch between *AR* and *RR*.

I am unsure as to whether this conclusion is warranted for the following reasons. I accept that any condition representing *AR* > *RR* should lead to the fastest RTs given that it is the only condition where excess resources are available. Less obvious to me, however, is whether differences would be expected for RTs between the *AR* = *RR* and *AR* < *RR* conditions. In the match condition, participants are presumably devoting all of their resources to advertising processing, leaving minimal resources for responding to the secondary task. Accordingly, RTs should be slower than those observed in the *AR* > *RR* condition. Similarly, participants in the *AR* < *RR* condition should be allocating all of their resources to ad processing, which again would lower their RTs.

However, because participants are presumed to allocate all of their resources in both the *AR* = *RR* and *AR* < *RR* conditions, it would seem that RTs should be the same in these two conditions. Do you agree? If so, then what would be an appropriate interpretation of what each condition represents? For example, might one fairly conclude that the data indicate that *AR* always exceeded *RR*, but that the extent of this excess varied such that it was greatest in the *AR* > *RR* condition, smallest in the *AR* < *RR* condition, with the excess failing between these two extremes in the *AR* = *RR* condition? More generally, how might researchers attempt to establish the relative levels of *AR* and *RR* represented by a particular experimental condition? This is a critical issue because the degree to which a given data pattern supports the resource matching framework directly depends on what the conditions actually represent.

Professor Joan Meyers-Levy
University of Minnesota

Because it is impossible to directly assess the level of resources an individual truly has activated and has available for processing (because, of course, this is not directly observable), assessing individuals' RT to a secondary task would seem to provide some, albeit perhaps incomplete or nondefinitive, insight into this issue. An observation that under particular conditions significant differences occur in people's RT for performing a secondary task can serve as an indicator of the degree of difficulty such persons encounter in conjuring up and recruiting resources to perform the secondary task. Therefore, to the extent that RTs are longer in one condition than another, it follows that individuals are experiencing more difficulty identifying or freeing up resources for performing the secondary task, and, thus, available resources are to some degree less than plentiful. Hence, if one wishes to assert that one condition represents $AR < RR$, whereas another represents $AR = RR$, a necessary condition is that longer RT must occur in the so-called $AR < RR$ than in the $AR = RR$ condition, because even though it also may be difficult and time consuming to identify and recruit resources for the secondary task in the $AR = RR$ condition, it should be even more so in the more resource challenged $AR < RR$ condition (i.e., it should take longer to identify or free up such resources).

At the same time, observing significant differences in such RT across any three or more conditions only indicates that relative differences exist in ease of recruiting resources in these conditions, which by itself does not necessarily mean (as you note) that the three conditions truly represent ones in which $AR < RF$, $AR = RR$, and $AR > RR$. As such, it seems that observing such significant RT differences may represent a necessary but not sufficient condition to make such a claim. What is needed further is additional evidence of some unique pattern of outcomes on some other measure such as persuasion, which follows from the particular logic of resource matching theory and how this outcome measure in question should be affected if $AR$ are in fact greater, equal, or less than $RR$.

Although I can appreciate your desire for some other means of substantiating or verifying the levels of available relative to required resources, I am unable to offer any suggestions about how researchers might provide more definitive evidence. Indeed, I doubt whether this is possible, given that the levels of available relative to required resources can never be observed directly and thus can be inferred only indirectly via the effects they exert on necessarily indirect measures.

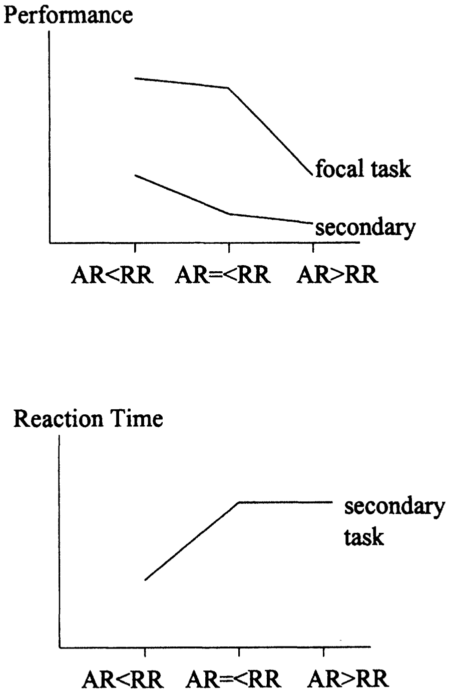


FIGURE 5 Relations between available and required resources.

Editor: Perhaps $RR$ needed to have been defined in the article more clearly. Perhaps this questioner took $RR$ to mean the resources required to complete the focal (advertising processing) task (which I think is the typical understanding); whereas the authors of the article conceptualized or operationalized $RR$ as the resources required to complete both the focal and secondary tasks. The questioner's perspective would predict that $AR = RR$ ought to yield results on the focal task performance different from those in $AR < RR$ (i.e., in the $AR = RR$ condition, the participant would have just sufficient resources to complete the primary task, but insufficient in the $AR < RR$ condition) but similar performance on secondary task requirements (i.e., in both the $AR = RR$ and $AR < RR$ conditions, the participant would not have had the opportunity to fully process the second task). (It may help to envision these equalities and inequalities through the use of Figure 5.) Even if the questioner and authors were defining $RR$ similarly, it is important to remember that no matter how $AR$ or $RR$ are operationalized, each is experienced as a cognitive, not directly observable process, so calibrations will be rough, exhibiting individual differences, though hopefully sufficient for comparing across conditions.

I think the issue can be resolved more clearly if, as Meyers-Levy suggests, a pattern of results is sought across multiple dependent measures. Distinctions may be refined and the process understood better if the researchers built an argument by simultaneously examining measures of performance on both the primary and secondary tasks, and RTs on at least the latter. If, say, we operationalize "resources available" simply as time allocation, it does seem puzzling that RT for $AR < RR$ could be less than for $AR = RR$, given that, when the allotted time is over and a response is required of the participant, the $AR = RR$ will have just completed the focal task, whereas the $AR < RR$ will not have done so, but they must also yield a response due to the time constraint, say 1,500 msec, so RTs for respondents in both of these conditions would be at ceiling. Hence, the RTs may be predicted to be $AR > RR$ (<) $AR = RR$ (=) $AR < RR$ (e.g., 1,000 msec < 1,500 msec = 1,500 msec). That is, the time constraint puts a ceiling on how high the RTs can be; if there were no time constraint, then the RTs in $AR = RR$ might be expected to be less than those in $AR < RR$, but it is the very time constraint that defines $AR$. (Of course, duration is not the only variable available for manipulating resources.)

The questioner states that the data in the article indicated that RTs were ordered: $AR > RR$ (<) $AR = RR$ (<) $AR < RR$. Perhaps we might conclude that the conditions, however conceptualized, were manifest operationally as, *AR* >>> *RR*, *AR* >> *RR*, and *AR* > *RR*. That is, available resources exceed those required in all conditions, differing in a matter of degree in which the calibration was not that fine (it is not clear how to do so error free), but the theoretical inquiry was one of relative comparisons, so was this result sufficient?

If the researcher would also examine some performance indicator, memory, or extent of processing data, then presumably performance on the focal task should follow the order suggested by the writers of the article referred to in the question—that is, *AR* > *RR* (>) *AR* = *RR* (>) *AR* < *RR*. That is, in the first condition, participants can succeed on the primary task and do so before the time constraint. In the middle (matching) condition, participants can succeed on the primary task, just completing it during the time allocated. In the insufficient resources condition, when time is up, participants' performance is incomplete, or they respond with less accuracy, given that the focal task would not have been completed. Therefore, even if RTs were equal on the secondary task for the latter two conditions, the processes may be inferred as different if data on a performance construct were also analyzed.