Analysis of Variance in Clinical Aphasiology

Robert H. Brookshire

Variance, a statistical term for variability, is central to the analysis of variance method and represents the average dispersion around the mean of scores in a group. Table 1 shows how variance is calculated. To calculate variance, subtract the mean for the group from each individual's score, square that result, add the squared scores together, and divide that sum by the number of scores that were squared in the first place, or more commonly, by $N - 1$, the degrees of freedom.

The data come from measuring the height of 10 hypothetical adults randomly selected from a hypothetical crowd inside the Little Big Dollar Store in Ottumwa, Iowa, and another 10 randomly selected adults from a group gathered in front of the Holiday Inn in Kayenta, Utah. You can see that not everyone is the same height because height ranges from 53 to 81 inches, and the average heights for the groups are 66.2 and 68.6 inches, respectively. The third column shows how each score's deviation from the mean is calculated and the fourth column shows these deviations squared and added together. The variance is derived by dividing the total sum of squares by the number of scores in the total. Note also that standard deviation is the square root of variance.

So why aren't our 10 people each 67.4 inches tall, the overall mean for the group? Which is another way of saying, "What are the sources of variability in height for this group?" In any randomly selected group of people, there are a multitude of sources of variability in height, age, sex, genetic profile, and so forth. All these unmeasured variables contribute to the dispersion of scores around the mean. The experimental variables manipulated in the experiment are another source of dispersion. In the height example, differences in location of subject recruitment is such a source.

The unmeasured variables are sometimes called *nuisance variables* because they add noise and confusion to the data, making it hard to see the effects

TABLE 1. HOW VARIANCE IS CALCULATED

GROUP 1

Subj.	*Height*	*Hgt.*-M	*(Hgt.*-M*)*^{2}
1	70	3.8	14.44
2	62	-4.2	17.64
3	69	2.8	7.84
4	54	-12.2	148.84
5	72	5.8	33.64
6	66	-0.2	0.04
7	78	11.8	139.24
8	64	-2.2	4.84
9	59	-7.2	51.84
10	68	1.8	3.24
Total	662	0	421.6

Mean = 66.2
Variance (S^2) = 42.16
Standard Deviation (S) = 6.49

GROUP 2

Subj.	*Height*	*Hgt.*-M	*(Hgt.*-M*)*^{2}
1	76	7.4	54.76
2	62	-6.6	43.56
3	73	4.4	19.36
4	53	-15.6	243.36
5	78	9.4	88.36
6	67	-1.6	2.56
7	81	12.4	153.76
8	64	-4.6	21.16
9	58	-10.6	112.36
10	74	5.4	29.16
Total	686	0	768.4

Mean = 68.6
Variance (S^2) = 76.84
Standard Deviation (S) = 8.74

of experimental manipulations. The noise and confusion caused by nuisance variables is called *error* or *error variance* in the analysis of variance. It represents the overall effects of all the uncontrolled nuisance variables that affect the scores.

Error variability comes from two primary sources. The first source is the effects of uncontrolled (and in many cases unmeasured) differences among subjects. In the case of aphasic subjects, differences in aphasia severity, time postonset, general health, mood, blood-sugar level, and any of a multitude of other differences can cause variability in performance across subjects.

The second source of error variability is variability in experimental procedures from subject to subject or from test occasion to test occasion. Differences in scoring criteria, unreliability in scoring, differences in timing of stimulus delivery, and variability in instructions or feedback are a few of the large number of potential procedural sources of error variability.

The most common, but messiest, way of dealing with variables that contribute to error variability is random assignment of subjects to groups or levels of treatment so the effects of nuisance variables will be distributed equally across groups or levels. However, the error variability is not removed, nor, in most cases, is it reduced by random assignment. It is still there with the potential to affect the interpretation of results. A better way of dealing with nuisance variables is to control them—to hold them at a constant level, so they do not affect the measures. This is why we attempt to match subjects on characteristics that may affect their performance in experimental tasks, and why we attempt to keep experimental procedures exactly the same from subject to subject and from test occasion to test occasion.

Another way of dealing with nuisance variables is to make them into experimental (independent) variables. Consider what happens when a nuisance variable is changed into an independent variable. Suppose that half of the 10 people in each group are men and half are women, and the shortest 10 are women and the tallest 10 are men. Now the two groups of 10 can be partitioned into four groups of 5 each. Table 2 shows the data for Group 1 partitioned by sex.

Now there are two means—one for men and one for women. And the variance of scores is reduced from 42 inches for the mixed-sex group to 13 inches and 18 inches, respectively for men and women. Partitioning the group decreased the dispersion in scores, made the mean a better predictor of individual scores, and increased the likelihood of getting a significant effect from any treatments introduced. The movement of variability from unexplained to explained variability is summarized in Table 3 for analyses of variance for the unpartitioned group and the groups partitioned by sex. The variability attributable to error in the unpartitioned group (signified by the number 1,190) is broken into three parts in the partitioned group, with about three-fourths of the original error now attributable to the new main effect and the new interaction.

It is important to remember that analysis of variance results are not data. They are opaque in terms of portraying the magnitude of the effects of independent variables. Consider the height by groups by sex analysis of variance in Table 2. Neither the sum of squares nor the mean square for groups gives any indication that group 1 is 2.4 inches shorter, on the average, than group 2, and neither the sum of squares nor the mean square for

TABLE 2. DECREASING ERROR VARIANCE BY PARTITIONING A LARGE GROUP INTO TWO HOMOGENEOUS GROUPS

GROUP 1: MALE

Subj.	*Height*	*Hgt.*-M	(*Hgt.*-M)^2
1	70	-1.4	1.96
2	69	-2.4	5.76
3	72	0.6	0.36
4	78	6.6	43.56
5	68	-3.4	11.56
Total	357	0	63.20

Mean = 71.4
Variance (S^2) = 12.64
Standard Deviation (S) = 3.55

GROUP 2: FEMALE

Subj.	*Height*	*Hgt.*-M	(*Hgt.*-M)^2
1	62	1.0	1.0
2	54	-7.0	49.0
3	66	5.0	25.0
4	64	3.0	9.0
5	59	-2.0	4.0
Total	305	0	88.0

Mean = 61.0
Variance (S^2) = 17.60
Standard Deviation (S) = 4.19

the interaction gives any indication that males in group 1 are 10.4 inches taller than females. Analysis of variance tables or graphs (or analysis of variance results described in the text of a paper) only supplement means, standard deviations, ranges, or other measures of effect size—it never substitutes for them.

Furthermore, the probability value assigned to an *F*-ratio (or any test statistic) has no dependable relationship to the size of the experimental effect it has been used to test. A probability value of $p < .05$ for one comparison and a probability value of $p < .001$ for another does not mean that the difference between the means in the second comparison is larger than the difference in the first. You can get a highly significant *F*-ratio for several reasons: when the difference between means is large, when the variability of scores is small, and when the number of observations is large.

The second major source of variability across observations in an experiment is the effects of the independent variables. To show how the vari-

TABLE 3. ANALYSIS OF VARIANCE SUMMARY TABLES FOR UNDIVIDED SAMPLE (HEIGHT BY GROUPS) AND SAMPLE DIVIDED BY SEX (HEIGHT BY GROUPS BY SEX)

HEIGHT BY GROUPS

Source of Variance	*SS*	*df*	*MS*	*F*	*p*
Groups	28.80	1	2,880.00	0.43	.52
Error	1,190.00	18	66.11		
Total	1,218.80				

HEIGHT BY GROUPS BY SEX

Source of Variance	*SS*	*df*	*MS*	*F*	*p*
Groups	28.80	1	28.80	1.48	.24
Sex	845.00	1	845.00	43.44	.01
Group by Sex	33.80	1	33.80	1.74	.20
Error	311.20	16	19.45		
Total	1,218.80				

ability attributable to the effects of independent variables is analyzed, I designed a hypothetical experiment and analyzed the results with analysis of variance. The experiment is a study of the effects of three aphasia treatment approaches—individual treatment by speech-language pathologists, group treatment by speech-language pathologists, and television-based treatment in which patients watched reruns of *Wheel of Fortune* for two hours per day, five days per week. Subjects were aphasic patients meeting all the standard selection criteria. Twenty patients received each treatment, and each treatment group was composed of 10 fluent and 10 nonfluent patients.

Figure 1 shows the effects of treatment. (Note that I'm not providing an analysis of variance table.) The figure shows individual treatment was better than group treatment, which was better than television treatment.

Figure 2 shows the means for groups according to aphasia type. It shows that fluent patients did better than nonfluent patients. These two graphs summarize the results of the main effects in the analysis of variance. The main effects are the effects of the independent variables in the experiment, and there are always as many main effects in the analysis of variance as there are independent variables in the experiment. But in any experiment with more than one independent variable, the main effects don't always tell the whole story. Sometimes the effects of one independent variable depend on the level of another independent variable. In our example the effects of treatment might depend on the group to which the aphasic patients belonged. Such an interdependency is called an interaction.

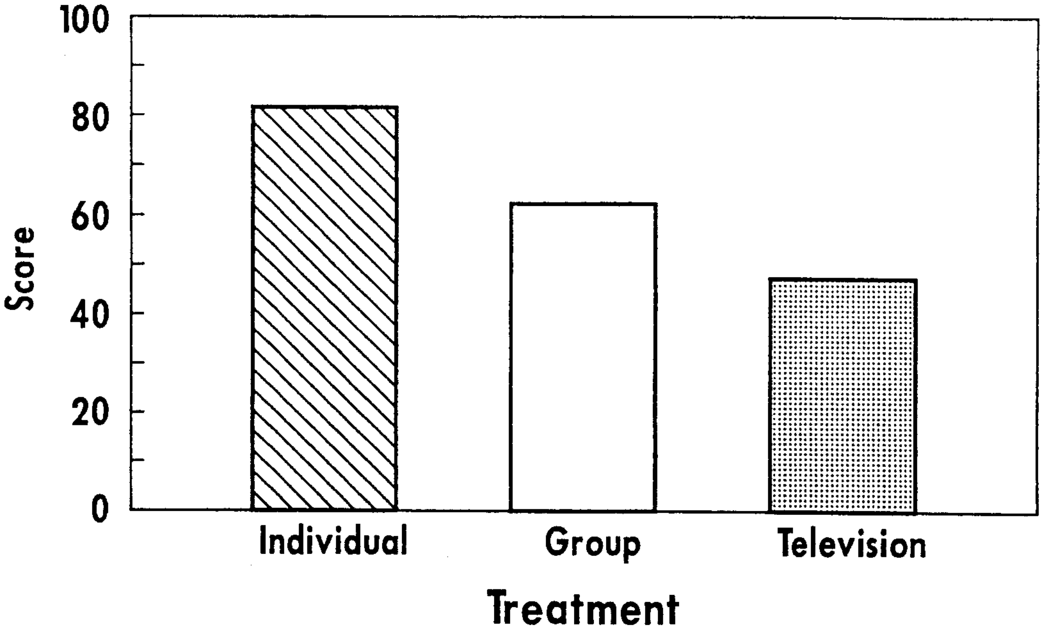

Figure 1. Effects of treatments for hypothetical treatment study.

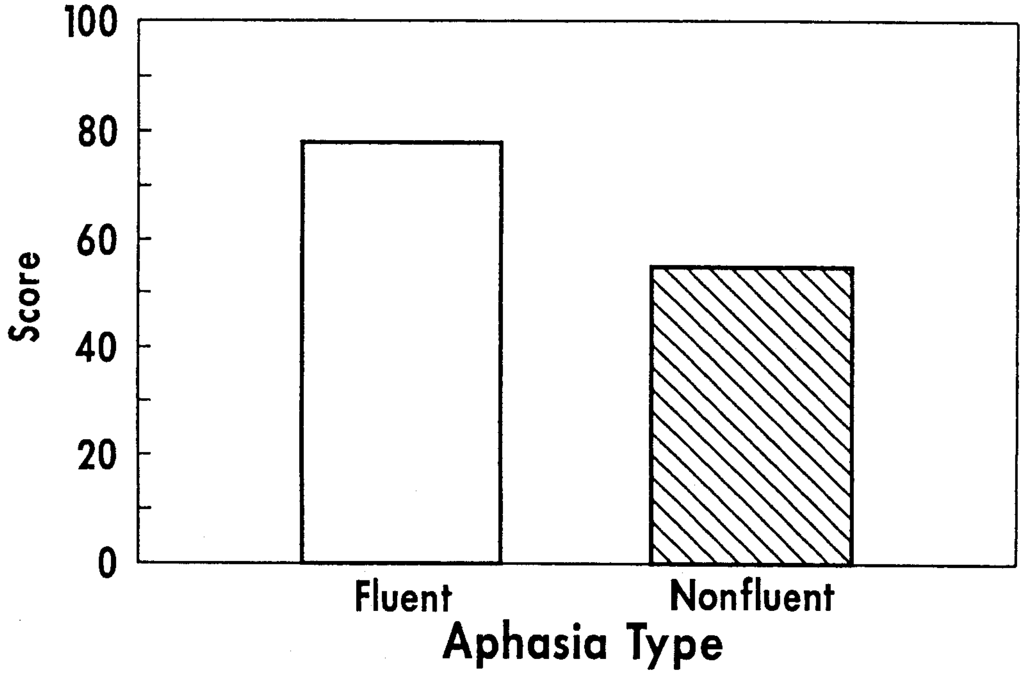

Figure 2. Effects of aphasia type for hypothetical treatment study.

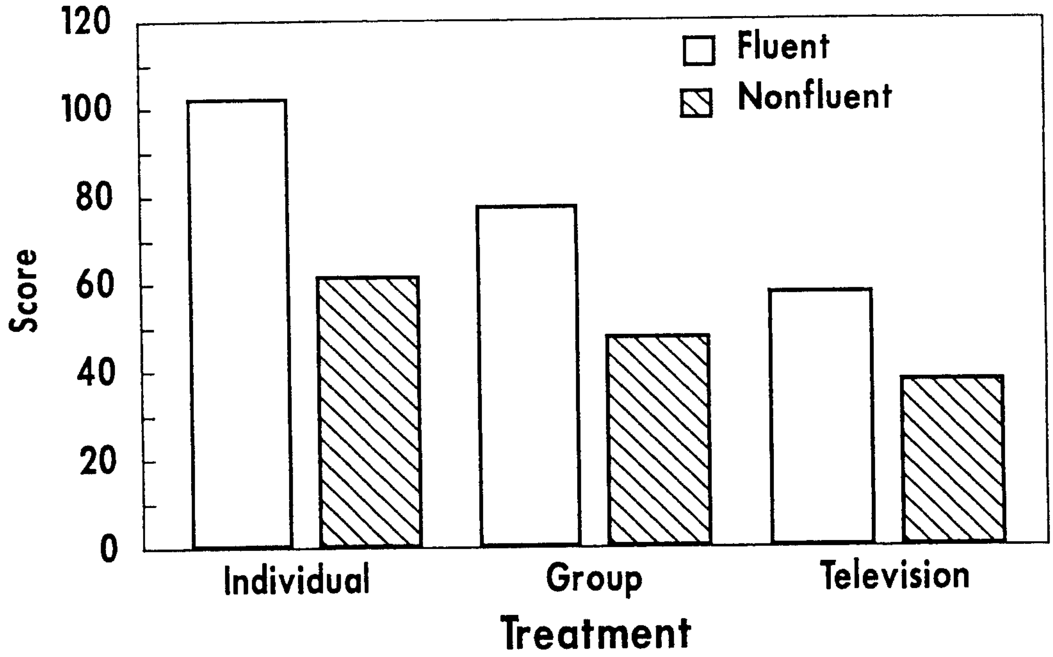

Figure 3. Bar graph of aphasia type by treatment effects in hypothetical treatment study. There is no interaction between the two factors.

Figure 3 helps to see if there was an interaction between the two independent variables in this study. It shows that nonfluent aphasic patients always did worse than fluent patients, regardless of which treatment they received. Said another way, individual treatment was always better than group treatment, which was always better than television treatment regardless of whether subjects were fluent or nonfluent. This is what we mean when we say that there was no interaction between subject groups and treatment types. A convenient way to visualize interactions is to graph them with line graphs.

Figure 4 shows a line graph for the interaction we are presently examining. If there is no interaction, the lines in the graph should be close to parallel. Marked deviations from parallel lines suggest the presence of an interaction. The lines in Figure 4 are essentially parallel, confirming that treatments and groups do not interact.

Table 4 shows how the results of the analysis look in an analysis of variance table. As illustrated, the effects of the different treatments were significant as were the effects of groups, and there was no interaction between the two. Because there was no interaction, we can draw conclusions based simply on analysis of the main effects of the independent variables. This often involves what are called follow-up tests, which will be addressed after we consider a different set of data for this experiment.

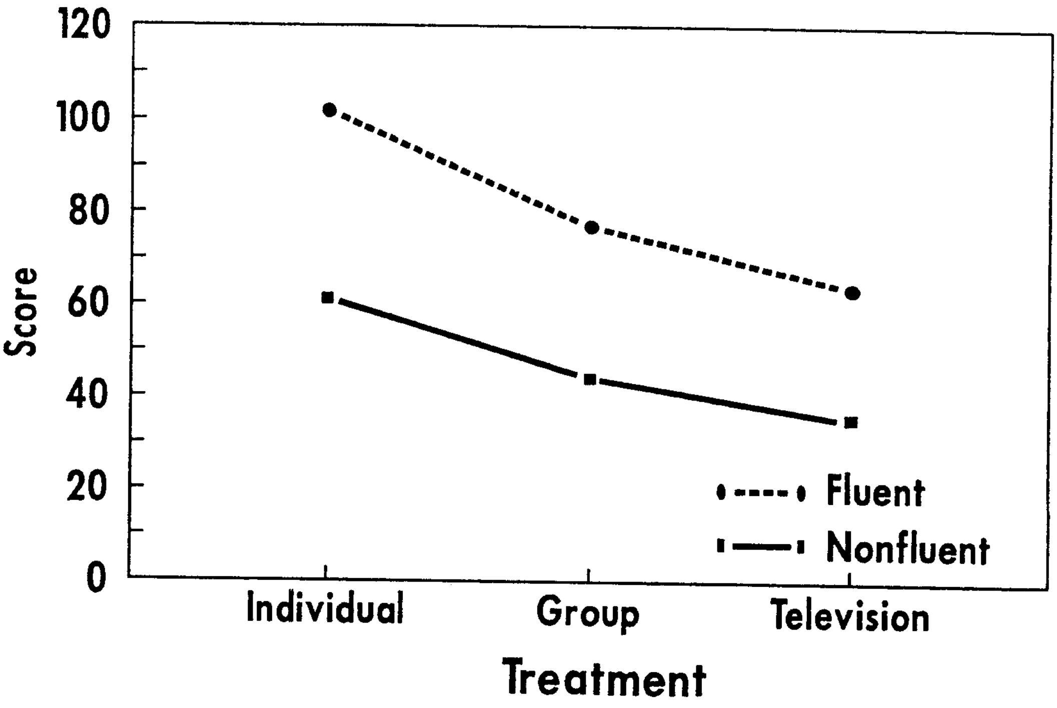

Figure 4. Line graph of group by treatment effects shown in Figure 3. The lines are essentially parallel, showing that no interaction is present.

TABLE 4. ANALYSIS OF VARIANCE SUMMARY TABLE FOR DATA SET 1

Source of Variance	*SS*	*df*	*MS*	*F*	*p*
Treatment	11,289.44	2	5,914.72	325.65	.01
Aphasia Type	13,741.06	1	13,741.06	756.54	.01
Treatment by Type	51.58	2	25.79	1.42	.25
Error	980.80	54	18.16		

Figure 5 shows the effects of treatments for a new set of data. The effects of treatment are similar to those for the first data set—individual treatment is better than group treatment, which is better than television treatment.

Figure 6 shows the effects of aphasia type. The effects of aphasia type are also similar to those for the previous data set—nonfluent patients do worse than fluent patients. At this point the effects of treatments and aphasia types appear identical to those for the first data set. However, when the interaction between treatments and groups is examined, we find that something has changed (see Figure 7). For this data set, the effects of aphasia type depend on which treatment was received. Fluent

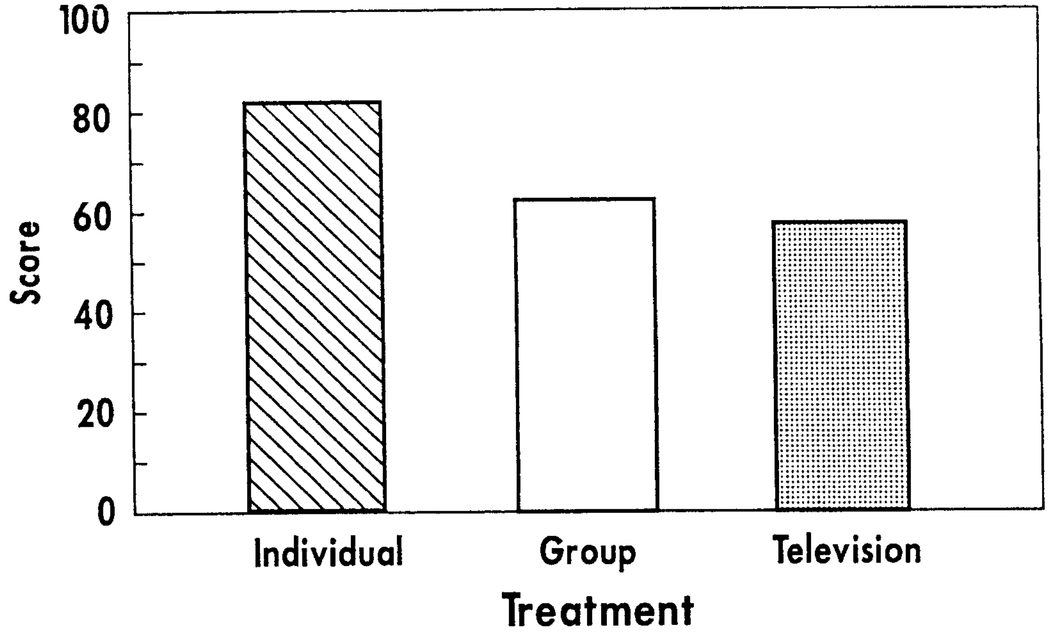

Figure 5. Effects of treatments: second hypothetical treatment study.

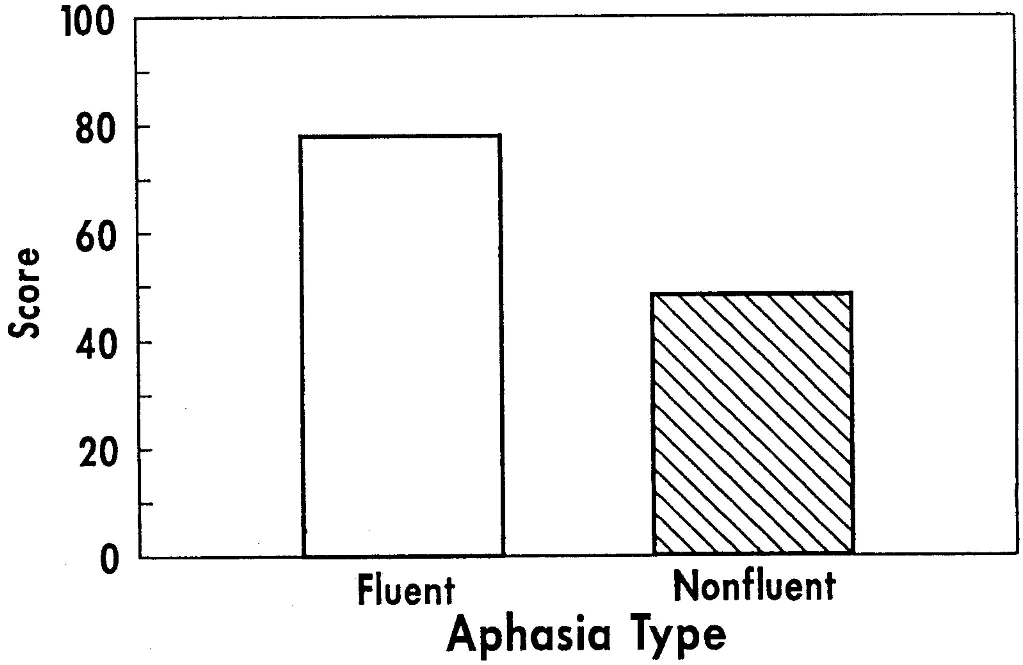

Figure 6. Effects of aphasia type: second hypothetical treatment study.

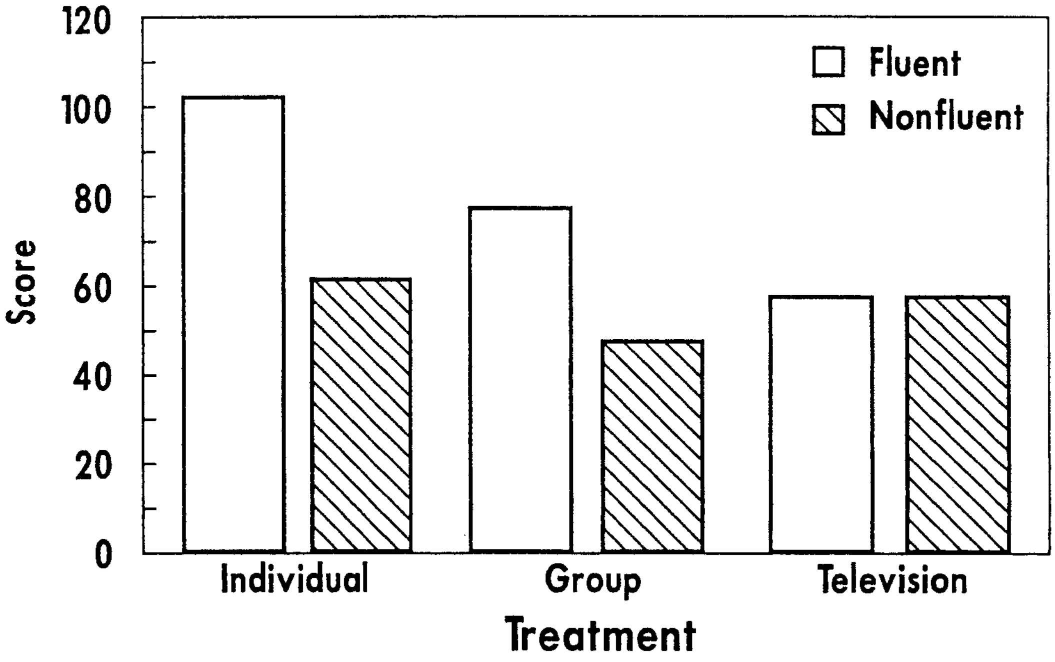

Figure 7. Interaction of treatment and aphasia type: Data set 2. The effects of treatment depend on aphasia type.

subjects did better than nonfluent subjects when treatment was either individual or group, but when treatment was television there was no difference between them. Now we cannot make a simple, generic statement about the effects of aphasia type on patients' performance, because the effect of aphasia type depends on which treatment the patients received. Figure 8 shows the same interaction in a line graph. The lines are no longer parallel—they converge in the third treatment condition, confirming the presence of the interaction.

The point of all of this is that when an analysis of variance yields a significant interaction between or among main effects, conclusions cannot be drawn from the main effects alone and discussion of such main effects must *always* be qualified by taking note of the interactions which affect them. Interactions are explained by follow-up tests, just as main effects involving two or more levels of an independent variable are.

Consider follow-up tests using the present study as an example. To interpret significant main effects and interactions, calculate the significance of differences between means, called pair-wise comparisons. Consider the significant effect of aphasia type first. Because there are only two levels of aphasia type there is only one difference between means to consider—the difference between fluent and nonfluent subjects. Examination of the means or a look at the graph for this effect tells us that

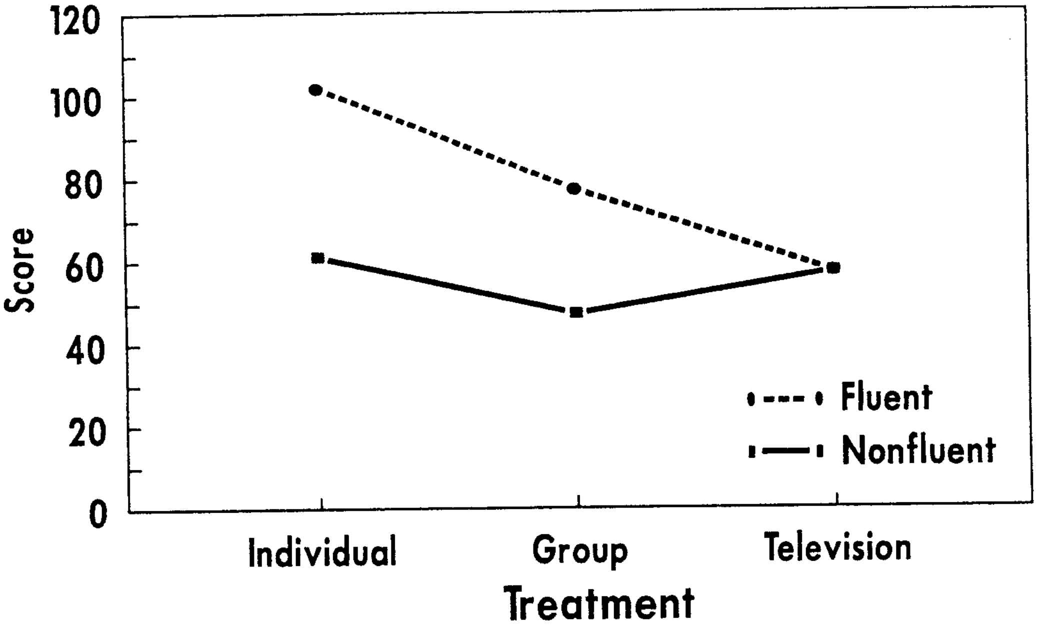

Figure 8. A line graph of the interaction shown in Figure 7. The lines converge, showing that an interaction is present.

fluent aphasic subjects did better than nonfluent subjects, and the significant main effect in the analysis of variance tells us that they did significantly better.

Now consider the main effect for treatments. Three kinds of treatment were involved in the experiment: individual, group, and television. The significant main effect for treatments tells us only that at least one of the three differed significantly from the others; it doesn't tell us which ones. Which means that of the possible differences between pairs of means (1 vs. 2, 1 vs. 3, and 2 vs. 3) one, two, or perhaps all three represent significant differences. Now consider the interaction. In the experiment there were two groups and three treatment conditions, so there are nine potential pairwise comparisons among the six means.

If one made all nine comparisons by running nine *t* tests, the outcome of the tests would be compromised by a phenomenon called inflated Type 1 error. Type 1 error (sometimes called *alpha* error and symbolized by α) refers to the probability of concluding that two means differ significantly when, in fact, they do not. An α level of .05 means that no more than 5 times out of 100 will one falsely conclude that two means differ. The probability of Type 1 error increases with multiple tests because the tabled α levels for tests such as the *t* test are set for a single comparison. Running multiple tests of significance increases the likelihood of falsely rejecting the null hypothesis in the same way as the likelihood of getting

TABLE 5. EFFECTS OF MULTIPLE COMPARISONS ON TYPE 1 ERROR

# Comparisons	$\alpha = .05$	$\alpha = .01$	$\alpha = .001$
1	.050	.010	
2	.098	.020	
3	.143	.030	
4	.186	.039	
5	.227	.049	
6	.265	.059	
7	.302	.068	
8	.337	.077	
9	.370	.086	
10	.401	.096	.010
20		.149	.020
30			.030
40			.039
50			.049

at least one red marble from a box containing 95 white marbles and 5 red ones increases with the number of times one draws a marble from the box. Table 5 shows what happens when one makes multiple, independent comparisons at a given Type 1 error level, and how the actual Type 1 error escalates as the number of comparisons increases.

One way to deal with this problem is to set α for each test at a conservative level in order to bring the overall Type 1 error down. There are a dozen or more procedures for making multiple comparisons while controlling the overall Type 1 error. Table 6 shows six popular ones. Multiple-comparison procedures such as these differ in their conservativeness; some require larger differences between means before calling the differences significant.

The conservativeness of multiple-comparison procedures has important effects on their *power*. The power of a test can be defined loosely as its ability to correctly reject the null hypothesis when the null hypothesis is, in fact, false. The general rule is the more conservative a procedure is, the less powerful it is. Multiple-comparison procedures designed for making comparisons that were planned before the data were acquired usually are more powerful than those for evaluating comparisons suggested by examination of the data. Procedures that restrict comparisons to a subset of means usually are more powerful than those permitting all possible comparisons.

Multiple-comparison procedures differ in terms of their power, relative to a set of individual *t* tests, run without controlling overall Type 1 error. Some procedures, e.g. the *t* test, Dunn's test, Tukey's test (HSD),

TABLE 6. RELATIVE POWER OF SELECTED MULTIPLE-COMPARISON PROCEDURES

	Number of Comparisons			
Procedure	*2*	*3*	*4*	*5*
t-test	1.00	1.00	1.00	1.00
Dunn test	.87	.87	.87	.87
Newman-Keuls test	1.00	.83	.75	.71
Tukey HSD	.71	.71	.71	.71
Tukey WSD	.83	.76	.73	.71
Scheffé test	.62	.62	.62	.62

and Scheffé's test have the same power regardless of the number of comparisons made (Howell, 1987). The Newman-Keuls procedure and Tukey's HSD procedure are most powerful when a small number of comparisons is made. Tukey's HSD and Scheffé's test are the least powerful on the average; however, they maintain their power even when a large number of comparisons is made. Dunn's test may appear more powerful than the others, but that's because it's designed for evaluating preplanned comparisons.

So, which multiple-comparison procedure should one use? The following guidelines, based on Howell's (1987) recommendations, are reasonable.

- If tests are planned in advance and you want to run only one or possibly two comparisons, use a standard *t* test.
- If you wish to run several (three to seven) preplanned comparisons, use Dunn's test.
- If you wish to run two or three post-hoc comparisons, use Newman-Keuls.
- If you wish to run three to seven post-hoc comparisons, use Tukey's WSD.
- If you wish to run more than seven post-hoc comparisons, use Scheffé's test.

Besides choosing an appropriate multiple-comparison procedure, you can do other things to maintain Type 1 error at reasonable levels while maximizing the power of the tests. One way is to plan comparisons in advance. The experimental questions addressed usually make some comparisons important, while leaving others merely interesting or even trivial. Deciding in advance which comparisons directly address the experimental questions posed, and limiting statistical tests to those comparisons, permits you to use multiple-comparison procedures for preplanned comparisons that are more powerful than those for post-hoc comparisons.

The second way to control Type 1 error is to limit the number of comparisons made. Evaluating a subset of all possible comparisons permits you to use more powerful procedures such as Newman-Keuls or Tukey's WSD, rather than less powerful but more permissive procedures such as Scheffé's test.

The third way to limit Type 1 error is to make comparisons orthogonal. Making comparisons orthogonal means that whether or not one of the comparisons is significant has no effect on whether any other comparison in the set is significant. A major problem with orthogonal contrasts is they severely limit both the number and kind of comparisons among means that are permitted. Orthogonal contrasts are much less widely used than they were 10 or 20 years ago, partly because of their unforgiving nature and partly because efficient and relatively powerful procedures for evaluating multiple, nonorthogonal comparisons are now available.

The last aspect of analysis of variance to be considered concerns normality of distribution and homogeneity of variance. As is well known, analysis of variance is based on two assumptions: observations are drawn from normally distributed populations, and variances of the groups are equivalent. Moderate deviations from normality and moderate disparity in variance across groups usually do not greatly compromise the analysis of variance if sample sizes are equal. However, when the distribution of scores is grossly abnormal and sample variances are grossly disproportionate, tests of significance may be seriously in error.

This is an important issue in disciplines such as ours, where the performance of an impaired sample often is compared with that of an unimpaired sample. It is not unusual that the normal group performs with few or no errors, while the impaired group has high error rates. Consequently, the scores of the impaired subjects have dispersion and often approximate a bell-shaped distribution. On the other hand, the scores of the non-brain-damaged subjects often have no distribution because they all "topped out" on the test. In such a situation, analyzing the performance of non-brain-damaged subjects with an analysis of variance would be inappropriate because there is no variance to analyze. Even in less extreme cases, when scores for a normal group of subjects are tightly clustered and the distribution of scores is sharply peaked, it is not a good idea to run an analysis of variance including both the normal and impaired groups. Instead, a separate analysis of variance for each group may be appropriate if each group has reasonable dispersion of scores and if the distributions are not too wildly peaked or skewed.

ACKNOWLEDGMENTS

This research was supported by the Research Service, Minneapolis Veterans Affairs Medical Center, and by VA General Medical Research.

REFERENCES

Howell, D. C. (1987). *Statistical methods for psychology* (2nd ed.). Boston: PWS.