TECHNICAL AUDIOLOGY

Report TA No. 93
December 1979

ASSESSMENT OF PERCEIVED SOUND QUALITY IN HIGH FIDELITY SOUND-REPRODUCING SYSTEMS

Alf Gabrielsson, Sten-Åke Frykholm and Björn Lindström

Report TA No 93
Dec. 1979

Assessment of perceived sound quality in high fidelity sound-reproducing systems

Alf Gabrielsson, Sten-Åke Frykholm and Björn Lindström

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

Tel: 46-8-11 66 60

Karolinska Institutet
Technical Audiology
KTH, S-100 44 Stockholm

Report TA No. 93
December 1979

ASSESSMENT OF PERCEIVED SOUND QUALITY IN HIGH FIDELITY SOUND-REPRODUCING SYSTEMS

Alf Gabrielsson, Sten-Åke Frykholm and Björn Lindström

ABSTRACT

The sound reproduction of five selected high fidelity systems was rated in eight perceptual scales ("Softness", "Clearness/Distinctness", "Fullness", "Nearness", "Brightness", "Feeling of space", "Loudness", "Hissing/Disturbances") and two evaluative scales ("Fidelity", "Pleasantness") by different categories of listeners in two experiments, differing in the way of presenting the stimuli and in the way of doing the ratings. Ratings were also made concerning an "ideal" sound reproduction. The results indicate satisfactory reliability and validity of the rating scales and certain consistent relations between the perceptual scales and the evaluative scales. There were obvious differences between the results from subjects used to high fidelity sound reproduction and subjects not used to such reproduction. The consequences for continued research are discussed.

INTRODUCTION

Extensive research on perceived sound quality of sound-reproducing systems was described in a series of reports from Technical Audiology during the 1970s and recently summarized in three journal papers (Gabrielsson 1979a; Gabrielsson & Sjögren 1979a, 1979b). Multivariate analysis techniques were used to find out and interpret the meaning of relevant dimensions in perceived sound quality. The combined results of many experiments suggested the following perceptual dimensions: "Clearness/Distinctness", "Sharpness/Hardness - Softness", "Brightness - Darkness", "Fullness- Thinness", "Feeling of space", "Nearness", "Disturbing sounds", and "Loudness" (Gabrielsson & Sjögren 1979a). The relations of these perceptual dimensions to physical characteristics of the systems and to overall evaluative judgments were explored.

The validity of the suggested dimensions has to be checked in continued experiments. In the present investigation this is done by using the dimensions for assessment of perceived sound quality in some high fidelity systems. Two experiments are described, differing in certain methodological aspects. The common purposes for both of them are to study the reliability of ratings in the suggested dimensions, their capacity to differentiate between different systems, and their relations to overall evaluations of the systems - all this for different categories of listeners. Moreover certain other questions regarding stimulus presentation and rating procedures are studied to gain experiences for future listening tests aiming at evaluations of sound-reproducing systems from the consumer's point of view.

METHODS

The performance of five high fidelity systems, reproducing five different music programs, was rated in the above-mentioned dimensions and in two evaluative scales by different categories of listeners. In the first experiment the music programs were rather short, and each program x system combination was rated three times in each dimension by each subject. In the second experiment the music programs were considerably longer, but each program x system combination was rated only once in each dimension by each subject. In both experiments the subjects also made certain judgments concerning the "ideal" sound reproduction and concerning the importance of the various dimensions for the overall impression of the sound quality.

Experiment 1

Stimuli and listening conditions

The stimuli were five different music programs presented stereophonically over each of five different high fidelity systems. There were thus 25 program x system combinations.

Five high fidelity systems available on the Swedish market in 1978 were selected or composed. Each system included a turntable, an amplifier, and two loudspeakers. Three of the systems (systems A, B, and E) were "music centers", that is, the turntable and the amplifier (and tuner) are built together into one unit. The unit and the associated two loudspeakers are sold together under a certain name. System D consisted of separate units sold together. System C was composed for the present investigation by combining a selected turntable, a selected amplifier and selected loudspeakers, all of them considered to be of very high technical quality and with a considerably higher price than for the other systems. The systems are described in the following scheme. Approximate prices are given in Swedish crowns. There were magneto-dynamic pickups in all systems.

System	Price	Turntable	Amplifier	Loudspeakers
A	3400	idler type	2 x 40 W at 4 ohms	3-way, direct-radiating, floor position, 4 ohms
B	2000	idler type	2 x 18 W at 4 ohms	2-way, direct-radiating, book-shelf system, 4 ohms
C	16000	directdriven	2 x 80 W at 8 ohms	3-way, "omnidirectional" floor position, 8 ohms
D	3800	directdriven	2 x 20 W at 8 ohms	2-way, "omnidirectional", floor position, 8 ohms
E	2400	belt-driven	2 x 30 W at 4 ohms	2-way, direct-radiating, book-shelf system, 4 ohms

The electrical frequency response of pickup, pre-amplifier and power amplifier was measured. The total frequency response from pickup to loudspeaker terminals for each setup was within ±2 dB 50 - 18000 Hz. The frequency response and the non-linear distortion of the loudspeakers are shown in Figure 1.

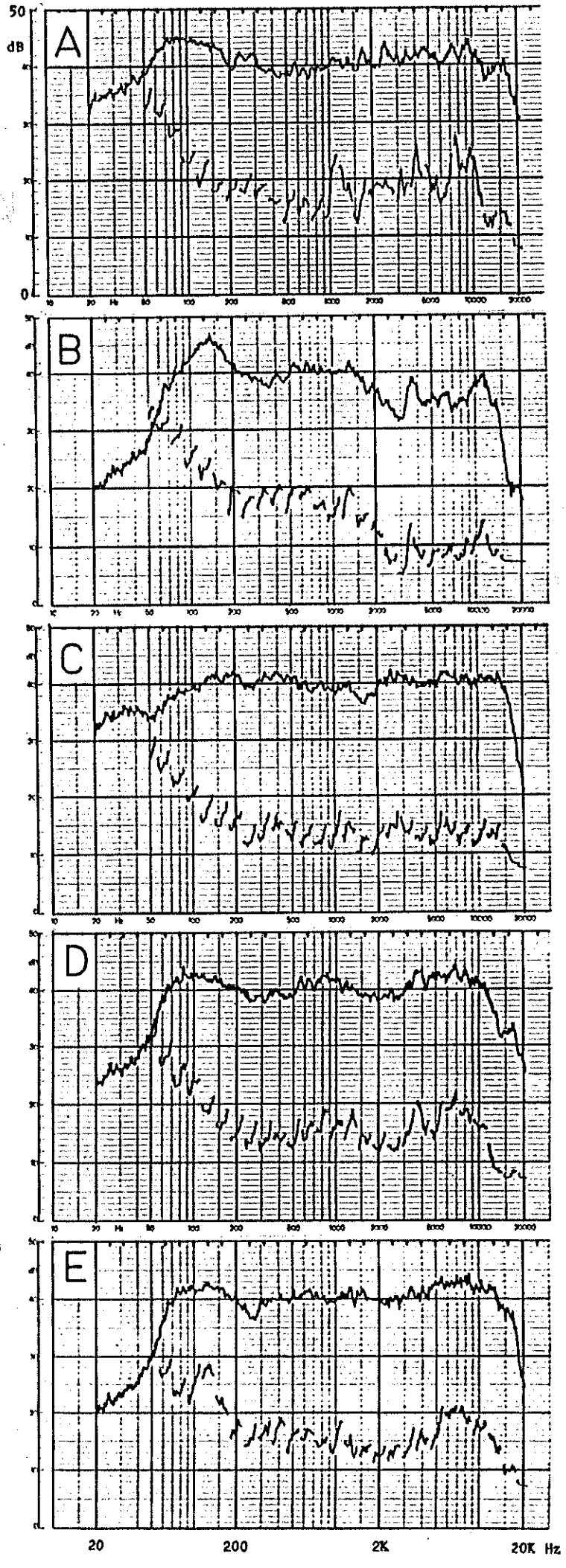


Figure 1. Loudspeaker responses for system A-E,measured in a reverberation chamber. The two curves represent the freqency response and the sum of the second and third harmonic. Test signal: white noise fed through a 30Hz wide bandpass filter. Zero level: 50dB rel. 1pW for the freqency response curve. Zero level: 30dB rel. 1pW for the distortion curve.

The music programs were the following:

1. Organ, the beginning of J.S. Bach's Toccata in D minor performed by Daniel Chorzempa. Recorded in a big cathedral. Sound level about 82 - 94 dB(A). Gramophone record: "Bättre ljud", issued by The Swedish Hi-Fi Institute.

2. Piano, Des Abends by Robert Schumann, performed by Käbi Laretei on the grand piano. Recorded in a broadcasting studio. Sound level about 70 - 82 dB(A). Gramophone record: PROPRIUS, PROP 7793.

3. Singer, the Swedish folk tune Kristallen den fina sung by Marianne Mellnäs partly accompanied by a flute. Recorded in a school music auditorium. Sound level about 70 - 87 dB(A). Gramophone record: LYRICON, LRC6 (LP, 45 rpm).

4. Orchestra, Excerpt from the end of The Firebird Suite by Stravinsky, performed by the Stockholms Philharmonic Orchestra. Recorded in the Concert Hall of Stockholm. Sound level about 85 - 92 dB(A). Gramophone record: LJUD, issued by The Swedish Hi-Fi Institute.

5. Jazz band, Switch in time by Nestico, performed by "Symfonikernas jazzband". Recorded in a school music auditorium. Sound level about 80 - 91 dB(A). Gramophone record: LJUD (as above).

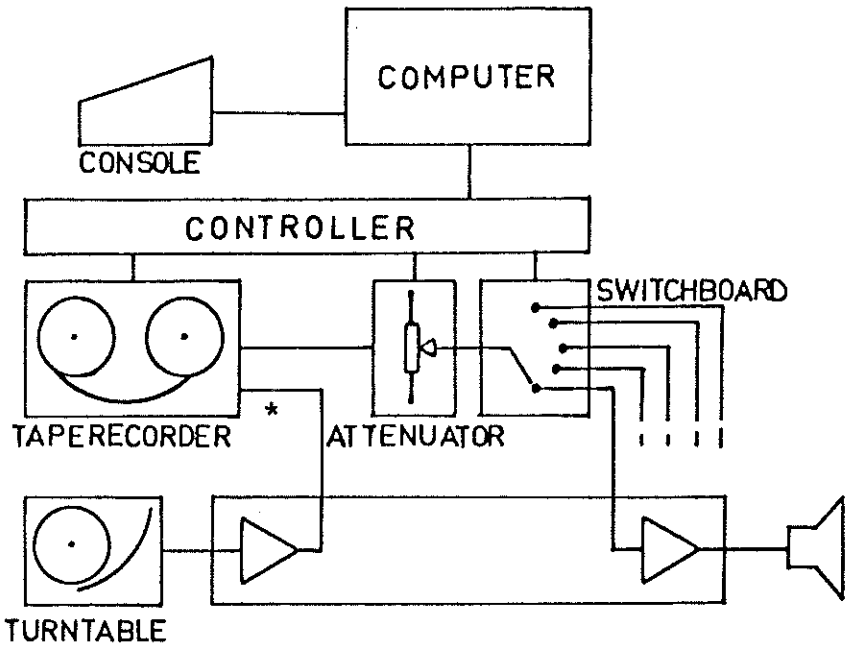
Each program lasted for about one minute from the beginning of the respective piece/recording representing "musically homogeneous" sections and also fairly homogeneous within themselves with regard to perceived loudness. The sound levels given above refer to level settings made by two experienced listeners to represent a "true-to-nature" level of the respective piece, as reproduced by system C and measured by a precision sound level meter at the listener's position in the listening room.

The reproduction to be judged by the listeners should refer to the sound reproduction of the total system, that is, the chain of turntable, amplifier, and loudspeakers in the respective system. (In our earlier experiments only the reproduction of various loudspeakers/headphones was judged, the preceding links in the reproduction chain being constant.) To avoid many practical problems associated with the handling of records on turntables in a busy experiment (rapid shifting between different programs in randomized order etc, see below), the reproduction of each program by the turntable in the respective system was recorded on tape. The tape recorder (a modified Revox B77) was directly connected to the "tape-out" connector with appropriate termination-network on the respective system. All recordings were made to obtain full modulation of the tape (Ampex Grand Master). The 25 tape recordings (five programs reproduced by five different turntables) were played back at a speed of 38 centimeters per second. The order of the 25 recordings on the tape was randomized. To check that the tape-recorded programs did not perceptually differ from the programs played directly from the respective turntables, the

experimenters and another experienced subject listened to the respective gramophone record and the corresponding tape recording in rapid succession. The listeners could not tell whether they were listening to the original record or the tape recording.

It is very important that the perceived loudness is about the same for each of the systems reproducing a certain program. To accomplish this, the acoustical output of the respective systems in the listening room was equalized by adjusting a white noise signal to an approximately equal output of the systems as measured in the octaves 250, 500, 1000, and 2000 Hz, and in broadband mode (dB, A-weighted) by a precision sound level meter placed in the listener's position. Furthermore one of the authors (A.G.) and another experienced subject listened in several long sessions and made further adjustments of the listening level for each system, checking that the perceived loudness was approximately equal for all five systems within each of the five programs.

During the experiment the signal from the tape-recorder was fed into a computer-controlled attenuator and a computer-controlled switchboard to the respective systems. As the tape was fully modulated, the computer fed the attenuator with appropriate signals to set the decided listening levels for each program x system combination, see the block diagram in Figure 2.



* used in the recording session.

Figure 2. Block diagram of the apparatus used in Experiment 1.

The presentation order of the program x system combinations was

randomized differently for each session (see Procedure). This necessitated much winding/rewinding of the stimulus tape. The finding of the correct section on the tape at each new presentation according to the actual randomization was handled by the computer. This was accomplished by using the remote control facility of the Revox B77 H.S., the optical "end of tape" sensor of the recorder and by splicing transparent tapes between the recorded stimuli. Given the correct starting position a computer program counted the number of transparent tapes and gave instruction to the tape recorder to wind or re-wind to get to the position specified by the computer program. To reduce the effect of differences in time for winding/rewinding from different parts of the tape, the computer program paused to make the inter-stimulus intervals approximately equal (about half a minute).

All systems (except for the loudspeakers) and other technical equipment were situated in a control room adjacent to the listening room. The computer was situated in another building, and the controlling apparatus for the experiment was designed to communicate with the computer through a dial-up telephone link. The actual sound in the listening room was monitored all the time by the experimenter through a playback system. This system could also be used for communication between the experimenter and the subjects.

The listening room was the same as in earlier experiments (Gabrielsson, Rosenberg & Sjögren 1971, 1974). The room is slightly rectangular, (510 x 445 centimeters), and the loud-speakers are placed along one of the long sides, hidden by acoustically transparent curtains. Two listener positions were used, quite near to the middle of the opposite long side, and separated by an acoustically transparent curtain.

To place five pairs of stereo loudspeakers in optimum positions is not easy in any listening room. If the loudspeakers in a pair are too far apart, there will be a perceptual "hole" in the middle. If they are too close, this gives a rather "narrow" sound picture. If they are somewhat asymmetrically placed with regard to the center axis of the listeners, this may give an "off-center" direction impression, which may be remembered by the subject and unduly affect his ratings for this reason. And so on.... No ideal solution was found in the present case. In the solution finally decided upon, the loudspeakers on either side were admittedly crowded too much together. This may affect their performance in various ways. It may especially be suspected that the properties of the "omnidirectional" loudspeaker in the "best" system (system C) were not made justice due to these position problems. Since this investigation mainly deals with methodological problems concerning the rating scales, this point is not very critical here. However, the position of several pairs of stereo loudspeakers for listening test purposes is a big problem to be considered in the continued work.

Subjects

Three different categories of subjects were used.

1) "Hi-Fi group". The seven members in this group (all males, 26-42 years old) were recruited by means of a questionnaire distributed to some 100 members of an association of "high fidelity enthusiasts". The selected seven subjects all had advanced high fidelity equipment for listening at home. They also fairly often listened to "live" music (attending some sort of concert once per month). All but one had some experience of performing music.

2) "Non Hi-Fi group". The seven members of this group (four males and three females, 23-43 years old) were also recruited by means of a questionnaire aiming at finding people with minimum experience of high-fidelity listening as well as of listening to "live" music. The selected seven subjects all had simple and/or old-fashioned listening equipment (transistor radios, cassette tape recorders, cheap and 10-15 years old stereo equipments etc). They hardly ever attended concerts (answers ranging from "never" to "five times a year"). None of them performed music.

3) "Music group". The five members of this group (three males, two females, 17-42 years old) were selected to represent an "intermediate" between the above-mentioned two groups. They had rather simple listening equipment at home (however, better than in the "Non Hi-Fi" group). They were very frequent concert visitors, and all of them performed music on one or more instruments. It was very hard to find subjects fitting into this intended category of people used to listening to "live" music but not used to listen to modern high fidelity equipment. This is the reason why this group contained only five members instead of seven as intended.

Preferences for types of music were rather mixed in all three groups. All subjects were tested for normal hearing (less than 20 dB hearing loss 250-8000 Hz, ISO R389), and were paid for their participation.

Rating scales, procedure

Each subject judged the perceived sound quality in each of the 25 program x system combinations on ten different rating scales. Eight of these scales refer to the eight perceptual dimensions found in analyses of earlier experiments (see Introduction). They were now labelled as follows: "Softness" (Swedish: "Mjukhet"), "Clearness/Distinctness" ("Tydlighet/Renhet"), "Fullness" ("Fyllighet"), "Nearness" ("Närhet"), "Brightness" ("Ljushet"), "Feeling of space" ("Rymdkänsla"), "Loudness" ("Ljudstyrka"), and "Hissing/Disturbances" ("Brus/Störningar"). Each scale was graded from 10 to 0, 10 meaning the maximum and 0 the minimum of the respective property. Furthermore definitions were given for the scale steps 9, 7, 5, 3, and 1 according to the principle that 9 represented "very much" of the quality in question (for instance, "very soft"), 7 "rather much" ("rather soft"), 5 a "midway position", 3 "rather much" of the "opposite" quality (for instance, "rather sharp" in opposite to "soft"), and 1 "very much" of the opposite quality ("very sharp"). These

definitions were explicitly given for each of the scales on the lists that the subjects used for doing their ratings, see the left part of Figure 3. The meaning of the different scales was not defined further but left for the subject's own interpretation (except for the scale "Hissing/Disturbances", see Instruction below).

There were also two scales for a summarizing evaluation of the perceived sound quality, namely "Fidelity" (Sw. "Naturtrohet", literally "True-to-natureness") and "Pleasantness" ("Behaglighet"), both of them also used in earlier research. They were also graded from 10 to 0 with definitions of the end points and certain scale steps as seen below in the instruction and in the right-hand part of Figure 3. (The English version of this figure gives straightforward translations of the Swedish words which may not be the best ones for English usage.) The main part of the instruction was as follows:

"You will listen to different pieces of music as they sound when they are played over various equipments for sound reproduction. You listen to one piece at a time that is played over one of the equipments. Each presentation lasts for about one minute, and during this time you shall judge the sound reproduction (how it sounds) on eight different scales which you find on your list. The scales are thus SOFTNESS, CLEARNESS/DISTINCTNESS, FULLNESS, NEARNESS, BRIGHTNESS, FEELING OF SPACE, LOUDNESS, and HISSING/DISTURBANCES.

Each scale is graded from 10 to 0. 10 designates the maximum and 0 the minimum of the respective property. For instance, if we look at the scale for SOFTNESS, 10 designates that the sound has "maximum softness" (the highest imaginable softness), while 0 designates "minimum softness" (the lowest imaginable softness). In the same way 10 and 0 are defined in the other scales, for instance, "maximum (highest imaginable) clearness/distinctness" - "minimum (lowest imaginable) clearness/distinctness", or "maximum (highest imaginable) fullness" - "minimum (lowest imaginable) fullness" etc.

The scale "HISSING/DISTURBANCES" refers to extraneous sounds in the reproduction, that is, sounds that do not belong to the music as, for example, hissing or other kinds of disturbing sounds. 10 on this scale denotes that there are so much disturbances that the music is not heard at all. 0 means that there are absolutely no disturbances.

Within each scale there are also definitions for 9, 7, 5, 3, and 1. For instance, on the SOFTNESS scale 9 means "very soft", 7 "rather soft", 5 "midway position", 3 "rather sharp", and 1 "very sharp". The corresponding definitions are given in the other scales. That these numbers have got special definitions does not mean that you should use them more frequently than the others. Use any number on the scale that you think is the best in each single case. Put a cross (X) in the ring for the number you choose! You have about one minute for the judgments in the eight scales. You will get training in several preliminary trials. Remember that your judgments shall refer to the sound reproduction, not to the music as such!

There is also a second list with two scales: FIDELITY and PLEASANTNESS. The maximum level 10 on the FIDELITY scale denotes a perfect fidelity, that is, the music sounds exactly in the same way as if you listened to the music in the room where it was originally performed. 9 denotes "very good fidelity", 7 "good fidelity", 5 a "midway position", 3 "bad fidelity", and 1 "very bad fidelity". Of course, it may be difficult to judge the fidelity when you have not heard the music in the original situation - but you have to imagine how the music sounded in the room where it was recorded.

The scale for PLEASANTNESS refers to how pleasant/nice you think the sound reproduction is (how pleasant/nice it sounds), irrespective of the fidelity. 10 denotes "maximum pleasantness" (the nicest imaginable), 9 "very pleasant" etc as you can see on the scale. Remember that it is the pleasantness of the sound reproduction you shall judge, not how pleasant you think the music is. In this case with only two scales the music will last only for about half a minute. When you have written your judgments about fidelity and pleasantness you may add further free comments to your judgments on the same list." The instruction was given both orally (tape-recorded) and in written form and was supplemented by various practical points. The subjects also got a list with short information about the music programs: composer, performers, and rooom where the recording took place.

There were 25 program x system combinations to be judged on all ten scales, and this was repeated three times to investigate the reliability of the ratings. Each subject therefore took part in three sessions of about two hours each (including instruction, preliminary trials, and various follow-up questions; there was a break for coffee or tea in the middle of each session). Each session was attended by one or two subjects. Each subject had the same position (armchair) in all his three sessions. The sessions were on different days.

The judgments on the eight perceptual scales were separated from the judgments on the two evaluative scales. Thus each of the 25 cases appeared twice in each session: one time with about one minute's duration during which the eight scales should be completed, and one time with about half a minute's duration (= the first half of the respective music program) during which the two evaluative scales were completed. This means that there were in fact 50 presentations (25 combinations x 2 times) during each session, and the order of these 50 presentations was randomized, differently for every new session. The reason for separating the two groups of scales was that it should not be possible for the subject to look at his ratings in the eight perceptual scales and somehow cognitively "derive" what his ratings in fidelity and pleasantness "should be" as a consequence of his ratings in the eight perceptual scales.

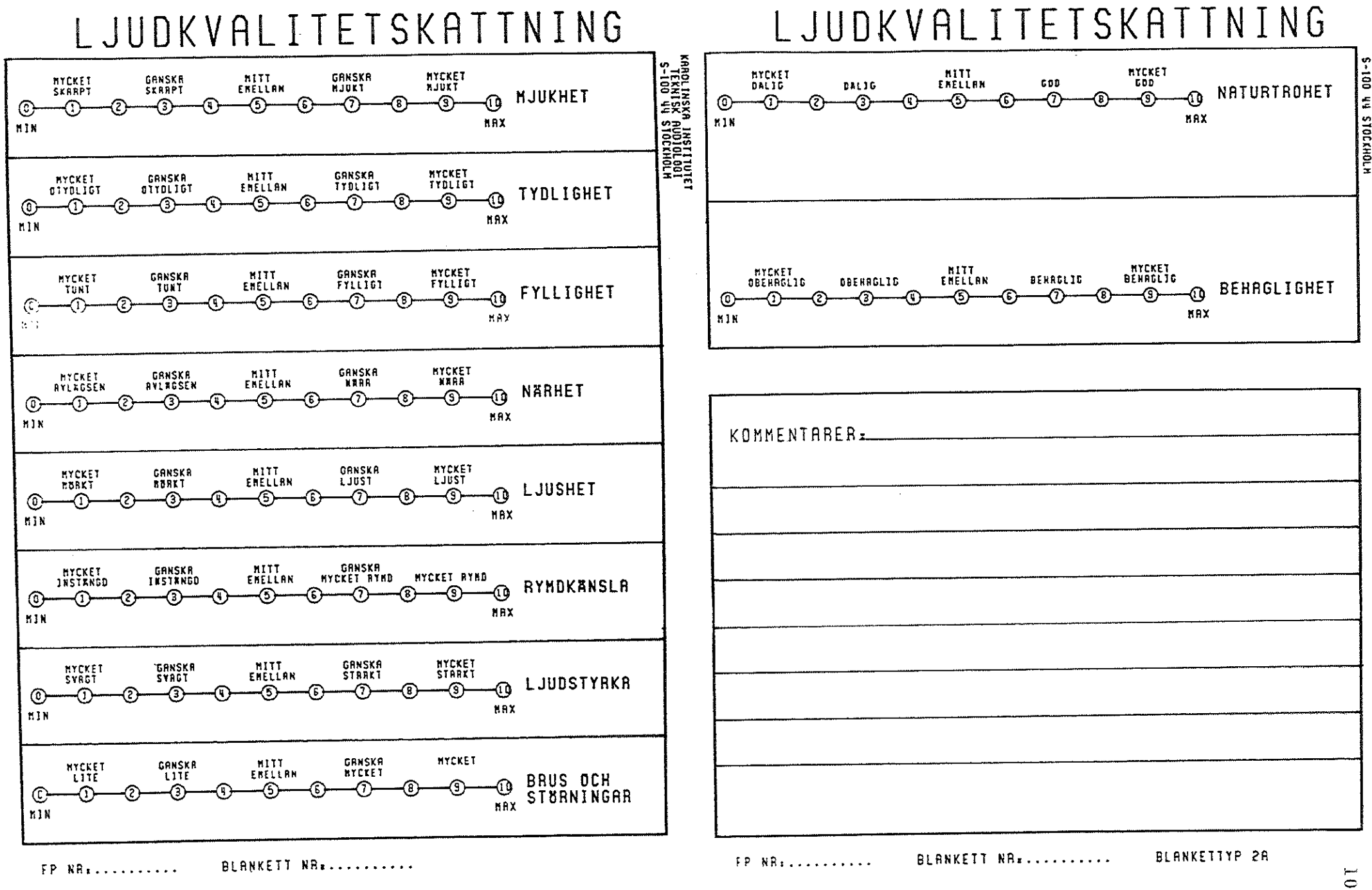
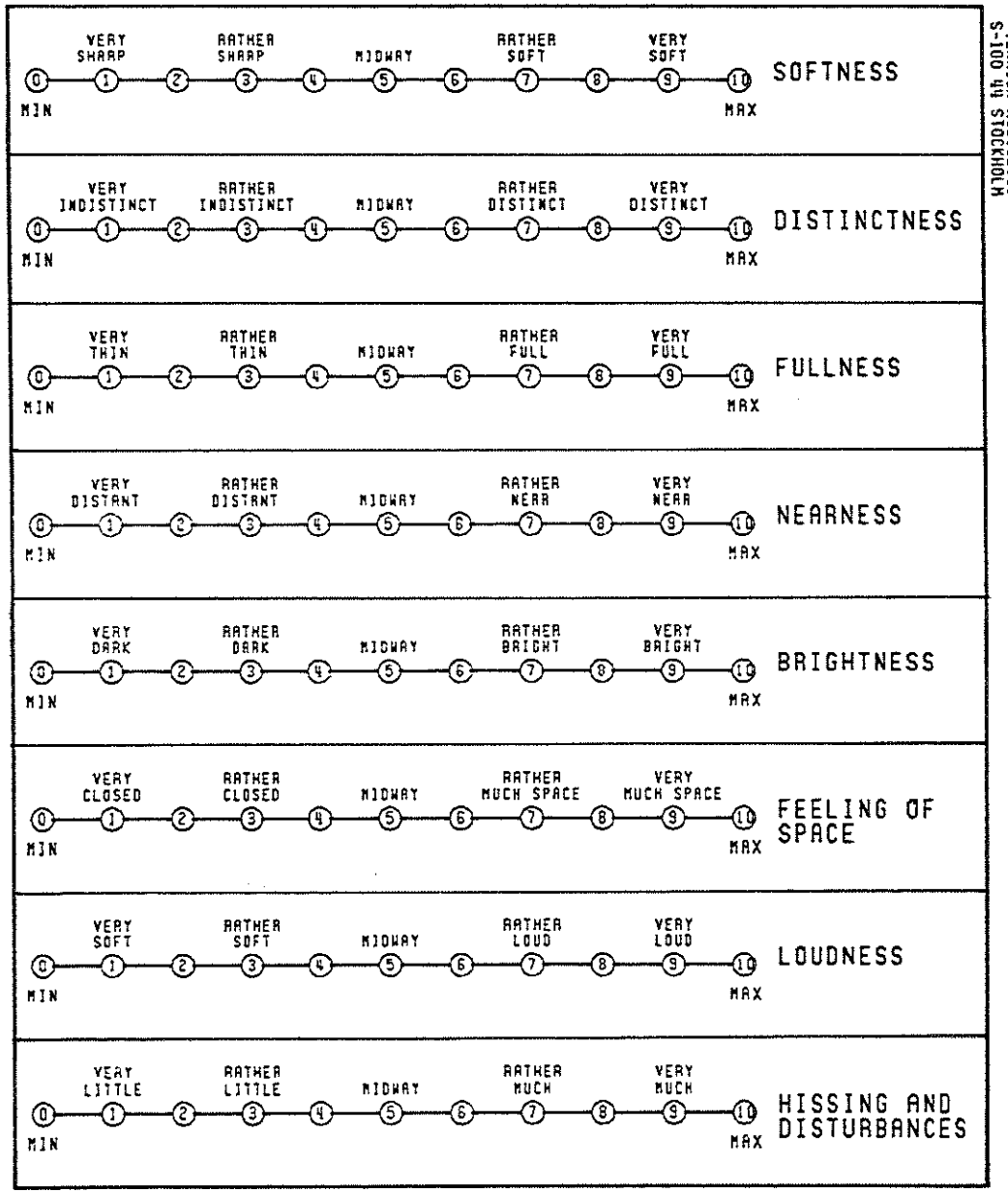


Figure 3. Example of forms for sound quality rating.

RATING OF SOUND QUALITY

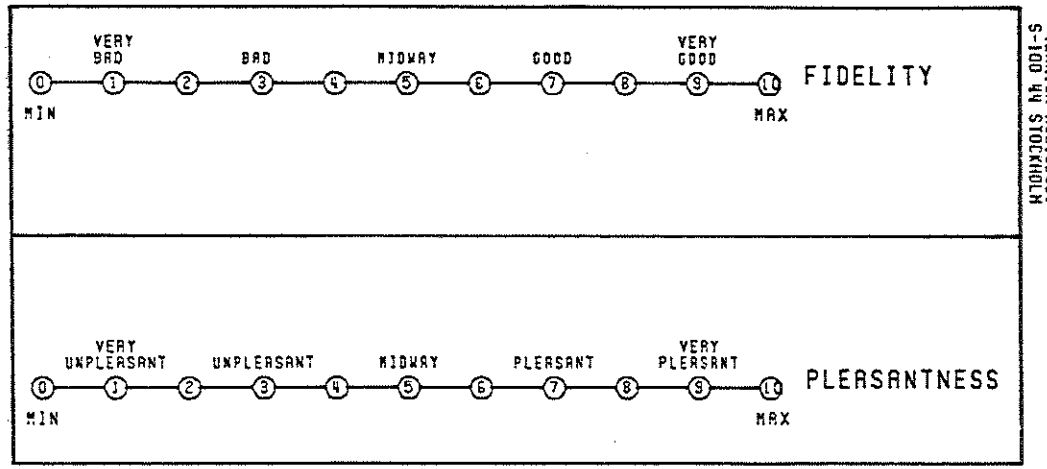
KAROLINSKA INSTITUTET
TEKNISK AUDIOLOGI
S-100 44 STOCKHOLM



FP NR:..... FORM NR:..... FORMTYPE 1A

RATING OF SOUND QUALITY

KAROLINSKA INSTITUTET
TEKNISK AUDIOLOGI
S-100 44 STOCKHOLM



COMMENTS:

FP NR:..... FORM NR:..... FORMTYPE 2A

In each session the subject had 50 lists to fill in, 25 of the type shown in the left part of Figure 3 and 25 of the type shown in the right-hand part of Figure 3. The order of the eight scales was randomized for each new list (Figure 3 shows one single example only). In the first session ten preliminary trials were made. In the second and third sessions the subject started by reading the instruction and had four preliminary trials.

When all ratings were completed in the third (last) session, the subjects were asked to make another type of rating. They should again use lists with the eight perceptual scales but now use them to "prescribe" how the sound reproduction of each music program should be to sound "true-to-nature" on one hand, and "pleasant" on the other hand - that is, to define an "ideal" sound reproduction as regards fidelity and pleasantness in terms of the eight perceptual scales. No stimuli were given during the completion of this task. Finally the subjects answered some questions about the importance of the various scales and some other related questions.

Data treatment

The data treatment generally follows the principles described in Gabrielsson (1979b). Data were analyzed both for each single subject separately and jointly for all subjects within the same listener category ("Hi-Fi", "Non Hi-Fi", and "Music" groups). The ratings for each judgment scale were displayed in programs (rows) x systems (columns) matrices, including the mean rating for each system in average over all programs, and the mean rating for each program in average over all systems. Various forms of analysis of variance were applied to investigate the effects of the systems, the programs, and interactions between these factors, and to compute indices for the reliability of the ratings.

Further multiple regression analyses were used to investigate if the ratings in the two evaluative scales ("Fidelity" and "Pleasantness") would be possible to describe as a weighted linear function of the ratings in the perceptual scales. Factor analysis was used to investigate whether the set of eight perceptual scales possibly could be interpreted in terms of some few "basic" factors. The computer programs used were BMDO8V for analysis of variance, BMDO2R for analysis of regression, and BMDO8M for factor analysis. For general orientation about regression analysis see Hays (1973) and about factor analysis Gorsuch (1974).

Experiment 2

In Experiment 1 the stimuli were rather short (about 1 minute), but each case was judged three times. The judgements in the eight perceptual scales were separated from the judgements in the two overall evaluative scales. This may be said to represent a well controlled situation for doing judgements about perceived sound quality. On the other hand it is not, of course, very typical for a "real" situation, when a consumer listens to various systems in a shop or at home and tries to compare and evaluate different systems. In the second experiment an attempt was made to come closer to such a situation, while still retaining control over various factors. Thus the music programs were considerably lengthened, and only one listening (judgement) was made for each case. The judgements in the percepual scales and the evaluative scales were not separated but made at the same time. Futhermore the stimuli were more realistic in the sense that they were played on the respective turntables (no tape recordings as in Experiment 1).

Since much in the methods is similar to those in the preceding experiment , only the differences in comparison with Experiment 1 are described below.

Stimuli and listening conditions

The stimuli were the 25 program x system combinations as earlier. The difference was that each program was considerably longer by simply taking the whole pieces of music as they appeared on the respective gramophone records. The organ piece was thus the whole Toccata (2 minutes and 40 seconds) and the piano piece by Schuman 4 minutes and 40 seconds. The pieces for symphony orchestra and jazzband were both 1 minute and 40 seconds(as recorded on that record), but were both presented twice in immediate succession. The program with solo singer could be extended to 1 minute and 25 seconds but was also repeated in immediate succession. The extension of the music programs to represent "whole" pieces of music necessarily means that the listening time for different programs will vary, and that the programs are not so homogeneous within themselves as in the shorter examples used in Experiment 1.

The technical setup of the equipment for the stimulus presentation was as follows. Five copies of each gramophone record were obtained, and care was taken to get equal copies from the same stamper. During the experiment the records were continuously rotating, and the experimenter carefully put down the pickup on the correct track on the record. The signal passed through the RIAA-section of the amplifier and was then tapped at the "tape-out" connector. The signal then passed through a computer- controlled switchboard, an attenuator (the same as in the first experiment), a fader operated by the experimenter and another switchboard, and then back to the power amplifier of the actual sound system. The reason for using a manual fader was to avoid noise arising from the lowering and lifting of the pickup. A block diagram of the

test equipment is shown in Figure 4.

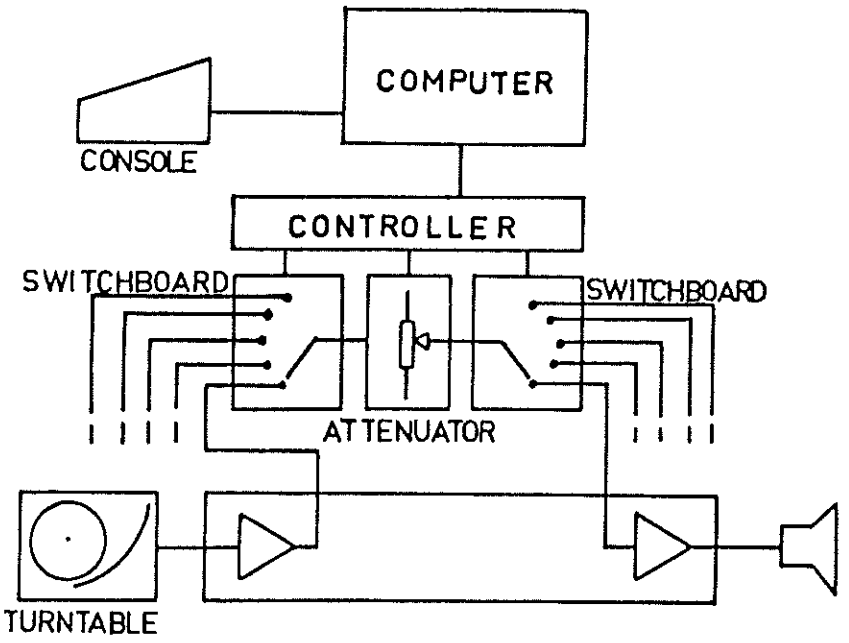


Figure 4. Block diagram of the apparatus used in Experiment 2.

The perceived loudness of the different systems was approximately equalized within each program in the same way as for Experiment 1.

Subjects

Due to the difficulties of getting subjects to the "Music" group in Experiment 1, this category of listeners was now omitted. Thus there were only two groups of new subjects, recruited in similar ways as those described for Experiment 1.

1) "Hi-Fi group". The seven members of this group were all males, 22 - 48 years old. They all had advanced equipment for high fidelity reproduction and were rather frequent concert visitors (in average 1-2 times per month). Three of them had some experience of musical performance.

2) "Non Hi-Fi group". This group consisted of one male and six females, 23 - 43 years old. They had very simple or old-fashioned equipment for sound reproduction and very seldom or never attended concerts (however, one subject attended concerts twice a month). Two of them had a little experience of musical performance.

Procedure

The rating scales and the instruction were the same as in Experiment 1. The difference was that the subjects were informed that the music sections lasted three to four minutes, and that they had the eight perceptual scales and the two evaluative scales simultaneously present for each of the program x system combinations. The list for each case thus looked like Figure 3 (the eight scales to the left, the two evaluative scales plus space for comments to the right). There were 25 program x system combinations to be rated once by each subject. However, to get a certain indication of the reliability of the ratings, five of these combinations actually appeared twice. Thus there were in all 25 + 5 = 30 cases, which were presented in a random order, differently for each new session. Each subject took part in two sessions of about two hours each. There were five preliminary trials in the first session, and three in the second session. The sessions were on different days.

RESULTS

Experiment 1

The results of the data analyses for single subjects are only briefly described under Individual data below. They are followed by detailed presentations of the results from the different groups (Group data).

Individual data

A complete matrix showing the ratings by all 19 subjects for all 25 system x program combinations in each of the 10 judgement scales is given in an appendix, which is available on request.

Two different indices for intra-individual reliability (the reliability of each single subject's ratings) were computed: MS_{within} (MS_w), which is an expression of the error variance for the respective subject, and $\underline{r}_{within}$($\underline{r}_w$), which puts the error variance in relation to the "true variance" represented by the effects of the systems, the programs, and the interaction between systems and programs (Gabrielsson 1979b). The lower the MS_w (lowest limit =0, that is, no error variance at all) and the higher $\underline{r}_w$ (lowest limit =0, highest limit =1.00), the better the reliability.

		Softness		Clearness		Fullness		Nearness		Brightness	
		MS_w	r_w	MS_w	r_w	MS_w	r_w	MS_w	r_w	MS_w	r_w
Non Hi-fi	Median	2.04	.51	1.51	.60	1.76	.46	1.40	.38	1.79	.45
	Range	1.16-4.23	.01-.88	0.99-4.09	.15-.82	1.32-6.13	.00-.74	0.92-3.25	.00-.67	1.03-2.07	.15-.73
Hi-fi	Median	1.11	.57	1.49	.60	1.11	.59	0.83	.17	0.84	.41
	Range	0.57-1.32	.08-.75	0.55-1.61	.38-.87	0.77-2.65	.22-.80	0.32-1.43	.00-.92	0.64-2.01	.00-.64
Music	Median	1.76	.36	1.25	.15	0.61	.20	0.87	.18	0.72	.47
	Range	0.63-3.63	.00-.71	0.64-2.71	.00-.63	0.43-2.51	.13-.71	0.24-1.84	.00-.50	0.68-2.23	.27-.86

		Feeling of space		Loudness		Disturbances		Fidelity		Pleasantness	
		MS_w	r_w	MS_w	r_w	MS_w	r_w	MS_w	r_w	MS_w	r_w
Non Hi-fi	Median	1.68	.60	0.55	.82	1.84	.80	2.16	.67	1.97	.75
	Range	0.85-5.45	.25-.76	0.27-1.03	.52-.89	0.83-3.23	.70 .90	1.71-3.69	.49-.85	0.96-4.35	.54-.81
Hi-fi	Median	0.77	.59	0.35	.87	0.81	.83	1.44	.65	0.95	.73
	Range	0.64-1.24	.00-.85	0.21-0.41	.50-.93	0.25-2.63	.11-.97	0.36-2.52	.28-.90	0.51-1.47	.45-.87
Music	Median	1.03	.29	0.41	.89	1.00	.84	1.20	.40	1.19	.60
	Range	0.23-2.99	.00-.43	0.21-0.87	.41-.92	0.59-3.00	.72-.96	0.84-2.89	.21-.74	0.60-2.99	.32-.85

TABLE I. Median and range of the MS_{within} and r_{within} indices for intra-individual reliability over all subjects within each of the three listener groups. See text for further explanation.

Table I presents a survey of the intra-individual reliabilities for each of the three listener groups with regard to all ten judgement scales. The upper row within each group gives the median value and the lower row the range of the MS_w and $\underline{r}_w$ indices considered over all subjects in the respective group. (The median was preferred to the arithmetic mean, since there were in general one or two subjects with highly departing indices in each group, especially in the "Non Hi-Fi" group). Comparing the three different groups, it is apparent that the median for MS_w is in general much lower for the "Hi-Fi" group and the "Music" group than for the "Non Hi-Fi" group. The range for MS_w is smaller for the "Hi-Fi" group than for the other groups. With regard to $\underline{r}_w$ the "Hi-Fi" group and the "Non Hi-Fi" group are rather similar, while the corresponding values for the "Music" group are lower for certain scales. Comparing the reliabilities for the different rating scales it is seen that the highest reliability occurs for "Loudness" and "Hissing/Disturbances".

Another observation (not shown in Table I) was that there were in general more members of the "Hi-Fi" group, for which the statistical test on differences between the systems was significant, than there were in the other groups.

Group data

Differences between systems, programs and listener categories

Table II shows the mean rating over all subjects in the respective group for all 25 program x system combinations within each of the 10 rating scales. The bottom margin within each matrix gives the mean rating for each of the systems in average over all five programs. The right-hand margin gives the mean rating for each program in average over all five systems. The value in the lower right-hand corner is the grand mean of all ratings in the respective matrix.

Visual inspection of all these matrices reveals that most means lie within the range 3.0 - 8.0. The differences within each row or each column of a matrix are thus not so large. To be interpreted, however, they must be put in relation to the error variance in the ratings. An analysis of variance was therefore performed on all ratings in each scale, for each of the groups, with systems (S), programs (P), listeners (L) and interactions between these factors as the sources of variation plus the "within cell" variation. The systems and the programs were considered as fixed variables, the listeners as a random variable and thus the mixed model was used in the analysis (Gabrielsson 1979b).

SOFTNESS

NON HI-FI GROUP	A	B	C	D	E	Mean	HI-FI GROUP	A	B	C	D	E	Mean
ORGAN	4.3	4.1	5.7	4.4	4.1	4.5		5.1	5.0	5.6	5.2	5.1	5.2
PIANO	5.4	5.6	6.1	6.1	6.2	5.9		5.7	6.1	6.6	6.3	5.9	6.1
SONG	5.7	6.4	6.6	6.0	5.8	6.1		5.0	6.0	5.4	5.1	5.4	5.4
ORCH	3.8	4.5	5.0	4.6	4.4	4.4		4.6	5.5	5.1	5.2	4.9	5.1
JAZZ	4.8	4.6	4.8	4.2	5.4	4.8		5.7	6.3	5.7	5.5	5.1	5.7
Mean	4.8	5.0	5.6	5.1	5.2	5.1		5.2	5.8	5.7	5.5	5.3	5.5

CLEARNESS/DISTINCTNESS

NON HI-FI GROUP	A	B	C	D	E	Mean	HI-FI GROUP	A	B	C	D	E	Mean
ORGAN	5.2	5.4	6.3	5.6	6.0	5.7		5.8	4.9	6.1	6.9	6.5	6.0
PIANO	5.6	5.9	6.3	6.5	6.1	6.1		5.7	5.2	6.1	6.3	6.3	5.9
SONG	6.7	7.2	7.3	6.9	6.7	7.0		6.4	6.2	6.8	6.4	7.0	6.6
ORCH	6.1	6.0	6.6	6.8	5.3	6.2		6.3	6.2	7.2	7.0	7.2	6.8
JAZZ	6.9	6.7	6.7	6.7	6.8	6.8		6.5	5.7	7.0	7.2	6.8	6.6
Mean	6.1	6.2	6.6	6.5	6.2	6.3		6.2	5.6	6.6	6.8	6.7	6.4

FULLNESS

NON HI-FI GROUP	A	B	C	D	E	Mean	HI-FI GROUP	A	B	C	D	E	Mean
ORGAN	6.1	6.4	7.2	6.3	6.0	6.4		6.0	5.1	6.2	6.6	6.0	6.0
PIANO	5.1	5.0	5.7	5.5	5.1	5.3		5.7	5.4	6.2	6.0	6.3	5.9
SONG	5.9	7.1	7.0	6.7	6.2	6.6		5.4	6.3	6.2	6.1	6.7	6.2
ORCH	6.0	5.9	6.4	7.0	5.4	6.1		6.0	6.0	6.9	6.8	6.6	6.5
JAZZ	6.8	6.3	6.3	7.0	6.5	6.6		6.4	5.3	6.3	6.4	6.3	6.2
Mean	6.0	6.1	6.5	6.5	5.8	6.2		5.9	5.6	6.4	6.4	6.4	6.1

NEARNESS

NON HI-FI GROUP	A	B	C	D	E	Mean	HI-FI GROUP	A	B	C	D	E	Mean
ORGAN	6.0	6.5	6.4	6.3	6.3	6.3		5.8	5.5	6.1	6.5	5.8	5.9
PIANO	6.1	6.1	6.0	5.7	5.9	6.0		6.1	6.0	6.1	6.4	6.1	6.2
SONG	6.2	6.6	6.8	6.0	6.4	6.4		6.0	6.0	6.1	5.7	6.0	6.0
ORCH	6.3	6.1	6.2	6.0	5.3	6.0		6.3	6.0	6.8	6.7	6.5	6.5
JAZZ	6.5	6.5	6.5	6.7	6.4	6.5		6.6	6.0	6.5	6.4	6.3	6.4
Mean	6.2	6.4	6.4	6.1	6.1	6.2		6.2	5.9	6.3	6.3	6.1	6.2

BRIGHTNESS

NON HI-FI GROUP	A	B	C	D	E	Mean	HI-FI GROUP	A	B	C	D	E	Mean
ORGAN	4.1	4.4	5.2	4.4	5.1	4.6		5.9	5.9	5.8	6.1	6.6	6.0
PIANO	6.0	5.8	5.4	6.1	6.2	5.9		6.0	5.1	5.8	5.8	6.0	5.7
SONG	5.9	5.8	5.5	6.0	6.0	5.8		6.5	5.7	6.6	6.5	6.6	6.4
ORCH	5.3	4.8	5.5	5.7	6.1	5.5		6.2	5.8	6.4	6.4	6.8	6.3
JAZZ	4.8	5.2	5.4	5.2	5.9	5.3		6.5	5.5	6.6	6.3	6.1	6.2
Mean	5.2	5.2	5.4	5.5	5.9	5.4		6.2	5.6	6.2	6.2	6.4	6.1

FEELING OF SPACE

NON HI-FI GROUP	A	B	C	D	E	Mean	HI-FI GROUP	A	B	C	D	E	Mean
ORGAN	5.9	6.7	7.6	6.6	6.5	6.7		6.0	6.6	6.8	7.0	7.2	6.7
PIANO	5.4	5.6	5.2	5.2	5.5	5.4		5.4	5.4	5.8	5.8	5.9	5.6
SONG	6.1	7.1	6.9	6.4	6.0	6.5		6.0	6.2	6.7	6.3	6.5	6.3
ORCH	6.4	6.5	6.5	6.7	6.0	6.4		6.5	6.6	6.9	7.1	7.2	6.8
JAZZ	6.8	6.8	6.6	6.9	6.7	6.7		7.0	6.5	7.2	7.0	6.7	6.9
Mean	6.1	6.6	6.6	6.4	6.1	6.3		6.2	6.3	6.7	6.6	6.7	6.5

LOUDNESS

NON HI-FI GROUP	A	B	C	D	E	Mean	HI-FI GROUP	A	B	C	D	E	Mean
ORGAN	7.1	7.7	6.9	6.8	7.0	7.1		6.8	6.8	6.9	7.0	7.1	6.9
PIANO	5.3	5.2	5.0	5.0	5.2	5.1		5.4	5.0	5.3	5.0	5.4	5.2
SONG	5.8	6.0	5.7	5.6	5.6	5.7		5.3	5.6	5.5	5.5	5.6	5.5
ORCH	7.9	7.4	7.3	7.3	6.9	7.3		7.2	6.8	7.3	7.1	6.9	7.1
JAZZ	7.1	7.0	6.9	7.0	6.9	7.0		6.6	6.6	6.4	6.7	6.7	6.6
Mean	6.6	6.7	6.3	6.3	6.3	6.5		6.2	6.2	6.3	6.3	6.3	6.3

HISSING/DISTURBANCES

NON HI-FI GROUP	A	B	C	D	E	Mean	HI-FI GROUP	A	B	C	D	E	Mean
ORGAN	5.7	6.0	4.6	6.0	6.3	5.7		4.9	4.4	4.4	5.0	4.5	4.6
PIANO	8.1	6.3	6.7	7.1	7.2	7.1		6.3	4.9	5.1	5.4	5.9	5.5
SONG	4.8	3.0	3.5	3.5	4.4	3.9		5.2	3.0	4.0	3.7	3.6	3.9
ORCH	3.6	3.6	3.3	2.6	3.8	3.4		2.9	2.6	2.1	2.4	2.3	2.5
JAZZ	2.9	3.1	2.7	2.9	3.1	2.9		2.4	2.4	2.3	3.0	2.7	2.6
Mean	5.0	4.4	4.2	4.4	5.0	4.6		4.4	3.5	3.6	3.9	3.8	3.8

FIDELITY

NON HI-FI GROUP	A	B	C	D	E	Mean	HI-FI GROUP	A	B	C	D	E	Mean
ORGAN	5.4	5.3	6.3	5.2	4.9	5.4		5.4	5.0	6.5	6.7	6.3	6.0
PIANO	3.6	5.1	4.3	4.9	5.0	4.6		5.1	5.4	5.9	6.3	6.4	5.8
SONG	5.7	7.8	6.3	6.2	6.1	6.4		5.8	6.3	7.0	6.8	7.1	6.6
ORCH	6.3	6.2	6.5	6.2	5.7	6.2		7.0	6.2	6.9	6.9	6.9	6.8
JAZZ	7.0	7.4	7.6	7.0	6.8	7.2		6.1	5.5	6.9	6.7	6.7	6.4
Mean	5.6	6.4	6.2	5.9	5.7	6.0		5.9	5.7	6.6	6.7	6.7	6.3

PLEASANTNESS

NON HI-FI GROUP	A	B	C	D	E	Mean	HI-FI GROUP	A	B	C	D	E	Mean
ORGAN	4.9	4.1	5.7	4.6	4.2	4.7		5.3	4.9	6.0	6.3	5.9	5.7
PIANO	3.3	5.3	4.6	4.4	5.0	4.5		4.8	5.4	5.5	6.0	5.8	5.5
SONG	6.0	7.9	6.5	6.3	6.2	6.6		5.2	6.0	6.3	6.3	6.7	6.1
ORCH	5.2	6.2	5.8	5.8	5.1	5.6		6.5	6.5	6.9	6.5	6.6	6.6
JAZZ	6.4	6.7	7.0	6.4	6.0	6.5		6.4	5.7	6.8	6.7	6.9	6.5
Mean	5.2	6.0	5.9	5.5	5.3	5.6		5.7	5.7	6.3	6.4	6.4	6.1

TABLE II. Mean ratings over subjects for the 25 program x system combinations, for systems in average over program(bottom margin), and for program in average over systems(right-hand margin) in Experiment 1.

SOFTNESS

MUSIC GROUP	A	B	C	D	E	Mean
ORGAN	5.8	4.5	6.2	6.3	5.3	5.6
PIANO	6.3	6.5	6.5	6.8	6.6	6.5
SONG	5.5	6.1	5.9	5.5	5.3	5.6
ORCH	6.0	5.9	5.9	5.7	6.0	5.9
JAZZ	5.8	5.7	5.9	6.1	5.5	5.8
Mean	5.9	5.8	6.1	6.1	5.7	5.9

CLEARNESS/DISTINCTNESS

MUSIC GROUP	A	B	C	D	E	Mean
ORGAN	6.5	5.9	6.9	6.9	6.5	6.5
PIANO	7.2	6.6	6.8	7.0	6.7	6.9
SONG	7.1	7.2	7.5	6.9	7.1	7.2
ORCH	7.5	6.7	7.9	7.4	7.3	7.3
JAZZ	7.4	7.1	7.3	7.1	6.9	7.2
Mean	7.1	6.7	7.3	7.1	6.9	7.0

FULLNESS

MUSIC GROUP	A	B	C	D	E	Mean
ORGAN	7.2	6.9	7.7	7.3	6.7	7.2
PIANO	6.7	6.7	6.5	6.2	6.3	6.5
SONG	6.4	7.1	7.0	6.5	6.9	6.8
ORCH	7.3	7.0	7.9	7.3	6.5	7.2
JAZZ	7.1	6.9	7.2	7.1	6.7	7.0
Mean	6.9	6.9	7.3	6.9	6.6	6.9

NEARNESS

MUSIC GROUP	A	B	C	D	E	Mean
ORGAN	6.9	7.5	7.3	6.8	7.1	7.1
PIANO	6.9	7.1	7.1	6.7	6.7	6.9
SONG	6.9	7.3	7.3	6.7	6.7	7.0
ORCH	7.2	6.7	7.5	6.8	7.0	7.0
JAZZ	7.3	7.3	6.9	6.9	6.9	7.1
Mean	7.0	7.2	7.2	6.8	6.9	7.0

BRIGHTNESS

MUSIC GROUP	A	B	C	D	E	Mean
ORGAN	6.3	5.1	4.9	4.9	6.0	5.4
PIANO	5.5	4.9	4.8	5.9	5.6	5.3
SONG	6.4	5.3	6.1	6.3	6.2	6.1
ORCH	6.1	4.9	6.3	6.0	5.9	5.8
JAZZ	6.1	5.2	5.9	5.1	5.6	5.6
Mean	6.1	5.1	5.6	5.6	5.9	5.7

TABLE II. Continued.

FEELING OF SPACE

MUSIC GROUP	A	B	C	D	E	Mean
ORGAN	6.9	6.6	7.3	7.3	6.8	7.0
PIANO	6.6	6.7	6.3	6.4	6.4	6.5
SONG	6.9	7.4	7.1	6.7	6.7	7.0
ORCH	7.3	6.9	7.1	7.1	6.9	7.1
JAZZ	7.7	6.9	7.3	7.0	6.7	7.1
Mean	7.1	6.9	7.0	6.9	6.7	6.9

LOUDNESS

MUSIC GROUP	A	B	C	D	E	Mean
ORGAN	7.8	7.6	7.2	7.7	7.1	7.5
PIANO	5.9	5.9	5.6	5.3	5.4	5.6
SONG	5.4	6.0	5.9	6.0	5.7	5.8
ORCH	7.7	7.0	7.2	7.5	6.4	7.1
JAZZ	7.1	7.2	6.8	6.8	6.5	6.9
Mean	6.8	6.7	6.5	6.6	6.2	6.6

HISSING/DISTURBANCES

MUSIC GROUP	A	B	C	D	E	Mean
ORGAN	5.5	5.7	4.6	5.4	5.6	5.3
PIANO	6.8	5.9	6.6	6.9	6.7	6.6
SONG	5.8	2.9	4.4	3.9	5.1	4.4
ORCH	2.7	2.3	1.9	1.9	2.1	2.2
JAZZ	2.2	1.7	1.9	1.8	2.3	2.0
Mean	4.6	3.7	3.9	4.0	4.4	4.1

FIDELITY

MUSIC GROUP	A	B	C	D	E	Mean
ORGAN	6.8	5.2	7.2	7.3	6.8	6.7
PIANO	5.1	6.5	6.4	6.1	6.6	6.1
SONG	6.3	6.3	6.7	6.3	6.9	6.5
ORCH	8.0	6.7	7.9	7.9	7.5	7.6
JAZZ	7.5	7.2	7.7	7.3	6.9	7.3
Mean	6.7	6.4	7.2	6.9	7.0	6.8

PLEASANTNESS

MUSIC GROUP	A	B	C	D	E	Mean
ORGAN	5.8	3.9	6.4	5.7	6.3	5.6
PIANO	4.7	6.4	5.5	5.9	6.1	5.7
SONG	5.4	5.9	5.9	5.6	5.9	5.7
ORCH	6.7	6.5	7.8	7.1	7.1	7.1
JAZZ	7.1	6.5	6.7	6.7	6.8	6.8
Mean	5.9	5.8	6.4	6.2	6.4	6.2

Table III summarizes the results of these analyses of variance and the supplementing statistical tests. It shows for each scale and each group which F tests were significant at (at least) .05 significance level. When the F test for systems was significant, it was followed by Tukey's HSD test to see which specific systems were different. The respective HSD values are included in Table III. If the difference between the means for any two systems (as given in the bottom margin of the matrices in Table II) exceeds the corresponding HSD value (denoted HSD_S), this difference is statistically significant at .05 level. Combined inspection of the means in Table II and the significant differences according to Table III leads to the following conclusions regarding systems and interactions between systems and programs (differences between programs are only briefly discussed, since they mainly reflect differences between the music programs as such and/or the recordings).

"Softness": For the "Non Hi-Fi" group system C is softer than A and B. For the "Hi-Fi" group, however, system B is softer than A (the difference between C and A does not reach the HSD value, but the difference is in the same direction as for the "Non Hi-Fi" group). In both these groups the piano and/or the song program seem to be softer than (certain of) the other programs.

"Clearness/Distinctness": For the "Hi-Fi" group systems C, D and E are clearer than B. There are tendencies in the same direction (although not statistically significant) in the other two groups.

"Fullness": For the "Non Hi-Fi" group systems C and D have more fullness than system E. For the "Hi-Fi" group, however, all these three systems (C, D and E) have more fullness than system B.

"Nearness": For the "Hi-Fi" group systems C and D sound nearer than system B.

"Brightness": For the "Non Hi-Fi" group system E is brighter than systems A, B and C, in the "Hi-Fi" group E is brighter than system B, and in the "Music" group system E and system A are brighter than system B.

"Feeling of space": For the "Non Hi-Fi" group systems B and C give more feeling of space than systems A and E. For the "Hi-Fi" group, however, systems C, D and E give more feeling of space than systems A and B. In all three groups the piano program scores lower in feeling of space than the other programs, which seems natural, since it was recorded in a studio.

"Loudness": As described under Stimuli and listening conditions an attempt was made to equalize the perceived loudness of the different systems. However, in the "Non Hi-Fi" and "Music" groups the F test for systems is significant, and for all three groups there is a significant interaction between systems and programs. A detailed inspection of the matrices for "Loudness" in Table II reveals that the differences between the systems within each single program is generally small for

	SOFTNESS			CLEARNESS			FULLNESS			NEARNESS			BRIGHTNESS			FEELING OF SPACE			LOUDNESS			DISTURBANCES			FIDELITY			PLEASANTNESS		
	NHF	HF	M	NHF	HF	M	NHF	HF	M	NHF	HF	M	NHF	HF	M	NHF	HF	M	NHF	HF	M	NHF	HF	M	NHF	HF	M	NHF	HF	M
S	x	x			x		x	x			x		x	x	x	x	x		x		x	x	x	x	x	x	x	x	x	x
P	x	x				x			x				x				x		x	x	x	x	x	x	x		x	x		x
S x P																			x	x	x	x	x	x			x	x		x
L	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
S x L		x			x	x		x	x			x		x	x							x			x	x			x	
P x L	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
S x P x L												x			x								x			x	x		x	x
HSD_S	0.60	0.60			0.86		0.69	0.80			0.40		0.50	0.59	0.80	0.50	0.34		0.36		0.37	0.75	0.46	0.66	0.83	0.90	0.66	0.69	0.87	0.48
r_b	.79	.73	.51	.61	.74	.44	.67	.62	.54	.06	.50	.00	.70	.68	.68	.67	.81	.27	.96	.97	.93	.96	.95	.96	.83	.75	.80	.87	.77	.83
MS_w	2.33	1.06	1.73	2.00	1.17	1.35	2.39	1.26	1.04	1.69	0.88	0.97	1.71	1.04	1.05	2.09	0.88	1.25	0.63	0.33	0.50	1.99	1.03	1.41	2.41	1.36	1.60	2.15	0.98	1.53

x = significant at .05 level. S = systems. P = programs. L = listeners.

TABLE III. Significant differences obtained in the different rating scales for different groups (NHF = "Non Hi-fi", HF = "Hi-fi", and M = "Music" group), and values for HSD, r_b and MS_w as explained in the text.

the "Hi-Fi" group with some exception for the orchestra program. In the "Non Hi-Fi" group and the "Music" group there are some more pronounced differences, but the pattern is rather dissimilar between these two groups. One common result for both of them is, however, that system A sounds loudest and system E least loud for the orchestra program; the difference beween these systems in loudness is as big as 1.0 (7.9-6.9) in the "Non Hi-Fi" group and 1.3 (7.7-6.4) in the "Music" group (in the "Hi-Fi" group it is only 0.3 units). That the programs differ in loudness is not surprising, of course. The means for programs in loudness generally reflect the different sound levels for the different programs as given under Stimuli and listening conditions.

"Hissing/Disturbances": For all three groups system A is judged to have more disturbances than (certain of) the other systems. This is very probably due to a rather pronounced rumble from its turntable. There is an interaction with programs, however: this disturbance (rumble) is perceived especially at the programs with lower sound levels (piano and song), while it is more or less masked in the louder programs (organ, orchestra, jazz). For the "Non Hi-Fi" group and the "Music" group also system E is more disturbing than certain other systems. Here, too, there is an interaction with programs so that this effect of system E is more pronounced at certain programs, for instance, at the song program (partly due to a mechanical defect in the turntable resulting in a "click sound" which became especially prononced in combination with the 45 rpm recording of the song program).

For all three groups there were significant differences between programs: there are less disturbances for the orchestra and jazz programs than for the other programs. The main explanation is that hissing/disturbances are effectively masked in the orchestra and jazz programs due to the high sound level and a "compact" broad frequency range in these programs, while they are more noticed in the other programs with lower sound level and/or more discretely spaced frequencies.

"Fidelity": For all three groups there are significant differences among the systems - in quite different ways, however. For the "Non Hi-Fi" group system B is rated highest and significantly better than system A. In the other two groups, however, system B is rated lowest, significantly worse than systems C, D, and E in the "Hi-Fi" group and worse than C in the "Music" group. System B is the cheapest system and probably most similar to the type of listening equipment that was familiar to the members of the "Non Hi-Fi" group. It is especially highly rated for the piano and the song programs by this group (suggesting an interaction, that did not reach the .05 level, however), possibly because disturbances in these recordings are "masked" in system B due to its relatively narrower frequency range (Figure 1). A similar (significant) interaction is seen in the "Music" group, where the difference between the best and the worst system (C and B, respectively) is big for the organ and orchestra programs, but small or non-existent for the piano and song programs.

As regards the programs, the orchestra and/or the jazz program are rated highest in average over all systems. This is probably due to, among other things, the relative absence of hissing/disturbances as mentioned above.

"Pleasantness": The situation is similar in this scale. The "Non Hi-Fi" group rates systems B and C highest, significantly better than A and E. In the "Music" group, however, C and E are rated highest, both of them better than A and B. In the "Hi-Fi" group the difference between any two systems falls quite short of the HSD value. However, the mean of systems C, D, and E together is significantly higher than the mean of A and B together (as tested with Scheffé's S method for complex comparisons, see Kirk 1968).

There are interations with programs, however, which modify the picture. For example, in the "Non Hi-Fi" group system B is rated highest among the systems for piano, song, and orchestra (especially for song), but lowest among the systems for the organ program. In the "Music" group there is something similar: although B is rated lowest in average over all programs it is rated highest among the systems for the piano and song programs.

Besides regarding the effects of systems and programs Table III also includes information on the effects of different listeners and interaction with listeners. The listener variable is significant in all cases. This means simply that listeners differ in their "mean position" on the respective scale, which is of less interest in the present context. There are certain cases with a significant system x listener interaction, and further a significant program x listener interaction in all cases but one. These interations generally mean that the effects of systems (and programs, respectively) somehow vary with different listeners. There are also some few significant three factor interactions (systems x programs x listeners). The interpretation of these interations can be made by means of the complete data matrix available in the special appendix mentioned earlier. However, they are not discussed further here.

The index $\underline{r}_{between}$ ($\underline{r}_b$) in Table III is an expression for the inter-individual reliability of the ratings, that is, how well listeners within each group agree in their ratings (lowest limit =0, highest limit =1; Gabrielsson 1979b). Very high indices occur for "Loudness" and "Disturbances", and they are also relatively high in most other cases except for the "Non Hi-Fi" group and the "Music" group in "Nearness" and for the "Music" group in "Feeling of space".

It is evident, however, that a big part of the agreement between the listeners refers to agreement concerning the characteristics of the programs (for instance, that the programs differ in loudness, disturbances etc). To study the relative importance of the programs, the systems, and their interaction one may compute indices for how much of the variance in the data is due to differences between programs, differences between systems, and to the interaction

(Gabrielsson 1979b). The details are not shown here, but the main results of these computations indicate that for the "Non Hi-Fi" group and the "Music" group differences between programs accounted for more of the variance than what differences between systems did. This was also the case for the "Hi-Fi" group in certain of the scales. However, the proportion of variance accounted for by differences between systems was higher for the "Hi-Fi" group than for the other groups in all scales (except in "Loudness", where it "should" be no difference because of the attempt to equalize the perceived loudness beforehand). This is in agreement with the fact that the "Hi-Fi" group shows significant differences between systems in all scales but "Loudness", while the other groups have fewer significant differences between systems but more significant differences as regards the programs (see Table III).

The MS_w values in the bottom of Table III represent an average of the variance between the three ratings made per each case (see Procedure) considered over all program x system combinations and all subjects in the respective group. This variance is in general smallest for the "Hi-Fi" group, next smallest for the "Music" group, and highest for the "Non Hi-Fi" group. In other words, the "Hi-Fi" group is the most stable and the "Non Hi-Fi" group the least stable in repeated ratings of the same stimuli. This is reminiscent about the results concerning MS_w in Table I (these two MS indices are computed in different ways and not directly comparable but still related).

Relations between perceptual scales and evaluative scales

A number of different correlation and regression analyses were performed to investigate the correlation between each perceptual scale and the two evaluative scales ("Fidelity" and "Pleasantness"), and to see if the ratings in the evaluative scales could be conceived of as a weighted linear function of the ratings in the perceptual scales. The input data to the analyses were varied as follows:

1) The mean ratings for each program x system combination as given in the matrices of Table II. This gives a total of 25 cases (5 programs x 5 systems). A problem with this type of input data is that they reflect the combined effects of programs and systems. It would be desirable to get rid of the effects of the programs to see the effects of the systems clearer. This was tried in the following two alternatives.

2) Input data were the mean ratings for each system in average over the programs, that is, the data given in the bottom margin of the matrices of Table II. However, this gives only five cases to do the computations from.

3) The program effects were eliminated from the means used in alternative 1 by taking the difference between the respective mean and the mean for the corresponding program (given in the right-hand margin in the matrices in Table II). For example, in the "Softness" scale for the "Non Hi-Fi" group the mean ratings for systems A - E at the organ program was 4.3, 4.1,

5.7, 4.4, and 4.1 (see in Table II). The mean of these is 4.5 (as given in the right-hand margin for the organ program). The above-mentioned differences will thus be -0.2 for system A (4.3 - 4.5), -0.4 for B (4.1 - 4.5), 1.2 for C (5.7 - 4.5) etc.

Table IV shows the correlations between the perceptual scales and the two evaluative scales obtained at the solutions according to alternatives 1 (labelled "Original" in Table IV) and 3 ("Adjusted"). Although there are certain differences between these two solutions, they are still rather similar. The scales "Clearness/Distinctness", "Fullness", and "Feeling of space" are in most cases highly positively correlated with both evaluative scales. Likewise there is a positive correlation as regards "Nearness" (except for the "Music" group), while there is a definite negative correlation for "Hissing/Disturbances". Concerning "Brightness" there are moderately high positive correlations for the "Hi-Fi" group but almost zero correlations for the other groups. For "Softness" the picture is varying.

			So	Cl	Fu	Ne	Br	Fe	Lo	Di	Mu
NON HI-FI	Original	Fidelity	-.08	.63	.69	.54	-.07	.73	.46	-.91	.94
		Pleasantness	.22	.75	.60	.48	.11	.57	.13	-.83	.96
	Adjusted	Fidelity	.51	.49	.46	.30	-.01	.59	.16	-.76	.86
		Pleasantness	.53	.43	.40	.24	-.15	.54	-.05	-.78	.89
HI-FI	Original	Fidelity	-.35	.86	.79	.52	.63	.59	.26	-.50	.91
		Pleasantness	-.27	.83	.76	.61	.52	.68	.41	-.74	.95
	Adjusted	Fidelity	-.08	.81	.77	.58	.48	.70	.34	-.18	.88
		Pleasantness	.04	.74	.81	.52	.33	.62	.23	-.36	.91
MUSIC	Original	Fidelity	.09	.58	.59	.08	.23	.61	.47	-.70	.85
		Pleasantness	.23	.61	.41	.03	.20	.45	.24	-.71	.87
	Adjusted	Fidelity	.61	.56	.38	-.10	.18	.43	.04	-.12	.72
		Pleasantness	.59	.47	.31	-.09	.19	.36	-.18	-.29	.78
NON HI-FI		Fidelity	.30	.87	.57	.29	.21	.25	.24	-.92	.96
		Pleasantness	.37	.66	.61	.15	-.15	.37	.13	-.86	.95
HI-FI		Fidelity	-.26	.83	.47	.58	.47	.70	.12	-.51	.90
		Pleasantness	-.15	.75	.42	.71	.42	.60	.13	-.74	.93
MUSIC		Fidelity	.50	.34	.32	-.04	.21	-.38	-.15	-.69	.82
		Pleasantness	.40	.11	.42	.42	.04	-.24	-.18	-.71	.89

TABLE IV. Product moment correlation between perceptual scales and the "Fidelity" and the "Pleasantness" scales in Experiment 1.
Upper matrix: from ratings on real reproductions.
Lower matrix: from differences between ratings on real and on "ideal" reproductions. See text for further explanation.
So:Softness Cl:Clearness Fu:Fullness Ne:Nearness Br:Brightness Fe:Feeling of space Lo:Loudness Di:Disturbances Mu:Multiple correlation

For the adjusted values, however, there is fairly high positive correlation in the "Non Hi-Fi" and the "Music" groups but around zero for the "Hi-Fi" group. For "Loudness" finally there are practically zero correlations at the adjusted values for the "Non Hi-Fi" and "Music" groups, but slightly positive correlations for the "Hi-Fi" group. On the whole the pattern of correlations is very similar for both evaluative scales ("Fidelity" and "Pleasantness"), and these two are highly positively inter-correlated (generally of the order +.80 to +.95).

The appearance of the multiple regression functions is somewhat varying (not shown here). In general the two or three first variables entering into the function result in a multiple correlation with each of the evaluative scales of the order +.75 - +.90. Adding the remaining variables increases the multiple correlation to about .85 to .95 (see right-hand column in Table IV). The first variable in the function is in general "Hissing/Disturbances" for the "Non Hi-Fi" group but "Clearness/Distinctness" or "Fullness" for the "Hi-Fi" group. For the "Music" group both of these alternatives appear as well as "Softness".

Table V presents the results from the ratings concerning an "ideal" sound reprooduction, that is, how the sound reproduction "should be" in terms of "Softness", "Clearness/Distinctness", "Fullness" etc to sound "true-to-nature" or "pleasant". The ratings were made for each of the five music programs, and there are a few examples that the ratings differ considerably between different programs (for example, in the "Feeling of space" dimension for the "Non Hi-Fi" group). On the whole, however, the pattern is similar over all programs and all categories of subjects.

Looking at the mean ratings over all programs (the bottom row of each matrix) it is evident that a "true-to-nature" sound reproduction shall have much of "Clearness/Distinctness", "Fullness", "Feeling of space", and "Nearness" (the means for these scales are in general higher than 7.0). It shall have rather much of "Loudness" (means 6.2 - 6.8), be somewhat over the middle position in "Softness" and "Brightness" (means 5.2 - 6.2), and have very little of "Disturbances". With regard to a "pleasant" reproduction the tendencies are about the same. There are some differences, however: it should be somewhat more of "Softness" (means 5.9 - 6.8) but somewhat less of "Nearness" (means 6.0 - 6.9) and "Loudness" (means 5.4 - 6.4) than in the "true-to-nature" reproduction.

With regard to possible differences between the groups it is noted, among other things, that the "Hi-Fi" group wants somewhat less "Softness" but somewhat more "Clearness" than the other two groups.

			FIDELITY								PLEASANTNESS							
			So	Cl	Pu	Ne	Br	Fe	Lo	Di	So	Cl	Pu	Ne	Br	Fe	Lo	Di
EXP 1	NON HI-FI	Organ	6.0	8.1	8.9	7.4	4.3	9.0	7.1	1.1	6.4	8.4	8.4	6.6	4.4	9.0	6.0	0.4
		Piano	6.4	9.1	7.6	8.4	6.1	6.4	6.1	0.6	7.4	8.4	7.7	6.9	5.4	6.7	5.6	0.3
		Song	6.9	8.7	7.4	7.4	5.3	6.4	5.9	0.6	8.0	9.0	8.1	7.1	5.4	7.4	5.6	0.7
		Orch.	6.1	8.3	8.6	7.0	5.0	8.6	7.4	0.6	6.7	8.6	8.6	6.6	5.0	9.3	6.0	0.1
		Jazz	5.0	7.9	8.0	8.6	5.3	8.0	7.7	1.3	5.6	8.1	8.1	6.7	5.1	8.6	6.1	0.9
		Mean	6.1	8.4	8.1	7.8	5.2	7.7	6.8	0.8	6.8	8.5	8.2	6.8	5.1	8.2	5.9	0.5
	HI-FI	Organ	5.3	8.9	7.4	6.6	5.4	8.3	6.9	0.3	5.6	8.4	7.6	6.7	5.4	8.3	6.3	0.1
		Piano	5.3	9.3	6.7	7.7	5.6	7.1	6.0	0.3	5.9	9.1	7.4	7.4	5.7	6.7	5.9	0.1
		Song	6.3	8.9	7.1	6.9	6.0	6.9	6.0	0.4	7.0	9.1	7.0	6.9	5.7	7.0	5.9	0.1
		Orch.	5.1	8.6	7.7	6.6	5.7	7.9	7.3	0.4	6.1	9.1	8.0	6.6	5.6	7.9	6.6	0.1
		Jazz	4.3	8.6	7.6	7.3	5.9	7.7	7.9	0.4	5.1	8.9	7.9	7.0	5.9	7.9	7.1	0.1
		Mean	5.3	8.9	7.3	7.0	5.7	7.6	6.8	0.4	5.9	8.9	7.6	6.9	5.7	7.6	6.4	0.1
	MUSIC	Organ	6.2	8.2	7.8	6.0	6.4	7.8	6.2	1.0	7.0	8.6	7.4	6.0	5.8	7.2	5.4	1.0
		Piano	6.0	8.4	7.4	7.2	6.4	6.2	5.6	1.2	6.4	8.6	7.4	6.2	5.8	6.2	5.0	1.0
		Song	6.0	7.8	7.6	6.4	5.8	6.8	5.4	1.2	8.0	7.4	7.4	6.2	5.8	6.4	4.8	1.0
		Orch.	4.8	8.6	8.0	7.0	6.0	8.0	7.0	1.2	6.4	8.8	7.2	5.8	6.2	7.2	6.0	1.0
		Jazz	5.6	8.2	7.8	6.8	6.4	7.6	7.0	1.2	6.4	8.6	7.6	5.6	5.8	7.6	5.8	1.2
		Mean	5.7	8.2	7.7	6.7	6.2	7.3	6.2	1.2	6.8	8.4	7.4	6.0	5.9	6.9	5.4	1.0
EXP 2	NON HI-FI	Organ	5.9	8.1	9.0	8.4	4.6	9.4	7.1	1.1	6.6	7.9	8.9	7.7	4.3	8.9	6.3	1.1
		Piano	8.1	9.3	7.6	9.0	6.1	7.6	6.1	0.9	8.4	9.0	7.7	8.3	6.1	8.0	6.4	0.7
		Song	7.1	8.3	6.9	7.1	6.0	7.1	5.7	1.9	7.9	9.1	8.4	8.4	5.4	8.3	5.9	1.0
		Orch.	6.3	8.1	8.6	8.7	4.9	9.1	7.1	1.6	7.0	8.7	9.0	8.3	4.9	9.1	6.6	0.7
		Jazz	5.8	7.9	8.1	8.6	5.4	8.4	7.1	1.9	6.7	8.6	8.6	8.1	4.9	8.6	6.1	1.1
		Mean	6.6	8.3	8.0	8.4	5.4	8.3	6.6	1.5	7.3	8.7	8.5	8.2	5.1	8.6	6.3	0.9
	HI-FI	Organ	5.0	9.1	8.0	5.0	5.3	6.9	6.6	0.3	6.7	8.3	7.7	5.0	4.7	6.4	5.7	0.3
		Piano	5.7	8.3	6.7	6.0	5.7	5.3	5.0	0.3	6.9	7.9	7.1	5.7	5.4	5.3	4.6	0.1
		Song	7.1	8.1	6.6	5.6	5.1	5.7	5.1	0.3	7.3	7.9	7.0	5.7	5.4	5.9	4.9	0.3
		Orch.	6.4	8.4	7.7	4.4	5.1	7.0	6.6	0.4	6.9	8.1	7.6	4.7	5.3	7.0	6.1	0.6
		Jazz	5.0	8.1	8.1	5.4	5.6	6.4	6.9	0.3	5.7	8.0	7.7	5.6	5.4	6.1	6.4	0.2
		Mean	5.8	8.4	7.4	5.3	5.4	6.3	6.0	0 3	6.7	8.0	7.4	5.3	5.2	6.1	5 5	0.3

TABLE V. Ratings concerning the "ideal" value in each perceptual scale to give perfect "Fidelity" and maximum "Pleasantness" in Experiment 1 and Experiment 2.
So:Softness Cl:Clearness Fu:Fullness Ne:Nearness Br:Brightness Fe:Feeling of space Lo:Loudness Di:Disturbances

Similar tendencies appear in the answers to a question that the subjects should rank order the eight perceptual scales with regard to their importance for "Fidelity" and for "Pleasantness". Table VI presents the resulting rank orders (averaged over the subjects within each group), where 1 = most important and 8 = least important. With regard to differences in the rankings referring to "Fidelity" and to "Pleasantness" it is noted that "Softness" gets higher ranking for "Pleasantness" than for "Fidelity", while it is the opposite situation as regards "Nearness". Concerning differences between the groups it is seen that the "Hi-Fi" group attaches most importance to "Clearness/Distinctness" and "Fullness", while the other two groups puts (absence of) "Disturbances" first followed by "Clearness".

EXPERIMENT 1

	Fidelity			Pleasantness		
	NHF	HF	M	NHF	HF	M
Softness	7	6	8	4.5	5	3.5
Clearness	2	1	1.5	2	1	2
Fullness	4	2	6.5	6	2	5
Nearness	5	5	4.5	7	6	6
Brightness	8	8	6.5	8	7	7
Feeling of space	3	3	3	3	4	8
Loudness	6	7	4.5	4.5	8	3.5
Disturbances	1	4	1.5	1	3	1

EXPERIMENT 2

	Fidelity		Pleasantness	
	NHF	HF	NHF	HF
Softness	7	6.5	4	2
Clearness	1	1	1	1
Fullness	2	3.5	2	5
Nearness	5	8	7	7
Brightness	8	5	8	8
Feeling of space	3	2	6	4
Loudness	6	6.5	5	6
Disturbances	4	3.5	3	3

TABLE VI Ranking of the perceptual scales with regard to their importance for "Fidelity" and "Pleasantness" in Experiment 1 and in Experiment 2.

The ratings concerning the "ideal" sound reproduction were put in relation to the ratings concerning the real sound reproductions in this experiment by means of another regression analysis. The purpose was again to see if the evaluative scales can be conceived of as a weighted linear function of the perceptual scales. In this case, however, the data used as input to the regression analysis were the differences between the ratings of the real reproductions and the corresponding ratings of the "ideal" reproduction. An example will illustrate the procedure.

The mean rating of the "Non Ni-Fi" group concerning the "Fidelity" of system A's reproduction of the organ program was 5.4 (check in Table II). The "ideal" sound reproduction in "Fidelity" corresponds, by definition, to 10. The difference between the real and the "ideal" reproduction is thus 5.4 - 10 = -4.6. This difference may be thought of as a function of the differences between the real and the "ideal" reproduction with regard to each of the perceptual scales. In our example the difference between system A's reproduction and the "ideal" reproduction with regard to "Softness" is found to be 4.3 (found in Table II) - 6.0 (found in Table V) = -1.7. With regard to "Clearness" the difference is found to be 5.2 - 8.1 = -2.9 etc through all scales.

The result of this type of correlation/regression analysis is shown in the lower part of Table IV, where it can be easily compared with the results from the earlier described correlation/regression analyses. In fact the results are so similar, that the conclusions given above for the earlier analysis apply here, too, with two slight modifications:

a) There appears a negative correlation between "Feeling of space" and the evaluative scales in the "Music" group.

b) The first three variables in the regression functions result in a multiple correlation with the evaluative scales of the order .89 - .94 in the "Non Hi-Fi" and the "Hi-Fi"groups, .78 - .84 in the "Music" group. The first variable entering into the function is "Disturbances" for the "Non Hi-Fi" and the "Music" groups and "Clearness/Distinctness" for the "Hi-Fi" group. Entering all eight variables into the function gives a multiple correlation of .90 - .96 for the "Non Hi-Fi" and "Hi-Fi" groups and .82 - .89 for the "Music" group (Table IV, right-hand column).

Relations between perceptual scales

The scales in this experiment represent a very reduced number of scales, which were obtained by means of multivariate analyses (such as factor analysis) on a big number of scales in earlier investigations. It may still be interesting, however, to see if the present eight perceptual scales could be represented in a still more condensed structure. Factor analysis (component analysis) was therefore applied to the correlations between the scales over 175 cases (5 programs x 5 systems x 7 subjects in each group), over 25 cases (5 programs x 5 systems, the input data were the means over subjects given

in Table II), and over 25 cases in which the program effects were eliminated in the same way as described above for the regression analyses (alternative 3).

The results are rather varying in the different analyses. The three scales "Clearness/Distinctness", "Fullness" and "Feeling of space" occur almost always together in one factor, sometimes also "Nearness". The other scales appear in various combinations or occupying a factor of their own. For instance, "Softness" sometimes appears as opposite to "Loudness", sometimes as opposite to "Brightness", "Nearness" may appear in combination with "Loudness", "Hissing/Disturbances" as contrast to "Clearness/Distinctness" or to "Softness" etc.

Experiment 2

Individual data

A complete matrix of all individual ratings is given in the special appendix mentioned under Experiment 1.

In this experiment each program x system combination was judged only once except for five combinations that were judged twice. It is thus not possible to compute the reliability indices MS_w and $\underline{r}_w$ over all 25 program x system combinations but only over five of these. In these five cases the MS_w values are in most cases lower for the "Non Hi-Fi" group than for the "Hi-Fi" group (unlike the situation in Experiment 1). On the other hand the $\underline{r}_w$ values are in most cases higher for the "Hi-Fi" group. Like the situation in Experiment 1 there are far more significant differences regarding the systems for the members of the "Hi-Fi" group than for the members of the "Non Hi-Fi" group.

Group data

Differences between systems, programs, listener categories

Table VII shows the mean rating over all subjects in the respective group for all 25 program x system combinations in each of the 10 rating scales. Most means lie in the range 3.0 - 8.0. An analysis of variance was performed on all ratings in each scale with systems, programs, listeners, and interactions between these factors as sources of variation. As in Experiment 1 the listeners were considered as a random variable, and the mixed model was used in the analysis. The results are summarized in Table VIII. Combined inspection of Tables VII and VIII leads to the following conclusions:

SOFTNESS

	NON HI-FI GROUP						HI-FI GROUP					
	A	B	C	D	E	Mean	A	B	C	D	E	Mean
ORGAN	4.6	4.9	5.9	4.3	5.0	4.9	4.8	6.7	5.9	6.4	5.1	5.8
PIANO	6.1	6.2	6.3	5.3	6.7	6.1	5.6	5.5	6.6	5.6	4.7	5.6
SONG	4.7	6.4	5.0	5.9	5.4	5.5	4.9	4.4	4.8	5.4	5.0	4.9
ORCH	4.0	5.4	5.4	4.4	4.6	4.8	3.6	7.0	5.1	5.4	5.1	5.3
JAZZ	4.4	6.1	5.6	4.6	5.1	5.2	3.7	6.6	5.9	5.0	4.5	5.1
Mean	4.8	5.8	5.6	4.9	5.4	5.3	4.5	6.0	5.6	5.6	4.9	5.3

CLEARNESS/DISTINCTNESS

	NON HI-FI GROUP						HI-FI GROUP					
	A	B	C	D	E	Mean	A	B	C	D	E	Mean
ORGAN	6.0	4.9	7.0	6.9	7.3	6.4	5.8	5.0	7.0	6.9	7.7	6.5
PIANO	5.9	7.1	5.7	6.4	7.0	6.4	5.7	5.0	6.3	5.9	6.9	5.9
SONG	5.6	7.1	6.3	6.4	7.6	6.6	5.7	3.9	6.6	5.1	6.3	5.5
ORCH	7.6	6.9	7.9	6.6	7.0	7.2	6.7	4.6	7.3	6.1	6.7	6.3
JAZZ	7.1	7.7	7.7	6.0	7.5	7.2	5.1	4.1	5.6	5.7	5.5	5.2
Mean	6.4	6.7	6.9	6.5	7.3	6.8	5.8	4.5	6.5	5.9	6.6	5.9

FULLNESS

	NON HI-FI GROUP						HI-FI GROUP					
	A	B	C	D	E	Mean	A	B	C	D	E	Mean
ORGAN	7.2	7.4	7.1	7.0	7.4	7.2	6.3	5.6	7.1	7.1	5.9	6.4
PIANO	6.9	6.1	6.1	5.6	6.6	6.3	5.4	5.9	6.4	5.9	6.3	6.0
SONG	5.9	7.4	6.6	6.9	6.6	6.7	4.7	4.7	6.2	5.1	6.4	5.4
ORCH	6.9	7.0	7.3	7.0	6.4	6.9	5.4	4.7	6.4	4.9	6.0	5.5
JAZZ	7.0	6.9	8.0	7.1	6.8	7.2	5.9	5.9	5.9	6.6	6.2	6.1
Mean	6.8	7.0	7.0	6.7	6.8	6.8	5.5	5.3	6.4	5.9	6.2	5.9

NEARNESS

	NON HI-FI GROUP						HI-FI GROUP					
	A	B	C	D	E	Mean	A	B	C	D	E	Mean
ORGAN	6.6	6.4	7.4	7.3	7.3	7.0	5.8	4.3	5.1	5.6	5.6	5.3
PIANO	6.3	6.5	5.4	5.9	6.7	6.2	7.0	5.5	6.3	6.7	6.4	6.4
SONG	6.4	7.3	6.8	7.0	6.6	6.8	5.4	5.9	6.1	5.7	6.4	5.9
ORCH	8.0	6.9	7.6	6.9	6.6	7.2	6.4	4.0	6.4	5.1	4.7	5.3
JAZZ	7.0	7.4	6.7	6.3	7.1	6.9	6.4	4.1	5.0	6.0	4.5	5.2
Mean	6.9	6.9	6.8	6.7	6.9	6.8	6.2	4.8	5.8	5.8	5.5	5.6

BRIGHTNESS

	NON HI-FI GROUP						HI-FI GROUP					
	A	B	C	D	E	Mean	A	B	C	D	E	Mean
ORGAN	4.2	4.3	4.6	4.9	4.4	4.5	6.3	4.9	5.9	5.0	6.9	5.8
PIANO	5.1	5.6	5.1	6.3	5.1	5.5	5.3	3.9	4.1	5.9	5.9	5.0
SONG	5.7	5.9	5.4	5.9	5.7	5.7	6.4	4.9	6.9	5.3	6.3	5.9
ORCH	5.7	4.6	5.3	5.6	5.9	5.4	6.7	3.7	6.7	5.7	6.3	5.8
JAZZ	5.0	4.6	4.4	4.4	5.0	4.7	6.9	3.7	4.7	4.7	5.2	5.0
Mean	5.2	5.0	5.0	5.4	5.2	5.1	6.3	4.2	5.7	5.3	6.1	5.5

FEELING OF SPACE

	NON HI-FI GROUP						HI-FI GROUP					
	A	B	C	D	E	Mean	A	B	C	D	E	Mean
ORGAN	7.2	7.1	7.6	7.4	7.9	7.4	5.9	6.4	6.9	6.9	6.7	6.5
PIANO	5.9	6.1	5.0	5.4	6.3	5.7	4.6	4.2	6.1	5.1	5.1	5.0
SONG	6.6	6.6	6.3	6.3	6.7	6.5	6.6	4.7	6.8	6.3	6.4	6.2
ORCH	7.0	6.6	8.0	6.6	6.3	6.9	5.3	4.4	6.4	5.1	5.9	5.4
JAZZ	6.9	7.3	7.4	7.0	7.1	7.1	5.1	5.0	5.0	5.9	5.2	5.2
Mean	6.7	6.7	6.9	6.6	6.9	6.7	5.5	5.0	6.2	5.8	5.9	5.7

LOUDNESS

	NON HI-FI GROUP						HI-FI GROUP					
	A	B	C	D	E	Mean	A	B	C	D	E	Mean
ORGAN	8.4	7.1	7.4	7.3	7.6	7.6	7.5	6.4	6.4	7.0	7.0	6.9
PIANO	5.6	5.3	5.3	5.1	5.1	5.3	5.1	5.3	5.3	5.4	5.4	5.3
SONG	5.4	5.9	5.7	5.9	5.6	5.7	5.1	5.7	5.1	5.0	5.3	5.2
ORCH	7.6	6.6	7.1	6.9	6.6	6.9	6.7	5.7	6.9	6.1	5.7	6.2
JAZZ	7.1	6.6	6.9	7.1	6.7	6.9	6.1	5.9	6.4	6.3	5.9	6.1
Mean	6.8	6.3	6.5	6.5	6.3	6.5	6.1	5.8	6.0	6.0	5.9	6.0

HISSING/DISTURBANCES

	NON HI-FI GROUP						HI-FI GROUP					
	A	B	C	D	E	Mean	A	B	C	D	E	Mean
ORGAN	6.4	4.9	4.4	6.6	5.7	5.6	5.5	4.0	3.7	3.1	3.6	4.0
PIANO	8.6	4.9	7.1	6.9	6.3	6.7	7.3	5.1	6.9	6.9	5.4	6.3
SONG	5.7	3.0	4.6	5.1	4.0	4.5	7.4	5.0	4.9	5.4	4.9	5.5
ORCH	4.2	2.8	2.7	3.0	2.3	3.0	5.0	2.6	3.9	3.0	2.6	3.4
JAZZ	2.7	2.3	2.0	2.3	2.0	2.3	3.9	2.4	2.9	2.9	2.7	3.0
Mean	5.5	3.6	4.2	4.8	4.1	4.4	5.8	3.8	4.5	4.3	3.8	4.4

FIDELITY

	NON HI-FI GROUP						HI-FI GROUP					
	A	B	C	D	E	Mean	A	B	C	D	E	Mean
ORGAN	6.3	6.1	6.9	7.0	7.1	6.7	4.6	5.1	6.0	6.3	6.9	5.8
PIANO	4.4	6.4	4.7	5.4	6.6	5.5	3.9	4.4	5.9	5.0	5.9	5.0
SONG	6.0	6.9	5.9	6.0	6.9	6.3	4.4	4.0	6.2	5.0	4.7	4.9
ORCH	6.6	7.1	6.6	5.9	6.1	6.5	5.0	4.0	7.7	5.4	5.7	5.6
JAZZ	7.1	7.1	6.7	6.9	7.2	7.0	3.4	3.3	4.6	4.4	5.0	4.1
Mean	6.1	6.7	6.1	6.2	6.8	6.4	4.3	4.2	6.1	5.2	5.6	5.1

PLEASANTNESS

	NON HI-FI GROUP						HI-FI GROUP					
	A	B	C	D	E	Mean	A	B	C	D	E	Mean
ORGAN	5.0	5.6	6.4	5.6	6.0	5.7	4.9	5.3	5.7	5.1	6.7	5.5
PIANO	3.4	6.1	3.6	4.6	5.1	4.6	2.7	4.5	5.1	4.9	6.0	4.6
SONG	5.1	7.0	5.6	5.4	6.7	6.0	3.7	2.9	5.3	4.7	4.1	4.1
ORCH	5.4	6.7	6.9	6.1	5.9	6.2	4.3	5.6	7.0	5.9	5.7	5.7
JAZZ	6.3	6.6	6.4	6.1	6.4	6.4	3.7	4.4	4.4	4.7	5.1	4.5
Mean	5.1	6.4	5.8	5.6	6.0	5.8	3.9	4.5	5.5	5.1	5.5	4.9

TABLE VII. Mean ratings over subjects for the 25 program x system combinations, for systems in average over program(bottom margin), and for program in average over systems(right-hand margin) in Experiment 2.

	Softness		Clearness		Fullness		Nearness		Brightness		Feeling of space		Loudness		Disturb.		Fidelity		Pleasantness	
	NHF	HF	NHF	HF	NHF	HF	NHF	HF	NHF	HF	NHF	HF	NHF	HF	NHF	HF	NHF	HF	NHF	HF
Systems		X		X				X		X					X	X		X	X	X
Programs								X		X		X	X	X	X	X				X
Systems x Programs			X					X		X				X	X					
Listeners	X		X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Systems x Listeners	X				X	X			X	X										
Programs x Listeners	X	X	X	X	X		X	X	X		X		X	X	X	X	X	X	X	X
HSD for Systems		1.10		1.30				0.99		1.11					0.97	0.93		1.30	0.80	1.00
r_b	.42	.45	.41	.52	.00	.26	.00	.65	.30	.78	17	.36	.87	.83	.90	.87	.14	.61	.61	.56

X= significant at .05 level

TABLE VIII. Significant differences obtained in the different rating scales for different groups (NHF="Non HI-FI", HF="HI-FI"), and values for HSD and r_b

"Softness": For the "Hi-Fi" group systems B, C, and D are all softer than system A, and system B is also softer than system E. In the "Non Hi-Fi" group the systems are not significantly different; however, there are similar tendencies in their data.

"Clearness/Distinctness": For the "Hi-Fi" group system B is less clear/distinct than all other systems. For the "Non Hi-Fi" group there is an interaction between programs and systems. It is seen in Table VII that which system is "best" and "worst", respectively, in this scale varies from program to program. For instance, system B is worst at the organ program, but best at the piano and the jazz programs. System C is best for the orchestra program, but worst for the piano program.

"Fullness": There are no significant differences for any of the groups. However, the data for the "Hi-Fi" group are reminiscent of the corresponding data in Experiment 1, namely that systems C, D, and E seem to have more "fullness" than systems A and B.

"Nearness": For the "Hi-Fi" group system B sounds less near than systems A, C, and D. There is an interaction with programs, however: system B is least near for the organ, piano, orchestra, and jazz programs, but third in nearness as regards the song program. For system A the situation is quite opposite: it sounds nearest for all programs but for the song program.

"Brightness": For the "Hi-Fi" group system B sounds darker than all other systems. There is also an interaction between programs and systems. Although system B sounds darkest for all programs, this is especially pronounced for the orchestra and jazz programs. Which system sounds brightest varies from program to program.

"Feeling of space": There are no significant differences between the systems for either group. However, the data for the "Hi-Fi" group are similar to the corresponding data in Experiment 1, namely that systems C, D, and E seem to give more feeling of space than systems A and B. As regards the programs the piano program gives less feeling of space than the organ program, which seems natural.

"Loudness": There are no significant differences between the systems for either group but significant differences between the programs in general agreement with the different sound levels for the different programs. However, there is an interaction between programs and systems for the "Hi-Fi" group. While the systems all sound about equally loud for the piano, song, and jazz programs, there are differences of 1.1 unit for the organ program (system A 7.5, systems B and C 6.4) and of 1.2 unit for the orchestra program (system C 6.9, systems B and E 5.7).

"Hissing/Disturbances": In both groups system A is more affected by disturbances than all other systems probably due to pronounced rumble in its turntable. In the "Non Hi-Fi" group

there is also an interaction with programs of the same type as discussed in Experiment 1, namely that this disturbance (rumble) is more noticed for the programs with lower sound levels (especially for the piano program). Furthermore the programs differ in disturbances in the same way as discussed in Experiment 1.

"Fidelity": For the "Hi-Fi" group systems C and E are more "true-to-nature" than systems A and B. On the whole there is in this group a clear distinction between systems C, D, and E on one hand (the better ones) and A and B on the other hand (the worse ones). In the "Non Hi-Fi" group there are no significant differences between the systems. However, as in Experiment 1 system B is rated highest (together with E), in fact as the best system for the song and orchestra programs.

"Pleasantness": The situation is similar in this scale. For the "Hi-Fi" group systems C and E are more pleasant than A and B, and system D more pleasant than A. In the "Non Hi-Fi" group, however, systems B and E are both more pleasant than A, system B also more pleasant than D especially due to its high values for the piano and song programs. As regards the programs the piano program is rated significantly lowest in the "Non Hi-Fi" group, probably due to hissing/disturbances as discussed earlier.

Besides the effects of systems and programs there are also significant differences between listeners and significant interactions between systems and listeners and between programs and listeners as briefly discussed in connection with Experiment 1. (Statistical note: The correct error term for these factors is in fact MS_W. Since there is no MS_W in this experiment, MS_{SPL} was used as error term, which may introduce a certain bias in these statistical tests; see Gabrielsson 1979b.)

The index r_b for inter-individual reliability is high for both groups in "Loudness" and "Hissing/Disturbances". It is relatively high for the "Hi-Fi" group in most other scales, but rather or very low for the "Non Hi-Fi" group. As in Experiment 1 it was also found that for the "Non Hi-Fi" group differences between programs accounted for more of the variance in the data than did differences between systems. This was also the case for the "Hi-Fi" group concerning "Loudness" and "Hissing/Disturbances". However, the proportion of variance accounted for by differences between systems was higher for the "Hi-Fi" group than for the "Non Hi-Fi" group in all scales (but "Loudness"). It is also seen in Table VIII that the "Hi-Fi" group has significant differences between systems in all scales except "Fullness" and "Feeling of space" (and "Loudness" in which there should be no difference), while the "Non Hi-Fi" group has significant differences between systems only in two scales.

Relations between perceptual scales and evaluative scales

Correlation and regression analyses were performed in the same way as in Experiment 1, as well as the analysis of the ratings

concerning the "ideal" sound reproduction.

The last-mentioned ratings appear in the lower part of Table V. As in Experiment 1 there are a few examples that the ratings in a certain scale may differ among programs, for instance, as regarding "Softness" in both groups. On the whole the tendencies over programs and subject categories are similar to those found in Experiment 1 and are not repeated here. A different result compared to Experiment 1 is that the groups differ concerning "Nearness" and "Feeling of space": for the "Non Hi-Fi" group both scales have a high mean rating (8.2 - 8.6), while the corresponding means are considerably lower for the "Hi-Fi" group (5.3 - 6.3). In Experiment 1 the corresponding values for the "Hi-Fi" group were higher (6.9 - 7.6). The reason for this discrepancy is presently obscure.

The ranking of the eight perceptual scales with regard to their importance for "Fidelity" and "Pleasantness" appears in the lower part of Table VI. The rankings are similar to those in Experiment 1. However, in the present experiment both groups (also the "Non Hi-Fi" group) rank "Clearness/Distinctness" to be most important both for "Fidelity" and for "Pleasantness".

The different correlation/regression analyses are summarized in Table IX. "Clearness/Distinctness" is highly positively correlated with both evaluative scales as is "Feeling of space" in most cases. "Nearness" shows relativly high positive correlations to both evaluative scales in the "Non Hi-Fi" group but not in the "Hi-Fi" group. For "Fullness" the situation is about the opposite. "Brightness" is positively correlated to both evaluative scales for the "Hi-Fi" group, but has about zero correlation for the "Non Hi-Fi" group. For "Softness" there is a positive correlation for the "Non Hi-Fi" group (except for the original values) but around zero correlation for the "Hi-Fi" group. For "Loudness" there are rather low correlations, positive or negative. For "Hissing/Disturbances" finally there are in general high negative correlations with both evaluative scales. Although these results may seem complex, they are in most respects similar to those in Experiment 1.

			So	Cl	Fu	Ne	Br	Fe	Lo	Di	Mu
NON HI-FI	Original	Fidelity	-.19	.62	.45	.66	-.33	.73	.48	-.65	.87
		Pleasantness	-.10	.64	.46	.61	-.11	.64	.34	-.84	.93
	Adjusted	Fidelity	.33	.68	.01	.54	.08	.48	-.18	-.53	.85
		Pleasantness	.51	.62	.20	.38	.07	.50	-.27	-.85	.95
HI-FI	Original	Fidelity	.03	.84	.45	.16	42	.63	.31	-.06	.88
		Pleasantness	.21	.61	.35	-.25	.19	.39	.39	-.41	.82
	Adjusted	Fidelity	-.11	.78	.62	.27	.39	.77	.17	-.19	.92
		Pleasantness	.15	.51	.42	-.13	.14	.57	-.12	-.44	.88
NON HI-FI	----------	Fidelity	.59	.79	-.01	.49	.10	.26	.32	-.64	.88
		Pleasantness	.46	.64	-.03	.53	.42	.49	.40	-.85	.93
HI-FI	----------	Fidelity	.00	.82	.37	.38	.45	.40	.35	-.08	.87
		Pleasantness	.19	.54	.20	.04	.28	.08	.11	-.43	.81

TABLE IX. Product moment correlation between perceptual scales and the "Fidelity" and the "Pleasantness" scales in Experiment 2.
Upper matrix: from ratings on real reproductions.
Lower matrix: from differences between ratings on real and on "ideal" reproductions.
So:Softness Cl:Clearness Fu:Fullness Ne:Nearness Br:Brightness Fe:Feeling of space Lo:Loudness Di:Disturbances Mu:Multiple correlation

The appearance of the multiple regression functions is somewhat varying in the different cases. The first three variables entering into the function generally result in a multiple correlation of the order .80 - .90. Adding the remaining variables increases the multiple correlation a little bit to the figures shown in the right hand column of Table IX. The first variable entering into the function is in general "Clearness/Distinctness" for the "Hi-Fi" group, the same variable or "Hissing/Disturbances" for the "Non Hi-Fi" group.

Relations between perceptual scales

Factor analyses were performed in the same way as in Experiment 1. Again the results were rather varying. The examples given for Experiment 1 may apply here too.

DISCUSSION

Reliability and validity of the rating scales

The main purpose with this investigation was to check the suitability of the eight perceptual scales and the two evaluative scales with regard to ratings of perceived sound quality. This may be done by studying the reliability of the ratings in the respective scales, the capacity of the scales to differentiate between systems, and the relations between the perceptual scales and the evaluative scales.

Concerning reliability it is apparent in both experiments that the ratings of the "Hi-Fi" group are more reliable than for the "Music" group and, especially, for the "Non Hi-Fi" group. Previous experience of high fidelity sound reproduction and/or of listening to "live" music are obviously important factors for attaining good reliability. In certain scales, "Loudness" and "Hissing/Disturbances", the reliability was high for all groups. Of course, these scales are more familiar to most listeners than the other scales like "Softness" etc. However, a detailed analysis showed that the high reliability in these scales very much depended on obvious differences between the programs rather than between the systems. The same type of analysis also showed that the amount of variance accounted for by differences between systems was generally higher for the "Hi-Fi" group than for the other groups. Referring to proposed criteria for acceptable reliability (Gabrielsson 1979b), all "Hi-Fi" subjects fulfill these criteria but only half of the "Music" subjects and one or two of the "Non Hi-Fi" subjects.

Concerning the capacity of the scales to differentiate between systems this capacity is more or less demonstrated as seen directly in the matrices with mean ratings (Tables II and VII) and especially in the survey of statistically significant differences (Tables III and VIII). It is evident that the "Hi-Fi" group differentiated between the systems in more scales than the other groups did. Since there were also significant differences between programs, the scales are also able to differentiate between music programs/recordings. Of course they may also be said to differentiate between listeners with different listening experiences as the "Non Hi-Fi", "Hi-Fi", and "Music" groups here.

The validity of the perceptual scales with regard to overall evaluations in terms of "Fidelity" and "Pleasantness" may be studied by means of the correlations/regressions in Tables IV and IX, further in the ratings about the importance of the various perceptual scales (Table VI), and in the ratings concerning the "ideal" sound reproduction (Table V). The importance of certain scales is very obvious, especially for "Clearness/Distinctness"and (absence of) "Hissing/Disturbances". However, all scales seem more or less correlated with the evaluative scales (there are only some few examples approaching zero correlation, and these hold only for one of the groups of for a specific data treatment). In general there is a multiple correlation of the order .80 - .90 already for

two or three of the perceptual scales and often well above .90 when all eight scales are included. Which two or three variables enter first into the regression function varies in different cases. However, "Clearness/Distinctness", and "Hissing/Disturbances" often appear among those first variables.

The validity of the evaluative scales, "Fidelity" and "Pleasantness", is just taken for granted here but is supported by the data on their reliability and capacity to differentiate between different systems. "Fidelity" and "Pleasantness" are in general highly inter-correlated. The correlation is not quite perfect, however, as noted earlier in the text (for instance, "Pleasantness" requires more of "Softness" than what "Fidelity" does).

The above evidence points in positive direction as regards the suitability of the rating scales. However, there are indications that they could be improved in various respects. There are in general significant differences between different listeners in their mean position on each scale, and there are also often significant interactions between listeners and systems, listeners and programs, or between all these three factors. Although such results are not uncommon in various experiments involving judgments of some kind, they are signals to give further consideration to the definition of the scales and the grades within them. Most scales here were given no other definition than that given by its name and the labels attached to different scale steps. In the case of "Hissing/Disturbances" it may be necessary to give more detailed specification referring to various kinds of disturbances (possibly different subjects attach different weight to different types of disturbances). A related question is whether the rather low reliability of the "Non Hi-Fi" subjects could be improved by using other types of scales.

Differences between systems and listener categories

A comparison of differences between systems found in both experiments may be made as follows:

"Softness": In general system C or B is rated softest, while system A is least soft (sharpest).

"Clearness/Distinctness": In general one of systems C, D, or E is rated as the most clear/distinct, while A or B is the least clear.

"Fullness": In general system C is rated highest in fullness. The "Hi-Fi" subjects rate system B to have least fullness, while the result on this point varies for the other groups.

"Nearness": The "Hi-Fi" subjects rate system B to sound least near. Otherwise the results are varying.

"Brightness": In most cases system E or system A is rated as the brightest, while system B sounds darkest.

"Feeling of space": In most cases system C gives most feeling of space.

"Loudness": The attempt to equalize the loudness beforehand was mainly successful with regard to the "Hi-Fi" group (the persons doing this equalization may also be characterized as "Hi-Fi" subjects) but not quite for the other groups as described earlier under Results.

"Hissing/Disturbances": System A was generally rated as having most disturbances. There were interactions between programs and systems in that the disturbances were more noticed for programs with lower sound level and/or discretely spaced frequency content.

"Fidelity": There was a marked difference between the "Hi-Fi" and "Music" groups on one hand and the "Non Hi-Fi" group on the other hand. The first-mentioned two groups considered one of systems C, D or E to give the most "true-to-nature" reproduction, while system A and B were clearly inferior. The "Non Hi-Fi" group, however, considered system B to be at least as good as systems C - E (possibly even better). Since system B is probably the system which sounded most similar to the type of listening equipment that the "Non Hi-Fi" subjects were used to hear, this result is a striking evidence for the importance of earlier listening experiences. This has been demonstrated before, for example, already by Kirk (1956) and Kötter (1968).

"Pleasantness": The same difference in results between the "Hi-Fi" and "Music" groups on one hand and the "Non Hi-Fi" group on the other hand holds also here. The preference for system B among the "Non Hi-Fi" subjects is even more marked in this scale.

As seen in Tables II and VII the systems rated highest in "Fidelity" by the respective groups get mean ratings (averaged over programs) in the range 6.1 - 7.2, that is, usually somewhat below the rating as "good" (=7) in the "Fidelity" scale (see Figure 3). The highest ratings occurring for single program x system combinations lie within 7.1 - 8.0, thus somewhat above the "good" position. In the "Pleasantness" scale the corresponding values are generally somewhat lower: the highest mean ratings (averaged over programs) lie within 5.5 - 6.4, well below the rating as "pleasant" (=7), and the highest ratings for single program x system combinations within 6.9 - 7.9.

The relations of the perceptual and evaluative scales to the physical characteristics of the different systems are no primary concern in this investigation. Some points may be suggested referring to the data on frequency response and non-linear distortion given in Figure 1. The (relative) "sharpness" of system A may be related to high non-linear distortion at higher frequencies, while the "softness" in systems B and C may partly depend on falling frequency response towards higher frequencies (as in B) or a flat frequency response and low distortion (as in C). The higher "clearness/distinctness" in systems C - E is probably related

to their broader frequency response and/or less non-linear distortion as compared to systems A and B. The same properties are probably of importance for explaining the ratings concerning "Fidelity" and "Pleasantness", at least for the "Hi-Fi" and "Music" groups. However, with regard to the "Non Hi-Fi" group the narrower frequency range and the falling frequency response of system B may be important, since this undoubtedly reduces the amount of perceptible hissing and similar disturbances and thus makes the reproduction more "true-to-nature" and/or "pleasant" in this respect.

With regard to the prices of the systems it is noted that the "Hi-Fi" and "Music" subjects usually prefer more expensive systems (C, D or E; however, not A), while the "Non Hi-Fi"subjects thus prefer the cheapest system. System C can hardly be said to be preferred to D or E in spite of its considerably higher price. It should be remembered, however, that the positions of the loudspeakers were not optimal (see Stimuli and listening conditions), and far-reaching conclusions should thus be avoided.

Comparison of the results in both experiments

In Experiment 1 the subjects rated each stimulus three times and made the ratings in the perceptual and in the evaluative scales separately from each other. In Experiment 2 each stimulus was considerably longer in duration but was rated only once and with the perceptual and evaluative scales given at the same time. A detailed comparison of the results in the two experiments was made with regard to differences between systems and relations between perceptual and evaluative scales, separately for the "Non Hi-Fi" group and the "Hi-Fi" group. The differences between systems found in both experiments may be compared by means of the mean ratings over programs given in Tables II and VII, by the information about significant differences between systems given in Tables III and VIII, and by simply comparing the rank order of the systems within each scale in both experiments (that is, the rank order of the mean ratings for the systems over programs as given in Table II and VII). The combined evidence of these comparisons may be summarized as follows:

a) The absolute difference between corresponding mean ratings for systems in both experiments are in general small. Differences exceeding 1.0 unit occur in only 2 cases out of 50 for the "Non Hi-Fi" group and in 12 cases out of 50 for the "Hi-Fi" group. Most of these last-mentioned 12 cases occur for the "Fidelity" and "Pleasantness" scales: the mean ratings for the "Hi-Fi" group in Experiment 1 lie generally about 0.8 - 1.8 units higher than for the "Hi-Fi" group in Experiment 2. The full reason for this is not quite clear. A detailed analysis showed, however, that there were two "high-raters" among the "Hi-Fi" subjects in Experiment 1, whose mean ratings on those two scales were about 2.0 - 3.0 units higher than for the other members in this group.

b) It is more important to compare which significant differences there were between the systems in each experiment and the

rank order of the systems in each experiment. As regards the "Hi-Fi" group the pattern of significant differences and the rank orders of the systems are clearly similar in both experiments (Spearman's rank order correlation coefficient varied between 0.67 to 0.90 over all ten scales). For the "Non Hi-Fi" group, however, there are considerable differences in both these respects between the two experiments. There were significant differences between systems in eight of the ten scales in Experiment 1 but only in two scales in Experiment 2, and the rank orders of the systems in the different scales often vary considerably between these two experiments (Spearman's rank order correlation coefficient varied between 0.15 to 0.70 over all ten scales).

A related distinction between the two listener groups also appears in a comparison of the inter-rater reliabilities in the two experiments, see Table III and VIII (the $\underline{r}_b$ index). It can be expected for purely statistical reasons that the $\underline{r}_b$ values should be lower in Experiment 2 than in Experiment 1 (due to the way of computing this index, see Gabrielsson 1979b). It is apparent, however, that this reduction of $\underline{r}_b$ in Experiment 2 is more obvious for the "Non Hi-Fi" group than for the "Hi-Fi" group.

As regards differences between systems the conclusion is thus that with experienced listener groups, such as the "Hi-Fi" groups here, the procedures used in Experiments 1 and 2 give similar results. With "Non Hi-Fi" groups, however, there may be rather big differences in results.

It is possible, of course, that the differences in results for the "Non Hi-Fi" groups would be decreased by using a bigger number of subjects, especially in the group doing only one rating per case (Experiment 2). Conversely, there would probably have appeared bigger differences in the results for the "Hi-Fi" groups, if the number of subjects had been less than the seven ones used here. It is always an advantage to have a "big" number of observations, either in terms of number of subjects or in terms of repeated ratings of the same stimulus by each subject. The alternative with repeated ratings offers certain statistical advantages for significance testing and for studying the intra-individual reliability. On the other hand repeated ratings may be boring and lower the motivation of the subjects. According to recommendations in Gabrielsson (1979b) there should be at least four subjects doing at least two ratings per case to satisfy various statistical criteria in listening tests. If the subjects do only one rating per case the number of subjects should be doubled (that is, at least eight subjects, thus seven subjects as in Experiment 2 here is a little below this limit). However, as also said in the same reference, statistical criteria should be supplemented with other relevant knowledge about the reliability of the available subjects, the characteristics of the systems and programs in the test etc. Continued experience of listening tests will generally provide the investigator a safer basis for decisions about the proper design of a test. The evidence from this investigation says

that with the present context of programs and systems it may be enough with seven experienced subjects doing one rating per case. Continued work is recommended to broaden the basis for more general conclusions.

With regard to the relations between the perceptual scales and the evaluative scales, these relations are very similar in both experiments as seen by comparing the results given in Tables IV and IX. There are some differences as regards "Nearness" and partly regarding "Fullness" and "Feeling of space". It may thus be tentatively concluded that it does not matter much whether the ratings in the perceptual and the evaluative scales are made simultaneously together of if they are separated.

Concluding comments

The results of this investigation should be considered in relation to its context of certain selected programs and systems. Other programs and other systems should be used in future experiments. Two especially important points in the continued work are the question about the positions of the loudspeakers in the listening room and questions about possible revisions of the rating scales.

ACKNOWLEDGEMENTS

The authors want to express their gratitude to Lennart Persson, Ann-Cathrine Lindblad, Sven Jalmell, Styrbjörn Sjögren, Harry Johnsson, Björn Hagerman, and Lennart Fahlén for much assistance in the setup and data treatment of the experiments, and to Gerhard Elger for valuable discussions and much active help in the planning. The investigation was initiated and supported by The National Swedish Board for Consumer Policies.

REFERENCES

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

Gabrielsson, A. (1979b). Statistical treatment of data from listening tests on sound-reproducing systems. Reports from Technical Audiology, Karolinska Institutet, Stockholm No. 92.

Gabrielsson, A., Rosenberg, U. & Sjögren, H. (1971). Judgments and dimension analyses of perceived sound quality of sound-reproducing systems. I. Reports from the Psychological Laboratories, University of Uppsala, No. 115.

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

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

Gabrielsson, A. & Sjögren, H. (1979b). Perceived sound quality of hearing aids. Scand Audiology 8, 159-169.

Gorsuch, R.L. (1974). Factor analysis. W.B. Saunders, Philadelphia.

Hays, W.L. (1973). Statistics for the social sciences (2nd ed.). Holt, Rinehart & Winston, New York.

Kirk, R.E. (1956). Learning, a major factor influencing preferences for high fidelity reproducing systems. J Acoust Soc Am 28, 1113-1116.

Kirk, R.E. (1968). Experimental design. Procedures for the behavioral sciences. Brooks/Cole, Belmont, California.

Kötter, E. (1968). Der Einfluss übertragungstechnischer Faktoren auf das Musikhören. Arno Volk, Köln.