

COMPLEX ANALYSES OF VARIANCE: GENERAL PROBLEMS*

BERT F. GREEN, JR.

LINCOLN LABORATORY, MASSACHUSETTS INSTITUTE OF TECHNOLOGY

AND

JOHN W. TUKEY

PRINCETON UNIVERSITY

Problems in applying the analysis of variance are discussed. Emphasis is placed on using the technique to understand the data. The scale of the dependent variable is important for the analysis. Crossed and nested categories must be recognized. The error terms in the analysis depend on whether the classes of each independent variable are (1) all out of a few or (2) a few out of many. To simplify the analysis, mean squares should be aggregated with their error term when they are less than twice its size. An illustrative example is discussed in detail.

The framework of experiment and the additive decomposition of observations that are associated with the analysis of variance, form what is perhaps our single most powerful methodological technique. Kogan [14] has recently reviewed the use of this technique in psychological research. His review shows that the analysis of variance has often been used with simple experimental designs involving only two or three factors, but that instances of its use in complex experiments are rare. It is our purpose to discuss some of the problems that arise in complex analysis of variance, and to suggest methods for dealing with these problems. To be discussed in the present paper are (1) the choice of the dependent variable to be analyzed, (2) the shape of the analysis, (3) the choice of the proper error term, and (4) the aggregation and pooling of mean squares.

Throughout the discussion we shall emphasize what may be considered the major purposes of the analysis of variance: to provide a simple summary of the variation in the experimental data, and to indicate the stability of means and other meaningful quantities extracted from the data (and thus to make more precise our understanding of *how much* has been learned from the experiment). Many investigators believe that the sole purpose of the analysis of variance is to provide statistical tests of significance, and some seem to equate these to tests of meaningfulness. We hope to counteract such views by showing how the analysis of variance can be used to summarize

*Prepared in connection with research sponsored in part by the Office of Naval Research, in part by the Educational Testing Service, and in part by Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology with the joint support of the U. S. Army, Navy, and Air Force.

the data effectively and to help in understanding what "goes on" in the experimental situation. While we shall rely on the conventional F -test to give some guidance, the primary function of the analysis of variance is to help the investigator understand his data. As such, it may need to be used more than once on the same data. As such, it deserves guidance from graphs and other devices for seeking understanding. It should not be an end in itself.

The Illustrative Example

The psychophysical experiment that serves as the illustration for the discussion is that used by Johnson and Tsao [12] as a methodological example. The detailed observations are in Johnson's textbook [11]. In the experiment, difference limens for weights were measured by a method of continuous change. An aluminum pail was attached by a lever system to a ring on the subject's finger. One of seven weights—100, 150, 200, 250, 300, 350, or 400 grams—was placed in the pail. By controlling a flow of water into the pail, the experimenter increased the load in the pail at one of four constant rates—50, 100, 150, or 200 grams per 30 seconds—until the subject reported a change in pull. The difference limen, DL , in grams, was measured by the amount of water added.

Four men and four women served as subjects. Two of each sex were congenitally blind, the others had normal vision. Age was partially controlled; all women were 33, while one blind and one normally sighted man were 21, and the two remaining men were 25.

Two practice trials were given at the start of the first session. In the experiment the order of presentation of the weight-rate combinations was randomized. For each weight-rate combination, five determinations of the DL were made for each subject. The mean of these five measurements was used in the analysis. The entire experiment was carried out on each of two days, one week apart.

Choice of the Dependent Variable

In setting up an analysis of variance, we must first decide what scale to use for the dependent variable. We want to choose a scale that will yield the simplest relations with the independent variables. By simplest relations we mean, for example, fewer important interactions, and larger main effects relative to the error variance. A change of variable that nearly removes a particular main effect also usually leads to very revealing results. Secondly, we would like the dependent variable to have approximately homogeneous variance within cells of the design. (Indeed, if nothing else could be altered, some attention might be given to making the within-cells distribution moderately normally distributed.)

Past experience with similar data often suggests the appropriate scale

to use. In other cases the results of one analysis may suggest a reanalysis with a transformed dependent variable. In Johnson and Tsao's experiment, the *DL* might be expressed as a difference in weight, a squared difference in weight, a ratio of weights, or a logarithm of a ratio of weights. It might also be expressed as a response time. In fact, analysis of the experiment will show that the response-time measure leads to a strikingly simple account of the data.

The Shape of the Analysis

The next step in setting up an analysis of variance is to specify for what main effects, interactions, etc., it would be appropriate to compute mean squares. In many texts the simplest experimental designs are described as "single classification," "double classification," "triple classification," etc. In this terminology, the present experiment has six classifications: sex, sight, weight, rate, date, and person. The relations among these classifications determine what interactions are available.

The most basic relation between two classifications is *crossing*. Two classifications are completely crossed if each value of one classification is paired with each value of the second classification. Thus rate and weight are completely crossed in the present experiment, since each selected weight appears with each selected rate. Sex and sight are similarly crossed. In this experiment, not only does each combination occur, but every combination occurs an equal number of times—the experiment is balanced. Completeness and balance of crossing greatly facilitate both analysis and interpretation of results. They are often built into planned experiments, although they occur much less frequently in unselected situations. Height and weight are incompletely crossed in human populations. Such combinations as height 3 feet, weight 300 pounds, or height 7 feet, weight 30 pounds are nonexistent, or at least very unlikely. When all variables are crossed in the experiment, all combinations of variables may be represented by interaction terms in the analysis.

A second common relation between classifications is *nesting*. One classification is nested in another classification, if the inner has meaning only within a single class of the outer. Thus person is nested in sex, since any particular person will be of a particular sex. When one classification is nested in a second, one cannot average over the outer classification without automatically averaging over the inner, or nested, classification. Nesting occurs, for example, in psychological experiments when subjects are divided into groups. The subjects are thus nested in the groups. The analysis includes a main effect and a set of interactions (again with other variables) for subject within group. In many replicated experiments, the replication is a special case of a nested classification. The replication is nested in the combination of all other variables in the experiment, since one must average over repli-

cations before averaging over any other classification. In other replicated experiments, replication is crossed with some or all of the other variables.

In Johnson and Tsao's experiment sex, sight, weight, rate, and date are all crossed, while person is nested in the combination of sex by sight. One cannot obtain such interactions as person $\times$ sex, person $\times$ sight $\times$ rate, etc., since one cannot average over sex or sight without first averaging over persons. There are such interactions as weight $\times$ person, but this is actually the weight $\times$ person (sex $\times$ sight) term, and is interpreted as the interaction of weight with "persons within the sex-sight combinations." Since the analysis of such situations is less familiar than for the completely crossed case, some sample computational formulas are given in the appendix.

Many experiments exhibit only combinations of crossing and nesting. The shape of the analysis can always be determined by identifying those means that are meaningful. In other experiments, more complicated relationships may occur. In cross-nesting, a variable which could be crossed with another is forced to nest within it instead. In complete or partial confounding, two or more variable are "mixed-up" together to a greater or less extent.

Variance Components and the Proper Error Term

There has been some confusion about the appropriate error term against which the main effects and interactions should be evaluated. Should all mean squares, or *lines*, be tested against the residual term, or should some lines be tested against one or more of the interaction terms? The solution to this problem depends on a choice of a model for the experiment. The analysis of variance is based on a model in which the dependent variable consists of the sum of several contributions; these contributions depend upon the nature of the independent variable and the way in which the experiment was performed. There is no definite model which is appropriate for all instances, even for all instances of what may seem, at first glance, to be exactly similar experiments. However, many models can be regarded as special cases of a single, flexible, general model.

In general, models differ in the assumptions that are made about interactions, and about the nature of the variables, i.e., whether the values at which each variable appears are fixed, random, or of some intermediate nature. It has been shown ([22]; [6]; [3], pp. 366 ff; [25]) that there is no need to make restrictive assumptions about interactions in order to obtain the most important results—those that determine the correct error term. The fixed and random cases can be interpreted as the extreme cases of sampling of actual values from a (usually finite) population of potential values (see same references). Thus models for the analysis of variance have been greatly expanded since the appearance in 1947 of the, now classical, paper by Eisenhart [8]. Discussion of this development may be found in Mood [15], Kempthorne [13],

Anderson and Bancroft [1], Crump [7], Wilk and Kempthorne [26], Scheffé [19], and Bennett and Franklin [3].

To illustrate and explain the general model, an auxiliary example from the field of psychological testing has been chosen. A number of persons $i = 1, 2, \dots, p$ are selected at random from a population of P persons, and a number of possibly equivalent tests $j = 1, 2, \dots, t$ are selected at random from a population of T tests. Each selected test is given to each selected person. Two scores are obtained for each test by a split-half method.

A model will be presented for these data that includes populations of P persons, T tests and two halves. One can then treat the experimental persons, and tests, as samples from the populations of persons, and tests. When a sample exhausts the population, the corresponding variable is *fixed*; when the sample is a small (i.e., negligible) part of the population the corresponding variable is *random*. The model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

where

- μ = common or general contribution,
- α_i = contribution specific to person i ,
- β_j = contribution specific to test j ,
- γ_{ij} = contribution specific to the combination of person i on test j ,
- ϵ_{ijk} = contribution specific to the combination of person i , test j , and half k ,

and where i runs from 1 to P (over the whole population of persons), j runs from 1 to T , and k runs from 1 to 2.

It is assumed here that *half* is nested in the combination of test and person, and is essentially a replication within this combination. If each test has been split into two equivalent subtests, this is the reasonable assumption. If the halves are not really equivalent, so that, for example, a certain sort of individual might do systematically better on half 1 of test j than on half 2 of the same test, while a second sort would do the reverse, then this assumption would be unreasonable. In the latter case, one would have to add an interaction between persons and "halves within tests." If the halves were the first and second halves of speeded tests, then one would also have to add a main effect of half, and to allow persons to interact directly with "half." For the present example suppose that neither of these complications has arisen.

It is reasonable to require that the average of γ_{ij} over all i in the population does not depend on j . This can be done since any apparent dependence on j is certainly specific for the test (in this population) and is represented by β_j . Notice that it would not make good sense to make a similar requirement for the means of γ_{ij} over only the selected persons, since such a require-

ment would force one to let the β 's depend on the sample of persons selected. In order for the model to apply to the population, the β 's cannot depend on any particular sample from the population. Similarly, one requires that the average of γ_{ij} over all T tests, $j = 1, 2, \dots, T$, used and unused, should not depend on i .

Since the ϵ 's represent the error variation in the model, hopefully and for simplicity, assume that they represent truly erratic error, and can themselves be treated as a sample of $2PT$ values from an infinite population. For the sake of the analysis, it is assumed that there is a very large—infinite—number of ways to split each test into equivalent halves. This assumption is closely related to the assumption that the halves are really equivalent.

Now define the variance components for persons, tests, interactions, and between-halves as follows:

$$\begin{aligned}\sigma_\alpha^2 &= \frac{1}{P-1} \sum_i (\alpha_i - \alpha_-)^2 \\ \sigma_\beta^2 &= \frac{1}{T-1} \sum_j (\beta_j - \beta_-)^2, \\ \sigma_\gamma^2 &= \frac{1}{(P-1)(T-1)} \sum_{ij} (\gamma_{ij} - \gamma_{--})^2, \\ \sigma_\epsilon^2 &= \frac{1}{2PT-1} \sum_{ijk} (\epsilon_{ijk} - \bar{\epsilon})^2,\end{aligned}$$

where the summations are over whole populations, and not merely over the indices appearing in the sample, and where “—” subscripts refer to averages over the corresponding populations, that is,

$$\alpha_- = \frac{1}{P} \sum_i \alpha_i, \quad \beta_- = \frac{1}{T} \sum_j \beta_j, \quad \gamma_{--} = \frac{1}{PT} \sum_{ij} \gamma_{ij}$$

$\bar{\epsilon}$ = average of ϵ in an infinite population.

With these definitions, Tukey [22], Cornfield and Tukey [6], Bennett and Franklin [3], and Wilk [25] have shown that in repetitions of the whole procedure, each time independently selecting new samples of persons, and of tests, the average values of the mean squares found by analysis of variance

$$\begin{aligned}\text{for persons:} \quad & \sigma^2 + 2\left(1 - \frac{t}{T}\right)\sigma_\gamma^2 + 2t\sigma_\alpha^2 \\ \text{for tests:} \quad & \sigma^2 + 2\left(1 - \frac{p}{P}\right)\sigma_\gamma^2 + 2p\sigma_\beta^2 \\ \text{for interaction} \quad & \sigma^2 + 2\sigma_\gamma^2; \\ \text{between halves} \quad & \sigma^2\end{aligned}$$

Here, as in many simple situations, the variance components σ^2 , σ_α^2 , σ_β^2 , σ_γ^2 are defined as the mean squares that would correspond to the analysis of variance of *all* the possible values in the model. In this example, an infinite number of halves of each of the T tests applied to each of the P persons (without practice effect) define this analysis.

Clearly, the average mean squares depend on the relation of t to T , and of p to P . If $t = T$, and $p = P$, one has a completely fixed model, since the sample has exhausted the population. Repetitions would lead to exactly the same set of values for the α 's and β 's. In this case, the proper error term for the main effects is the between-halves mean square. If t and p are, respectively, very much smaller than T and P , so that t/T and p/P are essentially zero, one has a random model so far as persons and tests are concerned. Here repeated sampling would lead to different values of the α 's and β 's. The proper error term for the main effects is now the interaction mean square.

One might, alternatively, have a mixed model, with $p/P = 0$, and $t = T$. The average mean squares would now be

$$\begin{array}{ll} \text{for persons:} & \sigma^2 + 2t\sigma_\alpha^2; \\ \text{for tests} & \sigma^2 + 2\sigma_\gamma^2 + 2p\sigma_\beta^2; \\ \text{for interaction} & \sigma^2 + 2\sigma_\gamma^2; \\ \text{between halves:} & \sigma^2 \end{array}$$

To inquire whether there are differences among the persons is equivalent to inquiring whether $\sigma_\alpha^2 = 0$. If it vanishes, then the average mean square for persons equals the average mean square between halves. The correct F -test is thus the ratio of the two corresponding mean squares.

To inquire whether there are differences among the tests is inquiring whether $\sigma_\beta^2 = 0$. If it vanishes, the average mean square for tests equals the average mean square for interaction. Thus the correct F -test has an *interaction* mean square, and not a *between-halves* mean square in the denominator.

The difference in the denominators seems strange at first, but on further consideration it becomes very meaningful. When one speculates about the equivalence of tests for the whole population of persons, one finds that all the tests have been applied to only a few out of many persons. Thus if persons and tests interact, another sample of persons would give systematically different comparisons of tests. The interaction mean square must be used. Things are quite different indeed when comparing persons. Every test has been applied to each selected person; redrawing the sample of tests would make no difference since one still uses the same tests. Here one should use the between-halves mean square.

In rare cases t/T and p/P might be finite fractions, say 0.8 and 0.1. Then the appropriate error term for the main effects would be a corresponding

weighted average of the interaction and between-halves mean squares. (Bennett and Franklin [3] and Cornfield and Tukey [6] give the explicit formulas.) Similar situations occur in examining the mean square for one variable when the interactions of this variable with at least two other variables are significant. In these cases, the F -distribution is surely an approximation, and a modified technique is needed to find an appropriate denominator. The procedure used in the analysis presented below is described in the Appendix.

Except for scales, i.e., quantitative variables (whose treatment must be postponed) and certain rather rare cases, the classification in an experiment will usually be of two types.

- (i) A finite number of classes, each of which is represented in the experiment.
- (ii) A large finite, or infinite, number of classes, only a few of which are represented in the experiment.

When we have decided the type of each of the classifications in the experiment—fixed (i) or random (ii)—then we can determine the average mean square for each line in the analysis. The rule is that the average mean square for a line includes the variance components for the interaction of that line with *random* classifications or combinations of *random* classifications. For example, if the classifications are labeled A, B, C, D, E , etc., and if C is fixed and D is random, then the AvMS for AB would include variance components for ABD but not for ABC or $ABCD$. If both C and D are random, it would include ABC , ABD and $ABCD$. The same rule applies for nested variables: if C is nested in B then the AvMS for AB would include the variance component ABC if C is random, but not if C is fixed. Note carefully that the types of A and B are *irrelevant* when choosing the error term for AB .

Variance Components in the Illustrative Example

In the difference-limen experiment, person is clearly random, and sex is clearly fixed. Although we could conceive of a continuum on which “complete sight” and “total blindness” are extreme, we naturally assume that in this experiment attention is directed at these extremes as special cases, so it is best to consider sight as fixed.

In the case of date we must beware of a practice effect. Insofar as the two dates bring in weather conditions, nearness to a week-end, etc., they may be reasonably well represented as a sample of two from a large number, but insofar as practice or learning is concerned they are two very particular stages. Since only two practice trials were given before the experiment was begun, there is no reason to suppose that the practice effect is so small as to be unmeasurable. The importance of this effect depends on the extent to which the habits, skills, and attitudes developed during the first session were retained over the intervening week between the two sessions. In any

event, all our data refers to subjects in this general state of moderate practice, and the main effects of all variables except date must, therefore, refer to moderate practice. Since it seems likely that the contribution of the date-to-date difference to the effects and interactions of other variables is more likely to be irregular than systematic, we decide to treat these two dates as a small sample from many possible dates.

Rate and weight present problems. We apparently need a third type of variable, a scale, a few categories of which are selected for the experiment. These values have very probably been spaced at regular intervals along the scale, so we cannot interpret them strictly as a small random sample from many possible classes. On the other hand, we are not interested only in the particular selected classes. Instead, these classes are used to represent the continuous variable in the analysis. Two different procedures can be adopted in analyzing classifications that represent scales. One is simple and familiar, but requires a somewhat cavalier interpretation of the variance components; the other is unfamiliar, but more rigorous and effective.

The simple technique is to treat scales as categories, using the "few out of many" model. This technique will show whether the scale has any significant main effects or interactions, since the null hypothesis is not violated by the technique. However, the size of the variance components will depend on the range of selected values from the scale. For example, if there is a monotonic relation between the scale and the variable being analyzed, the variance components will usually increase as the range studied increases. The variance components for scales, then, indicate the relative importance of the particular range of the scale included in the analysis. In order to discover "what is going on," the investigator must return to the data, perhaps graphing the observed relation of the response selected for analysis to the scale. The analysis of variance essentially tells him which graphs are important.

The rationale for treating scaled variables as random rather than fixed is far from well established. Bennett and Franklin [3] felt that they had to adopt one policy or the other, and elected to treat scaled variables as fixed. Wilk and Kempthorne [26] likewise treat scales as fixed. The present authors chose to treat them here as random. Surely the last word has not yet been spoken on this issue.

While psychologists have usually treated scales as categories in analyzing complex experiments, there is a more rigorous procedure in which the nature of the relationships between scales and the analyzed variable are an integral part of the analysis. Orthogonal polynomials may, for example, be used to fit curves describing the relationships. The analysis separates the smooth effects, represented by the fitted curve (and perhaps summarized by the coefficients of the polynomials) from the erratic effects, represented by the variation around the fitted curve (see Grant [9]). A detailed exposition of

curve-fitting techniques will be postponed, since such a discussion in the present context would introduce complexities that might obscure the main points of the argument. Consequently, at present rate and weight are treated as random classifications, although we recommend against this procedure in general.

The model for the difference limen experiment can now be represented by Table 1, which shows, schematically, the average mean square for each line in the analysis. The independent variables are denoted S = sex, I = sight, P = person, R = rate, W = weight, and D = date. Whenever P is written, $P(SI)$ is understood, since persons are nested in the SI cells, so that the presence of a P implies the presence of SI . In our schematic presentation we have omitted the σ^2 factors, and also the numerical coefficients. In each case the coefficient is the number of observations for any set of fixed values of all the variables represented by the component. For example, the coefficient of σ_{RP}^2 is the number of observations for a fixed person at a fixed rate, which is 7 weights $\times$ 2 dates, or 14. Thus, the notation RP in the AvMS column indicates 14 σ_{RP}^2 . The error term for each line is immediately apparent from the table of AvMS's.

Aggregation and Pooling

In complex experiments there are many mean squares. If it is possible to obtain a simple description of the experimental results, many of these mean squares will not be significantly different from their error terms. Some investigators prefer to pool the nonsignificant mean squares in order to get a more stable error term for further significance tests. In our example of the split-half tests, the pooling problem comes up when the interaction mean square and the between-halves mean square are comparable. In any event, a more stable error term would be obtained by pooling the sums of squares for interactions and for between-halves and dividing by the pooled number of degrees of freedom. If the interactions variance component were zero, this pooled error would be appropriate. If the interactions variance component were large, then pooling would yield too large an error term for testing lines not "above" the interaction term, and too small an error term for testing lines "above" the interaction term. When the interaction mean square is near the between-halves mean square, the proper balance of these alternatives—the choice to pool or not to pool—is difficult.

Two extreme points of view are prevalent, one optimistic and the other pessimistic. The optimistic extremists refuse to admit the existence of the interactions variance component unless an F -test applied to interactions and between-halves mean squares surpasses a chosen level of significance. In the absence of significance they pool. This policy is extremely hard to justify. The pessimistic extremists refuse to admit that the interactions mean square is ever zero, so they never pool.

TABLE 1

Average Mean Squares (AvMS) and
Coefficients (DIV) for DL Experiment

Line	DIV	AvMS (Schematic)*
Lines Not Involving P, S, or I		
R	112	RWDP+RWP+RDP+RP+RWD+RW+RD+R
W	64	RWDP+RWP+WDP+WP+RWD+RW+WD+W
D	224	RWDP+RWP+WDP+DP+RWD+RD+WD+D
RW	16	RWDP+RWP+RWD+RW
RD	56	RWDP+RDP+RWD+RD
WD	32	RWDP+WDP+RWD+WD
RWD	8	RWDP+RWD
Lines Involving S But Not I		
S	224	RWDP+RWP+RDP+WDP+RP+WP+DP+P+RWDS +RWS+RDS+WDS+RS+WS+DS+S
RS	56	RWDP+RWP+RDP+RP+RWDS+RWS+RDS+RS
WS	32	RWDP+RWP+WDP+WP+RWDS+RWS+WDS+WS
DS	112	RWDP+RDP+WDP+DP+RWDS+RDS+WDS+DS
RWS	8	RWDP+RWP+RWDS+RWS
RDS	28	RWDP+RDP+RWDS+RDS
WDS	16	RWDP+WDP+RWDS+WDS
RWDS	4	RWDP+RWDS
Lines involving I but not S are omitted: I, RI, WI, DI, RWI, RDI, WDI, RWDI; they may be obtained from the corresponding line above involving S but not I by replacing S by I. E.g.,		
RWI	8	RWDP+RWP+RWDI+RWI
Lines involving SI in combination are omitted: SI, RSI, WSI, DSI, RWSI, RDSI, WDSI, RWDSI; they may be obtained from the corresponding line above involving S but not I by replacing S by SI and halving the value of DIV. E.g.,		
RWSI	4	RWDP+RWP+RWDSI+RWSI
Lines Involving P		
P	56	RWDP+RWP+RDP+WDP+RP+WP+DP+P
RP	14	RWDP+RWP+RDP+RP
WP	8	RWDP+RWP+WDP+WP
DP	28	RWDP+RDP+WDP+DP
RWP	2	RWDP+RWP
RDP	7	RWDP+RDP
WDP	4	RWDP+WDP
RWDP	1	RWDP

* Each set of letters represents a term consisting of a σ^2 with that set as subscript, appearing with a coefficient equal to the DIV for the line of that set. E.g., RWP signifies $2\sigma_{RWP}^2$ wherever it appears. Thus, in detail, the AvMS for RW is

$$\sigma_{RWDP}^2 + 2\sigma_{RWP}^2 + 8\sigma_{RWD}^2 + 16\sigma_{RW}^2$$

The subject has been reexamined by Paull [17] who investigated the power of alternative tests both analytically and empirically. He concluded that it is desirable to combine sums of squares and degrees of freedom when the F -ratio is less than twice the 50 percent point of the F distribution with the corresponding degrees of freedom. As a simpler rule of thumb that gives about the same results when each mean square has more than six degrees of freedom, he suggested pooling when the F -ratio is less than two. (Notice that an empirical rule is used, *not* a significance test.) Bozovich, Bancroft, and Hartley [2] have also studied the pooling problem in detail. Their recommendations are more precise and more complicated than Paull's. The differences appear to be critical mainly when there are very few degrees of freedom and when the exact level of significance is at issue. Since we are mainly concerned with experiments with several degrees of freedom in the error terms, and since we attach minor importance to exact significance levels, we favor using Paull's rule of thumb. When mean squares are combined according to some form of this rule, we shall speak of *aggregation* rather than *pooling*.

To describe the aggregation procedure, some definitions are needed. If the average mean square (AvMS) for one mean square, or line, contains all the terms in the AvMS for another line, the first line is *above* the second line, and the variance components that appear in the first AvMS and not in the second are in the *difference* of the two lines. If the two lines are aggregated, the variance components in their difference are *neglected*.

The rules for aggregation are: a line should be aggregated with a basic line only when (a) the first line is above the basic line, (b) the first mean square is less than twice the basic mean square, (c) no variance component in the difference of the two lines has been, or can be, shown by another application of (b) to be a variance component that should *not* be *neglected*.

The aggregation procedure starts with the highest-order interaction or the replication error term; the lines that can be aggregated with it according to these rules are found and aggregated with the initial (basic) line by pooling sums of squares and degrees of freedom. Next, we determine which lines can be aggregated with the highest-order interaction among those lines not already aggregated in the first step. We proceed in this manner throughout the table of mean squares, being particularly careful about rule (c). A detailed example of the aggregation process is given in the illustrative analysis below.

Once mean squares have been aggregated according to this generalization of Paull's rule of thumb, we are in a position to consider whether further pooling is in order. It is conceivable either to keep the once-aggregated lines carefully separate, or to test the significance of aggregated lines, and pool nonsignificant lines with their error terms. The choice between these actions must depend on our purpose in making the analysis.

If our purpose is to identify those lines whose contributions demonstrably cannot be constant, then it does not matter. In any event the aggregated values, not the aggregated and pooled values, should be used as error terms. (If our purpose is to make confidence statements about these contributions, the situation is the same.)

If our purpose is to summarize the apparent distribution of variability among the several contributions, assigning variability to only those contributions which are *unequivocally* (at the chosen significant level) not constant, then somewhat of an argument can be made for pooling after aggregation. For by doing this we force all the variability into "significant" contributions, which *may* sometimes be desirable. (The counterargument is that we know, in reality, each line corresponds to the contribution of *some* variability.)

If our purpose is to make a generally useful summary of the distribution of variability, then the aggregated values should serve our needs. There has been some compression in the interest of apparent simplicity, but not too much. If our purpose is to make a point estimate of an individual variance component, then we should not even aggregate. Even a failure to *prove* a variance component nonzero does not justify an estimate of zero. (We may well wish to supplement our point estimate with an interval estimate in any case. Currently available methods have been discussed by Bross [4], whose method has been slightly modified by Tukey [21], pp. 66–68.)

In stating estimated variance components, we are often trying to arrive at a simple (and at times rough) description of the data. From our point of view, all variance components are almost sure to be nonzero. However, many may be so small that a general overview of the data is more effective when they are omitted. In these situations estimates of variance components should show the relative importance of the variables. Statistical significance is secondary.

Kogan [14] notes that some investigators do not calculate high-order interactions, but lump them with the residual term for a composite error term. This saves some work, since the high-order interactions require lengthy calculations, but, as Kogan has pointed out, involves dangers. If some of the lumped interactions are large, the composite error term will be too large, so that precision may be sacrificed. Moreover, information about these large interactions will be lost, and we are less likely to inquire about their causes. A nice balance between extra work in analysis and possible loss of valuable information is required. It usually seems to pay to go at least somewhat further than has been the typical practice in the past. Going all the way is likely to be wiser than not going far enough.

If a complex analysis, which has not been taken all the way, has a number of mean squares notably smaller than the mean square for pooled error, the analysis *should be pushed further*. The original Johnson and Tsao analysis provides a clear example where this rule should be applied. Their error

mean square is 154 and, above this error term, there are 12 mean squares less than 30.

The Illustrative Example Again

An analysis of the difference-limen experiment will provide a detailed example of the aggregation and estimation procedures that we suggest for complex experiments. Following Johnson and Tsao, we shall first use the *DL*'s, in grams, as the dependent variable. Table 2 shows the sums of squares, degrees of freedom, original mean squares, and aggregated mean squares for the experiment. Each "line" is represented by the entries in corresponding positions of the various two-way tables—its designation being found by combining row and column labels.

In the aggregation process, begin with *RWDP*, the lowest line in the table, whose observed mean square is 26. Every combination of *RWD*, *RD*, or *RW* with any of *P*, *SI*, *S*, *I*, and "—" is less than 52. Thus aggregate all of these. *WP* (observed MS = 39), *DS* (observed MS = 9) and *DI* (observed MS = 4) are also less than 52, but cannot be aggregated unless one also aggregates *WDP* (observed MS = 74) and *WDS* (observed MS = 106) which lie between—and these are too large for the rule of thumb. It may seem too bad not to aggregate a 9 and 4, but we must be consistent.

Combining the 15 sums of squares and 15 sums of degrees of freedom,

$$1043 + 320 + \cdots + 61 + \quad + 205 + 527 = 6381,$$

$$18 + \cdots + 3 + \cdots + 18 + 18 = 312$$

The aggregated mean square is $6381/312 = 20$, which is labeled by *RWDP*, the label of the lowest line involved.

The lowest line now remaining is *WDP* with an observed mean square of 74. One can aggregate *WDSI*, *WDS*, *WDI*, *WD*, *WP*, and *WS* with this. (While *DS*, *DI*, and *D* are small enough, *DP* lies between with an observed mean square of 178.) The aggregated sum of squares and degrees of freedom are

$$943 + 405 + 1779 + 379 + 637 + 340 + 488 = 4971,$$

$$24 + 6 + 24 + 6 + 6 + 6 + 6 = 78$$

The aggregated mean square is $64 = 4971/78$.

The lowest lines now left are *DP*, *WSI*, *WI*, *W*, *RP*, no one of which is part of another.

DP can be aggregated with *DSI*, *DS*, *DI*, and *D* to yield

$$712 + 270 + 9 + 4 + 116 = 1111$$

$$4 + 1 + 1 + 1 + 1 = 8$$

and an aggregated mean square of 139.

TABLE 2
Initial Results of Analysis
(Weight and Rate Treated as Random)

Sums of squares				Degrees of freedom				Observed mean squares						
P	SI	S	I	-	P	SI	S	I	-	P	SI	S	I	
23,969	14,130	31,534	11,810		4	1	1	1	1	5,992	14,130	31,534	11,810	
R	3,970	2,216	5,720	2,083	51,077	12	3	3	3	331	739	1,907	694	17,024
W	943	1,406	405	2,089	1,909	24	6	6	6	39	234	68	348	318
D	712	270	9	4	116	4	1	1	1	178	270	9	4	116
RW	1,043	320	215	654	391	72	18	18	18	14	18	12	36	22
RD	139	60	61	14	61	12	3	3	3	12	20	20	5	20
WD	1,779	379	637	340	488	24	6	6	6	74	63	106	57	81
RWD	1,883	270	538	205	527	72	18	18	18	26	15	30	11	29

No other aggregation is possible at this step. We have reduced 39 lines to 15 and can reasonably proceed to the remainder of the analysis.

All aggregated analysis of variance tables will have the following columns.

Line = label of interaction or main effect.

DF = degrees of freedom.

MS = mean square.

AvMS = average mean square.

DEN = value of denominator, or error term for this mean square.

DDF = denominator degrees of freedom.

SIG = level of significance of F -ratio, $F = MS/DEN$; * = 5%, ** = 1%, *** = 0.1%.

DIV = coefficient of the variance component of this line in AvMS.

s^2 = estimated variance component = $(MS - DEN)/DIV$.

pooled s^2 = estimated variance component computed from pooled mean squares (insignificant MS's being pooled with their DEN's).

The final analysis is shown in Table 3. In the AvMS column appear only those variance components for which lines are present in the (aggregated) table. The main effects of Person and Rate predominate. The variance component for S is estimated to be large, but the mean square is not significantly larger than its error term. In addition to the main effects of P and R , there are some small, but significant interactions: RS , RP , W , WI , WSI , DP , and WDP .

At this point, it is interesting to compare our results with those of Johnson and Tsao. In the analysis of the difference limen experiment, Johnson and Tsao neglected the Persons variable. All Person interactions, and the main effect of Persons, were lumped in the error term. Thus the error term was made too large. On the other hand, all interactions were tested against this error term. Insignificant interactions were pooled with the error term to test the main effects. The significant lines according to their analysis were SIR , SI , SR , IW , IR , S , I , W , and R . These overlap our significant lines in only three cases (R , W , and WI).

In summarizing the results of the experiment, we must have a graph showing the mean limen for each person at each rate, since R , P , and RP are all significant. Other graphs are needed, but the implications of this graph, shown in Figure 1, force us to reconsider our analysis. For each person, the line connecting the four rate points is nearly linear, and, if extended, would nearly meet the origin. (We might well have noted this phenomenon before the analysis, and saved ourselves much labor! The implication is that the effect of Rate can be explained by assuming that each person in the experiment responds after a constant Time, regardless of the Rate.) If this is true, we would obtain a more clear-cut analysis by using response times, rather than different-limen in grams, as the dependent variable. It will be

TABLE 3

Analysis After Initial Aggregation

Line	DF	MS	AvMS (Schematic)	DEN	DDF	SIG	DIV	s ² [†]	Pooled s ²
I	1	11,810	RWDP + WDP + DP + RP + P + WI + RI + I	6639	4.8		224	23	
	1	31,534	RWDP + WDP + DP + RP + P + RS + S	7568	5.6		224	107	
SI	1	14,130	RWDP + WDP + DP + RP + P + RSI + WSI + SI	6570	4.7		112	68	
P	4	5,992	RWDP + WDP + DP + RP + P	450	17.5	***	56	99	200
R	3	17,024	RWDP + RP + R	331	12	***	112	149	149
RI	3	694	RWDP + RP + RI	331	12		56		
RS	3	1,907	RWDP + RP + RS	331	12		56	28	28
RSI	3	739	RWDP + RP + RSI	331	12		28	15	
RP	12	331	RWDP + RP	20	312	***	14	22	31
W	6	318	RWDP + WDP + W	64	78	***	64	4	4
WI	6	348	RWDP + WDP + WI	64	78	***	32	9	9
WSI	6	234	RWDP + WDP + WSI	64		**	16		11
DP	8	139	RWDP + WDP + DP	64	78		28		3
WDP	78	64	RWDP + WDP	20	312	***	4	11	11
RWDP	312	20	RWDP				1	20	20
	447							575	466

[†] The corresponding unaggregated s² are, in order, 24, 108, 33, 99, 137, 6, 28, , 23, 4, 9, 6, 4, 12, 26 (and 24 others totalling -10).

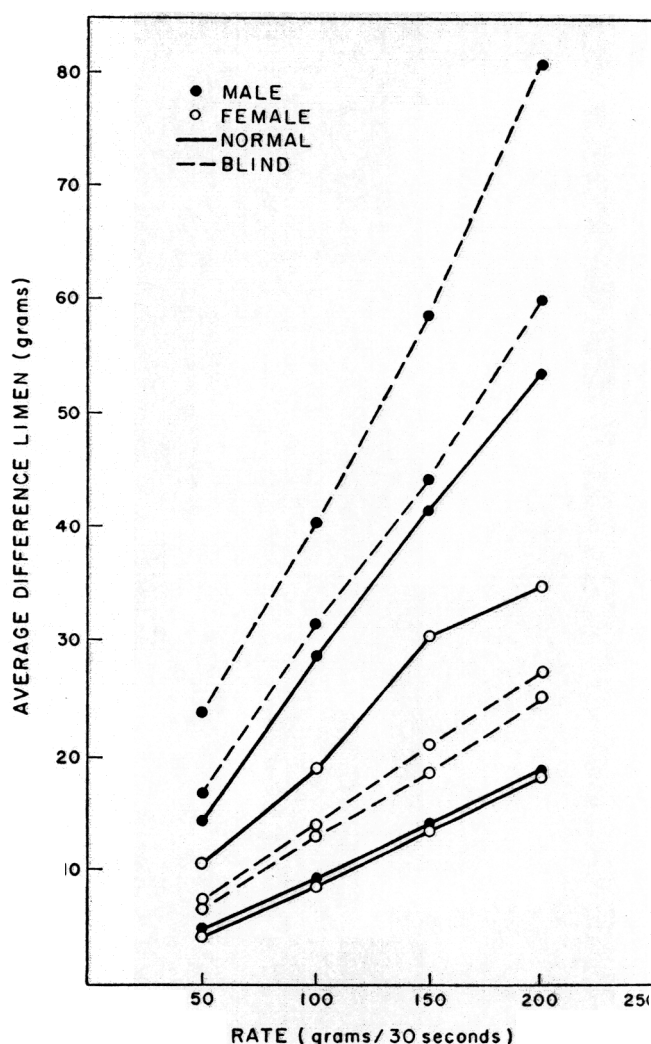


FIGURE 1
Average *DL* for Each Person-Rate Combination

instructive to compare an analysis of response times with the analysis of *DL*'s.

Reanalysis in Terms of Response Times

When the data were transformed to response times, it was found that the standard deviations were approximately linearly related to the means. This circumstance, which is usually found in latency data, suggests using the logarithm of response time as the dependent variable for the next analysis, so that the standard deviations would be more homogeneous. Table 4 shows the complete set of mean squares for the analysis of log times. Table 5 shows the final analysis-of-variance table.

The main effect of Person clearly predominates all other effects in this

TABLE 4

Mean Squares for Analysis 2 (Based on Log Response Times)					
	P	SI	S	I	-
-	27,894	34,021	76,860	37,852	-
R	62	87	40	130	36
W	48	147	19	539	
D	731	2091	489	36	
RW	22	25	21	21	45
RD	30	72	32	31	19
WD	63	60	63	40	39
RWD	28	10	25	13	16

analysis. Figure 2 shows the mean response times for each individual on each day. (Psychophysical experiments often show individual differences in sensitivity, which may be especially marked in unpracticed observers and can usually be reduced substantially by thorough training. The choice of naive or trained observers depends on the purpose of the experiment.)

The rate components were almost entirely eliminated by the trans-

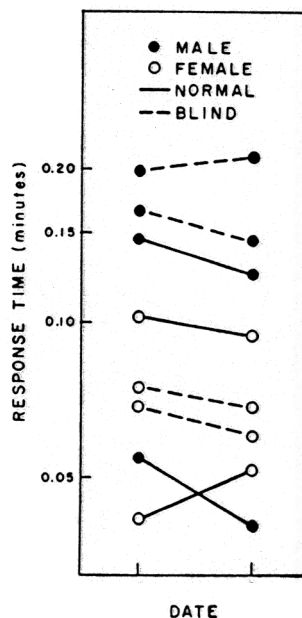


FIGURE 2

Response Time for Each Person on Each Date

(Arithmetic means of logarithms are equivalent to geometric means of times.)

TABLE 5

Analysis of Log Response Times (Analysis 2)

Line	DF	MS($\times 10^4$)	AvMS (Schematic)	DEN	DDF	SIG	DIV	s^2	Pooled s^2
I	1	37,852	RWDP + WDP + DP + RP + P + RI + WI + I	28,454	4.2		224	42	
S	1	76,860	RWDP + WDP + DP + RP + P + S	27,894	4		224	219	
SI	1	34,021	RWDP + WDP + DP + RP + P + RDSI + WSI + DSI + SI	29,505	4.4		112	40	
P	4	27,894	RWDP + WDP + DP + RP + P	658	7.8	***	56	486	652
RI	3	130	RWDP + RP + RI	58	21		56	1	
RP	21	58	RWDP + RP	24	309	***	14	2	3
W	6	551	RWDP + WDP + W	51	78	***	64	8	8
WI	6	539	RWDP + WDP + WI	51	78	***	32	15	15
WSI	6	147	RWDP + WDP + WSI	51	78	*	16	6	6
DSI	1	2091	RWDP + WDP + DP + DSI	624	7		56	26	
DP	7	624	RWDP + WDP + DP	51	78	***	28	20	27
WDP	78	51	RWDP + WDP	24	309	***	4		7
RDSI	3	72	RWDP + RDSI	24	309		14	3	3
RWDP	309	24	RWDP				1	24	24
447								899	745

The corresponding unaggregated s^2 are, in order, 45, 220, 21, 484, 1, 3, -1, 16, 12, 12, 24, 2, 9, 28 (and 25 others totalling -19).

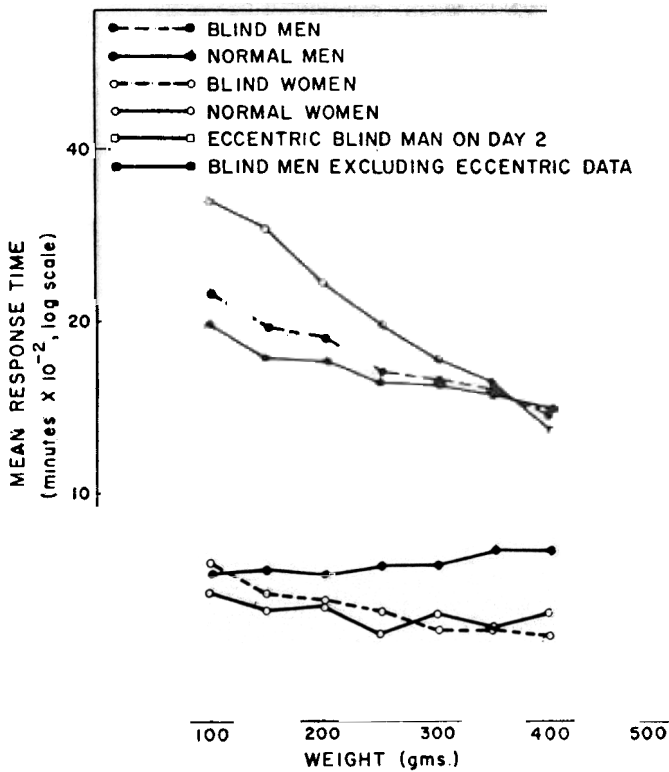


FIGURE 3
Average Response Time for Each Weight-Sex-Sight Combination

formation to response times. The conclusion that response time was virtually constant for the four rates of increase (for fixed person, weight, and date) has important implications for the method of continuous change. It suggests that one can get whatever difference limen one wants by choosing the presentation rate appropriately. Both Pal [16] and Grindley [10] found that *DL*'s were functions of presentation rate, though neither found the proportionality noted in this study. We also note that the effect of weight in the present study was not at all in accord with the Weber function. The peculiar effects of weight and rate indicate that the method of continuous change may be very different from the usual psychophysical methods.

Among the secondary effects, the person-date interaction is large (see Fig. 2). The weight, weight-sight, and weight-sex-sight lines are small but significant. Figure 3 shows the mean response time for weight-sex-sight combinations. The weight-date-person interaction is caused almost entirely by the atypical behavior of one individual, whose data are shown on Fig. 3. All other persons showed much less variation on the weight variable. Since the form of the atypical curve is similar to the form of many *DL* curves near the absolute threshold, some sort of pre-experimental adaptation effect might explain the anomaly.

Portrayal of Variability

We have emphasized that the true purpose of analysis of variance is increased understanding of the results of the experiment, and have made use of graphs of mean values to this end. Estimates of variability, too, are likely to require graphical presentation. The columns of estimated variance components in Tables 3 and 5, for example, can be pored over and gradually understood to a degree. But, if we include the unaggregated variance components, there are six sets, three from each analysis. We may wish to compare them, and will often find cumulative graphs, like Fig. 4, helpful in doing this.

No one line in Fig. 4 tells the whole story. The first three are based on *DL*'s in grams, and we have learned that this is unwise. The other three are based on logarithmic response times, and show no trace of the original apparent rate dependence. We can present the whole picture more completely, as in Fig. 5, if we take rate as estimated, from the pooled components in Table 3, to provide 32 percent of the variability, and divide up the remaining 68 percent among the pooled components in Table 5. We can gain effectiveness by using the two dimensions of the picture. (This also makes visibility directly related to standard deviation.)

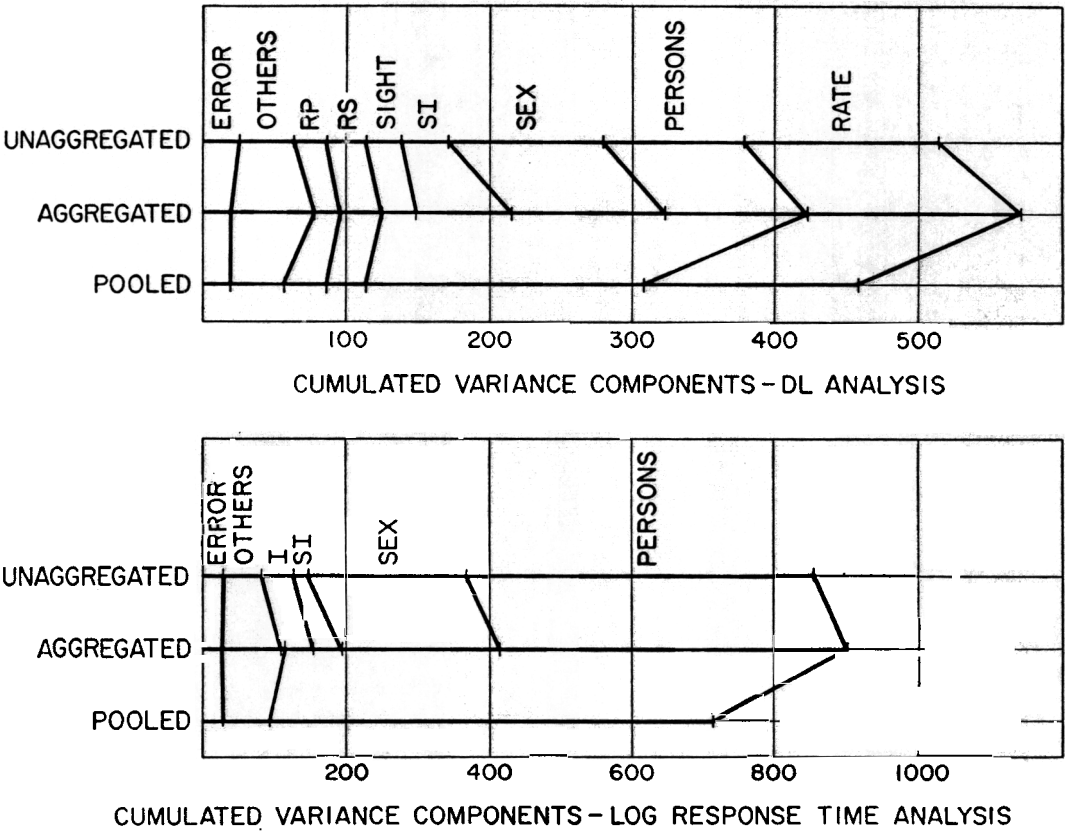


FIGURE 4
Cumulative Graphs of Variance Components

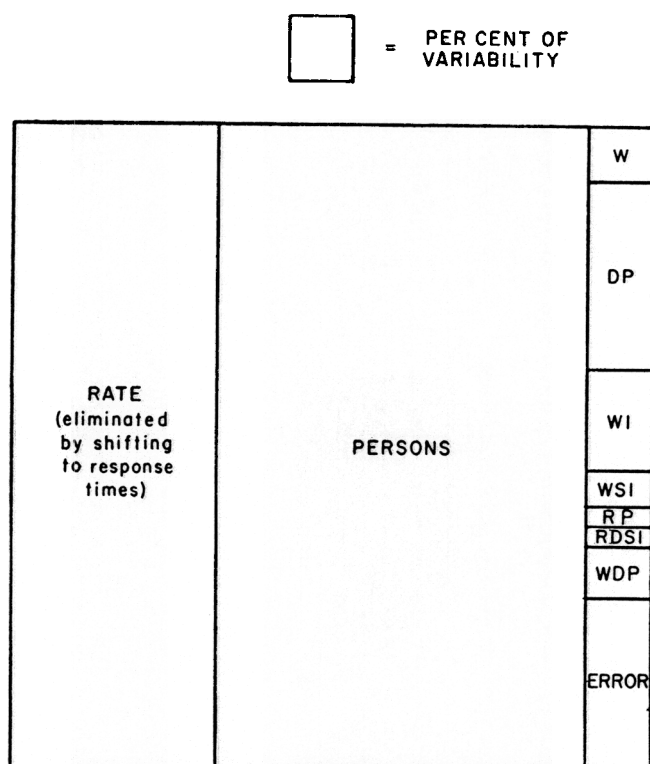


FIGURE 5
Relative Variability in *DL* Experiment

Summary

Application of the analysis of variance to complex experiments is discussed, with emphasis on the use of the analysis to understand the data. It is noted that the choice of the scale for the dependent variable is important. It is also important to recognize crossed and nested categories. The variance components model is used to specify the structure of the experiment and to determine the appropriate error term for each mean square. The error terms depend on whether the classes of each independent variable are (1) all out of a few or (2) a few out of many. A process of aggregation is described and illustrated to reduce the number of mean squares that need further study. The question of further pooling is discussed, and a graphical technique for portraying variance components is presented.

An illustrative example shows the utility of transforming the dependent variable. It illustrates nested classification and provides an example of the aggregation technique. The results of the experiment can be summarized very concisely: differences between individuals predominated; day-to-day variation and weight interactions were of secondary importance; for fixed weight, date, and person, response times were virtually constant, regardless of rate; one individual on one day behaved atypically.

*Appendix on Computation Formulas*1. *Sums of squares*

Computational formulas for all sums of squares can be written as shown below. Define:

$$n_R = 4, \quad n_W = 7, \quad n_D = 2, \quad n_P = 2, \quad n_S = 2, \quad n_I = 2;$$

$$N = n_R n_W n_D n_P n_S n_I = 448$$

(Here P represents persons within each SI combination. It does *not* represent PSI .) Let x represent the N individual values of the dependent variable, i.e., $x = x_{RWDPSI}$.

Define bracketed terms as in the following examples:

$$[SI] = \frac{n_S n_I}{N} \sum_{S,I} \left(\sum_{R,W,D,P} x \right)^2,$$

$$[RWD] = \frac{n_R n_W n_D}{N} \sum_{R,W,D} \left(\sum_{P,S,I} x \right)^2,$$

$$[DPSI] = \frac{n_D n_P n_S n_I}{N} \sum_{D,P,S,I} \left(\sum_{R,W} x \right)^2,$$

$$[RWDPSI] = \sum_{R,W,D,P,S,I} x^2$$

There is one bracketed term corresponding to each line in the analysis. There is also a bracketed term for the grand mean:

$$] = \frac{1}{N} \left(\sum_{R,W,D,P,S,I} x \right)^2.$$

With these definitions, the computational formulas for sums of squares (SS) for crossed classifications may be written as in the following examples.

$$SS_{RWD} = [RWD] - [RW] - [RD] - [WD] + [R] + [W] + [D] - [.]$$

$$SS_{SI} = [SI] - [S] - [I] + [.]$$

For the lines including persons (nested in SI), the sums of squares may be written as in the following examples.

$$SS_{P(SI)} = [PSI] - [SI],$$

$$SS_{DP(SI)} = [DPSI] - [DSI] - [PSI] + [SI],$$

$$SS_{RWP(SI)} = [RWPSI] - [RWSI] - [RPSI] + [RSI] - [WPSI] + [WSI] + [PSI] - [SI],$$

$$\begin{aligned}
 SS_{RWDP(SI)} = & [RWDP SI] - [RWDSI] - [RWPSI] + [RWSI] \\
 & - [WDPSI] + [WDSI] - [RDPSI] + [RDSI] \\
 & + [RPSI] - [RSI] + [WPSI] - [WSI] + [DP SI] \\
 & - [DSI] - [PSI] + [SI].
 \end{aligned}$$

2. Degrees of freedom for linear combinations of mean squares

The method used is adapted from Fairfield Smith [20], Welch [24], Satterthwaite [18], and Cochran [5]. If the error term for a particular mean square, DEN_A , is a sum (or difference) of other mean squares, e.g.,

$$DEN_A = MS_B \pm MS_C \pm MS_D + \dots$$

then the degrees of freedom for this error term, DDF_A , is obtained from the formula

$$\frac{(DEN_A)^2}{DDF_A} = \frac{(MS_B)^2}{DF_B} + \frac{(MS_C)^2}{DF_C} + \frac{(MS_D)^2}{DF_D} + \dots$$

Example: Consider line P in Table 3.

$$DEN_P = MS_{RP} + MS_{DP} - MS_{RWDP}$$

$$450 = 331 + 139 - 20.$$

$$\frac{(450)^2}{DDF_P} = \frac{(331)^2}{12} + \frac{(139)^2}{8} + \frac{(20)^2}{312}.$$

$$DDF_P = 17.5.$$

An alternative method, which has been proposed by Cochran [5] and which he believes to be safer, lacks flexibility, not being as appropriate for making interval estimates, or for comparing individual means.

3. Estimating variance components after pooling

Rather than recompute the entire table to find new pooled s^2 's, a formula that approximates the results of a reanalysis after pooling is

$$\begin{aligned}
 \text{pooled } s_a^2 &= s_a^2 + \sum_{\beta} \frac{DIV_{\beta}}{DIV_a} \frac{DF_{\beta}}{\text{Pooled } DF} s_{\beta}^2 \\
 &= s_a^2 + \frac{1}{DIV_a (\text{pooled } DF)} \sum_{\beta} DF_{\beta} (MS_{\beta} - DEN_{\beta})
 \end{aligned}$$

where the sum is over all β to be pooled with line A .

REFERENCES

- [1] Anderson, R. C. and Bancroft, T. A. *Statistical theory in research*. New York: McGraw-Hill, 1952.
- [2] Bozivich, H., Bancroft, T. A., and Hartley, H. O. Power of analysis of variance test procedures for certain incompletely specified models, I. *Ann. math. Statist.*, 1956, 27, 1017-1043.

- [3] Bennett, C. A. and Franklin, N. L. *Statistical analysis in chemistry and the chemical industry*. New York: Wiley, 1954.
- [4] Bross, I. Fiducial intervals for variance components. *Biometrics*, 1950, **6**, 136-144.
- [5] Cochran, W. G. Testing a linear relation among variances. *Biometrics*, 1951, **7**, 17-32.
- [6] Cornfield, J. and Tukey, J. W. Average values of mean squares in factorials. *Ann. math. Statist.*, 1956, **27**, 907-949.
- [7] Crump, S. L. The present status of variance component analysis. *Biometrics*, 1951, **7**, 1-16.
- [8] Eisenhart, C. The assumptions underlying the analysis of variance. *Biometrics*, 1947, **3**, 1-21.
- [9] Grant, D. A. Analysis-of-variance tests in the analysis and comparison of curves. *Psychol. Bull.*, 1956, **53**, 141-154.
- [10] Grindley, G. C. The variation of sensory thresholds with the rate of application of the stimulus. I. The differential threshold for pressure. *Brit. J. Psychol.*, 1936-37, **27**, 297-302.
- [11] Johnson, P. O. *Statistical methods in research*. New York: Prentice Hall, 1949.
- [12] Johnson, P. O. and Tsao, F. Factorial design in the determination of differential limen values. *Psychometrika*, 1944, **9**, 107-144.
- [13] Kempthorne, O. *The design and analysis of experiments*. New York: Wiley, 1952.
- [14] Kogan, L. S. Variance designs in psychological research. *Psychol. Bull.*, 1953, **50**, 1-40.
- [15] Mood, A. M. *Introduction to the theory of statistics*. New York: McGraw-Hill, 1950.
- [16] Pal, G. Limens for continuous change. *Indian J. Psychol.*, 1944, **19**, 155-160.
- [17] Paull, A. E. On a preliminary test for pooling mean squares in the analysis of variance. *Ann. math. Statist.*, 1950, **21**, 539-556.
- [18] Satterthwaite, F. E. An approximate distribution of estimates of variance components. *Biometrics*, 1946, **2**, 110-114.
- [19] Scheffé, H. Alternative models for the analysis of variance. *Ann. math. Statist.*, 1956, **27**, 251-271.
- [20] Smith, H. F. The problem of comparing the results of two experiments with unusual errors. *Jour. G.S.I.R.* (Australia), 1936, **9**, 211-212.
- [21] Tukey, J. W. Components in regression. *Biometrics*, 1951, **7**, 33-69.
- [22] Tukey, J. W. Interaction in a row-by-column design. Memo. Rep. No. 18, Statist. Res. Group, Princeton Univ., 1949.
- [23] Tukey, J. W. Variances of variance components: I. Balanced designs. *Ann. math. Statist.*, 1956, **27**, 722-736.
- [24] Welch, B. L. The specification of rules for rejecting too variable a product, with particular reference to an electric lamp problem. *J. roy. statist. Soc. Suppl.*, 1936, **3**, 29-48.
- [25] Wilk, M. B. The logical derivation of linear models for experimental situations and their use in selecting the appropriate error term in the analysis of variance, I, AV-3 (1953); II, On the analysis of a general randomized block type design, AV-4 (1954); III, (with O. Kempthorne) The expectations of mean squares for a general three-way classification, AV-7 (1954); IV, (with O. Kempthorne) Fixed, mixed, and random models, AV-8 (1954); V, Expectations of mean squares for a general three-way classification (Part 2), AV-9 (1954); VI, Interactions with experimental units, AV-10 (1954). Analysis of variance project reports, Statistical Laboratory, Iowa State College.
- [26] Wilk, M. B. and Kempthorne, O. Fixed, mixed, and random models. *J. Amer. statist. Assoc.*, 1955, **50**, 1144-1167.

Manuscript received 4/17/58

Revised manuscript received 11/6/59