KONGL CAROLINSKA MEDICO CHIRURGISKA INSTITUTET

TECHNICAL AUDIOLOGY

Report TA No. 92
November 1979

STATISTICAL TREATMENT OF DATA FROM LISTENING TESTS ON SOUND-REPRODUCING SYSTEMS

Alf Gabrielsson

Report TA No. 92
November 1979

STATISTICAL TREATMENT OF DATA FROM LISTENING TESTS ON SOUND-REPRODUCING SYSTEMS

Alf Gabrielsson

From the Department of Technical Audiology
Karolinska Institutet
KTH
S-100 44 Stockholm, Sweden

Tel: 46-8-11 66 60

Technical Audiology

Karolinska Institutet November 1979 TA 92

STATISTICAL TREATMENT OF DATA FROM LISTENING TESTS ON SOUND-REPRODUCING SYSTEMS

Alf Gabrielsson

ABSTRACT

This paper describes various procedures for the statistical treatment of data from listening tests on loudspeakers, which are performed in accordance with the recommendations given in the IEC Publication 268-13: Listening tests on loudspeakers. The data are supposed to be ratings in one or more selected scales pertaining to perceived sound quality. The following points are discussed:

a) construction of suitable data matrices,
b) use of analysis of variance and related tests,
c) tests for specific comparisons,
d) listening tests including extra variables,
e) estimating reliability of data,
f) critical issues in significance testing,
g) computations, and
h) presentation of results.

Some general comments and selected references conclude the paper. All tables, figures, and formulas appearing in the paper are also given separately in an enclosed appendix.

Table of Contents

ABSTRACT

1 INTRODUCTION....1

2 DATA MATRICES, DESCRIPTIVE STATISTICS....4
2.1 Individual data matrix....4
2.2 Group data matrix....5

3 ANALYSIS OF VARIANCE....8
3.1 ANOVA for individual data matrix....8
3.2 ANOVA for group data matrix....11
3.3 ANOVA when each subject makes only one rating per case....15

4 TESTS FOR SPECIFIC COMPARISONS....18
4.1 Planned independent comparisons....18
4.2 Planned non-independent comparisons....20
4.3 Non-planned comparisons....21
4.4 Specific comparisons within single programs....23

5 LISTENING TESTS INCLUDING EXTRA VARIABLES....25
5.1 One extra variable....25
5.1.1 Individual data....25
5.1.2 Group data....27
5.1.3 Tests for specific comparisons....29
5.1.4 One extra variable only for certain programs....30
5.2 Two extra variables....31
5.3 Some alternative designs ("split-plot" designs)....31
5.3.1 Example of "split-plot" design....32
5.3.2 Further applications of the "split-plot" design....34
5.3.3 "Split-plot" design for many loudspeakers....36
5.3.4 "Split-plot" design versus "repeated measurements" design.37

6 RELIABILITY....38
6.1 Intra-individual reliability....38
6.1.1 Visual inspection....38
6.1.2 Statistical significance....38
6.1.3 Variance "within cells"....39
6.1.4 Reliability of mean ratings....39
6.1.5 Proportion of variance accounted for....41
6.1.6 Conclusions regarding intra-individual reliability....43
6.2 Inter-individual reliability....43
6.2.1 Visual inspection....44
6.2.2 Statistical significance....44
6.2.3 Reliability of mean ratings....44
6.2.4 Proportion of variance accounted for....45
6.2.5 Conclusions regarding inter-individual reliability....46

7 SOME CRITICAL ISSUES IN SIGNIFICANCE TESTING....47
7.1 Error risks, significance level, "power"....47
7.2 Computing "power",number of ratings, number of subjects...48
7.3 Assumptions underlying significance tests....50

8 COMPUTATIONS, COMPUTER PROGRAMS, TABLES....52

9 PRESENTATION OF RESULTS....53
9.1 Only descriptive statistics....53
9.2 Data treatment involving ANOVA....54

10 CONCLUDING COMMENTS, ACKNOWLEDGEMENTS....55

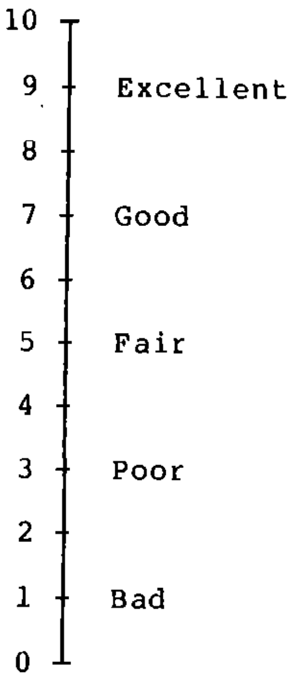
11 REFERENCES....56

1 INTRODUCTION

Perceived sound quality of sound-reproducing systems may be assessed by means of "listening tests". There are many varieties of such tests. In certain tests the listeners only make preference judgments, for instance, listening to pairs of systems and indicating which system they prefer in the respective pair. In more sophisticated tests the listeners rate the selected systems in a number of scales, which are assumed to reflect some important perceptual attributes of the reproductions. There is usually also some scale for the overall evaluation of the perceived sound quality, for instance, in terms of the "Pleasantness" of the reproduction, and/or the "Naturalness/Fidelity" of the reproduction. "Naturalness/Fidelity" refers to how well the original sound is reproduced, in other words how "natural" or "true-to-nature" the reproduction sounds.

In principle a listening test should be designed as an experiment, in which the experimenter varies the systems to be judged and has due control over various extraneous variables. Besides the systems he may also want to vary the programs (pieces of music, speech etc), or the positions of the systems, the positions of the listeners etc. Suitable experimental procedures are discussed in the IEC-Publication 268-13: Listening tests on loudspeakers (to be published), and examples of various procedures may be found in papers by Gabrielsson, Rosenberg & Sjögren (1974), Gabrielsson & Sjögren (1976, 1979), Gabrielsson (1979), and Gabrielsson, Frykholm & Lindström (1979).

This paper presents a discussion of the statistical treatment of data from listening tests. In accordance with the recommendations in the IEC-Publication it will be assumed that one of the rating scales is a "true-to-nature" scale going from 10 (denoting "a reproduction perfectly true-to-nature") down to 0 (denoting "practically no similarity at all with the original performance"), see Figure 1.



The number 10 denotes a reproduction which is perfectly true-to-nature.

The number 0 denotes a reproduction so bad that it has practically no similarity at all with the original performance.

FIGURE 1 "True-to-nature" rating scale.

This scale is assumed to represent an interval scale (equidistant scale steps). It is further supposed that the sound-reproducing systems are different loudspeakers, and that they are used for reproduction of some different programs (for instance, some different pieces of music). However, the statistical procedures outlined below may also be used for other rating scales (as "Pleasantness", "Clearness", "Softness", or what the case may be) constructed in similar ways. And, of course, the sound-reproducing systems are not necessarily loudspeakers but may be headphones, amplifiers, turntables, tape recorders etc, or combinations of such equipments.

Each loudspeaker should be rated for its reproduction of each program. The total number of stimuli will thus be the number of loudspeakers multiplied by the number of programs. The presentation order of all these loudspeaker x program combinations should be randomized, differently for each subject (listener). To increase (and to be able to estimate) the reliability of the ratings it is recommended that each subject makes at least two independent ratings for each combination. Furthermore each subject should make his ratings independently of the other subjects participating in the test.

The data obtained from the test are thus ratings on the 10 - 0 "true-to-nature" scale for all loudspeaker x program combinations. To get as much information as possible from these data it may be preferable to treat them intra-individually (that is, within each single subject) as well as inter-individually (that is, over all subjects within the same group of subjects). The statistical treatment may conveniently be divided into two steps:

1) Descriptive statistics. In this step the rating data are entered into suitable matrices to make them easily surveyable, and certain common statistics (e.g. arithmetic means) are computed. The data may also be displayed in graphical form. Visual inspection of the matrices, the graphs, and the computed statistics usually lead to certain conclusions about the loudspeakers under test.

This type of descriptive statistics should always be applied and generally presents no special difficulties. It is described in chapter 2 and in parts of chapter 9.

2) Inference statistics. In this step the rating data are analysed further by means of analysis of variance (ANOVA) and related procedures to test if differences between the ratings for different loudspeakers (and/or for different programs) are statistically significant or not, and if there is an interaction between loudspeakers and programs. ANOVA may also be used to estimate the reliability of the data, intra-individually and inter-individually.

The application of this type of statistical treatment is optional. It requires more work and more knowledge about statistics. On the other hand it usually enables the user to extract more detailed information from the data and to arrive at more definite conclusions. It is also easily generalized to more complex listening tests. It is described in chapters 3 - 8.

In order to illustrate the points stated above an example is given in the following, using real data from a listening test with four loudspeakers and five programs (Gabrielsson & Sjögren, 1976). Examples of listening tests including more variables (sound levels, positions etc) and alternative designs for listening tests are given in chapter 5. Questions about reliability, assumptions for the statistical tests, computer programs etc are treated in chapters 6 - 8, and guidelines for presentation of the results are given in chapter 9.

2 DATA MATRICES, DESCRIPTIVE STATISTICS

2.1 Individual data matrix

For a certain subject in this listening test the following data were obtained, see Table I. He made three ratings per each loudspeaker x program combination. These three values appear in the upper row of each cell, and their arithmetic mean (M) is given directly below.

	Loudspeaker				
Program	A	B	C	D	Means for programs
1	7 6 7 M=6.7	5 5 5 5.0	6 7 5 6.0	3 3 4 3.3	5.3
2	6 6 7 6.3	3 3 4 3.3	5 5 7 5.7	3 3 4 3.3	4.7
3	7 8 8 7.7	2 2 2 2.0	7 7 7 7.0	3 3 3 3.0	4.9
4	7 8 8 7.7	3 3 3 3.0	8 7 8 7.7	3 3 3 3.0	5.3
5	6 7 6 6.3	5 5 4 4.7	6 6 6 6.0	5 5 5 5.0	5.5
Means for loud-speakers	6.9	3.6	6.5	3.5	

TABLE I. Example of individual data matrix

Visual inspection of this matrix directly reveals much about the results. This subject shows a high stability in his ratings (=high intra-individual reliability) - the three ratings within each cell differ very little or are even the same in many cases. The mean ratings for the loudspeakers (over the five programs) appear in the bottom margin and indicate that loudspeakers A and C are superior to loudspeakers B and D. The mean ratings at each program (over the four loudspeakers) appear in the righthand margin and suggest that programs 2 and 3 are harder to reproduce in a "true-to-nature" way than the other programs. Looking at the means for the loudspeakers within each program it is easily seen that the differences between the loudspeakers vary from program to program and sometimes differ rather much from the corresponding differences between the means in the bottom margin. For instance, the average difference between loudspeakers A and B over all five programs is 3.3 (6.9 - 3.6), while the difference is as small as 1.6 (6.3 - 4.7) at program 5 and as big as 5.7 (7.7 - 2.0) at program 3. As regards loudspeakers B and D they get almost the same rating in the bottom margin (3.6 and 3.5, respectively). For program 1, however, there is a difference of 1.7 units between these loudspeakers (5.0 - 3.3). Still more examples could be added, but these may be sufficient to suggest that there is an interaction between loudspeakers and programs to be further studied.

2.2 Group data matrix

The data matrix for a group of subjects is constituted by a combination of individual data matrices. It may be represented in many different ways. One way is illustrated in Table II (on next page) for a group of four subjects, called subjects S, T, U, and V (real data, the data for subject S are the same as in Table I).

(It should be said at once that four subjects represent a minimum as regards the number of listeners, see further 7.2. Actually ten subjects were used in the real experiment, but only four of them are included in the present example to make it easier to follow the computations.)

For each loudspeaker x program combination there are twelve ratings, three per each subject. M_S denotes the arithmetic mean of the three ratings by subject S, M_T the same thing for subject T, and so on. M_g (g for group) denotes the arithmetic mean for the whole group of subjects. (These designations are written only in the upper-hand left case but are implicit in the other cases).

The means for loudspeakers in the bottom margin represent the mean ratings for the loudspeakers, averaged over programs and subjects. The means for programs in the right-hand margin represent the mean ratings at the different programs, averaged over loudspeakers and subjects. (If wanted, each subject's mean ratings for loudspeakers averaged over programs, and mean ratings at the programs averaged over loudspeakers, could be entered at appropriate places in the bottom margin and in the right-hand margin, respectively. They are omitted here not to overcrowd the matrix but appear, of course, in the individual data matrices.)

LOUDSPEAKER

Program	Subject	A	B	C	D	Means for programs
1	S	7 6 7 M_S=6.7	5 5 5 5.0	6 7 5 6.0	3 3 4 3.3	
	T	5 7 4 M_T=5.3	4 4 3 3.7	5 5 5 5.0	3 3 3 3.0	
	U	6 7 7 M_U=6.7	4 5 4 4.3	5 7 7 6.3	5 5 5 5.0	
	V	8 7 7 M_V=7.3	7 6 5 6.0	6 6 7 6.3	7 4 5 5.3	
		M_g=6.5	4.8	5.9	4.2	5.3
2	S	6 6 7 6.3	3 3 4 3.3	5 5 7 5.7	3 3 4 3.3	
	T	8 7 8 7.7	3 2 3 2.7	4 4 7 5.0	4 3 2 3.0	
	U	6 8 8 7.3	4 3 3 3.3	6 7 8 7.0	3 3 3 3.0	
	V	7 8 7 7.3	4 4 5 4.3	8 7 4 6.3	4 4 4 4.0	
		7.2	3.4	6.0	3.3	5.0
3	S	7 8 8 7.7	2 2 2 2.0	7 7 7 7.0	3 3 3 3.0	
	T	5 4 5 4.7	3 2 2 2.3	5 3 3 3.7	1 1 2 1.3	
	U	7 7 9 7.7	4 4 3 3.7	7 7 9 7.7	3 4 4 3.7	
	V	9 6 6 7.0	5 4 6 5.0	8 4 6 6.0	3 3 3 3.0	
		6.8	3.3	6.1	2.8	4.7
4	S	7 8 8 7.7	3 3 3 3.0	8 7 8 7.7	3 3 3 3.0	
	T	6 8 7 7.0	4 3 3 3.3	5 4 5 4.7	3 1 2 2.0	
	U	6 7 8 7.0	3 4 3 3.3	6 6 7 6.3	1 4 2 2.3	
	V	8 8 7 7.7	4 6 5 5.0	8 9 8 8.3	7 7 4 6.0	
		7.3	3.7	6.8	3.3	5.3
5	S	6 7 6 6.3	5 5 4 4.7	6 6 6 6.0	5 5 5 5.0	
	T	6 5 7 6.0	4 3 4 3.7	5 6 3 4.7	4 3 2 3.0	
	U	6 5 6 5.7	3 3 3 3.0	4 5 5 4.7	5 4 4 4.3	
	V	9 7 7 7.7	4 4 5 4.3	7 7 5 6.3	6 6 5 5.7	
		6.4	3.9	5.4	4.5	5.1
Means for loudspeakers		6.8	3.8	6.0	3.6	

TABLE II. Example of group data matrix for subjects denoted S, T, U, and V.

Visual inspection of this matrix leads to much the same conclusions as the inspection of Table I (which was the individual data matrix for subject S): loudspeakers A and C are superior to B and D, there are suggestions to interaction between loudspeakers and programs (compare the differences between the loudspeakers in the bottom margin with the corresponding differences between the loudspeakers within each program), and so on. However, a careful study of Table II reveals, as could be expected, that the results from this group of four subjects are somewhat different than the results from the single subject S in Table I. Inter-individual differences in the ratings may be observed, for instance, that subject V has a tendency to use higher rating values than the other subjects. There are also suggestions of interactions between loudspeakers and subjects, that is, the loudspeakers are rated differently by different subjects.

If wanted, Table II may be supplemented with a very condensed version of group data matrix like that shown in Table III. In this table all individual values are omitted, and only the arithmetic means for each program x loudspeaker combination are given (the M_g values in Table II) together with the means for the loudspeakers and for the programs in the margins.

		Loudspeaker				
		A	B	C	D	Means for programs
Program	1	6.5	4.8	5.9	4.2	5.3
	2	7.2	3.4	6.0	3.3	5.0
	3	6.8	3.3	6.1	2.8	4.7
	4	7.3	3.7	6.8	3.3	5.3
	5	6.4	3.9	5.4	4.5	5.1
Means for loudspeakers		6.8	3.8	6.0	3.6	

TABLE III. Condensed group data matrix.

Although this matrix is much easier to read, it gives no information about the dispersion of the ratings around the means. Some measure of the dispersion could be added, for instance, the standard deviation or the range of the individual means M_S,M_T etc within each program x loudspeaker combination. For various statistical reasons, however, none of these measures is quite satisfactory. Therefore a matrix like that in Table III should never be given alone; it must be supplemented with a more complete matrix of the type shown in Table II (see further 9.1).

The most effective way of taking the variation of the ratings into account is to apply so-called analysis of variance, which is described in the following chapter..

3 ANALYSIS OF VARIANCE

Visual inspection of data matrices may give sufficient information in many cases - especially if the results are interpreted with caution and are not meant to form the basis for some far-reaching conclusions. In the data used here certain results seem quite clear, especially that loudspeakers A and C are better than B and D. However, in other cases the results may not be so obvious. And even here one may be in doubt concerning certain questions, for instance, if there is a "true" difference between loudspeakers A and C as reflected in their means in the bottom margin of Table II (6.8 for A and 6.0 for C, a difference of 0.8 units).

To be able to extract more detailed information from the data and arrive at more definite conclusions it may be preferable to use statistical methods like analysis of variance (ANOVA) and related procedures for significance testing. ANOVA essentially means that the total variance in the data is split up into different components due to the different sources ("causes") of variation in a listening test, such as the loudspeakers, the programs, the subjects, and various types of interactions between these variables. The statistical tests make it possible to decide, with a certain probability, whether the differences in ratings between different loudspeakers, and/or between different programs, are "true" differences, or if they are due to chance variation. Similarly it is possible to decide whether there are some "true" interactions or not.

The rationale and the assumptions underlying these procedures are discussed in most texts on statistics and experimental design (for instance, Hays, 1973; Kirk, 1968; 1972; Winer, 1971), see also under 7 below.

3.1 ANOVA for individual data matrix

An ANOVA on the data in Table I (= the data for subject S) may be conveniently performed by any of many available computer programs for ANOVA (see 8 below). The results are presented in a summary table like Table IV below:

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	*F*	*p*
Loudspeakers (L)	148.93	3	49.64	177.29	<.01
Programs (P)	5.43	4	1.36	4.86	<.01
L x P	35.23	12	2.94	10.50	<.01
Within cell	11.33	40	0.28		
Total	200.92	59			

TABLE IV. Example of summary table for ANOVA on individual data matrix.

In this table the "mean squares" (MS) represent the most important information. Computationally they are obtained by dividing the "sum of squares" (SS, the sum of squared deviations around the corresponding mean, for instance, the deviations of the loudspeaker means from their common mean) by the "degrees of freedom" (*df*, an expression for the number of independent comparisons for the respective source). In this case the mean square for loudspeakers reflects the amount of variance due to differences between loudspeakers plus a certain "error" variance. Likewise the MS for programs reflects the amount of variance due to differences between programs plus error variance, and the interaction MS reflects variance due to interaction between loudspeakers and programs plus error variance. An independent estimate of the error variance is given by the "within cell" MS, which reflects the variance of the ratings within all cells of Table I. The error variance is thus an expression for the intra-individual variability. As noted at Table I, this subject was very stable in his ratings, and the estimated error variance is thus very low (0.28).

Consequently, the bigger the MS for a certain source is relative to the "within cell" MS (=error variance), the more likely it is that there are "true" (non-random) differences between the levels of the respective source (for instance, between different loudspeakers). This is formally tested by means of *F* tests, which means (in this case) that the corresponding MS is divided by the "within cell" MS. The resulting *F* values are given to the right in Table IV (for example, for loudspeakers equal to 49.64/0.28 = 177.29). These *F* values are compared to "critical values" at the respective degrees of freedom as given in tables of the *F* distribution, which appear in most textbooks in statistics. If the observed *F* value is higher than the "critical value", it is said to be significant at a certain selected probability (percentage) level.

Using a .01 (1%) significance level the critical value for the loudspeaker variable is found to be 4.31 (as found in an *F* table for 3 degrees of freedom in the numerator and 40 in the denominator; these are the degrees of freedom for loudspeakers and "within cell", respectively). The observed *F* value, 177.29, is far beyond the critical value and thus significant at .01 level. In the right-hand column of Table IV this is denoted by the expression *p* <.01, meaning that the probability (*p*) of getting the observed differences between the loudspeakers by chance alone is less than .01. Consequently the observed differences may be regarded as "true" differences.

For the program variable the critical *F* value is 3.83 (for 4 and 40 df) and for the interaction L x P it is 2.66 (df 12 and 40). In both cases the observed *F* values are higher than the critical values. Thus there are "true" differences between the programs, and a "true" interaction between loudspeakers and programs.

As used here the *F* test functions as an "overall" test. For example, *a significant F ratio* as regards loudspeakers tells only that there is at least one significant difference among all possible combinations of the loudspeakers (that is, the significant difference may be that between A and B, and/or that between A and C, and/or between A and D, B and C, B and D, C and D and/or between more complex combinations as, say, the mean of A+C versus the mean of B+D). To find exactly which difference(s) is (are) significant, it is necessary to perform tests for specific comparisons, see 4 below. On the other hand *a non-significant F ratio* for the loudspeakers would mean that *there is no significant* difference at all between the loudspeakers in any combination.

The interpretation of *a significant interaction loudspeakers x programs* can often be reasonably made by direct *inspection of* the data matrix. Statistical tests for this purpose are mentioned under 4 below. It is evident from Table I that the difference between loudspeakers A and B varies considerably from program to program, that the difference between the "best"and the "worst" loudspeaker is bigger for programs 2-4 than for program 5, etc. Generally it is important to study the meaning of a significant interaction, since it may give interesting information about the performance of the different loudspeakers at different types of program material.

It may be interesting to supplement the significance tests by estimations of the amount of variance accounted for by the different sources (that is, the loudspeakers, the programs, and the interaction between them), see 6.1.5.

3.2 ANOVA for group data matrix

Individual ANOVAS as described above may be performed for each subject in a listening test to get a detailed picture of each subject's ratings and also to provide estimates of intra-individual reliability (see 6.1). However, it is still more important to analyse the combined ratings from all subjects in a group by means of an ANOVA on a group data matrix like that in Table II. Besides loudspeakers and programs now also the subjects and various interactions including subjects enter as variables into the analysis.

The results from ANOVA on the group data in Table II are given in the following table:

Source of variation	SS	df	MS	F	p
Loudspeakers (L)	465.75	3	155.25	80.44	<.01
Programs (P)	11.94	4	2.99	0.84	-
Subjects (S)	105.25	3	35.08	46.16	<.01
L x P	47.36	12	3.95	3.09	<.01
L x S	17.34	9	1.93	2.54	=.01
P x S	42.86	12	3.57	4.70	<.01
L x P x S	45.98	36	1.28	1.68	>.01
Within cell	121.33	160	0.76		
Total	857.81	239			

TABLE V. Example of summary table for ANOVA on group data matrix (mixed model, case 2 below).

When computing the F values two cases must be distinguished:

1) The subjects used in the test constitute themselves the subject population, and the results from the test are strictly valid only for these subjects. The corresponding statistical model is called a fixed model. In this case all F values are computed by dividing the respective MS by the "within cell" MS (thus, for loudspeakers 155.25/0.76, for programs 2.99/0.76, for subjects 35.08/0.76, for L x P 3.95/0.76, and so on), and the critical F values are read from F tables at the respective degrees of freedom (for loudspeakers at 3 and 160 df, for programs at 4 and 160, for subjects at 3 and 160, for L x P at 12 and 160, and so on).

2) The subjects are randomly sampled from a certain defined population of subjects, to which one wants to generalize the results from the listening test. The four subjects in our example were randomly sampled from a society for high fidelity fans and could thus be considered as representative for the members of this society (actually ten subjects were used, but only four of them are included here to reduce the amount of data). The corresponding statistical model is called a mixed model, and it is this model that is used in the present case.

The *F* values are then computed as follows:

For loudspeakers the error term is the MS for the loudspeaker x subject interaction (MS_{LxS}), and *F* is obtained by MS_L / MS_{LxS}, that is, 155.25/1.93 = 80.44. Degrees of freedom for critical *F* values are 3 and 9, and the critical value at .01 level is found to be 6.99. Thus the observed *F* value is significant at .01 level.

For programs the error term is the program x subjects MS (MS_{PxS}), *F* is obtained by MS_P / MS_{PxS}, thus 2.99/3.57, which is less than 1.00 and consequently not significant. Df are 4 and 12, critical value 5.41.

For the loudspeaker x program interaction the error term is the triple interaction loudspeaker x program x subjects (MS_{LxPxS}). *F* is obtained by MS_{LxP} / MS_{LxPxS}, thus 3.95/1.28 = 3.09; df = 12 and 36, critical value 2.73 at .01 level. The interaction is significant at .01 level.

For the subjects and all interactions including subjects (L x S, P x S and L x P x S) the error term is the "within cell" MS. (In Table II a "cell" is the same as each combination of loudspeaker x program x subject. There are thus 4 x 5 x 4 = 80 cells, and there are three ratings in each cell - as in Table I.) *F* for subjects is 35.08/0.76 = 46.16 (critical value for df 3 and 160 is 3.92), for L x S 1.93/0.76 = 2.54 (critical value for df 9 and 160 is 2.53), for P x S 3.57/0.76 = 4.70 (critical value for 12 and 160 df is 2.31), and for L x P x S 1.28/0.76 = 1.68 (critical value for 36 and 160 df is about 1.77).

The term "mixed model" refers to the "mixing" of two "fixed" variables (loudspeakers, programs) and a "random" variable (subjects). If the subjects are not randomly sampled but themselves constitute the subject population, the subject variable is also "fixed", and consequently a "fixed model" should be used in the analysis.

The difference between these two situations is also the reason for using different error terms in the two models. For instance, the loudspeaker MS is divided by the "within cell" MS in the fixed model (in which generalization to other subjects is not possible), but by the loudspeaker x subject interaction MS in the mixed model (in which such generalization is possible, and where it is thus important to take a possible loudspeaker x subject interaction into consideration when testing differences between loudspeakers for statistical significance).

If a fixed model had been used in our example, the "within cell" MS (0.76) should be the error term for all F tests. In that case the same results are obtained as for the mixed model except for one thing, namely that the F test for programs would also be significant (2.99/0.76 = 3.93; critical value for df 4 and 160 is 3.45). The results may thus depend on which model is used in the statistical analysis. From a statistical point of view it is therefore important to consider whether the subjects are (or could reasonably be considered as) randomly sampled from a certain population or not. If they are, the mixed model should be used, and the results can be generalized to the population in question; if they are not, the fixed model is used, and the results hold only for the subjects used in the test.

In practice it is often difficult to achieve random sampling of subjects from a defined population. This does not necessarily preclude the possibility of generalizing the results. Even if the investigator has not achieved strict random sampling of his subjects, he might still find it reasonable to assume that his subjects are representative/typical for a certain population/category of listeners and thus have some justification for using a mixed model and tentatively generalize the results to the population in question. If this approach appears unreasonable, the investigator may prefer to use the fixed model for his analysis and refrain from generalizing his results on statistical grounds. However, he is free to consider a possible generalization on other, non-statistical, grounds. He may find, for instance, that his results agree with results from an earlier investigation with other subjects, or that the results agree with what may be expected from knowledge of the characteristics of the different loudspeakers etc. This type of generalization is thus not statistically based but justified by other types of information to be stated by the investigator.

The results in Table V may be briefly interpreted as follows:

Loudspeakers: There is at least one significant difference between the mean ratings for the different loudspeakers. Tests for specific comparisons can be made to clarify which differences are significant (see 4).

Programs: There are no significant differences between the mean ratings at the different programs (unlike the situation for the single subject S, see Table IV).

Subjects: There is at least one significant difference between the mean ratings for the different subjects. This means that different subjects tend to use somewhat different parts of the 10 - 0 scale. As noted earlier, subject V tends to have higher rating values than the other subjects (his mean rating averaged over all loudspeakers and programs was 6.0), while subject T goes in the opposite direction (mean rating 4.1, for subjects S and U the mean rating was in both cases 5.1). Such differences between subjects may be expected for several reasons and are relatively less important (sometimes a statistical test for the subject variable is not done at all). The main question regarding subjects is rather whether the differences between the loudspeakers are the same or not for different subjects, regardless of the subjects' "mean positions" on the 10 - 0 scale. The answer to this question is given by the *F* test for the L x S interaction, see below.

The interpretation of the significant interactions is only briefly discussed here. The L x P interaction shows up in various ways (see Table II), for instance, that the difference between loudspeakers A and B varies from program to program, that the difference between the "best" and the "worst" loudspeaker is different at different programs etc. The L x S interaction indicates that the ratings of different loudspeakers (in average over the programs) somehow differ with different subjects (please check! On the other hand a non-significant *F* value would mean that the differences between the loudspeakers are similar for all subjects). Likewise the P x S interaction suggests that the ratings at different programs (in average over the loudspeakers) vary with subjects. The L x P x S interaction was not significant here. If it happens to be so, this would require a fairly detailed inspection of the data to see its meaning (it might be, for example, that the meaning of an L x P interaction is different from subject to subject). The interpretation of interactions involving subjects is easier to do by means of individual data matrices like that in Table I.

How much work should be devoted to interpretations of significant interactions depends on the purposes with the test. A significant L x P interaction should be considered important to understand. In any case, one or more interactions involving the loudspeaker variable imply that the differences between the mean ratings of the loudspeakers (as given in the bottom margin of Table II) are not quite general, but somehow vary with different programs and/or subjects - and this may represent important information for future work in research or applications.

Estimations of the amount of variance accounted for by different sources of variation are discussed in 6.2.4.

3.3 ANOVA when each subject makes only one rating per case.

To increase the reliability of the mean ratings, and also to facilitate estimations of reliability (see 6), it was recommended in the introduction that each subject makes at least two independent ratings for each loudspeaker x program combination. In the example used above each subject made three ratings per combination.

However, in some listening tests it may happen that it is not possible to obtain more than one rating per combination by each subject. For instance, there may be so many loudspeakers and/or programs in the test, that it would be too tiring or lengthy for the subjects to listen more than once to each combination. Even in a smaller test it may happen that the available subjects do not have time for more than just one listening per combination etc. In an individual data matrix like that in Table I there will then be only one value per cell (and thus no need for computing a mean within each cell). In a group data matrix there will also be only one value per cell (cell = loudspeaker x program x subject combination), in other words there is only one value for each subject in each loudspeaker x program combination (and means like M_S, M_T, etc. in Table II do not appear; however, M_g would still appear).

ANOVA and related tests may still be performed but with certain modifications. Since there is only one value per cell, there can be no "within cell" variance, and this term therefore disappears in the summary table. The summary table for ANOVA on individual data matrix (compare Table IV) thus only contains loudspeakers (L), programs (P) and the loudspeaker x program interaction (L x P) as variation sources. The *F* test for loudspeakers and programs are both made using the interaction MS in the denominator, that is, for loudspeakers:MS_L / MS_{LxP} and for programs:MS_P / MS_{LxP}. An *F* test for the L x P interaction is not possible to do in this case.

The *F* tests for loudspeakers and programs may, however, be "biased" in a certain way. The correct denominator should be the "within cell" MS, which is an estimate of the error variance (see 3.1). Since there is no "within cell" MS now available, we use MS_{LxP} as the "best possible" substitute. However, MS_{LxP} is an estimate of variance due to interaction plus error variance (see again 3.1). If there is an interaction between loudspeakers and programs, the numerical value of MS_{LxP} will therefore be larger than what is actually due to error variance alone. Consequently it will be harder to get significant *F* ratios for the loudspeakers and the programs in such a case. A non-significant *F* test may thus become ambiguous: either there are in fact no "true" differences between the loudspeakers in average over the programs (between the programs in average over the loudspeakers, respectively), or the F test is not sensitive enough to detect the "true" differences because of the too big denominator.

In the summary table for ANOVA on group data matrix the "within cell" term likewise disappears (compare Table V). If the fixed model is used (see 3.2), all F tests (those for L, P, S, L x P, L x S and P x S) use MS_{LxPxS} as their denominator (and it is not possible to have an F test for the L x P x S interaction). Since MS_{LxPxS} is an estimate of the L x P x S interaction plus error variance, the F tests may be biased (if there is a "true" L x P x S interaction, the denominator will be "too large"), and non-significant F ratios may be ambiguous in analogy with what was described above. If the mixed model is used, the F tests for loudspeakers, programs and the loudspeaker x program interaction can still be performed as described in 3.2. However, one then usually refrains from making F tests for subjects and interactions involving subjects, since such F tests would require a "within cell" MS in the denominator.

ANOVA and F tests can thus be performed in listening tests in which there is only one rating per cell. However, certain tests may be insensitive and/or ambiguous, and certain other tests cannot be performed. With regard to the formulas for specific comparisons in chapter 4, these formulas can still be used with two modifications:

1) When an $MS_{within\ cell}$ appears in a formula, it should be replaced by the MS for that interaction that was actually used as denominator in the corresponding F test, and the degrees of freedom are those for this interaction. For example, in formula (1) in 4.1 $MS_{within\ cell}$ should be replaced by MS_{LxP} and the degrees of freedom are those for L x P in Table IV (that is, df = 12).

2) The value of n (= number of ratings in each cell) will be 1 in all formulas.

As seen above the statistical analysis is simpler and more conclusive if there are more than one rating per cell. And as regards reliability, it is intuitively clear that the mean of two or more ratings is more reliable than only one single rating. It is not even possible to get a statistical estimate of the intra-individual reliability, unless there are at least two ratings per cell (see 6.1). However, the inter-individual reliability may be estimated by procedures described in 6.2 (in formula (12) in 6.2.3 the SS and df for "within cell" are simply dropped from the formula).

Although all these circumstances point towards using more than one rating per case, there may still be defenses for not doing so. It is sometimes easier to get more subjects and have them do one rating per loudspeaker x program combination than to have a smaller number of subjects do more ratings per combination. Or it may be known from some earlier listening tests, that the subjects in general have satisfactory *intra*-individual reliability, so that one has confidence in their rating ability, even if they are allowed to make only one rating per case. Examples of listening tests both with three ratings per case and with one rating per case may be found in Gabrielsson, Frykholm & Lindström (1979).

If there are many loudspeakers and/or programs in a test quite another alternative is to have *some* subjects listen to all loudspeakers but only for one program, *some other* subjects also listen to all loudspeakers but with another program etc, that is, use one group of subjects per each program. This alternative is called a *"split-plot"* design and is discussed in many possible variants in 5.3. It may then be reasonable to have each subject do two (or more) ratings for those loudspeaker x program combinations he listens to.

4 TESTS FOR SPECIFIC COMPARISONS

A significant F value for loudspeakers means that there is at least one significant difference among all possible pair combinations of loudspeakers (A-B, A-C, A-D, B-C, B-D, and C-D in our case) or among more complex combinations (as the mean of A+B versus the mean of C+D etc). The F test does not tell how many or which of these differences are significant. This has to be studied by means of various tests for specific comparisons. The same line of reasoning applies to significant F values for the programs or for the subjects. However, we will limit the discussion to deal only with comparisons involving two loudspeakers at a time (A-B, A-C, A-D etc).

There are several alternative tests available for specific comparisons, and there is unfortunately no complete agreement among statisticians as to their use (see further in 4.2). However, the following procedures seem to be fairly commonly agreed upon. Four different situations will be discussed: planned independent comparisons (4.1), planned non-independent comparisons (4.2), non-planned comparisons (4.3), and specific comparisons within single programs (4.4).

4.1 Planned independent comparisons

Suppose that it was planned before the listening test (=before any data had been collected) that it for some reason was especially important to test whether loudspeaker A was better than loudspeaker B. This is called a planned (or a priori) comparison. This may be performed with a t test.

In the case of an individual data matrix like Table I the test formula would be:

$$(1) \qquad t = \frac{M_A - M_B}{\sqrt{2MS_{w.cell}/np}} = \frac{6.9 - 3.6}{\sqrt{(2 \times 0.28)/(3 \times 5)}} = 17.37$$

where M_A and M_B are the means for loudspeakers A and B given in the bottom margin of Table I, $MS_{within\ cell}$ is given in Table IV, n = number of replications (ratings) in each cell in Table I, and p = number of programs.

The degrees of freedom for this test are the same as for the "within cell" term in Table IV, that is, 40. The critical value for df = 40 at .01 level, one-tailed test (which is used when the test deals with a difference in a certain direction as here, that is, if A is better than B) is 2.42 as seen in a table for the t distribution. Since the observed t value is higher than the critical value, the difference is significant at .01 level. (If the interest was in testing whether it was any difference between A and B, regardless of direction, a two-tailed test should be used. This changes nothing in the formula, but the critical value will be different as found in a table for the t distribution.)

In the case of a group data matrix like that in Table II the same test would be for the mixed model

$$t = \frac{M_A - M_B}{\sqrt{2MS_{LxS}/nsp}} = \frac{6.8 - 3.8}{\sqrt{(2 \times 1.93)/(3 \times 4 \times 5)}} = 12.00 \qquad (2)$$

where M_A and M_B are found in the bottom margin of Table II, MS_{LxS} in Table V, n = number of replications (ratings) under each condition (each subject made three ratings for each loudspeaker x program combination), s = number of subjects, and p = number of programs. The degrees of freedom are the same as for L x S in Table V, that is, 9. The critical t value for df = 9 at .01 level, one-tailed test, is 2.82. Thus the difference is significant at .01 level.

In the fixed model the $MS_{w.cell}$ is used instead of MS_{LxS} and the degrees of freedom are those for the "within cell" term. For the same situation that would give

$$t = \frac{M_A - M_B}{\sqrt{2MS_{w.cell}/nsp}} = \frac{6.8 - 3.8}{\sqrt{(2 \times 0.76)/(3 \times 4 \times 5)}} = 18.75 \qquad (3)$$

Critical value for df = 160 at .01 level, one-tailed test, is about 2.35.

The t test procedure may be used for more than one planned comparison, if these comparisons are independent of each other.

Roughly this means that any two (or more) planned comparisons are "non-overlapping", do not have any loudspeaker in common. With four loudspeakers as here only two planned independent comparisons can be formed at a time, for instance, the comparisons A-B and C-D, or the comparisons A-C and B-D, or the comparisons B-C and A-D etc. It would not be possible to take, say, the two comparisons A-B and A-C because they have loudspeaker A in common and are thus not independent, nor would it be possible to take A-D and B-D because they have loudspeaker D in common, and so on.

The simple t test procedure is thus applicable for planned independent comparisons. With a certain modification it may also be used for planned non-independent comparisons.

4.2 Planned non-independent comparisons

Planned comparisons of interest to the investigator may of course include non-independent comparisons, for instance, the comparisons A-B and A-C, or A-B and B-C, or A-B, A-C, and C-D (three comparisons in the last example) etc.

In this case t tests can still be performed according to the formulas in 4.1, but the significance level should be distributed over the number of tests that are made. If we continue with the .01 level used for all earlier examples, and if we want to do two non-independent comparisons, each of these single tests should be performed with .01/2 = .005 significance level. If there were three non-independent comparisons, each single test should be made with .01/3 = .0033 level etc.

In case of .05 significance level two non-independent comparisons would be tested at .05/2 = .025 level, five non-independent comparisons at .05/5 = .01 level, and so on. In fact the significance level thus refers to the collection of all tests made (for instance, to the collection of two non-independent comparisons, or of five non-independent comparisons etc, whatever is the case).

This procedure (often called Bonferroni t statistics) does not change anything in the computations of the t values. The only difference is that the critical t value will be different, since the significance level for each single test is diminished. As an example suppose that it was planned to make the comparisons A-B and A-C for the data in the group data matrix (Table II). The two t tests are performed as shown for the A-B comparison in formula (2). For A-B the observed t value is 12.00, and for A-C it is 3.20 (computed from the same formula only replacing M_B with M_C). Since there are two non-independent tests, the significance level for each of them is .01/2 = .005, and the critical t value is found to be 3.25 (df=9, one-tailed) as compared with 2.82 for independent comparisons. The difference A-B is clearly significant, while the A-C difference is just on the limit of being significant (in such a borderline case one would probably look for some other relevant information as guideline for a conclusion).

Obviously it is harder to get significant t values for non-independent than for independent comparisons. And the more non-independent comparisons the more difficult it will be, since the significance level for the single tests diminish in proportion to the number of comparisons (and the corresponding critical t values thus increase). Consequently this procedure may get very insensitive to detect "true" differences as the number of comparisons is increased. Therefore it should only be used for a relatively small number of non-independent comparisons. For a big number of comparisons it may be better to use the procedure described below in 4.3.

A technical difficulty is that this procedure may require critical *t* values at significance levels which are not found in common tables for the *t* distribution (for instance, if a .05 level is distributed over three comparisons, each comparison should be performed at .05/3 = .017 level, which is not listed in common tables). There are ways of computing approximate critical *t* values, and it is also possible to use special tables by Dunn (reproduced in Kirk, 1968). The problem can also be avoided by not sticking so strictly to the commonly used significance levels of .05 or .01 (see 7.1). For instance, in the example above one could perform each of the single *t* tests at .01 level, which means that the significance level for the collection of these three tests would be (3 x .01) = .03.

The rationale behind these procedures and some in the following, is beyond the scope of this paper. Detailed discussions may be found in Kirk (1968, 1972), Hays (1973) and Winer (1971). Suffice it here to say that the discussions deal with considerations about planned and non-planned comparisons, about independent and non-independent comparisons, and about various conceptual units for the significance level. A general principle is that the more specific comparisons that are made, the less sensitive the corresponding tests will be for detecting "true" differences. Conversely, the fewer (and independent) comparisons that are made, the more sensitive the corresponding tests will be for detecting "true" differences (if there are any).

The *t* test procedures for planned comparisons described in 4.1 and 4.2 do not presuppose a significant overall *F* test. In fact they may be applied directly to the specific planned comparisons without a prior *F* test. However, since the underlying ANOVA procedure provides much valuable information regarding variances, estimates of reliability etc, it is still recommended to perform the ANOVA procedures as a first step.

4.3 Non-planned comparisons

If the investigator has no specific hypotheses or plans for doing certain selected comparisons, but simply wants to know if the loudspeakers in his test differ at all when rated on the 10 - 0 scale, the data are first analysed by ANOVA and *F* test. If the *F* test for the loudspeakers turns out to be not significant, the conclusion is that there are no differences between the loudspeakers in question. If the *F* test is significant, there is at least one significant difference somewhere among the loudspeakers, and he may want to know which specific difference(s) is (are) significant. In that case he would apply procedures for *non-planned* (*a posteriori* or *post-hoc*) comparisons.

There are various procedures for such comparisons. The one proposed here is known as Tukey's HSD (Honestly Significant Difference) procedure. In our case with four loudspeakers there are six possible pairwise comparisons (A-B, A-C, A-D, B-C, B-D, and C-D). Any of these is declared as significant, if the corresponding difference between means exceeds the computed value for HSD.

For the case with the individual data matrix (Table I) HSD is computed as follows:

(4) $$\mathrm{HSD} = \underline{q}_{.01,40}\sqrt{\mathrm{MS}_{\mathrm{w.cell}}/np} = 4.70\sqrt{0.28/(3 \times 5)} = 0.66$$

The value of $\underline{q}$ is given in tables of the "studentized range statistic". The value 4.70 here is found in such a table for the case of .01 significance level, df = 40 (these two values indicated in the suffices to $\underline{q}$ in the formula), and four means (loudspeakers). $\mathrm{MS}_{\mathrm{w.cell}}$ is found in Table IV, n = number of replications in each cell, and p = number of programs.

Looking at the means in the bottom margin of Table I, we conclude that the (absolute) difference between A and B, A and D, B and C, and between C and D all exceed the computed HSD of 0.66. Thus these four differences are significant, while the difference between A and C is not significant, nor the difference between B and D.

Applied to the group data matrix (Table II) and the mixed model case HSD is given by

(5) $$\mathrm{HSD} = \underline{q}_{.01,9}\sqrt{\mathrm{MS}_{\mathrm{LxS}}/nsp} = 5.96\sqrt{1.93/(3 \times 4 \times 5)} = 1.07$$

The $\underline{q}$ value is looked up for .01 level, df = 9, and four means. $\mathrm{MS}_{\mathrm{LxS}}$ is found in Table V, n = number of replications (defined as in formula (2)), s = number of subjects, and p = number of programs. Looking at the means in the bottom margin of Table II, we conclude that the differences associated with A-B, A-D, B-C, and C-D are bigger than HSD and thus declared as significant, while the differences associated with A-C and B-D are not significant.

In the case of a fixed model $\mathrm{MS}_{\mathrm{w.cell}}$ is used instead of $\mathrm{MS}_{\mathrm{LxS}}$ and the degrees of freedom are those for $\mathrm{MS}_{\mathrm{w.cell}}$ Thus:

(6) $$\mathrm{HSD} = \underline{q}_{.01,160}\sqrt{\mathrm{MS}_{\mathrm{w.cell}}/nsp} = 4.50\sqrt{0.76/(3 \times 4 \times 5)} = 0.50$$

In this case the difference associated with A-C is also significant besides the differences given above for the mixed model. The difference A-C is thus declared significant if the conclusions are restricted to the four subjects in the test (that is, using a fixed model), but not if the intent is to generalize from these subjects to the population from which they are drawn (that is, using a mixed model).

The tests performed by the HSD procedure are *two-tailed* tests, that is, they test for differences regardless of the direction of the differences. This is in line with the typical *a posteriori* situation, in which the investigator has no specific comparisons in mind but only wants to see if the loudspeakers differ at all. It may also be noted that the significance level in the HSD procedure refers to the *collection* of all tests (however, in a somewhat different way than for the Bonferroni *t* statistics).

4.4 Specific comparisons within single programs

The investigator may have planned to *compare certain loudspeakers within one of the programs* used in the listening test. These tests follow the principles described in 4.1 and 4.2. If he has planned to make a single comparison (say A-B) or two *independent* comparisons (for instance, A-B and C-D) within a certain program, he may use the *t* test procedure in 4.1. The formulas given there apply here too with two modifications: the means in the numerator should of course be the means for the respective loudspeakers within the selected program, and further the value of p (= number of programs) is taken away from the denominator. To take but one example: Assume that we planned to test whether loudspeaker C is better than D *at program 1* in the group data matrix (Table II), using mixed model. The test would be (compare formula (2)):

$$t = \frac{M_C - M_D}{\sqrt{2MS_{LxS} / ns}} = \frac{5.9 - 4.2}{\sqrt{(2 \times 1.93)/(3 \times 4)}} = 3.00 \qquad (7)$$

which turns out to be significant (critical value 2.82 for df=9, one-tailed as at formula (2)).

If the planned comparisons within the program in question are *non-independent*, procedures analogous to those of the Bonferroni *t* statistics in 4.2 are used, that is, with the significance level distributed over the number of comparisons.

If the ANOVA and *F* tests have revealed a significant interaction between loudspeakers and programs, this implies that the differences between the loudspeakers somehow vary from program to program. These variations are often obvious enough from simple visual inspection of the data matrix. For instance, in the group data matrix (Table II) it is evident that loudspeakers A and C are the best for all five programs, but that their superiority is most marked for programs 2-4 and somewhat less marked for programs 1 and 5. This information would probably be sufficient for the investigator (he could then try to understand why the difference between the loudspeakers is bigger for certain programs than for others). If, for some reason, he wants a formal statistical test on the differences between the loudspeakers within each of the programs, one simple way would be to compute Tukey's HSD with respect to the single programs (instead of over all programs as illustrated in 4.3). For the group data matrix and mixed model formula (5) can be used but taking away p (= the number of programs) under the square root sign. The resulting HSD is 2.38. The difference between any two loudspeakers within any of the programs must thus exceed 2.38 to be significant. Inspecting the data in Table II reveals that loudspeaker A is significantly different from B for all programs except for program 1, and from D for programs 2-4. Loudspeaker C is significantly different from B and from D for programs 2-4, but not for programs 1 and 5.

5 LISTENING TESTS INCLUDING EXTRA VARIABLES

Besides loudspeakers and programs an investigator may sometimes want to include more independent variables in his listening test. For instance, some or each of the programs may be presented at different sound levels, the positions of the loudspeakers and/or of the listeners may be varied etc. The more variables are included, the more complex the analysis will be - and the more important it is to use an efficient procedure like ANOVA and accompanying tests. The statistical procedures are in most cases rather straight-forward generalizations of those described earlier. In the following some examples of possible analyses are given for the case with one extra variable (5.1) and two extra variables (5.2).

When extra independent variables are included in a test, the total number of listening conditions may sometimes become so big that it would be too tiring or unpractical to have all subjects make ratings under all conditions. In such a case one alternative is to have some subjects take part in the test under certain listening conditions and have some other subjects participate under certain other conditions. Depending on the circumstances many different combinations may be considered, the subjects may be divided into still more sub-groups participating under different conditions, and so on. Many examples of such possibilities (so called "split-plot designs") and the related statistical procedures are given in 5.3.

5.1 One extra variable

The example given in this section uses differences in sound level as an extra variable (besides loudspeakers and programs). The same principles apply, of course, to any extra variable - for instance, differences in listening positions or differences in loudspeaker positions.

5.1.1 Individual data

In the listening test used as example here each program was actually presented at two different sound levels, called "high" and "low" in the following. The individual data matrix for subject S is given here:

	Loudspeaker								
	A		B		C		D		
Level	High	Low	High	Low	High	Low	High	Low	Means for programs
	7	6	5	6	6	7	3	6	
Program	6	6	5	5	7	6	3	4	
1	7	7	5	5	5	7	4	4	
	6.7	6.3	5.0	5.3	6.0	6.7	3.3	4.7	5.5
	6	8	3	4	5	7	3	6	
	6	7	3	4	5	7	3	5	
2	7	8	4	4	7	7	4	4	
	6.3	7.7	3.3	4.0	5.7	7.0	3.3	5.0	5.3
	7	8	2	4	7	5	3	3	
	8	8	2	3	7	6	3	3	
3	8	8	2	4	7	8	3	3	
	7.7	8.0	2.0	3.7	7.0	6.3	3.0	3.0	5.1
	7	7	3	4	8	7	3	3	
	8	3	3	3	7	7	3	3	
4	8	7	3	4	8	8	3	3	
	7.7	5.7	3.0	3.7	7.7	7.3	3.0	3.0	5.1
	6	8	5	4	6	8	5	7	
	7	7	5	5	6	7	5	5	
5	6	8	4	4	6	7	5	5	
	6.3	7.7	4.7	4.3	6.0	7.3	5.0	5.7	5.9
Means for levels	6.9	7.1	3.6	4.2	6.5	6.9	3.5	4.3	
Means for loud-speakers	7.0		3.9		6.7		3.9		

TABLE VI. Example of individual data matrix for listening test with three independent variables: loudspeakers, programs, and sound levels.

The data under the "High" level are the same as those in Table I (although written vertically here), so the new data are those under the "Low" level. The mean of the three ratings (replications) within each cell is written at the bottom of the respective cell.

Visual inspection of such a matrix may in many cases give enough information. However, to be able to look at more details and facilitate the conclusions an ANOVA may be performed giving the following result:

Source of variation	SS	df	MS	*F*	*p*
Loudspeakers (L)	262.43	3	87.48	171.53	<.01
Sound level (SL)	7.01	1	7.01	13.75	<.01
Programs (P)	10.08	4	2.52	4.94	<.01
L x SL	1.49	3	0.50	<1.0	-
L x P	45.12	12	3.76	7.37	<.01
SL x P	8.95	4	2.24	4.39	<.01
L x SL x P	14.38	12	1.20	2.35	>.01
Within cell	40.67	80	0.51		
Total	390.13	119			

TABLE VII. Summary table for ANOVA on data in Table VI.

As in Table IV all *F* tests are performed by dividing the respective MS by the "within cell" MS. The resulting *F* values are significant at .01 level for loudspeakers, sound levels, programs and the interactions loudspeakers x programs and sound levels x programs.

No detailed interpretation of these data is given here. Note that although there are significant *F* tests for loudspeakers, sound levels, and programs, there are also significant interactions (L x P and SL x P), which necessitate a careful inspection of the data matrix (it is obvious, for example, that the effect of different sound levels varies in rather complex ways for different cases).

5.1.2 Group data

The group data matrix for the four subjects can be written as Table II but doubled - that is, Table II represents the data at the "high" level, and consequently a similar representation must be made for the "low" sound level. These data are not given here, but the summary table for the corresponding ANOVA should have the following principal appearance:

Source of variation	SS	df	MS	*F*	*p*
Loudspeakers (L)		3			
Sound level (SL)		1			
Programs (P)		4			
Subjects (S)		3			
L x SL		3			
L x P		12			
L x S		9			
SL x P		4			
SL x S		3			
P x S		12			
L x SL x P		12			
L x SL x S		9			
L x P x S		36			
SL x P x S		12			
L x SL x P x S		36			
Within cell		320			
Total		479			

Table VIII. Schema for summary table in ANOVA on group data in listening test including loudspeakers, programs and sound levels as independent variables.

In analogy with what was said in 3.2 (especially regarding Table V) two cases must be distinguished when computing *F* values:

1) If a fixed model is used (that is, the subjects used in the test constitute the subject population and the results are valid only for these subjects), all *F* values are computed by dividing the corresponding MS by the "within cell" MS. (In this case a cell is constituted by each combination of loudspeaker x program x sound level x subject.)

2) If a mixed model is used (that is, the subjects are randomly sampled from a certain population to which the results are generalized), the *F* values are computed in analogy with what was described for the mixed model in connection with Table V. The MS for each of the three "fixed" variables (loudspeakers, sound levels and programs) are divided by the MS for their respective interactions with subjects:

$MS_L / MS_{L \times S}$

$MS_{SL} / MS_{SL \times S}$

$MS_P / MS_{P \times S}$

The MS for each of the two-factor interactions of these variables are divided by the MS for the respective three-factor interaction including subjects as the third factor:

$MS_{L \times SL} / MS_{L \times SL \times S}$

$MS_{L \times P} / MS_{L \times P \times S}$

$MS_{SL \times P} / MS_{SL \times P \times S}$

The MS for the three-factor interaction of the fixed variables is divided by the MS for the four-factor interaction with subjects as the fourth factor:

$MS_{L \times SL \times P} / MS_{L \times SL \times P \times S}$

Finally for all terms involving subjects (S, L x S, SL x S, P x S, L x SL x S, L x P x S, SL x P x S, and L x SL x P x S) the corresponding MS is divided by the "within cell" MS (that is, $MS_S / MS_{w.cell}$, $MS_{L \times S} / MS_{w.cell}$, and so on).

5.1.3 Tests for specific comparisons

Tests for specific comparisons in a listening test with three variables (loudspeakers, sound levels, programs) follow the principles described in chapter 4. The distinction between planned independent comparisons, planned non-independent comparisons, non-planned comparisons and comparisons within single programs applies here too. The formulas given in chapter 4 can be used with the following modifications:

a) The desired means (the two means involved in the comparison in question) are written in the numerator. (This does not apply to formulas for computation of HSD, since no means appear there.)

b) The MS term in the denominator (or under the square root sign in HSD formulas) should always be the error term in the corresponding *F* test. For individual data and for the fixed model with group data this is very simple, because the error term for all *F* tests is the "within cell" MS. For the mixed model with group data there are several different error terms as described in 3.2 and 5.1.2, and care must be taken to use the same MS in formulas for specific comparisons as for the corresponding *F* test - for example, in a specific comparison involving two loudspeakers the MS term will thus be the MS for loudspeaker x subjects interaction (MS_{LxS}).

c) The term to the right of the MS term in formulas 1-7 should equal the number of observations from which each of the actual means are computed. This number is thus obtained by multiplying the number of replications (designated n in these formulas) with the number of "levels" in each other variable over which the means in question are computed (that is, over the number of subjects and/or over the number of programs and/or over the number of sound levels).

The degrees of freedom for tests on specific comparisons are the same as for the MS term, when it is used as error term (denominator) in the corresponding F test.

An example is used to illustrate the above points: Suppose data were available for the group data matrix mentioned (but not given) in 5.1.2, and that the mixed model was applicable (subjects randomly selected from a population). Suppose further that we planned before the listening test to test whether loudspeaker C was better than loudspeaker D in average over all programs, sound levels, and subjects. That would result in a t test of the type given in formula (2), but with the actual means for C and D entered into the numerator. The MS term in the denominator should be MS_{LxS} (since this is the error term for F test on loudspeakers in the mixed model, see 5.1.2). The term to the right in the denominator would be the product of the number of replications x the number of subjects x the number of programs x the number of sound levels, which equals the number of observations from which each of the means in the numerator are computed. The degrees of freedom would be those for MS_{LxS} from the ANOVA procedure (Table VIII).

If this comparison between C and D was planned only for one of the programs, there would be no difference as regards the MS term (still MS_{LxS}) and degrees of freedom. The means have to be replaced by their respective means for this selected program, and the term to the right in the denominator would be the product of the number of replications x the number of subjects x the number of sound levels, which equals the number of observations from which these means are computed.

5.1.4 One extra variable only for certain programs

Assume that two different sound levels were used only for, say, two out of five programs in a listening test. This situation could be analysed in different ways. One way could be to regard the added levels for the two programs as new "programs", so that in fact the analysis is made as if there were seven "programs" (the five "real" programs plus two of them at another sound level). This would correspond to the data matrices of Table I (for a single subject) or Table II (for a group of subjects), and the ANOVA would take the form of that in Table IV or Table V, respectively.

For a more detailed analysis regarding the effects of different sound levels it may be interesting to make a separate ANOVA restricted to those two programs, which were presented at different levels. Such an analysis would thus include three variables: the loudspeakers, the (two) programs, and the sound levels and would be of the type shown in Table VII (for a single subject) or Table VIII (for a group of subjects).

Tests for specific comparisons would follow the principles described in chapter 4 and in 5.1.3.

5.2 Two extra variables

Besides including differences in sound level (used as example in 5.1) one may also include differences in the position of the listeners or of the loudspeakers. The analysis of data in such a listening test with two extra variables represents a straightforward generalization of the procedures described in 5.1.

An individual data matrix like that in Table VI must then be extended to include also the position variable, and the corresponding ANOVA likewise represents an extension of that in Table VII to include four independent variables (loudspeakers, sound levels, programs and positions) and all possible interactions between them. The analysis of group data would represent an extension of that in Table VIII to include five variables (loudspeakers, sound levels, programs, positions and subjects) and all possible interactions between those variables. The way of making *F* tests and making tests for specific comparisons follows the principles described in chapter 4 and in 5.1.3.

Still more variables may be included in a listening test (for instance, characteristics of the subjects as age, sex, hi-fi experience etc), and their effects may be investigated by analogous extensions of the ANOVA. The more variables included, the more complex the analysis will be - on the other hand, the more information may be obtained. Such questions must be given careful consideration when designing a listening test.

5.3 Some alternative designs ("split-plot" designs)

In all earlier examples it has been assumed that each subject listens to all combinations of loudspeakers x programs (as illustrated in Table II), or to all combinations of loudspeakers x programs x sound levels (Table VI) etc. This type of design is often called "repeated measurements on the same subjects", sometimes "randomized block factorial design".

If the number of variables in a listening test is increased (or if the number of loudspeakers and/or programs is very big), it follows that each subject gets more and more to do, and the test may be lengthy and tiring (even though it may be split up into several sessions separated in time). In such a situation an alternative may be to apply the principle of "repeated measurements on the same subjects" only to certain variables but not to other variables. This type of designs is often called "split-plot" designs (sometimes "mixed designs").

5.3.1 Example of "split-plot" design

As an example consider the situation that an investigator has selected four loudspeakers and five programs (resulting in a total of twenty loudspeaker x program combinations), but also finds it interesting to use two different listening positions in the room. If the "repeated measurements" principle is used, this doubles the work for each subject. An alternative could be to have some subjects listen to all loudspeaker x program combinations in one of these positions, and some other subjects in the other position. The "repeated measurements principle" would thus apply to loudspeakers and to programs but not to positions. An individual data matrix and corresponding ANOVA would still look like Tables I and IV. A group data matrix may be arranged as follows (labels in margins omitted):

		Loudspeaker																			
		A					B					C					D				
		Program					Program					Program					Program				
Position	Subject	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	S																				
	T																				
	U																				
	V																				
2	W																				
	X																				
	Y																				
	Z																				

TABLE IX. Group data matrix for a "split-plot" design in a listening test with loudspeakers, programs, and positions as independent variables.

Assuming that eight subjects are available, they are randomly divided into two groups of four members each, and it is also randomly decided which group should use which listening position. All eight subjects listen to all twenty loudspeaker x program combinations ("repeated measurement"), but each subject uses only one position ("non-repeated measurement", subjects S, T, U, and V have position 1, but subjects W, X, Y, and Z have position 2).

The summary table of an ANOVA for this case may have the following appearance:

Source of variation	SS	df	MS	F	p
Between subjects:		7			
Positions (PO)		1			
Subj. within groups		6			
Within subjects:		472			
Loudspeakers (L)		3			
L x PO		3			
L x subj.w. groups		18			
Programs (P)		4			
P x PO		4			
P x subj.w. groups		24			
L x P		12			
PO x L x P		12			
L x P x subj.w. groups		72			
Within cell		320			
Total		479			

TABLE X. Summary table for ANOVA of the "split-plot" design in Table IX.

The terms "Between subjects" and "Within subjects" in Table X are not necessary to include. They illustrate the two "parts" of this design, the "non-repeated measurements" part including the position variable and the "repeated measurements" part including loudspeakers and programs. The SS and df for "Between subjects" and "Within subjects" are simply the sum of the SS and df, respectively, for the sources following them in the table. It is assumed, as in earlier examples, that each subject makes three ratings for each loudspeaker x program combination.

The F tests are performed as follows:

1) If a fixed model is used (the results apply only to the used subjects), all F values are computed by dividing the corresponding MS by the "within cell" MS. (In this case a "cell" is constituted by each combination of loudspeaker x program x subject x position.)

2) If a mixed model is used (the subjects were sampled randomly from a population to which the results are generalized), the following computations apply:

MS_{PO} / $MS_{subj.w.groups}$

MS_{L} / $MS_{L \text{ x subj.w.groups}}$

$MS_{L \text{ x } PO}$ / $MS_{L \text{ x subj.w.groups}}$

MS_{P} / $MS_{P \text{ x subj.w.groups}}$

$MS_{P \text{ x } PO}$ / $MS_{P \text{ x subj.w.groups}}$

$MS_{L \text{ x } P}$ / $MS_{L \text{ x } P \text{ x subj.w.groups}}$

$MS_{PO \text{ x } L \text{ x } P}$ / $MS_{L \text{ x } P \text{ x subj.w.groups}}$

F tests for the remaining terms (L x Subjects w.groups, P x Subjects w. groups, and L x P x Subjects w. groups) may be made by dividing their respective MS by the "within cell" MS.

Tests for specific comparisons follow the principles described in chapter 4 and 5.1.3. The following examples apply for the mixed model. A test of the difference between two loudspeakers in average over programs, positions, and subjects would have $MS_{L \text{ x subj.w.groups}}$ as error term (and the corresponding df) divided by the product of the number of programs x the number of positions x the number of subjects within each group x the number of replications. A test of the difference between two loudspeakers at one of the listening positions would also use $MS_{L \text{ x subj.w.groups}}$ but divided by the product of the number of programs x the number of subjects within each group x the number of replications.

In the "split-plot" design used here (Table IX) it might be intuitively clear that the possible effect of different listening positions is not investigated as efficiently as regards differences between loudspeakers and between programs. Since there are different subjects in the two different listening positions, possible differences between the subject groups will be confounded with possible differences between listening positions. Tests on "repeated measurements" variables are in general more sensitive than tests on "non-repeated measurements" variables, since variation within individuals is typically smaller than variation between individuals.

5.3.2 Further applications of the "split-plot" design

The above analysis may be attractive to use in listening tests in which two (or even more) subjects listen simultaneously in the listening room. The subjects then necessarily have different listening positions and by the above analysis not only the effects of different loudspeakers and programs can be studied but also the effects of different listening positions. However, the test regarding the position variable may not be very sensitive.

If it is desired to increase the sensitivity, either the number of subjects should be increased or "repeated measurements" should be applied also for the position variable (in the latter case resulting in a design of the type described in 5.1).

There is, of course, no limitation to only two positions as used in our example. Any number of positions is possible to include in the "split-plot" design (see Table IX), but the more positions the more subjects are required. The number of subjects per position is minimum two, but in that case the statistical test is very insensitive (in practice it seems necessary to have at least 3-4 subjects per position).

There are several other possible applications of the "split-plot" design in listening tests. Just a few examples are suggested here:

1) In the design given in Table IX the position variable could be replaced by sound level (that is, subjects S - V listen to all loudspeaker x program combinations at one sound level, subjects W - Z at another sound level).

2) In the same table the program variable and the position variable could change places. That is, all subjects would listen to all loudspeakers and in all listening positions used ("repeated measurement" with regard to loudspeakers and positions), but different subjects would listen to different programs ("non-repeated measurement" as regards the program variable).

3) Still another possibility would be that all subjects listen to all loudspeakers ("repeated measurement"), but different subjects listen to different programs and in different positions ("non-repeated measurement" as regards programs and positions using, say, three subjects per program x position combination).

4) If both position and sound level are included in a listening test, one possibility could be to let all subjects listen to all loudspeaker x program combinations ("repeated measurement" in these two variables), but different subjects listen at different sound levels and in different positions ("non-repeated measurement" in those two variables with, say, at least three subjects per each sound level x position combination; compare this design with "repeated measurement" in all four variables as sketched in 5.2).

5) If a listening test is performed with subjects of different characteristics (as age, sex, degree of earlier experience of high-fidelity reproduction etc), and it is desired to see if these characteristics influence the ratings, a design analogous to that in Table IX may be used.

Simply replace the position variable in Table IX with, for instance, sex (subjects S - V males, subjects W - Z females), or "high-fidelity experience" (subjects S - V being "hi-fi enthusiasts", subjects W - Z "non hi-fi experienced people").

The analysis of the examples given in points 1, 2, and 5 above are analogous to the example given in 5.3.1 as regards ANOVA and F tests (there are some differences as regards tests for specific comparisons). Suitable analysis of the examples in points 3 and 4 may be found in the texts by Kirk (1968) or Winer (1971).

5.3.3 "Split-plot" design for many loudspeakers

Still another application would be to let all subjects listen to all loudspeakers ("repeated measurement"), but different subjects listen to different programs ("non-repeated measurement", the design given in Table II would thus be changed so that there were different subjects for each of the programs). This might be a possibility if so many loudspeakers are tested, that it would be unreasonable to have each subject listen to all loudspeaker x program combinations. In this case the summary table from ANOVA would look like this:

Source of variation	SS	df	MS	*F*	*p*
Between subjects:					
Programs (P)					
Subj. within groups					
Within subjects:					
Loudspeakers (L)					
L x P					
L x subj.w. groups					
Within cell					
Total					

TABLE XI. Summary table for ANOVA in "split-plot" design with "repeated measurements" as regards loudspeakers but "non-repeated measurements" as regards programs.

In a mixed model the following *F* tests apply:

MS_P / $MS_{subj.w.groups}$

MS_L / $MS_{L \text{ x } subj.w.groups}$

$MS_{L \text{ x } P}$ / $MS_{L \text{ x } subj.w.groups}$

The L x Subjects within groups interaction may be tested by dividing the corresponding MS by the "within cell" MS.

Tests for specific comparisons follow the general principles described in chapter 4 and in 5.1.3. A test of the difference between two loudspeakers in average over the programs and the subjects would have $MS_{L \times subj.w.groups}$ as error term (and the corresponding df) divided by the product of the number of programs x the number of subjects within each group x the number of replications. A test of the difference between two loudspeakers for a certain program has the same MS term but divided by the product of the number of subjects within each group x the number of replications.

5.3.4 "Split-plot" design versus "repeated measurements" design

The choice between a design with "repeated measurements" in all variables (as described in chapters 2 - 4, 5.1, and 5.2) and a "split-plot" design with "non-repeated measurements" in one or more variables (5.3) could be discussed at some length. In general the "repeated measurements" type of design should be preferred in listening tests due to its higher sensitivity and less complexity. There may be cases, however, where a "repeated measurements" design involves too much time and work for each single subject, and then a "split-plot" design of some type may be a solution. The "split-plot" design generally requires more subjects to attain enough sensitivity in the "non-repeated measurement" variable(s).

6 RELIABILITY

There are several ways of checking the reliability of the rating data from a listening test. The procedures suggested here utilize easily available information from the ANOVA, which makes possible estimates of reliability both for each subject individually (intra-individual reliability) and for all subjects together (inter-individual reliability).

6.1 Intra-individual reliability

Intra-individual reliability refers to the consistency of the ratings within each individual. A minimum of two ratings per each of the loudspeaker x program combinations (or other stimulus conditions) is necessary in order to estimate this consistency.

6.1.1 Visual inspection

A first simple check is made simply by visual inspection of the data matrix for the individual in question. This was illustrated in 2.1 in connection with the individual data matrix for subject S, see Table I. The three ratings in each cell of this matrix vary very little among themselves, which indicates a high reliability. This can also be seen in the data matrix for the same subject when the sound level was included as a variable, see Table VI. Other individual data matrices can be seen within Table II for subjects T, U, and V whose data also display a good reliability (however, not as high as for subject S).

6.1.2 Statistical significance

If ANOVA and *F* tests are performed, a significant *F* test for the loudspeakers may be taken as an indication of satisfactory reliability in the ratings concerning loudspeakers. The reason is that there is a very low probability of getting a significant *F* test solely by chance (see 3.1 and 7.1) - as, for instance, if the subject's ratings were made on random basis.

On the other hand a non-significant *F* test does not in itself imply irreliability. It may simply mean that the loudspeakers are about equally good, and the subject's ratings may be quite reliable/consistent as indicated by the $MS_{\text{within cell}}$ described below.

6.1.3 Variance "within cells"

The "within cell" MS obtained in the ANOVA on individual data reflects the variance of the ratings within all cells of the respective data matrix, that is, how much the subject varies in his ratings of the same stimulus at different occasions in the test - in other words, his "error variance" (see 3.1).

The "within cell" MS may therefore be taken as an indication of intra-individual reliability. The lower this value is, the better the reliability. For the data of subject S in Table I it was 0.28 as seen in Table IV, and for the same subject's data in Table VI it was 0.51 as seen in Table VII, (cell = loudspeaker x program combination in Table I, loudspeaker x program x sound level combination in Table VI). For subjects T, U, and V (in Table II) the "within cell" MS was 0.85, 0.65 and 1.25 respectively.

Where an acceptable upper limit for the "within cell" variance should be set may be discussed at some length. Considering the characteristics of the 10 - 0 "true-to-nature" scale and using information from other subjects and from other listening tests it is suggested that 1.50 may be set as an approximate upper limit. However, this should not be taken in an absolute sense and has to be considered in connection with other possibilities described in following sections.

6.1.4 Reliability of mean ratings

The procedure described here is taken from Winer (1962 p. 124, or 1971, p. 283) but applied to data within individual here (see also 6.2.3). It provides an estimate of the reliability of the mean of the ratings made for each of the stimuli (for instance, each loudspeaker x program combination).

For individual data in a matrix like Table I (subject S) and the corresponding ANOVA in Table IV this reliability index ($\underline{r}_w$, w for within) is computed as follows:

(8) $$\underline{r}_w = 1 - \frac{MS_{\text{within cell}}}{(SS_L + SS_P + SS_{LxP}) \;/\; (df_L + df_P + df_{LxP})}$$

Applied to the data in Table IV we obtain:

$$\underline{r}_w = 1 - \frac{0.28}{(148.93+5.43+35.23) \;/\; (3+4+12)} = 0.97$$

An interpretation of this reliability (adapted from Winer) would be that, if the listening test was repeated with this subject under the same conditions (including the subject's characteristics), the correlation between the mean ratings for each loudspeaker x program combination in the two listening tests would be approximately 0.97. For this subject, then, the correlation would be almost perfectly positive. For subjects T, U, and V $\underline{r}_w$ is 0.89, 0.93 and 0.79, respectively.

If more variables are included in the test, there will be more terms included in the denominator of formula (8). For example, to compute $\underline{r}_w$ for the data in Table VI the denominator will include SS and df for seven terms (see the corresponding ANOVA in Table VII): loudspeakers, sound levels, programs, and all possible interactions between these variables. For these data the reliability index will be:

$$\underline{r}_w = 1- \frac{0.51}{(262.43+7.01+10.08+1.49+45.12+8.95+14.38)/(3+1+4+3+12+4+12)}$$

$= 0.94$

It is obvious that the lower the "within cell" MS (the "error variance") is relative to the variation associated with the variables in the denominator, the higher $\underline{r}_w$ gets. If there is no "error variance" at all ($MS_{\text{within cell}}=0$) $\underline{r}_w$ will be 1.00. On the other hand, if $MS_{\text{within cell}}$ is as big as the expression in the denominator, $\underline{r}_w$ will be zero. (If $\underline{r}_w$ happens to be negative,it is set =0.)

Where to set an acceptable lower limit for $\underline{r}_w$ in listening tests is, of course, subject to discussion. On the basis of data from several listening tests it is suggested that 0.50 may be used as an approximate lower limit.

However, $\underline{r}_w$ cannot be the only basis for a decision about acceptable reliability or not. In fact $\underline{r}_w$ can sometimes be misleading. Suppose that in a listening test the loudspeakers in question are about equally good and thus get about the same ratings on the 10 - 0 scale. This means that the variation associated with the loudspeakers (and possibly also with the programs and the interaction loudspeakers x programs) will be low, and the denominator in formula (8) will be small. And so, even though the $MS_{within\ cell}$ happens to be low, $\underline{r}_w$ may be misleadingly low for the purpose of judging the subject's reliability. It is obvious that $\underline{r}_w$ is dependent on the specific context of loudspeakers and programs in the test.

Thus $\underline{r}_w$ should be considered together with $MS_{within\ cell}$. If $\underline{r}_w$ is $\geq .70$ and $MS_{within\ cell}$ is ≤ 1.50, the reliability may be considered as good. If $\underline{r}_w$ is medium high (say, .40 - .60) but $MS_{within\ cell}$ is ≤ 1.50, the reliability may still be satisfactory. However, if $\underline{r}_w$ is low and $MS_{within\ cell}$ is ≥ 1.50, the reliability is probably not satisfactory. In such cases the data should be scrutinized to find the possible reason(s) for the low reliability.

The meaning of $\underline{r}_w$ can be further analysed by computing another index as described in the following section.

6.1.5 Proportion of variance accounted for

Still another way of judging the reliability would be to consider how much of the variance in the subject's data is accounted for by the different sources of variation in the listening test. This is in itself interesting information, regardless of its use for estimations of reliability.

To compute the proportion of variance accounted for (designated by $\underline{\omega}^2$) by different sources in an individual data matrix as Table I the following formulas are used:

$$(9)\qquad \underline{\omega}_L^2 = \frac{SS_L - (df_L \times MS_{w.cell})}{SS_{total} + MS_{w.cell}}$$

$$(10)\qquad \underline{\omega}_P^2 = \frac{SS_P - (df_P \times MS_{w.cell})}{SS_{total} + MS_{w.cell}}$$

$$(11)\qquad \underline{\omega}_{LxP}^2 = \frac{SS_{LxP} - (df_{LxP} \times MS_{w.cell})}{SS_{total} + MS_{w.cell}}$$

The values to enter in these formulas are easily found in the corresponding ANOVA summary table, in this case Table IV. The proportion of variance accounted for by the loudspeaker variable is thus:

$$\underline{\omega}_L^2 = \frac{148.93 - (3 \times 0.28)}{200.92 + 0.28} = 0.74$$

For the program variable:

$$\underline{\omega}_P^2 = \frac{5.43 - (4 \times 0.28)}{200.92 + 0.28} = 0.02$$

Finally for the interaction loudspeakers x programs:

$$\underline{\omega}_{LxP}^2 = \frac{35.23 - (12 \times 0.28)}{200.92 + 0.28} = 0.16$$

Thus 74% of the variance in the data of subject S is accounted for by differences between the loudspeakers, 2% by differences between the programs and 16% by the interaction loudspeakers x programs. The remaining 8% represents the "error variance" estimated by $MS_{\text{within cell}}$.

If more variables are included in the listening test, the variance accounted for by these variables and all interactions can be estimated in analogy with formulas (9) - (11). For example, if the analogous computations are performed with regard to the data in Table VI, there will be seven computations, one for each of the seven first terms in Table VII in which the relevant values are found. In that case the variance accounted for by the differences between loudspeakers was about 67% and in total the seven sources accounted for about 84%, leaving 16% due to "error variance".

The rationale behind $\underline{\omega}^2$ and more computational formulas may be found in Hays (1973), Kirk (1968) and Vaughan & Corballis (1969).

The meaning of the $\underline{r}_w$ index in 6.1.4 may be further clarified by computing the $\underline{\omega}^2$ indices described here. For subject S $\underline{r}_w$ was as high as 0.97 due to his low "error variance" (numerator in formula 8) relative to the variance associated with differences between loudspeakers, differences between programs, and interaction loudspeakers x programs (denominator in formula 8). Among the last-mentioned sources it was now clarified that the variance associated with differences between loudspeakers was by far the biggest (74%, see above). In another example, however, it could happen that it is the differences between programs or the interaction between loudspeakers and programs which account for most of the "true" variance. $\underline{r}_w$ in itself does not make any distinction between these different alternatives (or still other alternatives). Therefore it may be elucidating to supplement $\underline{r}_w$ with the $\underline{\omega}^2$ indices described here. (For instance, if the proportion of variance accounted for by differences between programs is high, it may be an indication that the reliability of the ratings rather refer to differences between the programs than to differences between the loudspeakers.)

The amount of variance accounted for by different sources is apparently dependent on the context of stimuli. If there are obvious differences between loudspeakers (as in our example), the variance accounted for by differences between loudspeakers may be high. On the other hand, if the loudspeakers are about equally good, this variance may become quite low. The same line of reasoning applies, with due modifications, to programs, and to interaction between loudspeakers and programs. Therefore it is not very useful to decide upon acceptable limits for the respective proportions of variance. The $\underline{\omega}^2$ indices have to be considered in connection with $\underline{r}_w$ and $\overline{MS}_{\text{within cell}}$ as described in 6.1.4.

6.1.6 Conclusions regarding intra-individual reliability

Several ways of estimating the intra-individual reliability exist. The necessary data are found in the corresponding ANOVA summary table, and the computations are easily done. It seems that a combined consideration of all procedures described in 6.1.1 - 6.1.5 is to be recommended. In general these procedures should all lead to similar conclusions regarding the reliability of the subject's ratings.

If a subject's reliability appears to be unsatisfactorily low, a possible way to increase his reliability would be to give him more practice and/or to let him do more ratings per each case (for instance, four ratings per loudspeaker x program combination instead of three as illustrated here). If he still shows unsatisfactory reliability in his ratings, his data may be discarded. However, if the tendencies in his data seem to be in line with those from other, reliable subjects, it probably does not matter whether his data are included or not. Of course, the more subjects there are, the less influence an irreliable subject will have on the results.

6.2 Inter-individual reliability

Inter-individual reliability refers to the agreement between the ratings from different subjects.

(With regard to the "true-to-nature" scale there might sometimes be some problems concerning the inter-individual reliability. The reason is that the "true-to-nature" scale is probably a multidimensional one, that is, a composite of several separate perceptual dimensions such as "Distinctness", "Brightness", "Fullness" etc. It is possible that different subjects give different weights to different perceptual dimensions, when they judge the "true-to-nature" character of the reproductions in question. Consequently it may happen that, although the intra-individual reliability is high for each single subject, the inter-individual reliability is not as high, because different subjects use different judgment principles. However, if both intra- and inter-individual reliabilities are high, this indicates that the subjects are consistent within themselves and also agree between themselves.

This discussion does not apply, of course, to rating scales of unidimensional character.)

6.2.1 Visual inspection

A rough estimate of inter-individual reliability is gained by simple inspection and comparison of individual data matrices. It is evident, for example, that there is in general a good agreement between the ratings of subjects S, T, U, and V in the group data matrix shown in Table II.

6.2.2 Statistical significance

A significant F test for loudspeakers may indicate satisfactory reliability for the same reason as discussed in 6.1.2. On the other hand a non-significant F test does not in itself imply irreliability, see 6.1.2.

6.2.3 Reliability of mean ratings

This alternative is adapted from the discussion in Winer (1962 p. 124 or 1971 p. 283). For the group of four subjects given in Table II and with corresponding ANOVA in Table V a reliability index (r_b, b for between) for the agreement between subjects is computed as follows:

$$(12)\quad r_b = 1 - \frac{(SS_{LxS}+SS_{PxS}+SS_{LxPxS}+SS_{w.cell})/(df_{LxS}+df_{PxS}+df_{LxPxS}+df_{w.cell})}{(SS_L+SS_P+SS_{LxP})/(df_L+df_P+df_{LxP})}$$

(Note: SS_S is not included in the numerator, since it only reflects differences between subjects with regard to their "mean position" on the 10 - 0 scale, see comments in 3.2.)

As applied to the data in Table V, this reliability index is:

$$\underline{r}_b = 1 - \frac{(17.34+42.86+45.98+121.33)/(9+12+36+160)}{(465.75+11.94+47.36)/(3+4+12)} = 0.96$$

Apparently there is a high agreement between the four subjects. The result may be interpreted as follows: If the listening test was repeated with another random sample of four subjects from the same population (which was a society for hi-fi enthusiasts), the correlation between the mean ratings for each loudspeaker x program combination in the two tests would be approximately 0.96.

Due to the above considerations concerning the "composite" character of the "true-to-nature" scale it is less meaningful to set up a lower limit for acceptable <u>inter</u>-individual reliability. If $\underline{r}_b$ is as high as here, there is no problem. If $\underline{r}_b$ turns out to be less than, say, 0.50, it may be wise to scrutinize the data to see the reason(s), and/or to increase the number of subjects. It may also happen that $\underline{r}_b$ becomes low for the same reasons that may lead to a misleadingly low $\underline{r}_w$ (as discussed in 6.1.4).

6.2.4 Proportion of variance accounted for

For the <u>fixed model</u> the proportion of variance accounted for by the different loudspeakers, different programs and the loudspeaker x program interaction can be estimated by using formulas (9) - (11), respectively. The values to put into the formulas are found in the corresponding ANOVA summary table (in this case thus Table V). The proportion of variance accounted for by different subjects and by interactions involving subjects can be estimated by formulas analogous to those in formulas (9) - (11) - note how formulas (9) - (11) are built up and use the same principles as regards the subjects and the interactions involving subjects. Using data from Table V and assuming the fixed model, computations show that the differences between loudspeakers account for 54% of the variance, the differences between programs for only 1%, the interaction loudspeakers x programs for 4.4%, while the differences between subjects and the interactions involving subjects together account for 19.4%.

For the <u>mixed model</u> (the subjects are randomly sampled from a population) the corresponding computations are somewhat more complex. Formulas and procedures may be found in Vaughan & Corballis (1969, their Table 2, model <u>aBC</u>, where <u>a</u> corresponds to subjects, <u>B</u> to loudspeakers and <u>C</u> to programs). Using these formulas on the data in Table V (assuming mixed model) approximately the same results (percentages) are obtained as for the fixed model above. (Although such an agreement cannot be generally expected to occur, the simpler formulas for the fixed model may in many cases be enough to get at least a crude estimation of the corresponding results for the mixed model.)

6.2.5 Conclusions regarding inter-individual reliability

Due to the "composite" character of the "true-to-nature" scale and the possibility that different subjects use different judgment principles, it is less meaningful to decide upon acceptable lower limits for inter-individual reliability than for intra-individual reliability. However, the following considerations seem relevant:

1) If $\underline{r}_b$ is high, this points to a good agreement between the subjects in their ratings.

2) If $\underline{r}_b$ is only medium high or even lower, this should be a signal to compare the data from different individuals to see the possible reason(s). If there are obvious differences between different subjects in their ratings, and if these subjects each show a satisfactory intra-individual reliability, this may indicate different judgment principles for different subjects. To clarify the situation it is probably wise to increase the number of subjects, and/or to supplement the "true-to-nature" ratings with ratings in separate, unidimensional scales of relevance for the test.

3) However, if both inter- and intra-individual reliabilities seem to be unsatisfactory (that is, the subjects do not agree between themselves, nor are they consistent within themselves), this may indicate much "error" in the data. Steps should be taken to increase reliability by adding more ratings and/or more subjects (possibly discarding data from irreliable subjects).

With a truly unidimensional rating scale the situation is simpler, since the considerations under point 2 above become mainly irrelevant.

If estimations of the proportion of variance accounted for by different sources have been made (see 6.2.4), these data may also be useful for judging the inter-individual reliability. The more variance accounted for by loudspeakers, and/or programs, and/or loudspeaker x program interaction, and the less variance attributable to differences between subjects and/or interactions involving subjects, the better the inter-individual reliability. The relative amount of variance accounted for by differences between loudspeakers, differences between programs, and the interaction between loudspeakers and programs is, of course, interesting information in itself and may clarify the meaning of the $\underline{r}_b$ value in the same way as discussed for $\underline{r}_w$ in 6.1.5.

7 SOME CRITICAL ISSUES IN SIGNIFICANCE TESTING

The use of significance tests always involves certain error risks, commonly known as type I and type II errors. Further all significance tests are based upon a series of assumptions.

7.1 Error risks, significance level, "power"

As seen in the earlier examples, the result of a significance test is stated in probability terms. For instance, if an *F* test for the loudspeaker variable is significant at .01 level (also denoted 1% level), this means that the probability of getting the observed differences between the loudspeakers by chance alone is less than .01. Since this probability is very low, one generally concludes that there are "real" or "true" differences between the loudspeakers. The risk that this conclusion is wrong is at most .01 and defines the so-called "type I error" (the risk of concluding that the loudspeakers are different although they are not).

The risk of a "type I error" is regulated by the choice of significance level. The significance levels mostly used in behavioral and social science are .05 or .01 level. These are conventions, however, and ideally the choice of significance level should be considered in connection with the risk of making a "type II error". The "type II error" is the risk of concluding that there are no differences between loudspeakers, although there are in fact "real" differences between them. Obviously it is important in listening tests to avoid a "type II error". Therefore the statistical test should have a good "power" or "sensitivity" to detect really existing differences. The "power" of a statistical test is formally defined as 1- the probability for type II error.

Unfortunately an exact calculation of the probability of "type II error", and thus also of "power", is not as easy as for "type I error". In general the following rules apply. The lower the "type I error" is made, the higher the "type II error" will be (in other words, the lower the significance level is set, the bigger is the risk of not detecting "real" differences). Conversely, the smaller "error variance" there is in the data, and the bigger "real" differences there are between loudspeakers, the lower the "type II error" will be (= the higher "power" the statistical test will have). Further the "power" of a one-tailed test (= test for difference in a certain direction, for example, if one loudspeaker is better than another, see 4.1) is higher than for a two-tailed test (= test for difference regardless of direction).

For a given significance level (= probability of "type I error") the "power" should be as high as possible. The investigator can increase the "power" by decreasing the error variance, which is achieved by increasing the number of replications (the number of ratings under the same conditions, for instance, the number of ratings each subject makes per each loudspeaker x program combination) and/or the number of subjects. Some guidelines for achieving satisfactory "power" are given below.

7.2 Computing "power",number of ratings, number of subjects

There are various methods for computing the "power" of statistical tests, based on certain assumptions (Kirk, 1968; Hays, 1973). One of these methods gives a value for the minimum "power", if the largest difference among the actual means equals the standard deviation of the "errors" (= the square root of the error variance) times a multiplicative factor to be decided by the investigator (Kirk, 1968, p. 109). For instance, it is possible to compute the minimum "power", if the largest difference occurring among the mean ratings for a number of loudspeakers (such as the means in the bottom margin of Table II) would equal the size of the standard deviation of "errors" (multiplicative factor = 1.00), or would be equal to twice this standard deviation (multiplicative factor = 2.00), etc. An advantage with this method is that it does not require a direct numerical estimate of the error standard deviation but "only" an expression of differences among the means in units of the error standard deviation. If the computations would show that the minimum "power" would be unacceptably low, it is also possible to compute how much the number of replications and/or subjects should be increased to attain acceptable "power".

By using this method it is thus possible to compute _in advance_ (before the listening test is started) how many replications and/or subjects are necessary to achieve satisfactory "power". Adapting the method to apply to group data (as described in 2.2) and to the use of a fixed or a mixed model in the ANOVA on group data (see 3.2), the author made extensive computations concerning the "power" of the significance test on loudspeakers in listening tests with varying numbers of loudspeakers, programs, subjects, and replications. The computations require too much space to be shown here. Only the main conclusions for practical use are given in the following.

Computations have been made for the cases (_a_) that the largest occurring difference between the loudspeaker means is _equal to the error standard deviation_ (multiplicative factor = 1.0), and (_b_) that the largest occurring difference between the loudspeaker means is _equal to twice the error standard deviation_ (multiplicative factor = 2.0). "Satisfactory power" was defined as "power" $\geq$ .90, and the significance level was set to .05 or .01.

For case (a) above, the computations show that it is necessary to have at least eight subjects, doing at least two ratings per each loudspeaker x program combination to attain the desired "power". This case represents a rather "hard" criterion, that is, a largest occuring difference between the loudspeaker means not bigger than the error standard deviation should result in a significant *F* test for the loudspeakers.

For case (b), with the less severe criterion that a largest occuring difference between the loudspeaker means corresponding to twice the size of the error standard deviation should result in a significant *F* test, it is necessary to have at least four subjects, doing at least two ratings per each loudspeaker x program combination.

The above conclusions hold for listening tests with three to ten loudspeakers reproducing three to six programs. This probably covers the range of loudspeakers and programs used in most listening tests. If more than ten loudspeakers are used, it may be preferable to increase the number of subjects. The conclusions apply to both the fixed and the mixed model. If for some reason the subjects can do only one rating per each loudspeaker x program combination, it is suggested that the number of subjects in the above recommendations is doubled.

The investigator must decide for himself which of cases (*a*) or (*b*) above is most relevant with regard to his purposes. Of course, he can choose a criterion in between those two by simple interpolation (for instance, setting the multiplicative factor = 1.5 and using at least six subjects doing at least two ratings per loudspeaker x program combination). In the author's opinion, however, a less severe criterion than that represented by case (*b*) should not be permitted. Thus four subjects doing at least two ratings per loudspeaker x program combination should be considered as an absolute minimum.

The above recommendations are based on statistical considerations. Of course there may be situations in which other (non-statistical) factors may be important for deciding the number of subjects etc. With continued experience of listening tests an investigator probably develops a certain "intuitive feeling" for which number of loudspeakers, programs, subjects etc are needed to get the necessary precision and relevance in his test.

Since the above method makes use of the error standard deviation as a kind of unit for specifying "true" differences between loudspeakers, it is therefore desirable that the error variance is as low as possible - in other words that the subjects have a satisfactory reliability in their ratings. The recommendations concerning number of subjects and ratings above may therefore be supplemented by the recommendations about satisfactory reliability as discussed in chapter 6.

7.3 Assumptions underlying significance tests

Underlying ANOVA and accompanying statistical tests there are certain general assumptions as well as certain specific assumptions for various designs. These assumptions are listed in most texts on statistics and design (Kirk, 1968, 1972; Hays, 1973; Winer 1971). The general assumptions refer to such things as normal distribution and independence of "errors", "homogeneous error variances" and the like, while assumptions specific for various designs may deal with, for instance, the symmetry of variance-covariance matrices in designs with "repeated measurements".

If the assumptions are violated, this will affect the significance level and the "power" of the statistical tests, that is, the probabilities associated with the significance level and the "power" will not be exact but more or less approximate. It has been shown mathematically and by means of simulation experiments (see the references mentioned above) that violation of certain assumptions has very small effects on the significance level, while other assumptions may be more critical. Thus violations of the "normal distribution" assumption and of the "homogeneous error variances" assumption in general have very little effects as regards the significance level (for the last-mentioned assumption, however, it is important that there is the same number of ratings per each loudspeaker x program combination, as in Tables I or II.) The effects on "power" are generally harder to estimate. There are various ways of testing the validity of the assumptions and possibly transforming the data to better fit the assumptions. From a practical point of view, however, it is often doubtful whether these procedures are worthwhile.

With regard to listening tests two assumptions seem especially important:

1) The assumption concerning independent "errors" is generally very important. If it is violated, the probability statements related to the significance level may be very much in error. The detailed meaning of this assumption requires too much space to be explicated here. Suffice it to say that this assumption can be considered as fulfilled in listening tests if the presentation order of loudspeaker x program combinations is randomized. The randomization should be different for different subjects. For instance, in our example using four loudspeakers and five programs the order of the resulting twenty loudspeaker x program combinations should be randomized differently for each subject. Since there are three ratings per combination, there are in fact sixty presentations requiring three different randomizations of the twenty combinations for each subject. Of course, the randomization can be made with reference to all sixty presentations together requiring only one (but more extensive) randomization per subject.

Generally this assumption also implies that the ratings of each subject is independent of each other subject's ratings. It is thus necessary to somehow control that the subjects do not communicate or otherwise influence each other in connection with the test (which sometimes can be difficult to control). A special problem is that one sometimes have two (or even more) subjects simultaneously listening in the test. Although the order of the loudspeaker x program combinations in itself may be randomized, it will of course be the same randomization for these subjects. In such a case it may be doubtful whether the assumption of independent errors is quite fulfilled, and a certain caution in the interpretation of the results is recommended.

2) When doing ANOVA on group data(as illustrated in Tables V and VIII) and using the mixed model for the analysis (that is, the subjects are randomly sampled from a population), there is an intricate assumption about equality and symmetry of certain variance-covariance matrices. If this assumption is not met, the actual significance level may be somewhat higher than the nominal level (it will be too "easy" to get significant results). There are certain ways of circumventing this difficulty, none of which is quite satisfactory (one simple way is to use a lower significance level than planned). Moreover, since the validity of the assumption is often hard to evaluate, the situation is problematic and considerations of non-statistical character may be helpful (see below; note also that this assumption is not required if a fixed model is used).

Since it is impossible to be quite certain that all assumptions are strictly fulfilled in a set of data, the probability statements associated with significance tests should be regarded as approximations. In practice the conclusions from statistical tests always have to be considered in combination with other information of relevance for the investigated problem, for instance, if they agree with earlier results, if they are reasonable with regard to what is known about the loudspeakers' physical characteristics, and the like.

A brief comment may finally be made to the question concerning appropriate statistics for different types of scales. In the introduction it was assumed that the 10 - 0 "true-to-nature" rating scale represents an interval scale. If this assumption is not fulfilled, there may be a discussion about the appropriateness of the statistical procedures described in chapters 1 - 6. In fact the question about using ANOVA and related procedures for data on "lower" types of scales (for instance, ordinal scales) is a very much debated and unsettled problem. A review of this discussion may be found in Kirk (1972). From a practical point of view there are, however, no satisfactory alternative statistical procedures for the type of designs discussed in this report, so there is in fact no choice.

8 COMPUTATIONS, COMPUTER PROGRAMS, TABLES

The computations involved in ANOVA are conveniently performed by aid of a computer. There are many available computer programs for ANOVA, for instance, within "Biomedical Computer Programs" (BMD), "Statistical Package for the Social Sciences" (SPSS), "IBM Scientific Subroutine Package" (SSP), "International Mathematical & Statistical Libraries" (IMSL), and others. New programs appear now and then. Ask for a convenient program at your nearest program library.

If a computer is not easily available, the computations may be performed by means of electronic calculators (preferably equipped with memory functions). Computational schemas for ANOVA in various designs appear in many textbooks on statistics and experimental design, for instance, in Kirk (1968), Hays (1973), and Winer (1971). A lot of computational schemas for ANOVA and many other statistical areas appear in "Computational Handbook of Statistics" by Bruning & Kintz (1977). This book also includes lists of short computer programs for various applications of ANOVA (written in FORTRAN IV), which can be used if no other computer program is available.

Unfortunately there are certain differences between different programs and different books with regard to the terminology. Several examples in the above-mentioned books, and the examples used in this paper, provide possibilities to check that the correct program (or computational schema) is used. In the book by Bruning & Kintz (1977), there is also a table over various designs and the associated terminology as used in many textbooks on statistics and design. The use of a computer program facilitates the computations for ANOVA. The computations involved in tests for specific comparisons (chapter 4) and in estimation of reliabilities (chapter 6) are easier to do with your own calculator. It should be noted that the results of computations may be slightly different from different computer programs, depending on which precision is used in the respective programs or computers. The computations in this paper for ANOVA and for formulas (1) - (12) were made mostly using two decimals.

Tables for the distributions of $\underline{F}$, $\underline{t}$, and $\underline{q}$ test statistics are found in most textbooks on statistics and design, for instance, in Kirk (1968), Winer (1971), and Bruning & Kintz (1977).

9 PRESENTATION OF RESULTS

Two cases will be distinguished:

1) Only descriptive statistics are used (9.1)

2) The statistical data treatment involves ANOVA and accompanying statistical tests (9.2)

9.1 Only descriptive statistics

If the data treatment is limited to descriptive statistics as described in chapter 2, the main results can be presented in a condensed group data matrix like that in Table III. The corresponding data can also be displayed in graphical form, for instance, simply like this:

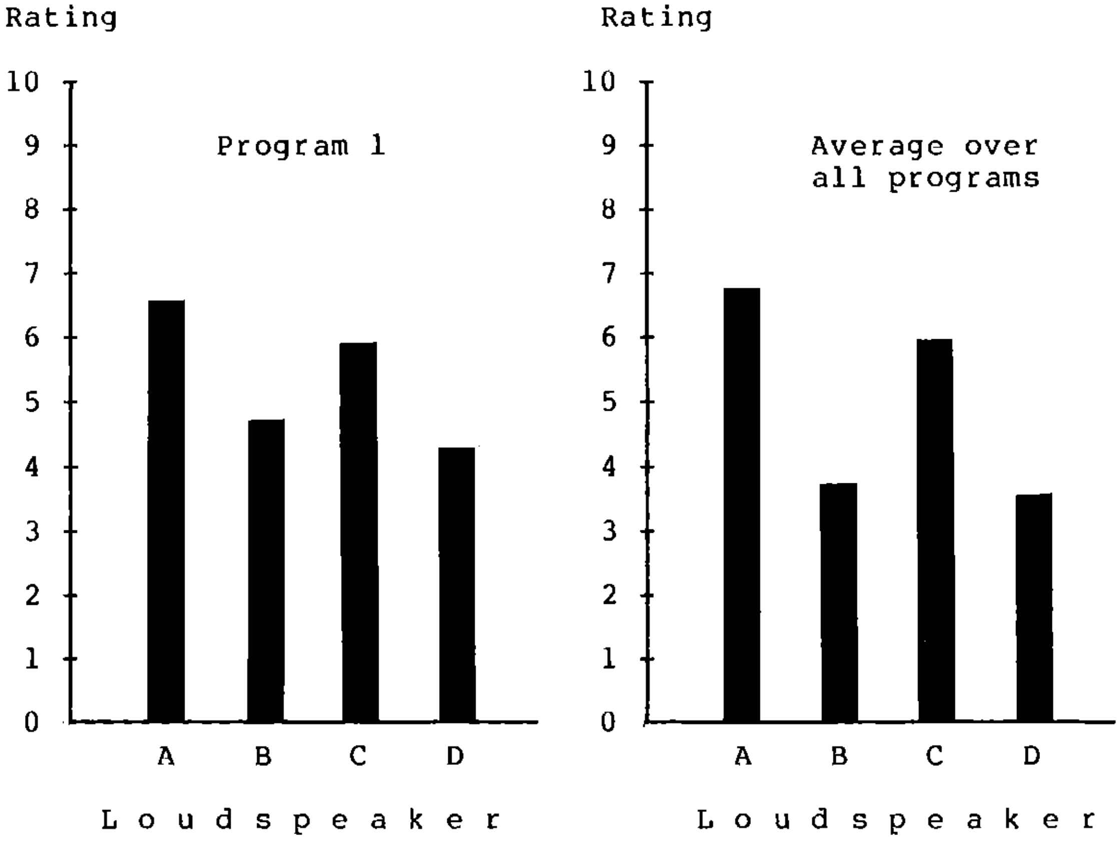


FIGURE 2. Example of graphical display of group means for loudspeakers, data from Table II. (Only means at program 1 and in average over all five programs are included here. Of course, more sophisticated figures can be made.)

However, it is also strongly recommended to present a complete group data matrix of the type shown in Table II. This allows the reader to look at the dispersions around the means, to get an impression of the reliability of the data, and to make further computations and analyses on the data, if he wants to do so. The investigator's own conclusions from his data analysis should, of course, be clearly stated.

9.2 Data treatment involving ANOVA

Group data matrices and graphs may be presented as described under 9.1 above. (Since information about the dispersions of the ratings will appear in the results from ANOVA, the contents of the group data matrix of type Table II may be reduced by omitting the replicated ratings of each single subject and only including the individual means and the group mean within each loudspeaker x program combination. For example, each loudspeaker x program combination in Table II would thus only contain the values denoted M_S, M_T, M_U, M_V (individual means), and M_g (group mean). The means for the loudspeakers and for the programs in the margins are, of course, retained.) From the *ANOVA* a summary table like that in Table V should be presented. It should be clearly stated whether a fixed model or a mixed model is used, and how the *F* values are computed. The conclusions from the ANOVA and *F* tests should be presented and related to the data in the group matrix.

A significant result for the loudspeaker variable may be further investigated by means of *tests for specific comparisons.* If so, the test(s) used should be stated, as well as the conclusions from them.

If *significant interactions* occur, they should be interpreted by means of the group data matrix and possibly by tests for specific comparisons. It is especially important to note the meaning of significant loudspeakers x programs interaction, loudspeakers x subjects interaction, and loudspeakers x programs x subjects interaction, since such interactions may present highly important information concerning the loudspeakers in addition to the result from the *F* test in the loudspeaker variable (either this is significant or not).

If the listening test includes more variables than loudspeakers and programs, and/or if a "split-plot" type design is used, the above recommendations still apply, but the group data matrix and the ANOVA summary table will take somewhat other forms as described in chapter 5.

Data on *intra-individual and inter-individual reliability* should be given using one or more of the possibilities described in chapter 6.

10 CONCLUDING COMMENTS, ACKNOWLEDGEMENTS

The procedures described in this paper may provide aids for the investigator to understand the meaning of the data in his listening test. Although the paper is rather long, there are in fact several simplifications in the text, and alternative methods could be proposed at several points. The paper is designed for the statistically not so well-trained investigator. If there is an expert in statistics available he should be consulted, preferably already in the planning of the listening test.

Throughout this paper a rating scale of the "true-to-nature" type has been assumed. However, most of the statistics would apply equally well to other types of rating scales which may be used in listening tests.

Finally it should be emphasized that the statistical treatment should be an aid for the investigator to understand the empirical meaning of his data. Statistics for its own sake is uninteresting. The investigator should make full use of his expert knowledge concerning loudspeakers and related things and not let himself get lost in a wealth of statistical tests. As said earlier, the results are sometimes so obvious from visual inspection of data that a more advanced data treatment is not worthwhile. In other situations, however, a judicious use of statistical methods may clarify a complex situation and provide important information for the future work. It is no easy thing to attain the proper combination of empirical investigator and statistician in one person.

The author wants to express his gratitude to professor Paul Seeger for valuable discussions in many statistical questions, to Anna-Greta Seger and Lennart Persson for all work with typing and editing this paper, and to Björn Hagerman and Björn Lindström for check-reading the manuscript.

11 REFERENCES

Bruning, J.L. & Kintz, B.L. (1977) Computational Handbook of Statistics (second edition). Scott, Foresman and Company, Glenview, Illinois.

Gabrielsson, A. (1979) Dimension analyses of perceived sound quality of sound-reproducing systems. Scand J Psychol 20, 159-169.

Gabrielsson, A., Frykholm. S-Å. & Lindström, B. (1979) Assessment of perceived sound quality in high-fidelity sound-reproducing systems. Reports from Technical Audiology, Karolinska Institutet, Stockholm, No. 93.

Gabrielsson, A., Rosenberg, U. & Sjögren, H. (1974) Judgments and dimension analyses of perceived sound quality of sound-reproducing systems. J Acoust Soc Amer 55, 854-861.

Gabrielsson, A. & Sjögren, H. (1976) Preferred listening levels and perceived sound quality at different sound levels in high fidelity sound reproduction. Reports from Technical Audiology, Karolinska Institutet, Stockholm, No. 82.

Gabrielsson, A. & Sjögren, H. (1979) Perceived sound quality of sound-reproducing systems. J Acoust Soc Amer 65, 1019-1033.

Hays, W.L. (1973) Statistics for the Social Sciences (second edition). Holt, Rinehart & Winston, New York.

IEC-Publication 268-13: Listening tests on loudspeakers (to be published)

Kirk, R.E. (1968) Experimental Design: Procedures for the Behavioral Sciences. Brooks & Cole, Belmont, California.

Kirk, R.E. (1972) Statistical Issues: A Reader for the Behavioral Sciences. Brooks & Cole, Belmont, California.

Vaughan, G.M. & Corballis, M.C. (1969) Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. Psychol Bull 72, 204-213.

Winer, B.J. (1962, 1971) Statistical Principles in Experimental Design (first and second edition). McGrawHill, New York.

TECHNICAL AUDIOLOGY
Karolinska institutet
KTH
S-100 44 STOCKHOLM

Report TA No. 92
APPENDIX

STATISTICAL TREATMENT OF DATA FROM LISTENING TESTS ON SOUND-REPRODUCING SYSTEMS
(Tables, formulas, figures)

Alf Gabrielsson

	Loudspeaker				
Program	A	B	C	D	Means for programs
1	7 6 7 M=6.7	5 5 5 5.0	6 7 5 6.0	3 3 4 3.3	5.3
2	6 6 7 6.3	3 3 4 3.3	5 5 7 5.7	3 3 4 3.3	4.7
3	7 8 8 7.7	2 2 2 2.0	7 7 7 7.0	3 3 3 3.0	4.9
4	7 8 8 7.7	3 3 3 3.0	8 7 8 7.7	3 3 3 3.0	5.3
5	6 7 6 6.3	5 5 4 4.7	6 6 6 6.0	5 5 5 5.0	5.5
Means for loud-speakers	6.9	3.6	6.5	3.5	

TABLE I. Example of individual data matrix

LOUDSPEAKER

Program	Subject	A		B		C		D		Means for programs
1	S	7 6 7	M_S=6.7	5 5 5	5.0	6 7 5	6.0	3 3 4	3.3	
	T	5 7 4	M_T=5.3	4 4 3	3.7	5 5 5	5.0	3 3 3	3.0	
	U	6 7 7	M_U=6.7	4 5 4	4.3	5 7 7	6.3	5 5 5	5.0	
	V	8 7 7	M_V=7.3	7 6 5	6.0	6 6 7	6.3	7 4 5	5.3	
		M_g=6,5		4,8		5,9		4,2		5,3
2	S	6 6 7	6.3	3 3 4	3.3	5 5 7	5.7	3 3 4	3.3	
	T	8 7 8	7.7	3 2 3	2.7	4 4 7	5.0	4 3 2	3.0	
	U	6 8 8	7.3	4 3 3	3.3	6 7 8	7.0	3 3 3	3.0	
	V	7 8 7	7.3	4 4 5	4.3	8 7 4	6.3	4 4 4	4.0	
		7,2		3,4		6,0		3,3		5,0
3	S	7 8 8	7.7	2 2 2	2.0	7 7 7	7.0	3 3 3	3.0	
	T	5 4 5	4.7	3 2 2	2.3	5 3 3	3.7	1 1 2	1.3	
	U	7 7 9	7.7	4 4 3	3.7	7 7 9	7.7	3 4 4	3.7	
	V	9 6 6	7.0	5 4 6	5.0	8 4 6	6.0	3 3 3	3.0	
		6,8		3,3		6,1		2,8		4,7
4	S	7 8 8	7.7	3 3 3	3.0	8 7 8	7.7	3 3 3	3.0	
	T	6 8 7	7.0	4 3 3	3.3	5 4 5	4.7	3 1 2	2.0	
	U	6 7 8	7.0	3 4 3	3.3	6 6 7	6.3	1 4 2	2.3	
	V	8 8 7	7.7	4 6 5	5.0	8 9 8	8.3	7 7 4	6.0	
		7,3		3,7		6,8		3,3		5,3
5	S	6 7 6	6.3	5 5 4	4.7	6 6 6	6.0	5 5 5	5.0	
	T	6 5 7	6.0	4 3 4	3.7	5 6 3	4.7	4 3 2	3.0	
	U	6 5 6	5.7	3 3 3	3.0	4 5 5	4.7	5 4 4	4.3	
	V	9 7 7	7.7	4 4 5	4.3	7 7 5	6.3	6 6 5	5.7	
		6,4		3,9		5,4		4,5		5,1
Means for loudspeakers		6,8		3,8		6,0		3,6		

TABLE II. Example of group data matrix for subjects denoted S, T, U, and V.

		Loudspeaker				
		A	B	C	D	Means for programs
Program	1	6.5	4.8	5.9	4.2	5.3
	2	7.2	3.4	6.0	3.3	5.0
	3	6.8	3.3	6.1	2.8	4.7
	4	7.3	3.7	6.8	3.3	5.3
	5	6.4	3.9	5.4	4.5	5.1
Means for loudspeakers		6.8	3.8	6.0	3.6	

TABLE III. Condensed group data matrix.

Source of variation	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	*F*	*p*
Loudspeakers (L)	148.93	3	49.64	177.29	<.01
Programs (P)	5.43	4	1.36	4.86	<.01
L x P	35.23	12	2.94	10.50	<.01
Within cell	11.33	40	0.28		
Total	200.92	59			

TABLE IV. Example of summary table for ANOVA on individual data matrix.

Source of variation	SS	df	MS	*F*	*p*
Loudspeakers (L)	465.75	3	155.25	80.44	<.01
Programs (P)	11.94	4	2.99	0.84	-
Subjects (S)	105.25	3	35.08	46.16	<.01
L x P	47.36	12	3.95	3.09	<.01
L x S	17.34	9	1.93	2.54	=.01
P x S	42.86	12	3.57	4.70	<.01
L x P x S	45.98	36	1.28	1.68	>.01
Within cell	121.33	160	0.76		
Total	857.81	239			

TABLE V. Example of summary table for ANOVA on group data matrix (mixed model, case 2 below).

	Loudspeaker								
	A		B		C		D		
Level	High	Low	High	Low	High	Low	High	Low	Means for programs
Program 1	7	6	5	6	6	7	3	6	
	6	6	5	5	7	6	3	4	
	7	7	5	5	5	7	4	4	
	6.7	6.3	5.0	5.3	6.0	6.7	3.3	4.7	5.5
2	6	8	3	4	5	7	3	6	
	6	7	3	4	5	7	3	5	
	7	8	4	4	7	7	4	4	
	6.3	7.7	3.3	4.0	5.7	7.0	3.3	5.0	5.3
3	7	8	2	4	7	5	3	3	
	8	8	2	3	7	6	3	3	
	8	8	2	4	7	8	3	3	
	7.7	8.0	2.0	3.7	7.0	6.3	3.0	3.0	5.1
4	7	7	3	4	8	7	3	3	
	8	3	3	3	7	7	3	3	
	8	7	3	4	8	8	3	3	
	7.7	5.7	3.0	3.7	7.7	7.3	3.0	3.0	5.1
5	6	8	5	4	6	8	5	7	
	7	7	5	5	6	7	5	5	
	6	8	4	4	6	7	5	5	
	6.3	7.7	4.7	4.3	6.0	7.3	5.0	5.7	5.9
Means for levels	6.9	7.1	3.6	4.2	6.5	6.9	3.5	4.3	
Means for loud-speakers	7.0		3.9		6.7		3.9		

TABLE VI. Example of individual data matrix for listening test with three independent variables: loudspeakers, programs, and sound levels.

Source of variation	SS	df	MS	F	p
Loudspeakers (L)	262.43	3	87.48	171.53	<.01
Sound level (SL)	7.01	1	7.01	13.75	<.01
Programs (P)	10.08	4	2.52	4.94	<.01
L x SL	1.49	3	0.50	<1.0	-
L x P	45.12	12	3.76	7.37	<.01
SL x P	8.95	4	2.24	4.39	<.01
L x SL x P	14.38	12	1.20	2.35	>.01
Within cell	40.67	80	0.51		
Total	390.13	119			

TABLE VII. Summary table for ANOVA on data in Table VI.

Source of variation	SS	df	MS	F	p
Loudspeakers (L)		3			
Sound level (SL)		1			
Programs (P)		4			
Subjects (S)		3			
L x SL		3			
L x P		12			
L x S		9			
SL x P		4			
SL x S		3			
P x S		12			
L x SL x P		12			
L x SL x S		9			
L x P x S		36			
SL x P x S		12			
L x SL x P x S		36			
Within cell		320			
Total		479			

Table VIII. Schema for summary table in ANOVA on group data in listening test including loudspeakers, programs and sound levels as independent variables.

		Loudspeaker																			
		A					B					C					D				
		Program					Program					Program					Program				
Position	Subject	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	S																				
	T																				
	U																				
	V																				
2	W																				
	X																				
	Y																				
	Z																				

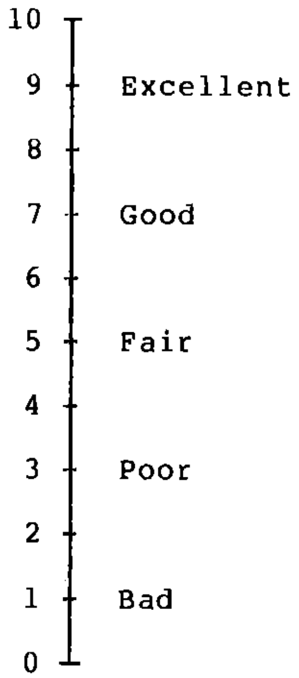
TABLE IX. Group data matrix for a "split-plot" design in a listening test with loudspeakers, programs, and positions as independent variables.

Source of variation	SS	df	MS	*F*	*p*
Between subjects:		7			
Positions (PO)		1			
Subj. within groups		6			
Within subjects:		472			
Loudspeakers (L)		3			
L x PO		3			
L x subj.w. groups		18			
Programs (P)		4			
P x PO		4			
P x subj.w. groups		24			
L x P		12			
PO x L x P		12			
L x P x subj.w. groups		72			
Within cell		320			
Total		479			

TABLE X. Summary table for ANOVA of the "split-plot" design in Table IX.

Source of variation	SS	df	MS	F	p
Between subjects:					
Programs (P)					
Subj. within groups					
Within subjects:					
Loudspeakers (L)					
L x P					
L x subj.w. groups					
Within cell					
Total					

TABLE XI. Summary table for ANOVA in "split-plot" design with "repeated measurements" as regards loudspeakers but "non-repeated measurements" as regards programs.



The number 10 denotes a reproduction which is perfectly true-to-nature.

The number 0 denotes a reproduction so bad that it has practically no similarity at all with the original performance.

FIGURE 1 "True-to-nature" rating scale.

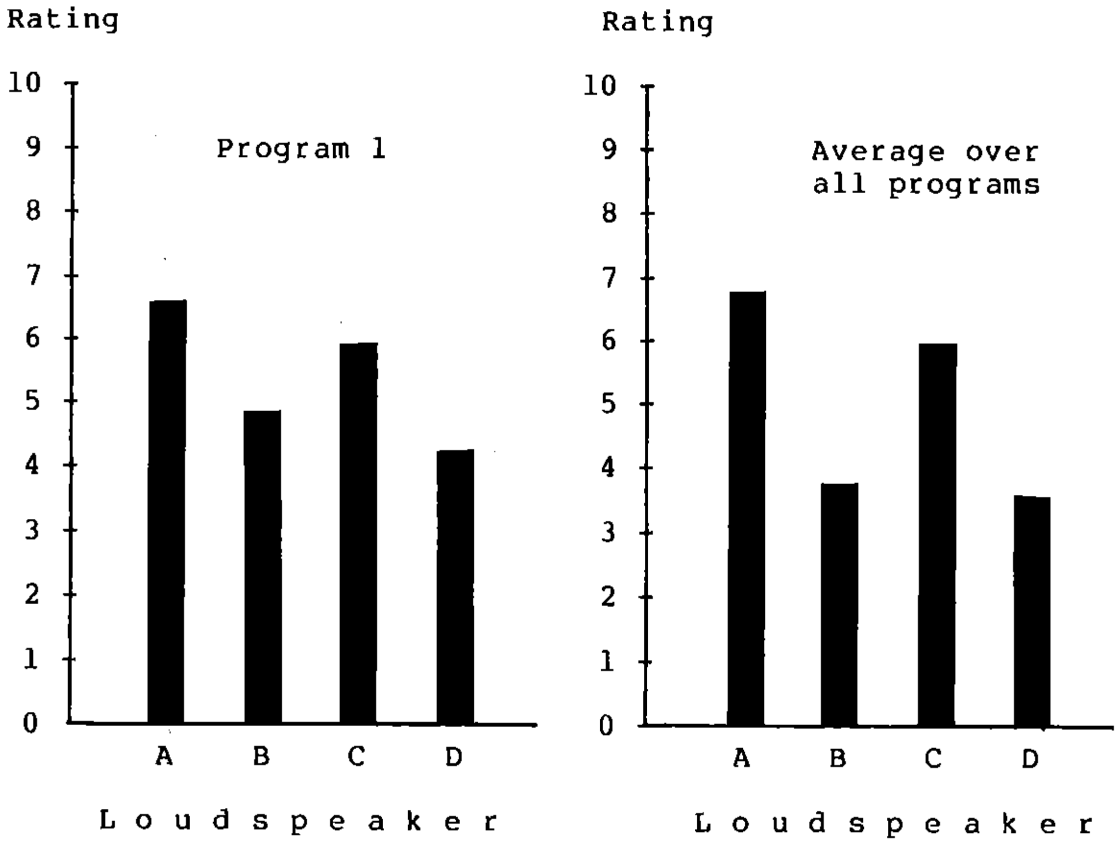


FIGURE 2. Example of graphical display of group means for loudspeakers, data from Table II. (Only means at program 1 and in average over all five programs are included here. Of course, more sophisticated figures can be made.)

FORMULAS

(1) $$\underline{t} = \frac{M_A - M_B}{\sqrt{2MS_{w.cell}/np}} = \frac{6.9 - 3.6}{\sqrt{(2 \times 0.28)/(3 \times 5)}} = 17.37$$

(2) $$\underline{t} = \frac{M_A - M_B}{\sqrt{2MS_{LxS}/nsp}} = \frac{6.8 - 3.8}{\sqrt{(2 \times 1.93)/(3 \times 4 \times 5)}} = 12.00$$

(3) $$\underline{t} = \frac{M_A - M_B}{\sqrt{2MS_{w.cell}/nsp}} = \frac{6.8 - 3.8}{\sqrt{(2 \times 0.76)/(3 \times 4 \times 5)}} = 18.75$$

(4) $$HSD = \underline{q}_{.01,40}\sqrt{MS_{w.cell}/np} = 4.70\sqrt{0.28/(3 \times 5)} = 0.66$$

(5) $$HSD = \underline{q}_{.01,9}\sqrt{MS_{LxS}/nsp} = 5.96\sqrt{1.93/(3 \times 4 \times 5)} = 1.07$$

(6) $$HSD = \underline{q}_{.01,160}\sqrt{MS_{w.cell}/nsp} = 4.50\sqrt{0.76/(3 \times 4 \times 5)} = 0.50$$

(7) $$\underline{t} = \frac{M_C - M_D}{\sqrt{2MS_{LxS}\ /\ ns}} = \frac{5.9 - 4.2}{\sqrt{(2 \times 1.93)/(3 \times 4)}} = 3.00$$

(8) $$\underline{r}_w = 1 - \frac{MS_{within\ cell}}{(SS_L + SS_P + SS_{LxP})\ /\ (df_L + df_P + df_{LxP})}$$

(9) $$\underline{\omega}_L^2 = \frac{SS_L - (df_L \times MS_{w.cell})}{SS_{total} + MS_{w.cell}}$$

(10) $$\underline{\omega}_P^2 = \frac{SS_P - (df_P \times MS_{w.cell})}{SS_{total} + MS_{w.cell}}$$

(11) $$\underline{\omega}_{LxP}^2 = \frac{SS_{LxP} - (df_{LxP} \times MS_{w.cell})}{SS_{total} + MS_{w.cell}}$$

(12) $$\underline{r}_b = 1 - \frac{(SS_{LxS}+SS_{PxS}+SS_{LxPxS}+SS_{w.cell})/(df_{LxS}+df_{PxS}+df_{LxPxS}+df_{w.cell})}{(SS_L+SS_P+SS_{LxP})/(df_L+df_P+df_{LxP})}$$

KAROLINSKA INSTITUTET
TEKNISK AUDIOLOGI (TA)

February, 1988.

Supplementary note to Report TA No. 92 (November, 1979) STATISTICAL TREATMENT OF DATA FROM LISTENING TESTS ON SOUND-REPRODUCING SYSTEMS (Alf Gabrielsson)

There are a few differences in terminology between this report (published in 1979) and the IEC report Publication 268-13: Listening tests on loudspeakers (Genève, 1985):

(1) The expression "true-to-nature scale" that appears in Report TA No. 92 (for instance, in Figure 1, page 2) should be replaced by "fidelity scale" as in the IEC report.

(2) The definitions of the number 10 and the number 0 in Figure 1 of the TA report should be replaced by the corresponding definitions given in the IEC report in Appendix A.

However, these changes in terminology do not in any way affect the description of the statistical procedures in the TA report.

In addition to the references given in the TA report further examples of the statistical treatment of data from listening tests, for the fidelity scale as well as for other rating scales, are given in the following papers:

Gabrielsson, A. & Lindström, B. (1985). Perceived sound quality of high-fidelity loudspeakers. Journal of the Audio Engineering Society, 33, 33-53.

Gabrielsson, A., Schenkman, B.N. & Hagerman, B. (1988) The effects of different frequency responses on sound quality judgments and speech intelligibility. Journal of Speech and Hearing Research, 31 (in press)

8