Escalation of Commitment and Nonorthogonal Analysis of Variance Revisited: A Comment on Schoorman, Bobko, and Rentsch

LEE A. KIRKPATRICK AND JOHN B. NEZLEK[1]
The College of William and Mary

Schoorman, Bobko, and Rentsch (1991) reanalyzed previously published data (Schoorman, 1988) and claimed that, in contrast to the original analysis, the data supported Schoorman's negative escalation hypothesis. We contend that this reanalysis was conceptually and technically flawed, and that Schoorman's original analysis yielded the correct (negative) conclusion. Schoorman et al. also discussed a variety of alternatives for conducting a factorial analysis of variance with unequal cell sizes, demonstrating that different procedures can lead to different statistical conclusions. However, their discussion is misleading and overlooks recent literature that has removed much of the mystery from the unequal-*n* problem. We offer a brief review of the fundamental problems, and their solutions, in analysis of variance with unequal *n*s.

In a recent article published in this journal, Schoorman, Bobko, and Rentsch (1991) reanalyzed previously published data (Schoorman, 1988) in which a theoretically expected interaction had been found to be nonsignificant. The reanalysis, which the authors regarded as more theoretically guided and more appropriate than the original analysis, yielded statistically significant results in support of Schoorman's (1988) hypotheses. In light of these discrepant findings, Schoorman et al. proceeded to discuss at length a variety of ways of conducting a factorial analysis of variance, demonstrating that different procedures for partitioning variance can lead to widely different statistical conclusions. We contend that the Schoorman et al. reanalysis was faulty both conceptually and technically, and that the nonsignificant interaction from the original analysis was in fact the appropriate test of Schoorman's hypothesis. Moreover, we regard their discussion of analysis of variance with unequal *n*s as confusing and potentially misleading, particularly with respect to issues that have been resolved clearly by previous researchers. Specifically, Schoorman et al. overlooked the substantial literature demonstrating the importance of considering cell and population weights in forming and testing hypotheses in the analysis of variance (e.g., Davidson & Toporek, 1977).

[1]Correspondence concerning this article should be addressed to either author at the Applied Social Psychology Research Institute, Department of Psychology, College of William & Mary, P.O. Box 8795, Williamsburg, VA, 23187-8795.

Journal of Applied Social Psychology, 1993, **23**, 2, pp. 98-106.

Nonorthogonal Analysis
nment on Schoorman,
entsch

JOHN B. NEZLEK[1]
n and Mary

analyzed previously published data
ast to the original analysis, the data
othesis. We contend that this reanal-
id that Schoorman's original analysis
rman et al. also discussed a variety of
of variance with unequal cell sizes,
id to different statistical conclusions.
overlooks recent literature that has
n problem. We offer a brief review of
in analysis of variance with unequal

urnal, Schoorman, Bobko, and
lished data (Schoorman, 1988) in
had been found to be nonsignifi-
regarded as more theoretically
inal analysis, yielded statistically
n's (1988) hypotheses. In light of
. proceeded to discuss at length a
alysis of variance, demonstrating
ariance can lead to widely differ-
t the Schoorman et al. reanalysis
ally, and that the nonsignificant
in fact the appropriate test of
ard their discussion of analysis of
potentially misleading, particu-
en resolved clearly by previous
overlooked the substantial litera-
onsidering cell and population
in the analysis of variance (e.g.,

ddressed to either author at the Applied
Psychology, College of William & Mary,

, pp. 98–106.

Testing Schoorman's Hypothesis

The study conducted by Schoorman (1988) was designed to examine the well-known escalation-of-commitment effect, according to which (for example) supervisors who promote employees subsequently rate those employees more favorably than supervisors who did not participate in the promotion decision (Bazerman, Beekun, & Schoorman, 1982). Schoorman (1988) hypothesized that a reverse or *negative escalation effect* would occur when supervisors *disagree* with promotion decisions made by a group. His data consisted of ratings of employees by supervisors who (a) either did or did not have input into the hiring or promotion decision of each employee, and (b) either did or did not agree with the group's promotion decision. The design and cell means for this 2×2 design are reproduced, following Schoorman et al. (1991), in Table 1. Schoorman predicted that supervisors who had input into and agreed with a promotion decision (Cell A) would show a positive escalation bias, and that supervisors who had input but disagreed with the decision (Cell B) would show a negative escalation bias, relative to supervisors who had no input (Cells C and D). Schoorman (1988) tested these hypothesis with a "traditional" analysis of variance and found the critical interaction effect to be nonsignificant, $F(1,347) = 3.51, p > .05$.

The Schoorman et al. reanalysis. Schoorman et al. (1991) argued, however, that a more appropriate test of this hypothesis would involve pooling the two no-input cells (C and D) and testing separate planned comparison of cells A and B, respectively, against the pooled C/D cell. According to their reasoning, the no-input groups were not expected to differ from each other, but both Cells A and B were expected to differ (in opposite directions) from the pooled C/D group. First, to justify pooling the groups they tested the difference between Cells C and D. After obtaining a nonsignificant result, they proceeded to test planned comparisons of Cell A versus pooled Cell C/D, and Cell B versus pooled Cell C/D. The tests of these contrasts were reported to be significant at the .01 and .001 levels, respectively. The authors concluded that both the positive and negative escalation hypotheses were supported.

Problems With the Schoorman et al. Reanalysis

In this section we argue that the reanalysis conducted by Schoorman et al. (1991) was faulty both conceptually and technically. We then discuss two alternatives for analyzing the data correctly and show that the data do not provide convincing statistical evidence for a negative escalation effect.

The pooling of Cells C and D. Conceptually, Schoorman et al.'s (1991) decision to combine Cells C and D of their design into a single group is

Table 1

Design and Cell Means From Schoorman (1988)

Input	Agree	
	Yes	No
Yes	Cell A 4.26 (142)	Cell B 3.47 (9)
No	Cell C 4.09 (195)	Cell D 3.86 (5)

Note. Adapted from Schoorman (1988) and Schoorman, Bobko, and Rentsch (1991). *N*s are in parentheses.

problematic because it rests on a strong assumption that employee ratings are similar in the two cells. No evidence was cited to support this assumption, and, in fact, there are good reasons to expect the means of these cells to differ. Supervisors were asked about the hiring or promotion decisions that led to employees being *in the jobs they now held.* Presumably, then, all of these hiring and promotion decisions were positive ones. Classifying a supervisor as having agreed with a promotion decision is therefore equivalent to classifying him or her as having *favored* promotion or hiring; supervisors classified as disagreeing with promotion decisions were those *opposed* to promotion/hiring. Assuming that agreement with promotion decisions is correlated with employee ratings, ratings should be higher on average in the agree cells (favored promotion) than in the disagree cells (opposed promotion). In other words, a main effect for the agree variable should be expected: The mean of Cell C should be greater than the mean of Cell D, and the mean of Cell A should be greater than the mean of Cell B. In light of this expectation, it is clearly inappropriate to pool the data from Cells C and D as if they together represented a homogeneous group.

The means reported for Cells C and D did indeed follow this pattern (see Table 1), but the difference between them was not significant according to a *t*-test. Schoorman et al. used this null finding as further justification for pooling the cells. However, this is a misguided strategy. Under the best conditions, null results from a significance test provide only weak support for the conclusion that the population means represented by two samples are equal. This is the age-old problem of "confirming the null hypothesis." The

n (1988)

Agree	
	No
	Cell B
	3.47
	(9)
	Cell D
	3.86
	(5)

d Schoorman, Bobko, and Rentsch

assumption that employee ratings was cited to support this assump- expect the means of these cells to iiring or promotion decisions that *ow held*. Presumably, then, all of /ere positive ones. Classifying a)tion decision is therefore equiva- *ored* promotion or hiring; super- tion decisions were those *opposed* ment with promotion decisions is hould be higher on average in the e disagree cells (opposed promo- igree variable should be expected: the mean of Cell D, and the mean of Cell B. In light of this expecta- lata from Cells C and D as if they).

id indeed follow this pattern (see vas not significant according to a ıding as further justification for guided strategy. Under the best test provide only weak support ıs represented by two samples are irming the null hypothesis." The

problem is exacerbated in the present case by low statistical power due to the small sample size ($n = 5$) in Cell D. The failure to find a statistically significant difference between Cells C and D cannot reasonably justify the pooling of the two cells.

Calculation of contrasts. Aside from the conceptual problem of assuming equivalence of Cells C and D, the method of constructing planned comparisons (contrasts) employed by Schoorman et al. was technically incorrect. Due to the manner in which Cells C and D were combined, these contrasts tested hypotheses that were highly dependent upon the specific cell sizes obtained in the sample. The first contrast, for example, tested whether the population mean corresponding to Cell A differs from a *weighted average* of the population means corresponding to Cells C and D, where the weights were a function of the sample sizes *that happened to occur in this particular study*. The aggregate C/D cell mean produced by Schoorman et al.'s method weighted the means of Cells C and D by their sample sizes, making the aggregate C/D cell virtually identical to Cell C. The aggregate mean against which Cells A and B were compared was 4.09, which is identical (within rounding error) to the mean of Cell C. (This can be determined from the "mean rating difference" in the top half of Table 1 in Schoorman et al., 1991.) Cell D contributed only 2.5% (5 of 200) of the cases to the aggregate C/D cell, so for all practical purposes Cells A and B were each compared to Cell C. The inappropriate influence of cell weights in this analysis can be further illustrated by observing what would have happened had the means of Cells C and D been the same, but their relative weights (i.e., cell sizes) been reversed: The mean of C/D aggregate cell would be 3.87 rather than 4.08. Different cell sizes would be inconsequential if the population means of Cells C and D were identical, but this strong assumption cannot be justified on either statistical or conceptual grounds.

The conceptual and statistical problems outlined above can be corrected easily by defining two contrasts somewhat differently. The most direct test of the positive escalation effect (Hypothesis 1 in Schoorman et al., 1991) would involve a comparison of Cell A versus Cell C: Among supervisors who agreed with promotion/hiring decisions, do those who had input rate employees more positively than those without input? Similarly, a direct test of the negative escalation hypothesis (Hypothesis 2) would involve a comparison of Cells B and D: Among supervisors who disagree with promotion/hiring decisions, do those with input rate employees less favorably than those without input? It should be clear that although the design has adequate power for testing the first hypothesis, there simply are not enough subjects in the disagree cells (B and D) to conduct a meaningful test of the second hypothesis. These comparisons, which each hold constant level of agreement in assessing the effect of input, directly reflect the specific hypotheses sug-

gested by the theory. Unfortunately, only one of them can be adequately tested with these data.

The omnibus Agree × Input Interaction revisited. An alternative way to conceptualize the primary hypotheses would be in terms of the *difference of differences* among cell means. The hypothesis to be tested would then be whether the difference between Cells A and C is equal to the difference between B and D. In other words, does the effect of input differ depending on whether the supervisor agreed or disagreed with the decision? The correct test of this hypothesis is the interaction term from a standard analysis of variance. This is the analysis originally conducted by Schoorman (1988), and the result was nonsignificant. Although inspection of the means clearly suggests an interaction, the null hypothesis of equal differences cannot be rejected with a high degree of confidence.

Analysis of Variance With Unequal Cell Sizes

Following the reanalysis of the Schoorman (1988) data, the remainder of the Schoorman et al. (1991) article was concerned with illustrating the diversity of results obtained by different analytic strategies when conducting analysis of variance with unequal cell sizes. This discussion reinvents an old wheel about which much has been written, and it does so in a confusing and misleading way. We feel it is important to try to resolve some of the confusion created by the Schoorman et al. discussion, as well as offer some recommendations based on the literature in this area.

First, it should be emphasized that every analysis illustrated by Schoorman et al. (1991) would produce exactly the same results if cell sizes were equal. The choice between an ANOVA or a regression program, and the order in which effects are tested, are irrelevant when cell sizes are equal. Differences among the analyses they conducted occurred simply because unless appropriate adjustments are made, main effects and interactions are not statistically independent when cell sizes are unequal. As a result, estimation of the effect of any one variable can depend on whether other variables are controlled (i.e., included) in the analysis. However, the fact that different strategies for partitioning variance may produce widely discrepant results should not be as troubling as Schoorman et al. (1991) seem to suggest: Once the null hypothesis of interest is identified, the correct analytic strategy can be determined unambiguously.

Determining the hypotheses of interest. When cell sizes are unequal in a factorial design, the first question that needs to be answered is "Why are cell sizes not equal?" Typically, unequal cell sizes simply reflect the fact that a researcher was for some reason unable to obtain equal cell sizes, and they do not reflect any meaningful state of affairs external to the analysis. Schoor-

man's (1988) data clearly fit this description. The relative distribution in a given population of agreement and disagreement with promotion decisions, and of involvement and noninvolvement with these decisions, did not figure into the formulation of the hypotheses; therefore, the relative sizes of the obtained samples should not have figured into the tests of these hypotheses. The correct analysis of these data should, in effect, approximate the results that would have been observed if equal cell sizes had in fact been obtained.

Once one has decided upon the null hypothesis of interest, the choice among analytic strategies is generally unambiguous. Most of the analyses presented by Schoorman et al. (1991) are simply wrong for the vast majority of purposes; they test null hypotheses about population means that simply do not correspond to the questions with which researchers are generally concerned. It seems to us that the discussion by Schoorman et al. creates potential for confusion by giving equal billing to all of these various approaches and presenting the reader with an unnecessarily bewildering array of options.

Unweighted means. When unequal cell sizes are not meaningful (the case in virtually all social psychological research), there is a standard, accepted procedure to test the effects in an analysis of variance with unequal cell sizes. This procedure is used by BMDP-2V, SYSTAT's MGLH/ANOVA, SAS's PROC GLM (Type III/IV sums of squares), and SPSS-PC/SPSSx's ANOVA ["Regression Approach": METHOD = UNIQUE or Option 9, depending on the version] and MANOVA [METHOD = SSTYPE(UNIQUE)]. Each of these programs conducts the correct analysis by default, except for the ANOVA program in SPSSx and SPSS-PC.

In a factorial ANOVA in which variance is partitioned into omnibus main effects and interactions, this approach weights all cell means equally and tests hypotheses about linear combinations of these means. Main effects are conceived in terms of particular linear combinations of means, namely *means of means*. In Schoorman's data, the main effect for the agree variable should test the null hypothesis that in the populations represented by the four cells, the mean of the means of Cells A and C is equal to the mean of the means of Cells B and D. The corresponding marginal sample means being compared are similarly calculated as means of sample means. Note that these unweighted marginal means are different from the weighted means reported by Schoorman et al. (1991): For Agree = Yes, the mean of the means = (4.26 + 4.09)/2 = 4.18; for Agree = No, the mean of the means = (3.47 + 3.86)/2 = 3.67. In this case the resulting values are quite similar to the unweighted means of 4.16 and 3.61, respectively, because Cells A and C are roughly equivalent in size (as are Cells B and D). In contrast, the difference between weighted and unweighted means is striking when the input variable is considered. The unweighted marginal mean for Input = Yes is (4.26 + 3.47)/2, or

3.87; for Input = No the value is (4.09 + 3.86)/2, or 3.98. Notice that this difference is not only much smaller than the difference between weighted means reported by Schoorman et al. (4.21 vs. 4.08, respectively), it is even reversed in direction! Because of drastic differences in cell sizes, comparison of the weighted means reported by Schoorman et al. for the two Input groups essentially amounts to a comparison of Cell A with Cell C. Cells B and D (where Agree = No) had little effect on the assessment of the input main effect due to their small *ns*.[2]

Testing main effects. Of all the analyses reported by Schoorman et al. (1991), *only one* correctly tests the null hypothesis about means of population means. This is the analysis reported at the top of their Table 4, based on a regression analysis using effect-coding and simultaneous entry.[3] This analysis yields a significant main effect for agreement and nonsignificant effects for input and for the Agreement by Input interaction. The nonsignificant input effect corresponds to the trivial difference between unweighted means calculated above (3.87 vs. 3.98), in contrast to the larger (and reversed) difference between weighted means reported by Schoorman et al. (4.21 vs. 4.08). As noted above, Schoorman also would have obtained the correct results had he used the default procedure in virtually any major analysis-of-variance program other than SPSSx ANOVA, or had he specified Option 9 in this program.

Testing interactions. The null hypothesis that the difference between the population means corresponding to Cells A and C is equal to the difference between the population means corresponding to Cells B and D (or, alternatively, A − B = C − D) is tested by the interaction term of this same analysis, and yields a nonsignificant result. Note that in contrast to main effects, the test of the interaction turns out to be identical in many of Schoorman et al.'s (1991) other analyses as well. This is because the interaction term in ANOVA is defined essentially in terms of the between-cell variability that remains after main effects have been accounted for, and all of these analyses do this correctly even though they yield different estimates of the main effects. Different results are obtained only if the interaction term is entered alone

[2]Unweighted marginal means can be requested in most major computer programs, again with the exception of SPSS ANOVA. The latter program will not print any means at all when the "Regression Approach" is requested, and will print only weighted marginal means under other options. In this case unweighted marginal means must be calculated by hand, as we have done in the present example.

[3]For technical discussions see Carlson and Timm (1974), Overall and Spiegel (1969), and Overall, Spiegel, and Cohen (1975). Ultimately, the issue is defined in terms of the null hypotheses being tested by each approach when *ns* are unequal. Carlson and Timm (1974) proved that the null hypothesis tested by procedures other than the one recommended here concern linear combinations of population means weighted by exceedingly complex (and, usually, theoretically meaningless) functions of sample sizes.

without controlling for main effects, in which case the interaction is hopelessly confounded with one or both main effects and the results are utterly uninterpretable. It is unfortunate that Schoorman et al. (1991) reported results for such analyses; they are incorrect and can serve only as a source of confusion.

Planned comparisons. As noted above, main effects can be conceptualized as planned comparisons involving linear combinations of population means, where the weights assigned to these means are determined by the investigator. For example, the main effect for agreement in the Schoorman (1988) data should be thought of as a test of the hypothesis that the mean of population means corresponding to Cells A and C is equal to the mean of population means corresponding to Cells B and D. If this is the null hypothesis posed by the investigator, as is almost invariably the case in theory testing, the analysis strategy discussed here yields the correct results.

On the other hand, the researcher is free to assign differential weights to these means and test other null hypotheses given a rationale for doing so. We agree wholeheartedly with Schoorman et al.'s general message about the importance of using planned comparisons to test specific hypotheses dictated by one's theory, although we have argued against the specific procedure they adopted. Omnibus main effects and interactions represent only one way of apportioning variance which, in a given application, may or may not correspond to the specific hypotheses of interest. In the case of Schoorman's (1988) data, however, the test of the interaction in the traditional ANOVA (with unweighted means) is correct for his purposes.

Alternative strategies. Our discussion has focused on the situation in which the researcher is interested in testing hypotheses about causal effects and in which unequal sample sizes are not theoretically meaningful. We have also maintained that this is nearly always the case in social-psychological research. Although there is an established solution to the problem of unequal cell sizes that do not reflect attempts to generalize to populations with specific parameters, there is some disagreement about what to do when weights are meaningful. The disagreement focuses not on the technical aspects of the problem (i.e., the mathematics of calculating F-ratios), but rather on the formulation of hypotheses and the extent to which statistical tests examine the hypotheses of interest. For example, if the goal of a study is to describe the relationships among variables in a particular population, and a sample is drawn specifically to reflect accurately the makeup of the population, it might be argued that weighted rather than unweighted means provide the more appropriate tests of the hypotheses of interest. Moreover, a researcher might want to assign specific weights to sample means to reflect a known population distribution that is not accurately represented in a sample. Such procedures are common in epidemiological research, for example,

when minority subpopulations are deliberately oversampled for statistical efficiency, and differential weights are later applied to reconstruct the known population distribution. BMDP'S PC-90 program BMDP-4V (see Dixon, 1990) provides such options.

Analysis of variance is a difficult topic and, as demonstrated by Schoorman et al. (1991), the wide variety of analytic options available can yield surprisingly diverse results. We agree that it is imperative that researchers report the specific strategy employed in their analyses; simply stating that a two-way analysis of variance was conducted is insufficient when sample sizes are unequal. However, recent advances in the statistical literature have removed much of the mystery that once surrounded the topic. Once the null hypotheses of interest have been clearly specified, choosing an analytic option is a straightforward matter.

References

Bazerman, M. H., Beekun, R. I., & Schoorman, F. D. (1982). Performance evaluation in a dynamic context: A laboratory study of the impact of prior commitment to the ratee. *Journal of Applied Psychology,* **67**, 873–876.

Carlson, J. E., & Timm, N. H. (1974). Analysis of nonorthogonal fixed-effects designs. *Psychological Bulletin,* **81**, 563–570.

Davidson, M., & Toporek, J. (1977). Analysis of variance with general cell weights. *Proceedings of the Statistical Computing Section of the American Statistical Association*, 174–179.

Dixon, W. J. (1990). *BMDP statistical software manual, Vol. 2.* Los Angeles: University of California Press.

Overall, J. E., & Spiegel, D. K. (1969). Concerning least squares analysis of experimental data. *Psychological Bulletin,* **72**, 311–322.

Overall, J. E., Spiegel, D. K., & Cohen, J. (1975). Equivalence of orthogonal and nonorthogonal analysis of variance. *Psychological Bulletin,* **82**, 182–186.

Schoorman, F. D. (1988). Escalation bias in performance appraisals: An unintended consequence of supervisor participation in hiring decisions. *Journal of Applied Psychology,* **73**, 58–62.

Schoorman, F. D., Bobko, P., & Rentsch, J. (1991). The role of theory in testing hypothesized interactions: An example from the research on escalation of commitment. *Journal of Applied Social Psychology*, **21**, 1338–1355.