

METHODOLOGICAL DEVELOPMENTS

Main Effects Analysis in Clinical Research: Statistical Guidelines for Disaggregating Treatment Groups

John S. Lyons
Northwestern University Medical School

Kenneth I. Howard
Northwestern University

Treatment outcome research generally relies on main effects analysis of variance to determine whether treatments are differentially effective. Bryk and Raudenbush (1988) developed a decision strategy for disaggregating treatment groups under conditions of heterogeneity of variance. There is, however, reason to consider disaggregating main effects even when this assumption is not violated. The potential statistical significance of disaggregation can be shown to be a function of the reliability of the dependent measure. With this reliability, residual variance can be partitioned into a systematic (individual differences) component and a random error component. It is then possible to calculate an F test of the ratio of these variances. When this F is statistically significant and the proportion of within-cell systematic variance to total variance is large, disaggregation should be undertaken to search for important individual or treatment difference variables (i.e., interactions).

Analysis of variance (ANOVA) models have been the standard statistical technique whereby the efficacy of treatment interventions have been evaluated in clinical research. These models use the F statistic, the ratio of variation across groups to variation within each group, to determine whether group members are more like each other on the dependent measure(s) or whether subjects are essentially indistinguishable on the basis of group membership.

ANOVA assumes, among other things, that the variation within each group is an estimate of the population variance of the dependent measure. Within-treatment variations are pooled across treatment groups to form this estimate. Thus, in main effects analysis, variation of the dependent measure is partitioned into that attributable to constant treatment effects and that attributable to other sources of variation. ANOVA generally refers to this second type of variation as "error."

There are some logical shortcomings of this approach when applied in practice. Pooling within-treatment variation across groups requires the assumption that each within-group variation is estimating the same parameter (the population variance). Bryk and Raudenbush (1988) have persuasively demonstrated that heterogeneity of variance, a common phenomenon in evaluation research (Light & Smith, 1971), may result from differential treatment effects within treatment groups (interactions). Bryk and Raudenbush (1988) suggest that when variance

heterogeneity is encountered, individual and treatment differences within the treatment group should be explored.

However, their approach also suggests that individual differences that interact with treatments may be present even when there is homogeneity of variance. First, treatments may have an overall significant main effect on group means in addition to important individual differences in treatment response. That is, even when the assumptions of homogeneity of variance are met and a statistically significant main effect for treatment is detected, individual differences may still be interacting with treatment effects. Regardless of the treatment, it would not be surprising to find that some individuals respond better to a specific intervention than do others and that this differential responsiveness can be predicted by other measured independent variables (e.g., patient, setting, therapist, and treatment characteristics).

Second, advances in clinical research, as well as changes in ethical guidelines given these advances, generally dictate that new treatments be compared with standard-of-care interventions. New treatments and the standard-of-care treatment could be expected to have important, albeit different, interactions with independent variables. There is no reason to expect an absence of differences in variation between or among groups to reflect the absence of important individual differences in treatment responsiveness. Even inactive treatments such as placebos are known to have effects on dependent measures (e.g., Elkin et al., 1989).

What is needed, therefore, is a decision rule that provides guidance regarding whether a main effects analysis is sufficient in order to fully understand the relative effects of different treatments. In order to accomplish this, it is useful to review how variance is partitioned in main effects analysis and to conceptualize the mean square error (MS_e) in terms of some of its components.

This work was partially supported by Research Grant R 01 MH42901 from the National Institute of Mental Health.

We gratefully acknowledge the comments of Albert Erlebacher and Merton S. Krause.

Correspondence concerning this article should be addressed to Kenneth I. Howard, Department of Psychology, Northwestern University, 102 Swift Hall, Evanston, Illinois 60208-2710.

In a one-way ANOVA, total variation is partitioned as follows:

$$SS_t = SS_b + SS_e.$$

That is, the total variation (SS_t) is equal to the sum of the variation between groups (SS_b) and the variation within groups (SS_e). The variation between groups is assumed to be created by differential treatment effects and random variation. The variation within groups is assumed to represent random variation (error). For the purposes of significance testing, these sums of squares are converted to mean squares (MS_b and MS_e) by dividing by the degrees of freedom for each.

The MS_e , however, has several distinct components (each with the same degrees of freedom). Consistent with classical (and modern) psychometric theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Gulliksen, 1950; Lord & Novick, 1968), each subject's score on any variable consists of a combination of a true score on that measure and the error of measurement. Following from this, the MS_e can be usefully partitioned into two components: the variation associated with the subjects' true scores on the dependent measure and the variation associated with error of measurement.¹ The variation of true scores would be expected to reflect "real" differences among subjects and should thus vary depending on systematic individual differences among subjects. The remaining variation, attributable to measurement error, should be a reflection of the unreliability of the measurement of the dependent variable. Thus the MS_e can be partitioned as follows:

$$MS_e = MS_{er} + MS_{es},$$

where MS_{er} represents measurement error and MS_{es} represents systematic variation. In this conceptualization, there are two basic sources of systematic variation (MS_{es}). First, reliable individual differences within groups can create systematic variation that would become part of the MS_{es} . Second, reliable treatment differences can lead to systematic variation (interaction) that would also become part of the MS_{es} .

Given that the MS_e can be further partitioned, the portion attributable to measurement error can be estimated, and a significance test can be performed allowing a decision rule regarding the utility of exploring systematic variation within each treatment group. The presence of measurement error in MS_e is a function of the reliability of the dependent variable. That is, to the degree that the dependent variable is measured with perfect reliability ($r = 1.00$; $MS_{er} = 0$), the variation within groups is a result only of systematic effects. To the degree that the dependent variable is measured with no reliability ($r = .00$; $MS_{es} = 0$), the MS_e is a result only of measurement error. It follows, then, that the MS_e not attributable to systematic individual (and treatment) variations can be expressed as the proportion of unreliable variation in the dependent variable ($1 - r$) times the amount of variation not accounted for by treatment effects (MS_e). Therefore

$$MS_{er} = (1 - r)MS_e,$$

where r is the reliability of the measure of the dependent variable. Note that r is a reliability coefficient (i.e., it is already a

proportion of variation estimate and therefore should not be squared; Nunally, 1976).

The amount of variation that is attributable to systematic variation is the product of the proportion of reliable variation in the dependent variable (r) and MS_e :

$$MS_{es} = (r)(MS_e).$$

Because MS_{er} and MS_{es} are both variances, the ratio of these variances can be expressed as follows:

$$F = (MS_{es}/MS_{er}) = [(r)(MS_e)]/[(1 - r)(MS_e)].$$

However, since MS_e is in both the numerator and the denominator, this F can be reduced as follows:

$$F = [(r)(MS_e)]/[(1 - r)(MS_e)] = [r]/[1 - r].$$

This F would be significant in situations where the systematic variation in MS_e was sufficiently larger than the variation due to error of measurement. Each component of MS_e variance is estimated with n observations. Therefore, the degrees of freedom for this F would be $(n - 1, n - 1)$.

Thus, the initial decision of whether it is useful to explore the possibility of individual and treatment variations within groups is a function of the reliability with which the dependent variable is measured. The more reliable the measure, the more likely systematic variation contributes significantly to within-treatment group variation (MS_e). With a reliability of zero, the F is 0. With a reliability of .50, the F is 1. As the reliability exceeds .75, the F will be above 3.0.

Given a reliable measure and an adequate sample size, the bulk of MS_e in main effects analysis will be attributable to systematic variation (between and within groups), and the F to disaggregate will generally be statistically significant. However, this statistic provides no information regarding the potential information value of disaggregating. If there are very strong main effects (or interactions in a factorial model) using a reliable dependent measure, the ratio of MS_{es} to MS_{er} would be large even though MS_{es} might be quite small. Thus, it is necessary to perform a second step: Estimate the proportion of the total variation that is attributable to additional systematic variation beyond the main effects (and specified interactions) of treatment. One can determine this proportion as SS_{es} divided by SS_e . However, the investigator has to determine the information value of such a proportion in any particular study. In power (and meta) analysis, it has become customary to use effect size for this determination. Cohen (1977) has suggested .50 as an indicator of a "moderate" effect size. This effect size translates into a proportion of SS_{es} to SS_e of .33. Thus, if the observed proportion of remaining systematic variation is less than .33, disaggregation may not be worth pursuing.

Examples From the Literature

In order to demonstrate the importance of considering the disaggregation of treatment groups in clinical research, the results of two well-known outcome studies are reviewed. In the

¹ Zimmerman and Williams (1986) partitioned MS_e using psychometric theory to study the relationship between reliability and power.

first example, there was no significant main effect, whereas in the second there was. For each of these studies, the F to disaggregate and the proportion of total variation attributable to systematic variations (other than treatment main effects) are presented in Table 1.

Elkin et al. (1989) in the National Institute of Mental Health Treatment of Depression Collaborative Research Program used two primary dependent measures—the Hamilton Depression Rating Scale (HDRS; Hamilton, 1967) and the Global Assessment Scale (GAS; Endicott, Spitzer, Fleiss, & Cohen, 1976). The reliabilities of the two measures were .93 and .83, respectively. Applying the first step resulted in an $F(151, 151)$ of 13.27 ($p < .001$) for the HDRS and an $F(151, 151)$ of 4.88 ($p < .001$) for the GAS. The second step, estimating the proportion of systematic variation in the residual, indicated that 93% of the observed total variation (among treatment completers) on the HDRS and 83% on the GAS were attributable to systematic variation (real individual differences in response to each treatment condition). In the absence of main effects, Elkin et al. (1989) did disaggregate on the basis of severity of depression and discovered clinically important treatment effects.

Kazdin, Bass, Siegel, and Thomas (1989) studied the treatment of child antisocial behavior and evaluated three treatments: a cognitive-behavioral intervention, a cognitive-behavioral intervention plus in vivo practice, and a relationship therapy intervention. They found significant treatment effects on the Child Behavior Checklist (CBCL; Achenbach & Edelbrock, 1983). The reliability of the CBCL was .92. The $F(82, 82)$ to disaggregate was 11.54 ($p < .001$). In the second step, it was found that 84% of the observed total variation in this measure was attributable to systematic variation other than treatment main effects. This is in spite of using the estimated posttest means after covarying pretest levels in determining the variation between treatments. Even in the presence of main effects, Kazdin et al. (1989) also conducted a secondary analysis in which they disaggregated their groups on the basis of initial severity (scoring outside the normal range), again discovering important additional effects.

Discussion

The immediate clinical purpose in conducting comparative treatment research is to decide to which treatment subsequent

Table 1
F to Disaggregate and Proportion of Variance Potentially Attributable to Individual Differences in Selected Studies

Measure	r	MS_b	MS_e	MS_{et}	MS_{et}	F
HDRS	.93	0.77	33.11	2.32	30.79	13.27**
GAS	.83	4.28	118.54	20.15	98.39	4.88**
CBCL	.92	13.31	3.26	0.26	3.00	11.54** ^b

Note. HDRS = Hamilton Depression Rating Scale. GAS = Global Assessment Scale. CBCL = Child Behavior Checklist.

^a $df = 151, 151$. ^b $df = 82, 82$.

* $p < .001$.

patients would be most beneficially assigned. This clinical goal must be kept clearly in mind when data are analyzed, so that the ability to assign patients to the best treatment is maximized. If one treatment is statistically significantly better (has a significantly higher mean outcome) than another, the clinician will be inclined to assign all subsequent patients to the better treatment (once it has been determined that this is indeed a reliable finding). And if no treatment is reliably significantly better than another, the clinician will be inclined, other things being equal, to assign subsequent patients indifferently to one or the other. All of this seems quite unexceptional and is quite logical if and only if a certain familiar and simple statistical model holds, if all the variance in outcomes within groups is residual variance (i.e., due to random error: patient sampling error, treatment application error, and outcome measurement error) and the groups' sample means are unbiased estimators of their respective population means. Here, we have provided a model to test for the existence of reliable individual differences in responsiveness within treatments.

Bryk and Raudenbush (1988) have provided a decision tree for studying individual and treatment variations in main effects research that suggests that when heterogeneity of variance is observed, the identification of individual difference variables followed by a corrected reestimation of main effects can be undertaken. However, even if the variance within treatment groups is homogeneous and regardless of whether or not main effects for treatment are found, in most circumstances it appears useful to disaggregate treatment groups and study potential sources of systematic variation.

Perhaps the most important conclusion from the above analysis is the demonstration that in most cases main effects analysis of treatment outcome is not likely to tell the whole story. An investigator should know, before undertaking an investigation, the extent to which this will be a problem. If the reliability of a dependent measure is zero, no differences among subjects are real. If the reliability is perfect, then all differences are real. Real differences among subjects should be attributable to systematic variations due either to individual differences or to variations in the application of treatments. Two examples from well-designed and controlled studies (Elkin et al., 1989; Kazdin et al., 1989) clearly demonstrate the potential information value in disaggregating main effects for treatment.

Bryk and Raudenbush (1988) state the importance of considering potential individual difference variables that might interact with treatment in the design of the study. Such consideration appears crucial. A decision on what variables to include in the measurement is a complicated one. A choice of too many variables threatens α , the probability of finding nonreproducible results. Failure to include important variables, however, limits the investigator's ability to predict differential treatment outcome. The growing consensus in treatment outcome research is not identifying what treatments work, per se, but rather what treatments work for which patients (DeAngelis, 1990). Our model of disaggregation demonstrates that there is a strong statistical rationale for this approach.

As one anonymous reviewer suggested, we have provided a further "argument for (a) why researchers ought to be using measures of high reliability, and (b) why authors ought to use

larger sample sizes—in both cases, they facilitate further searches for moderator and predictor variables.”

References

- Achenbach, T. M., & Edelbrock, C. S. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: University Associates in Psychiatry.
- Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*, 396–404.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.
- Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.
- DeAngelis, T. (1990, June). Value of medication versus psychotherapy debated. *APA Monitor*, p. 17.
- Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., Glass, D. R., Pilkonis, P. A., Leber, W. R., Docherty, J. P., Fiester, S. J., & Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program: General effectiveness of treatments. *Archives of General Psychiatry*, *46*, 971–983.
- Endicott, J., Spitzer, R. L., Fleiss, J. L., & Cohen, J. (1976). The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disturbance. *Archives of General Psychiatry*, *33*, 766–771.
- Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.
- Hamilton, M. A. (1967). Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology*, *6*, 278–296.
- Kazdin, A. E., Bass, D., Siegel, T., & Thomas, C. (1989). Cognitive-behavioral therapy and relationship therapy in treatment of children referred for antisocial behavior. *Journal of Consulting and Clinical Psychology*, *57*, 522–535.
- Light, R. J., & Smith, P. V. (1971). Accumulating evidence: Procedures for resolving contradiction among different studies. *Harvard Educational Review*, *41*, 429–471.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Nunally, J. (1976). *Psychometric theory*. New York: Wiley.
- Zimmerman, D. W., & Williams, R. H. (1986). Note on the reliability of experimental measures and the power of significance tests. *Psychological Bulletin*, *100*, 123–124.

Received December 7, 1990

Revision received April 12, 1991

Accepted April 12, 1991 ■