

Differences in Performance and Preference of Trained versus Untrained Listeners in Loudspeaker Tests: A Case Study*

Sean E. Olive, *AES Fellow*

Research & Development Group, Harman International Industries, Inc., Northridge, CA, 91329, USA

Listening tests on four different loudspeakers were conducted over the course of 18 months using 36 different groups of listeners. The groups included 256 untrained listeners whose occupations fell into one of four categories: audio retailer, marketing and sales, professional audio reviewer, and college student. The loudspeaker preferences and performance of these listeners were compared to those of a panel of 12 trained listeners. Significant differences in performance, expressed in terms of the magnitude of the loudspeaker F statistic F_L , were found among the different categories of listeners. The trained listeners were the most discriminating and reliable listeners, with mean F_L values 3–27 times higher than the other four listener categories. Performance differences aside, loudspeaker preferences were generally consistent across all categories of listeners, providing evidence that the preferences of trained listeners can be safely extrapolated to a larger population. The highest rated loudspeakers had the flattest measured frequency response maintained uniformly off axis. Effects and interactions between training, programs, and loudspeakers are discussed.

0 INTRODUCTION

Unfortunately, controlled listening tests in audio products are seldom performed by audio manufacturers, retailers, and the audio review press. The most common excuse is that the tests are too time consuming, expensive, or difficult to conduct. Among the few organizations that routinely perform controlled listening tests, it is common practice to use a small panel of highly trained expert listeners [1]–[6] on the basis that they are more reliable and discriminating in their judgements of various attributes of sound quality and preference. For example, Bech has reported that one trained listener can yield the equivalent statistical confidence of using seven untrained listeners [1]. If this is true, training listeners can save considerable time and money over the long term.

In sensory measurements of consumer products (such as food and wine), subjects must be highly trained to perform reliably a complex descriptive analysis of various perceptual attributes of the products being tested [7]. On the other hand, for preference testing, naive or untrained subjects representative of the targeted customer are pre-

ferred. This is because preference for most consumer products is influenced by demographic and socioeconomic factors [7]. Some audio marketing departments have argued that the same rationale for testing chocolate and chardonnay should be applied to preference testing of audio products, their underlying assumption being that different demographics have different tastes in sound quality. Another legitimate concern is that the training process itself may inherently bias listeners' preferences. They may become conditioned to prefer certain types of loudspeakers based on how they are trained and rewarded. Costly sonic improvements valued by a critically trained ear may be unappreciated by the average untrained listener. The money might be better spent on improving the product's cosmetics or increasing the marketing and advertising budgets. These are all valid arguments that challenge the wisdom and rationale for using trained listening panels for preference testing.

A third approach in selecting listeners is to solicit the opinions of the audio retailers who sell the products and the reviewers who write about them based on the assumption that their opinions largely determine a product's commercial success. The problem with this approach is that unless their opinions can be measured under the same controlled listening conditions, there is little chance that they will agree with themselves or with each other. In other

*Presented at the 114th Convention of the Audio Engineering Society, Amsterdam, The Netherlands, 2003 March 22–25; revised 2003 June 19.

words, the opinions on sound quality are biased by the influence of a number of nuisance variables [8], which include visual and psychologically related bias (such as size, brand, price, cosmetics) [9], listening-room acoustics [10]–[13], and loudspeaker placement [14], [15].

All of these arguments have been largely untested or supported with scientific data. With this in mind, a study was designed to answer the following hypothetical question. To what extent do the loudspeaker preferences of a group of untrained listeners, measured under identical listening conditions, agree with those of an expert panel of trained listeners? To answer this question we measured the loudspeaker preferences of 256 listeners with little or no formal training or experience in judging sound quality under controlled listening conditions. The untrained listeners included audio marketing and sales people, retailers, audio reviewers, and college students. The measured variations in loudspeaker preference and performance between the groups were compared to those of a panel of trained listeners. This study addresses directly some of the untested arguments against using trained listeners, which are essentially that their preferences are too biased and unrepresentative of a naive, untrained listener.

1 PREVIOUS WORK

To the best of the author's knowledge, no scientific studies have investigated the sound-quality judgements of trained listeners and compared them to those made by audio retailers, professional reviewers, and untrained listeners. However, a few studies have compared the judgements of experienced listeners versus inexperienced listeners and examined how training and hearing performance affect the listener's reliability in judging sound quality.

Kirk, in 1956, was one of the first to report the effects of listening experience and learning on loudspeaker bandwidth preferences among 210 college students [16]. He found that preference was dictated by the quality of the reproduction systems the student most commonly experienced. Most students preferred a severely band-restricted loudspeaker, and preferred the wider bandwidth loudspeakers only after repeated exposure to them over 6.5 weeks. By today's standards, these tests were not very well controlled. Besides the questionable linearity of the loudspeaker used, the recordings and phonographs were likely major sources of noise and distortion, which undoubtedly would have been the least audible and annoying on the band-limited loudspeaker.

More extensive work has been done by Gabrielsson and his colleagues [17]–[19], who investigated 12 listeners' judgements of sound quality among five different loudspeakers [17]. These 12 subjects were divided into three categories: listeners in general (L), musicians (M), and "hi-fi" subjects (H). All listeners had normal hearing and ranged in ages from 23 to 41. Listeners gave ratings based on loudspeaker fidelity, similarity, and various verbal descriptions. The reliability of ratings among groups was generally high, although the inexperienced group L tended to be less reliable than the other two groups and generally

awarded higher ratings to the poorer loudspeakers. All groups tended to vary in the weightings applied to certain dimensions. The experienced listeners gave greater weight to the "brightness" dimension compared to the inexperienced listeners, who gave more weight to "loudness."

Toole conducted a large-scale series of listening tests that involved 42 listeners and 37 loudspeakers over a period of 2 years [20], [21]. The ratings were given on a 0–10 point interval fidelity scale. This was the first large-scale study that examined the effect of hearing loss on the repeatability of the listener. As the mean hearing threshold below 1 kHz increased, the listeners' standard deviations in responses increased. Listeners with normal hearing had standard deviations of less than 1 interval.

Bech further explored the effects of hearing loss, listening experience, and training on a listener's ability to rate the fidelity of four different loudspeakers reliably [1]. He found no clear correlation with hearing loss and standard deviation in ratings, although he noted that his subjects were on average younger and all had normal hearing (<15 dB HL) compared to Toole's subjects. To explore the effects of training, Bech repeated a loudspeaker test (four loudspeakers and four programs) six times using 12 inexperienced listeners. He found that 65% of the subjects reached an asymptotic performance after only four listening sessions based on the magnitude of their error variance and their individual loudspeaker F statistic, which is defined hereafter. The remaining subjects reached their peak performance after seven to eight listening sessions. The difference in performance between a trained and an untrained listener seemed to disappear after about four to eight listening sessions. This finding agrees with the 50+ combined years of loudspeaker listening test experience of the author and Toole. Similar training effects have been reported using various computer-based listener training programs [3], [4], [22].

The issue of which metric is best for measuring listener performance has been examined in depth by Gabrielsson [23] and more recently by Bech [1]. While the use of standard deviations in response ratings represents the repeatability of a listener accurately, it fails to measure the effect size or the ability to discriminate among loudspeakers. For example, a listener using a very small range (for example a 0.5 rating) produces a low standard deviation score. The listener who recognizes the sonic signature of the loudspeakers, uses a larger scoring range, and replicates their ratings perfectly will also produce a standard deviation of 0. In this case the second listener is more useful to the experimenter because he is as reliable as the first subject but more discriminating. However, his usefulness is somewhat compromised if the judgements are no longer independent due to product recognition.

Another performance metric proposed by Gabrielsson et al. [17] is the intraindividual reliability index MS_w , which represents normalized within-cell error variance calculated from an individual analysis of variance (ANOVA). Gabrielsson noted that the reliability index value varies significantly depending on the attribute. The most reliable ratings were 0.53 (loudness) with the fidelity being 1.29. Ratings on spatial attributes were among the

least reliable (spaciousness 1.64, nearness 1.48). One benefit in using this metric is that it accounts for variance in a listener's ratings caused by other factors such as program and its interactions with loudspeakers.

However, if the main focus of interest is the effect of the loudspeaker on preference ratings, Bech argues that the individual loudspeaker F statistic F_L is a better choice [1]. F_L is the ratio of the loudspeaker effect (mean sum of squares for loudspeaker ratings) divided by the error variance (mean sum of squares of the residual). This metric accounts for the listeners' ability to discriminate between loudspeakers as well as their ability to repeat their ratings, expressed in the denominator. In the current study, listener performance is based on the magnitude of the loudspeaker F statistic F_L . The author uses this metric for selecting the best listeners based on their performance in various training tasks [5] and day-to-day performance in preference testing of audio products.

One of the issues with using F_L is the problem that occurs when the error variance is 0, resulting in an undefined value due to division by 0. This happens when a listener replicates his or her ratings perfectly in every trial. In this paper the author arbitrarily assigned a maximum F_L value of 2000. Only 16 of the 268 listeners (6%) achieved a 0 error variance, all occurring in the three-way loudspeaker test.

2 EXPERIMENTS

This section describes the experimental design of the tests, including the selection of loudspeakers, programs, listeners, physical test setup, and the experiment protocol.

Two different tests were repeated over the course of 18 months, involving a total of 268 different listeners and 36 listening groups. A total of 5,256 preference ratings were measured. The categorization and details of the different listening groups are discussed in section 2.8.

The two tests are referred to hereafter as the four-way test and three-way test. The four-way test involved multiple comparisons among four loudspeakers rated independently using four different programs. The tests comprised four trials conducted in the morning, followed by a repeat of the test in the afternoon for a total of eight trials. In the three-way test loudspeaker I was dropped from the test, and there were no repeats. Otherwise the two tests were identical in all aspects, including program material, playback level, and seating arrangements. One confounding factor was that the seating was not included as a variable in the design of the experiment. The 12 trained listeners listened alone, seated in the front row directly on axis to the loudspeakers, whereas the 256 untrained listeners

Table 1. Codes and descriptions of loudspeakers.

Loudspeaker Code	Description	MSRP (approx.)
B	Three-way dynamic	\$8,000
P	Four-way dynamic	\$10,000
M	Electrostatic/dynamic	\$11,000
I	Four-way dynamic	\$5,000

were assigned randomly to one of eight seats arranged in two rows. This would explain some of the differences in preference and intergroup reliability between the trained and untrained listeners, but not necessarily differences in their individual performances.

2.1 Loudspeakers and Measurements

The four loudspeakers used in both tests are shown in Table 1. Each loudspeaker is coded with a letter, since the brand name and model were not relevant to the aims of this study. The manufacturer's suggested retail price per pair (MSRP) ranges from approximately \$5000 to \$11,000. The loudspeakers were chosen because they are all widely available and compete against each other in the marketplace. Given the relatively high prices of the loudspeakers, they should in theory represent "state-of-art" designs in terms of technical and sonic performance. Indeed, all four models have received high accolades and recommendations from the audiophile press. In one magazine, two of the models (P and M) have received the highest performance category status possible (class A) for the past three years, and loudspeaker M was declared a "product of the year."

The loudspeaker measurements are shown in Appendix 1 (Fig. 9). Each loudspeaker was measured in the large Harman anechoic chamber at a distance of 2 m with 2-Hz frequency resolution. The chamber is anechoic down to approximately 60 Hz and has been calibrated down to 20 Hz. For each loudspeaker the set of curves represent (from top to bottom) the on-axis response, the spatially averaged ($\pm 30^\circ$ horizontal, $\pm 10^\circ$ vertical) listening window, the average early reflected sounds, and the calculated sound power response. The lower two curves represent the directivity indices derived from the early reflected sound and the total radiated sound power. Details of the anechoic chamber and measurement procedure are available in [24]. A discussion of these measurements and their correlation with listeners' preferences is presented in Section 3.13.

2.2 Program Selections

Table 2 lists the four program selections used in these tests. Each program was a short 20–30-second loop digitally extracted from a compact disc as a 16-bit, 44.1-kHz stereo wav format file. The programs were selected on the basis of their ability to reveal spectral and preferential dif-

Table 2. Program selection used in tests.

Program Code	Artist, Track, and Album
JT	James Taylor, "That's Why I'm Here" from "That's Why I'm Here," Sony Records.
LF	Little Feat, "Hangin' on to the Good Times" from "Let It Roll," Warner Brothers.
TC	Tracy Chapman, "Fast Car" from "Tracy Chapman," Elektra/Asylum Records.
JW	Jennifer Warnes, "Bird on a Wire" from "Famous Blue Rain Coat," Attic Records.

ferences between different loudspeakers in over 100 different loudspeakers in over 100 different listening tests and various listener training exercises.

2.3 Preference Scale

In each listening test, listeners were required to rate each loudspeaker on the interval preference scale defined in the listener instructions (see Appendix 2). The scale consists of 11 points ranging from 0 to 10, where the magnitude of the rating indicates the degree to which the listener likes or dislikes the sound quality of a loudspeaker. The distance between two loudspeaker ratings represents the magnitude of preference: separations of 2 or more points indicate a strong preference for the higher rated loudspeaker; 1 point difference, a moderate preference; and a 0.5 point difference represents a slight preference. These definitions were intended to encourage listeners to use the scale in a similar manner and to help make the scale linear, so that equal distances of separation between two loudspeakers imply the same thing no matter what part of the scale is used. Listeners were instructed not to give tied ratings.

2.4 Listening Room

All of the listening tests were conducted in the multi-channel listening lab (MLL) located at Harman International in Northridge, CA. The physical and acoustical characteristics of the listening room and its special features have been described extensively in [25]. Since the publication of this document, the walls and ceiling of the room have been finished with standard gypsum board to better simulate the surfaces found in domestic homes.

One unique feature of this listening room is the automated loudspeaker shuffler. The device permits fast (~3 second) positional substitution of four mono, stereo, or left-center-right sets of loudspeakers and effectively eliminates loudspeaker position as a variable in the test. This is important since the effect of loudspeaker position on the perceived sound quality has been shown to be a significant variable, at times larger than the effect between two different loudspeakers [14], [15].

Another benefit accrued from the loudspeaker shuffler is that the judgments of loudspeakers between trials or different program selections are truly independent. The control computer automatically shuffles the loudspeakers between trials and randomly assigns a letter code (A through D) to each loudspeaker. For multiple-comparison loudspeaker tests that do not employ a loudspeaker shuffler, the positions of the loudspeakers should be randomized between trials to reduce bias. The effects of this bias on the results of nonshuffled loudspeaker tests have been raised by Jason [26], [27]. Given that loudspeakers can be recognized easily or identified by differences in their physical positions in the room, there is an increased likelihood of measuring artificially high individual listener F_L values. Bech has argued the opposite, saying that the loudspeaker-position interactions are likely to increase the variations in a listener's ratings each time the positions of the loudspeakers are swapped [1]. Clearly having a loudspeaker shuffler eliminates the need to sort out the various

effects and biases that loudspeaker positions have on subjective ratings.

All control of the equipment including the switching of audio signals and loudspeakers, was computer automated through custom software. In these tests, listeners were required to enter their responses on a standard listening test form using a pencil. The ratings were later entered manually into the database server by the experimenter. This labor-intensive process has recently been eliminated by giving each listener a personal digital assistant device (PDA) that is wirelessly networked to the database server and control computer. In this way, listener data input and storage is completely automated and monitored. Real-time analysis of the data is also possible.

2.5 Playback Equipment

The program signals were reproduced from the hard disk on the control computer equipped with a digital sound card (SEK'D ProDif 96). The AES-EBU signal was fed to a digital switcher-distributor (Spirit 328 digital mixer) and converted to four analog signals using an eight-channel Studer D19 digital-to-analog converter. Precise level matching between loudspeaker was done by adjusting the trim controls on each analog output. Each loudspeaker was amplified with a Proceed AMP3 amplifier.

2.6 Level Adjustment

Each loudspeaker was level matched to within 0.1 dB (B-weighted) using pink noise fed to each loudspeaker. The calibrated microphone (AKG-CK62) was positioned at ear height over the middle front-row chair. Levels were calculated using SpectraLAB (version 4.32). The average playback level of the program selections in the listening room was 75 dB (B-weighted).

2.7 Test Procedure

All tests were performed double blind using monophonic (single-loudspeaker) comparisons. Before each test, listeners were given their instructions and were free to ask questions about the test procedure.

In both tests the program order was randomized. For each trial the control computer determined randomly the letter (A through D) assigned to each loudspeaker. Listeners were provided feedback through an LCD monitor that indicated the current loudspeaker being played.

Switching between loudspeakers in each trial was performed in a random sequence by the experimenter. The music was paused during the 3-second interval required to substitute the positions of the loudspeakers. Although the effect of this silent interval on the loudspeaker tests has not been investigated, different studies have shown that increasing the interstimulus interval impairs listeners' discrimination of pitch, loudness [28], and timbre [29]. While decreasing the interstimulus time gap is advisable for measuring much smaller audible differences (such as different high-quality audio codecs) the author has not found the 3-second time gap to be limiting factor in measuring loudspeakers. In fact Toole found that using positional substitution of loudspeakers in both stereo and monophonic comparisons lead to a lower error variance in the

listeners' ratings, despite the fact the method increased the interstimulus interval from almost 0 to 5 seconds [30]. The benefits of controlling the loudspeaker positional biases clearly outweigh any effects that result from increasing the interstimulus time interval to a few seconds.

The presentation time for each loudspeaker was typically equal to the length of the program loop (15–30 seconds) and shortened to 10–15 seconds toward the end of each trial. Switching continued until all listeners had entered a rating for each loudspeaker, at which point the next trial would begin. A trial typically lasted 3–5 minutes, with an entire session typically lasting 15–20 minutes.

For the four-way test listeners were told not to discuss their responses with one another until the end of the second session. All listeners were shown their results after the completion of the test.

2.8 Listeners

The 268 listeners were categorized according to their occupations (see Table 3). The table shows the number of listeners in each category and the percentage of listeners based on the total number. Note that the number of listeners in each category was not balanced. This is because the recruitment and selection of the untrained listeners was not a factor controlled by the experimenter. The listeners were all guests invited by the various Harman-brand marketing groups, including the retailers and audio reviewers. The students were unsolicited guests who were interested in visiting the facility. Factors such as age, gender, years of audio experience, and hearing loss were not measured or controlled, except for the trained listeners.

The first category (AR) was comprised of 215 audio equipment retailers, ranging from small privately owned boutiques to large audio retail chains located across North America. This group represented by a wide margin the largest percentage of the total listeners (80.2%).

The second group (S) consisted of 14 university students from two California universities. One group (CALP) consisted of undergraduate electrical/mechanical engineering students with an interest in audio engineering. The other student group (UC) was enrolled in programs preparing them for careers in music and recording industries. Based on personal observations it would be safe to say that the student group was the youngest group in this study and had the least amount of experience judging the sound quality of loudspeakers. As a group they represent 5.2% of the total sample size.

The third group (MS) consisted of field marketing and

sales people within Harman Consumer Group (HCG) and JBL Professional (JBL). This group had relatively more professional audio experience in evaluating sound quality compared to the students. However, none were members of the Harman-trained listening panel, and they had little experience in controlled listening tests. This group consisted of 21 listeners, or 7.8% of the sample size.

The fourth group (PR) consisted of six professional audio reviewers who review products for some of the most popular audio and home theater trade magazines. These members had considerable experience evaluating the sound quality of audio products but not necessarily under controlled listening test conditions.

The final group (T) included 12 members of the Harman-trained listening panel, including one trainee with broad-band hearing loss (mean of 38 dB HL between 250 Hz and 8 kHz) in one ear caused by a genetic mechanical defect in the ossicles of the middle ear. Post-hoc analysis of this listener's test results showed perfect (–1.0) negative correlation between loudspeaker preferences in the four-way and three-way tests. In other words, he completely reversed his order of loudspeaker preferences between tests, supporting Toole's finding that listeners with hearing loss are less reliable [20], [21]. This listener was not included in the final results. All other listeners were audiometrically normal (<15 dB HL at all audiometric frequencies between 250 Hz and 8 kHz) and had completed listener training successfully. All had participated in numerous controlled loudspeaker listening tests with experience ranging from 2 to 17 years. The mean age for the trained listeners was 36 years, ranging from 25 to 43 years.

3 RESULTS

In this section the results of the two listening tests are presented and discussed.

3.1 Statistical Analysis

The results of the four-way and three-way tests were analyzed separately using a repeated-measures analysis of variance (ANOVA). In both tests the dependent variable was preference rating.

The four-way test was analyzed as a 16 × 4 × 4 × 2 design where the within-subject fixed factors included loudspeaker (4 levels), program (4 levels), and session (2 levels; morning and afternoon). The between-subjects factor, group (16 levels), is a nominal variable representing the 16 different groups of listeners that participated in the test.

The repeated measures ANOVA for the three-way test consisted of a 20 × 3 × 4 design that included the between-subjects factor, group (20 levels), and the within-subject fixed factors, loudspeakers (3 levels) and program (4 levels). There was no afternoon repetition of the test so session was not a variable. The differences in the total numbers of listeners and listening groups between the two tests were not factors controlled by the experimenter. The selection and pooling of the 256 untrained listeners were based on the availability of guests invited to participate by various marketing departments within the company. A

Table 3. Category and number of untrained listeners according to occupation.

Occupation	Code	Count	Percent of Total
Audio retailers	AR	215	80.2
University students	S	14	5.2
Marketing and sales	MS	21	7.8
Audio reviewers	PR	6	2.2
Trained Harman Listeners	T	12	4.5
Total		268	100

complete factorial analysis was used in the ANOVA model with a significance level of 0.05 for all statistical tests. Inspecting the distribution of the mean loudspeaker ratings we found that the means were relatively normal and symmetrical except in the four-way test, where the means for loudspeakers P and I were negatively skewed, and the ratings for loudspeaker M were somewhat positively skewed. The deviations from normality were likely related to the untrained listeners' tendency to use the entire range of the preference scale, including the extreme end points. The tendency for subjects to use the entire scale in psychophysical judgement is described in Parducci's range-frequency theory [31]–[38]. Over many judgements, subjects tend to use all available categories defined on the scale an equal number of times. When closing in or spreading out the overall range of products, subjects will map their experience onto the available categories. Distortions in the scale tend to decrease as the range, number, and frequency of the stimuli increase. These three factors also influence the extent to which the ratings are biased by contextual effects (see Section 3.5).

Given that ANOVA is quite robust to deviations from normality, particularly when the sample size is quite large,

the probability of committing a type I error was considered remote. As a safeguard, a nonparametric analysis (both Friedman and Wilcoxon signed rank post-hoc tests) was performed on the data, and this led to the general results and conclusions found in the ANOVA tests.

3.2 ANOVA Summary Tables

Appendix 3 gives the ANOVA summary tables for the four-way and three-way tests. Table 4 as well as the Scheffe post-hoc test summary tables for the variable, loudspeaker (Table 5). In the following discussion the F value for each effect is given in the form

$$F(\text{DF}_{\text{source}}, \text{DF}_{\text{residual}}) = x, p \quad (1)$$

where $F()$ is the F statistic expressed as number x , $\text{DF}_{\text{source}}$ is the degrees of freedom of the factor or variable or interaction, $\text{DF}_{\text{residual}}$ is the degrees of freedom in the residual, and p is the level of significance.

3.3 Practical Significance and Effect Size

In practice, significant effects can be easily achieved in listening tests by using a very large number of listeners.

Table 4. ANOVA table for preference.

	DF	Sum of Squares	Mean Square	F Value	P Value	Lambda	Power
1. Four-way test							
Group	15	1969.183	131.279	5.185	<0.001	77.779	1.000
Subject (Group)	86	2177.316	25.318				
Session	1	1.072	1.071	1.048	0.3088	1.048	.164
Session * Group	15	13.571	0.905	0.884	0.5836	13.263	0.530
Session * Subject (Group)	86	87.993	1.023				
Program	3	9.710	3.237	4.689	0.0033	14.045	0.903
Program * Group	45	28.743	0.639	0.925	0.6109	41.645	0.919
Program * Subject (Group)	258	178.067	0.690				
Loudspeaker	3	9496.861	3165.620	230.954	<0.0001	692.863	1.000
Loudspeaker * Group	45	675.109	15.002	1.095	0.3256	49.254	0.966
Loudspeaker * Subject (Group)	258	3536.325	13.707				
Session * Program	3	1.702	0.567	0.920	0.4318	2.760	0.244
Session * Program * Group	45	19.747	0.439	0.711	0.9153	32.016	0.795
Session * Program * Subject (Group)	258	159.131	0.617				
Session * Loudspeaker	3	5.875	1.958	0.585	0.6252	1.755	0.167
Session * Loudspeaker * Group	45	121.377	2.697	0.806	0.8063	36.270	0.860
Session * Loudspeaker * Subject (Group)	258	863.383	3.346				
Program * Loudspeaker	9	107.217	11.913	3.816	<0.0001	34.342	0.996
Program * Loudspeaker * Group	135	654.398	4.847	1.553	0.0002	209.608	1.000
Program * Loudspeaker * Subject (Group)	774	2416.437	3.122				
Session * Program * Loudspeaker	9	44.929	4.992	1.866	0.0539	16.791	0.831
Session * Program * Loudspeaker * Group	135	378.575	2.804	1.048	0.3489	141.486	1.000
Session * Program * Loudspeaker * Subject (Group)	774	2070.992	2.676				
2. Three-way test							
Group	19	702.353	36.966	4.206	<0.0001	79.919	1.000
Subject (Group)	146	1283.095	8.788				
Program	3	16.025	5.342	9.989	<0.0001	29.968	0.999
Program * Group	57	30.207	0.530	0.991	0.4975	56.492	0.979
Program * Subject (Group)	438	234.208	0.535				
Loudspeaker	2	2234.033	1117.017	149.233	<0.0001	298.466	1.000
Loudspeaker * Group	38	722.270	19.007	2.539	<0.0001	96.495	1.000
Loudspeaker * Subject (Group)	292	2185.634	7.485				
Program * Loudspeaker	6	44.295	7.382	4.244	0.0003	25.465	0.986
Program * Loudspeaker * Group	114	243.610	2.137	1.229	0.0623	140.051	1.000
Program * Loudspeaker * Subject (Group)	876	1523.752	1.739				

However, does statistical significance have any important practical consequence when the difference in preference between two loudspeakers is less than a 0.5 rating, a slight preference?

The 5th edition of *Publication of the American Psychological Association* (2001) states: "...for the reader to fully understand the importance of your findings, it is almost always necessary to include some [measure of practical significance such as an] index of effect size or strength of relationship" [39].

According to Hurlburt, there is no universally used method for reporting the effect size in an experiment [39]. The raw effect size is the magnitude of an experimental result measured on the same scale used in the experiment. In this study the maximum raw effect size was 4.2 preference ratings for the variable loudspeaker (a very strong preference) and 3.4 preference ratings for the variable listening group. The effect size index d is a unitless measure that expresses the magnitude of an experimental result, and is widely used in many scientific journals. In repeated-measures tests, where more than two loudspeakers are compared, Hurlburt recommends calculating the maximum effect size index d_M as follows:

$$d_M = D_{\max} / \sqrt{2 * MS_{\text{residual}}} \quad (2)$$

Table 5. Scheffe table for preference; variable, loudspeaker.

	Mean Difference	Critical Difference	P Value	
1. Four-way test				
B, P	-1.917	0.302	<0.0001	S
B, M	2.382	0.302	<0.0001	S
B, I	-1.581	0.302	<0.0001	S
P, M	4.300	0.302	<0.0001	S
P, I	0.336	0.302	0.0214	S
M, I	-3.963	0.302	<0.0001	S
2. Three-way test				
B, P	-1.113	0.252	<0.0001	S
B, M	1.865	0.252	<0.0001	S
P, M	2.979	0.252	<0.0001	S

Note: Significance level 5%.

where D_{\max} is the largest raw difference in the means found between different levels of the independent variable, and MS is the residual mean sum of squares in the ratings. Fig. 1 shows the maximum effect size indices d_M calculated for each of the three independent variables (listening group, loudspeaker, and program) in both tests. According to Cohen [40], d_M values of 0.2, 0.5, and 0.8 represent small, medium, and large effects, respectively. In both the four-way and three-way tests the variable loudspeaker had a large effect on the preference ratings ($d_M = 0.8$ and 0.77) while listening group produced a medium effect ($d_M = 0.47$ and 0.64). Program had a very small effect ($d_M = 0.11$ and 0.25) in both tests.

The anticipated effect size has important practical implications on the design of an experiment. The number of listeners required needs to meet a certain level of statistical power. It is proportional to the effect size index divided by the residual error variance in judgements. In other words, fewer listeners are required as the reliability and differences between mean rating increases. For example, Cohen has shown that to achieve the same power (for example, 0.8) in a repeated-measures test that has two loudspeakers, the number of listeners required for tests with small, medium, and large effects indices is 196, 33, and 14 listeners [39]. Reducing the power can lower the number of subjects needed, but at the expense of an increased probability of making a type I error (that is, rejecting the null hypothesis when in fact it is true). A better solution for reducing the size of the subject pool is to train a few listeners who can produce more reliable and discriminating judgements of sound quality.

3.4 Main Effects

In both tests there was a highly significant difference in preference between the different loudspeakers; $F(3, 258) = 231.0, p < 0.0001$ for the four-way test, and $F(2, 292) = 149.2, p < 0.0001$ for the three-way test. A Scheffé post-hoc test performed at a significance level of 0.05 showed a significant difference in the means between all pairs of loudspeakers in both tests.

Other main effects that were statistically significant in both tests were listening group; $F(15, 86) = 5.2, p < 0.0001$ for the four-way test and $F(19, 146) = 4.2, p <$

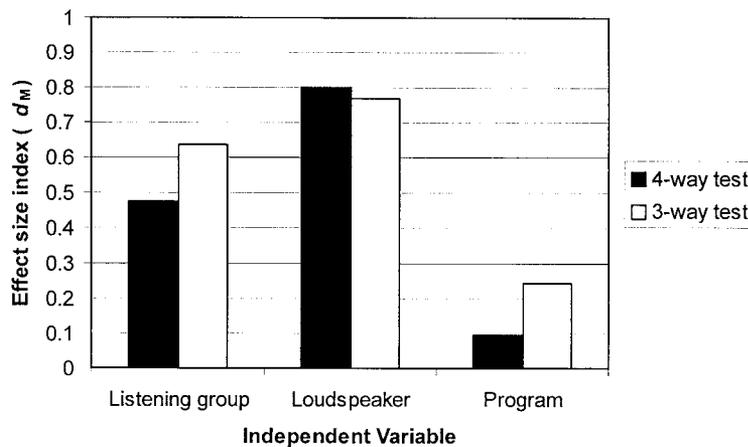


Fig. 1. Effect size index d_M in four-way and three-way tests for main independent variables: listening group, loudspeaker, and program.

0.0001 in the three way test.

Program was statistically significant in both tests; $F(3, 258) = 4.698$, $p = 0.0033$ for the four-way tests and $F(3, 438) = 9.99$, $p < 0.0001$ in the three-way tests. Details on the main effects and interactions are discussed in the following sections.

3.5 Loudspeaker Effects

The mean loudspeaker ratings and 95% confidence intervals are shown in Fig. 2 for both tests. Note that in all the graphs that follow, only the upper half of the 95% confidence interval is shown. The true mean lies somewhere between the upper bar and the same distance below the estimated mean, with a probability of 95%.

In the four-way tests the mean loudspeaker rating were 7.51 (loudspeaker P), 7.17 (loudspeaker I), 5.59 (loudspeaker B), and 3.21 (loudspeaker M). The difference in the means (0.34 preference) between loudspeaker P and I represents a slight preference. The difference in the mean ratings between these two loudspeakers and loudspeaker B (1.92 and 1.58) represents a moderate to strong preference. Both loudspeakers P and I were very strongly preferred over loudspeaker M based on the difference in the mean ratings (4.3 and 3.96). Loudspeaker B was strongly preferred over loudspeaker M (2.38 rating).

The mean loudspeaker rating in the three-way tests

were 7.07 (loudspeaker P), 5.96 (loudspeaker B), and 4.09 (loudspeaker M).

The rank orders of preference for loudspeakers P, B, and M were identical in both four-way and three-way tests. However, the relative magnitude of preference between loudspeakers P, B, and M was somewhat smaller in the three-way tests. The differences in the loudspeaker ratings were reduced between 0.51 and 1.3 preference ratings compared to those measure in the four-way tests.

A possible explanation for this difference is that a scaling effect occurred when the number of loudspeakers increased from three to four. The listeners may have expanded the separation and range in order to accommodate the additional loudspeaker. Another cause could be well-known context effect described earlier [31]–[38]. Contextual biases highlight a general principal in sensory judgment that human observers act like measuring instruments that constantly readjust themselves to the context or expected frame of reference. For example, how warm or cold 40° F feels depends on when (January versus July) or where (Arizona versus Alaska) the question is asked. Similarly, the preference and perceived attributes of a loudspeaker will be influenced by the context in which the judgment is made. A mediocre loudspeaker may receive higher ratings when it is compared against a group of weaker loudspeakers versus a group of stronger competi-

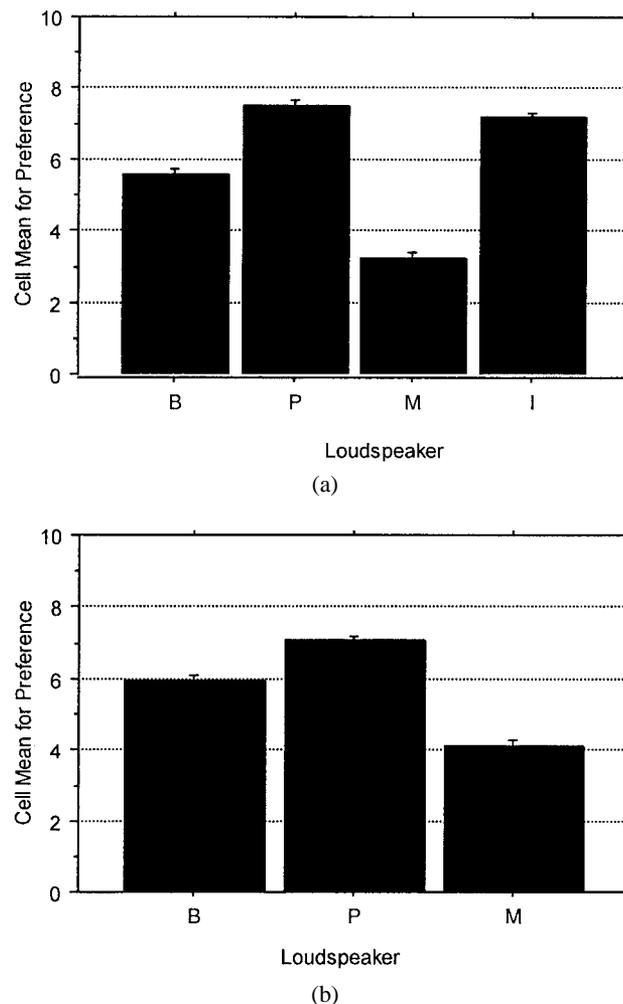


Fig. 2. Mean loudspeaker ratings and 95% confidence intervals. (a) Four-way test. (b) Three-way test.

tors. An otherwise neutral loudspeaker may be perceived as sounding “too dull” when it is compared against a predominantly “bright” group of loudspeakers, due to the contrast effect [41]. Therefore the range of loudspeakers tested affects the distribution of their ratings and where they fall on the scale.

In these tests the context effect would be as follows. In the four-way test the relative sonic similarities between loudspeakers P and I may have accentuated the sonic differences and deficiencies of loudspeakers B and M, and listeners accordingly adjusted the ratings of these loudspeakers downward.

The context or contrast effects can be minimized in listening tests by randomizing the order of presentation, using a large number of intervals, increasing the sample

size of test loudspeakers, and including anchors or references. Training the listeners may possibly reduce their susceptibility to context or range effects by creating a more stable sound-quality reference in their long-term memories. A comparison as rated by trained listeners of the two tests shows that the relative and absolute loudspeaker ratings did not change significantly. This suggests that they may have been less susceptible to range or context effects compared to the untrained listeners. More experiments are needed to test this hypothesis.

3.6 Listener Group Effect

Significant effects were reported earlier for the experimental variable, listening groups. The effects of listening groups on the mean preference ratings are plotted in Fig. 3

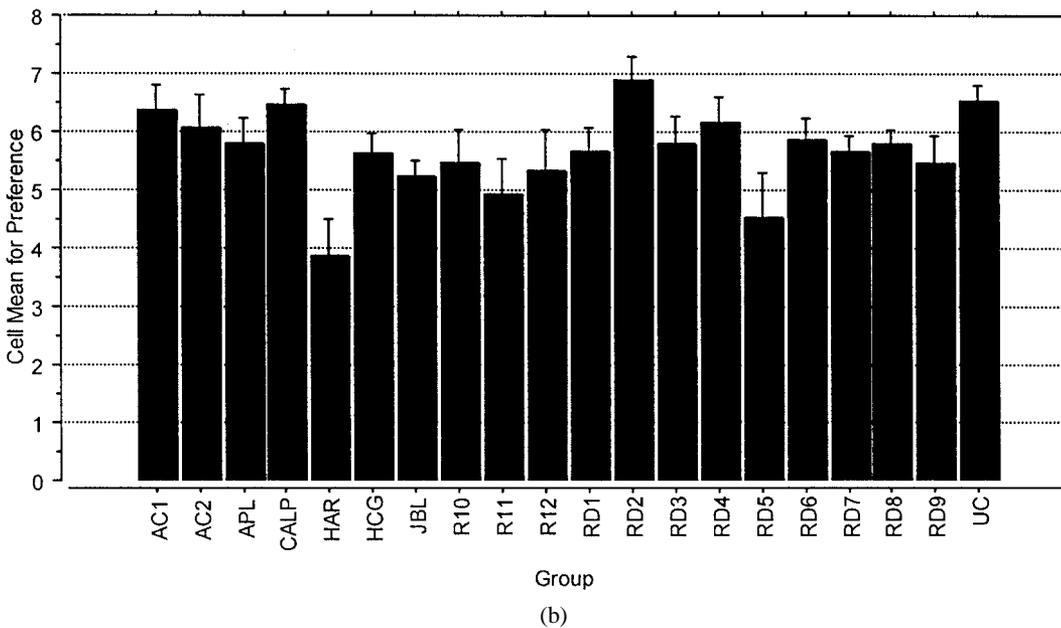
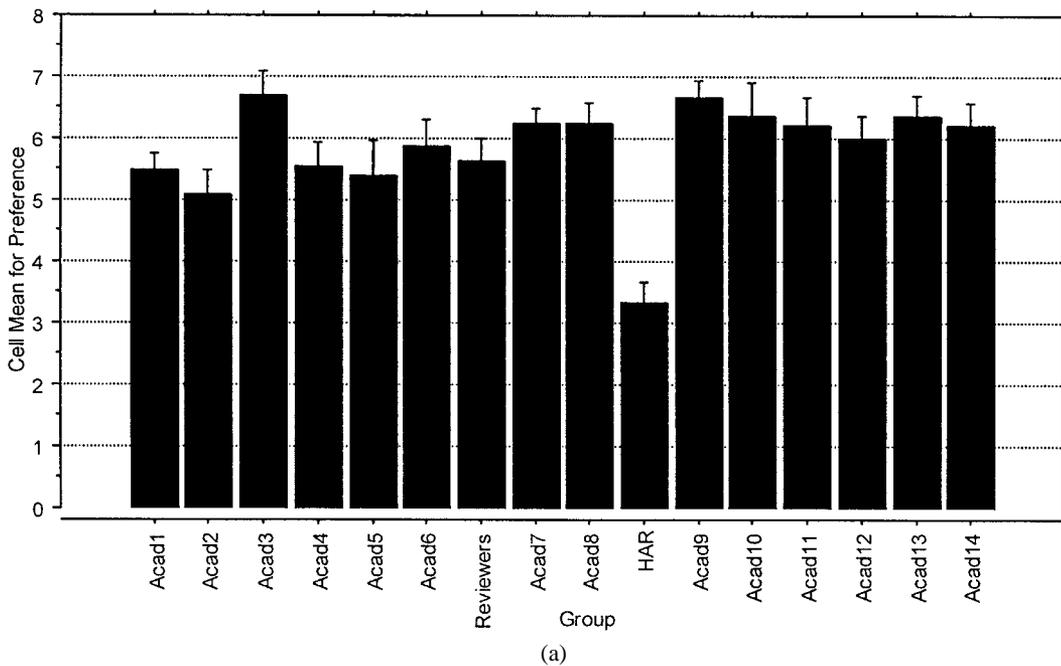


Fig. 3. Mean preference ratings and 95% confidence intervals for each listening group. (a) Four-way test. (b) Three-way test.

for the two tests. The graph indicates some significant variance among the different listening groups in their mean preference ratings. This implies that different groups used different parts of the preference scale. In both tests the trained listeners (HAR) gave the lowest mean preference ratings.

The mean preference ratings were also calculated as a function of the five different listener occupation categories. In the four-way test the trained-listener mean preference rating was 3.32 compared to 5.64 (audio reviewers) and 6.06 (audio retailers). In the three-way test the trained listeners' mean rating was 4.17 compared to 5.42 (marketing and sales), 5.72 (audio retailers), and 6.51 (students). If we assume that the scale is interpreted the same between the different categories of listeners, the students were generally the most pleased with the sound quality of the loudspeakers, whereas the trained listeners, on average found more things to complain about. One possibility is that trained listeners are generally more critical and difficult to please than untrained listeners. Gabrielsson et al. noted that experienced listeners tend to be more critical and give lower rating to poorer loudspeakers than inexperienced listeners [17]. The author's experience has been that trained listeners tend to use the same part of the preference scale whether they are judging small inexpensive

computer loudspeakers or very expensive "state-of-the-art" loudspeakers such as the ones used in these tests. These listeners put less value on the absolute ratings than they do in establishing meaningful differences between the ratings.

3.7 Program Effects

The mean preference ratings for the four-way and three-way tests are plotted in Fig. 4. Note that the order in which programs TC and JW are plotted is reversed between tests and that the scale has been zoomed in to highlight the small but significant effects program had on the preference ratings. Listeners on average gave higher preference ratings when the loudspeakers were auditioned using program JT, with lower ratings given for programs TC, JW, and LF. More noteworthy is that the rank order and relative magnitude of the preference ratings related to program was remarkably consistent across both listening tests.

There are a number of plausible reasons why program might influence the preference ratings. They include effects related to a listener's musical taste for certain programs as well as differences in the sonic fidelity of the recordings. Finally, certain programs may be better at revealing or concealing differences in the spectral, spatial,

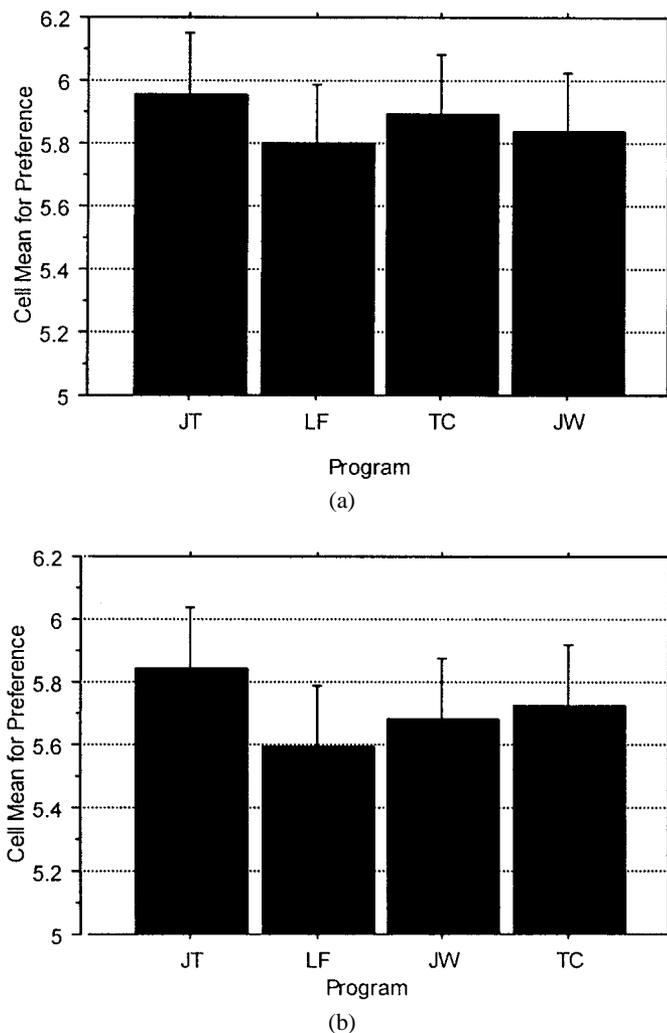


Fig. 4. Mean preference ratings and 95% confidence intervals for each program. (a) Four-way test. (b) Three-way test.

and nonlinear distortions that exist among the different loudspeakers.

3.8 Interaction Effects

There was an interaction effect between program and loudspeaker in the four-way and three-way tests; $F(9774) = 3.186, p < 0.0001$ and $F(6876) = 4.244, p = 0.0003$, respectively. Interaction effects between program, loudspeaker, and group were also found in the four-way test, $F(135, 774) = 1.553, p = 0.0002$.

In the three-way test an interaction was found between loudspeaker and group, $F(26, 22) = 3.292, p = 0.0458$. These interaction effects are discussed separately in more detail in the following sections.

3.9 Program–Loudspeaker Interactions

Fig. 5 shows the interactions between program and loudspeaker in the four-way and three-way tests. The interaction effect is largely isolated to interactions between loudspeaker B and program LF. In both tests the mean rating of loudspeaker B dropped almost 1 preference rating when auditioned with this program. The interactions between other loudspeakers and program were comparatively much smaller.

Program–loudspeaker interactions in listening tests have been reported widely in the literature [9], [11], [12], [14], [15], [42], [43]. Gabrielsson has attempted to explain the interactions by doing spectral analysis of the loudspeakers in the room while they were reproducing each program and looking for correlations with the subjective ratings [42]. It is the author’s experience that if the programs are selected carefully, well recorded, and spectrally homogeneous, the program interactions can be minimized.

3.10 Listener Group and Loudspeaker Interactions

The statistically significant interactions between group and loudspeaker are shown graphically in Fig. 6. Interaction effects are indicated by changes in the relative distances between the four horizontal lines, each representing the mean loudspeaker rating as a function of the listening group. In the four-way test, this largest deviation occurred between loudspeakers P and I, and to a lesser extent, between loudspeakers B and I. Some listening groups were better than others in their ability to discriminate between loudspeakers P and I, although the differences were seldom greater than 0.5 rating (slight preference).

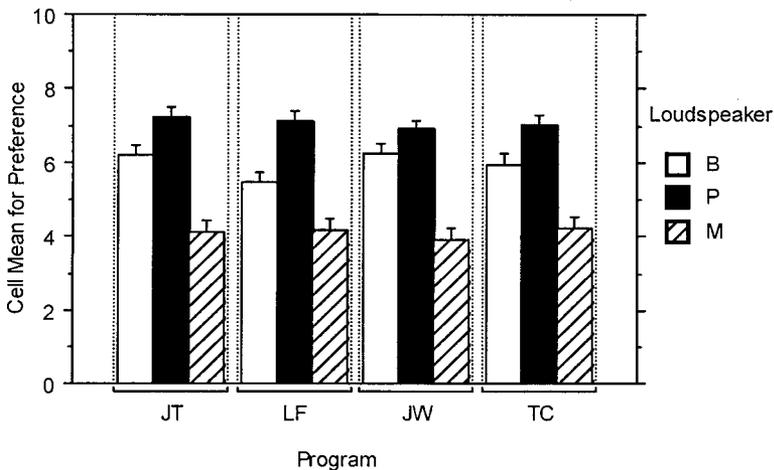
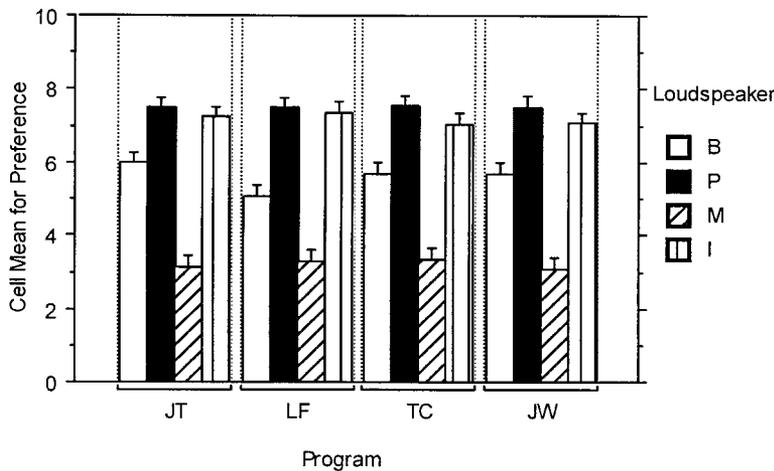


Fig. 5. Mean loudspeaker preference ratings and 95% confidence intervals for each loudspeaker as a function of program. (a) Four-way test. (b) Three-way test.

The trained listeners (HAR) had difficulty discriminating between loudspeakers I and P compared to other groups. One explanation for this could be related to difference in the seating positions between trained and untrained listeners. As mentioned previously, the same seat position was common to all trained listeners, who sat directly on axis to the loudspeakers. The untrained listeners were distributed among eight seats arranged in two rows. A hypothesis is that the audible differences between these two loudspeakers were more apparent for those untrained listeners who were off axis or in the second-row seating positions.

Apart from these interactions, the difference in the means between loudspeakers B, M, and I was remarkably

consistent across the 16 different listening groups in the four-way test.

In the three-way test the interaction between loudspeaker and group is much stronger. The reasons for this are not clear. Some of this interaction effect is traceable to the two student groups (UC and CALP). Both groups had difficulty forming preferences between loudspeaker, P and B reliably and, to a lesser extent, loudspeaker M. Some additional interaction variance comes from groups RD1 and RD2, who rated loudspeaker B the highest. Both these groups experienced an equipment failure during the test, where a cable became disconnected from the subwoofer in loudspeaker P, effectively removing all bass below 80 Hz from this loudspeaker.

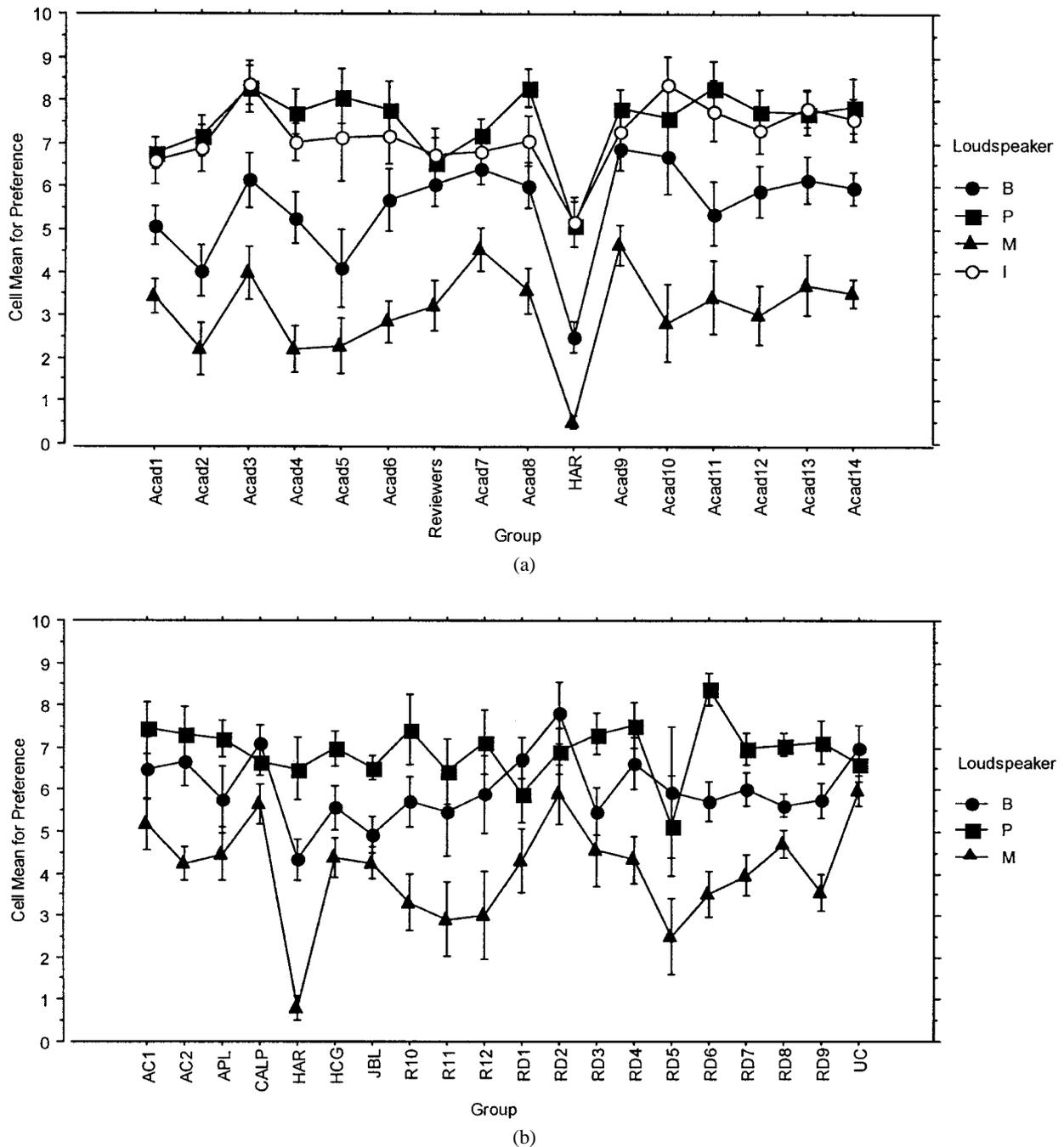


Fig. 6. Mean preference ratings and 95% confidence intervals for each loudspeaker as a function of listening group. (a) Four-way test. (b) Three-way test.

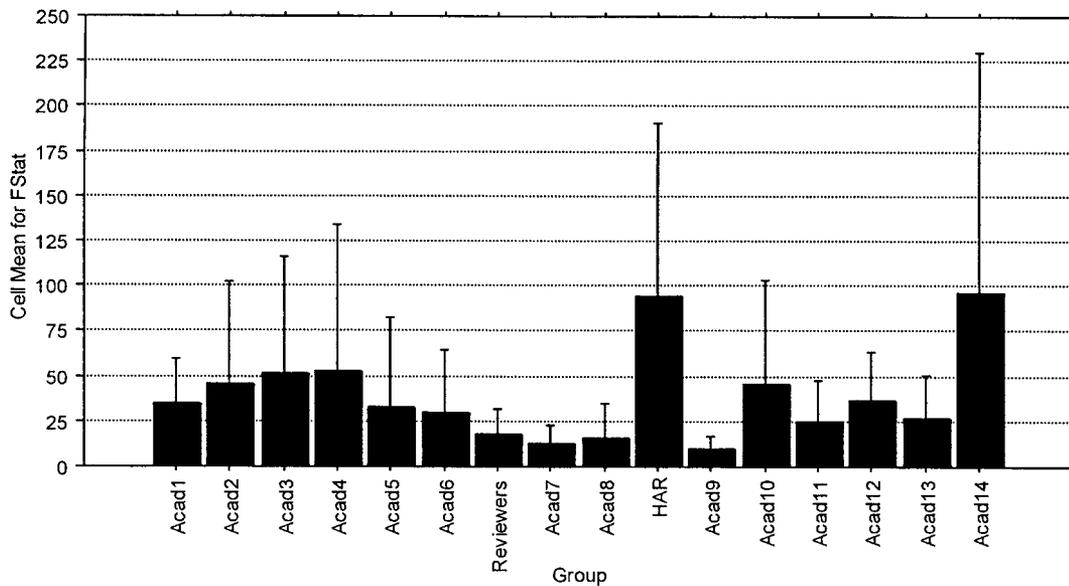
3.11 Performance among Different Listening Groups

The listener performance metric F_L described in Section 1 was calculated for each of the 268 listeners. This was done by performing a one-way ANOVA for each individual listener, where the independent factor was loudspeaker. F_L represents the mean sum of squares of loudspeaker ratings divided by the residual mean sum of squares, also known as the error variance.

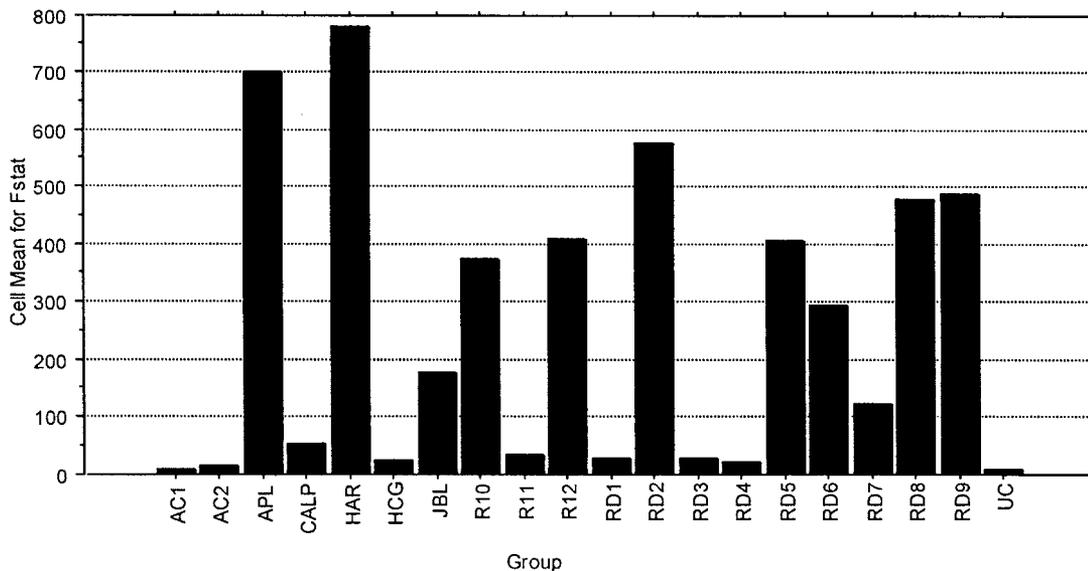
The average F_L values for each of the 36 different listening groups are plotted in Fig. 7 for the four-way and three-way tests. Due to the wide range of values in the three-way test the 95% confidence intervals are omitted in Fig. 7(b) to clarify better the differences between the different groups at the lower parts of the F_L scale. The confidence intervals in the three-way test were similar in pro-

portion to those found in the four-way test. In the four-way test the mean listening group F_L value was 34.5, ranging from a low of 10.06 (Acad9) to a maximum value of 96.21 (Acad 14), slightly higher than 94.6 for the trained listeners (HAR). The fairly large confidence intervals indicate higher standard deviations in listener performance within the group. However, given the relatively small number of listeners in each group, a high standard deviation could result if one of the listeners got a perfect score of 2000.

The same mean listening group F_L values are shown for the three-way test. In the three-way test the mean F_L value averaged across all listeners was 247.29 compared to 37.12 in the four-way test. The range of mean F_L values between groups is also much larger (8.14 to 781.32). This suggests two things. First, the three-way test presented an easier task in discriminating between the loudspeakers. Second, the variance in performance among the listening



(a)



(b)

Fig. 7. Mean loudspeaker F statistic as a function of listening group. (a) Four-way test. (b) Three-way test.

groups was much larger in the three-way test than in the four-way test.

Finally, the high F_L values of the nominally untrained groups Acad14 (four-way test) and APL (three-way test) relative to the trained listeners suggest that formalized training is not always a prerequisite for listeners to perform well in preference tests. These two groups of listeners apparently had sufficient experience, aptitude, and motivation to perform as reliably as a group of trained listeners. As the results show, this is an exception to the rule rather than the norm.

3.12 Occupation as a factor in Listener Performance

To examine listener performance in view of occupation more clearly, the mean listener F_L values were plotted as a function of occupation for both tests (see Fig. 8).

In the four-way tests the listener performance of the different categories based on the mean F_L values from highest to lowest was trained listeners (94.36), audio retailers (34.57), and audio reviewers (18.16). In the three-way test, the mean performance scores were: trained listeners (857.04), audio retailers (273.13), marketing and sales (112.15), and students (32.35).

(112.15), and students (32.35).

The performance of the trained panel is significantly better than the performance of any other category of listener. They are about three times better than the best group of audio retailers, five times better than the reviewers, and 27 times better than the students. The combination of training and experience in controlled listening tests clearly has a positive effect on a listener's performance. The students' poor performance is likely due to the student's lack of training and professional experience in the field of audio. The reviewers' performance is somewhat of a surprise given that they are all paid to audition and review products for various audiophile magazines. In terms of listening performance, they are about equal to the marketing and sales people, who are well below the performance of the audio retailers and trained listeners.

3.13 Correlation with Objective Measurements

The acoustic measurements of the loudspeakers first described in Section 2.1 are now discussed to determine the extent to which they correlate with the listening test results. The measurements of each loudspeaker are shown in Appendix 4 (Table 5) in the order (top to bottom) in

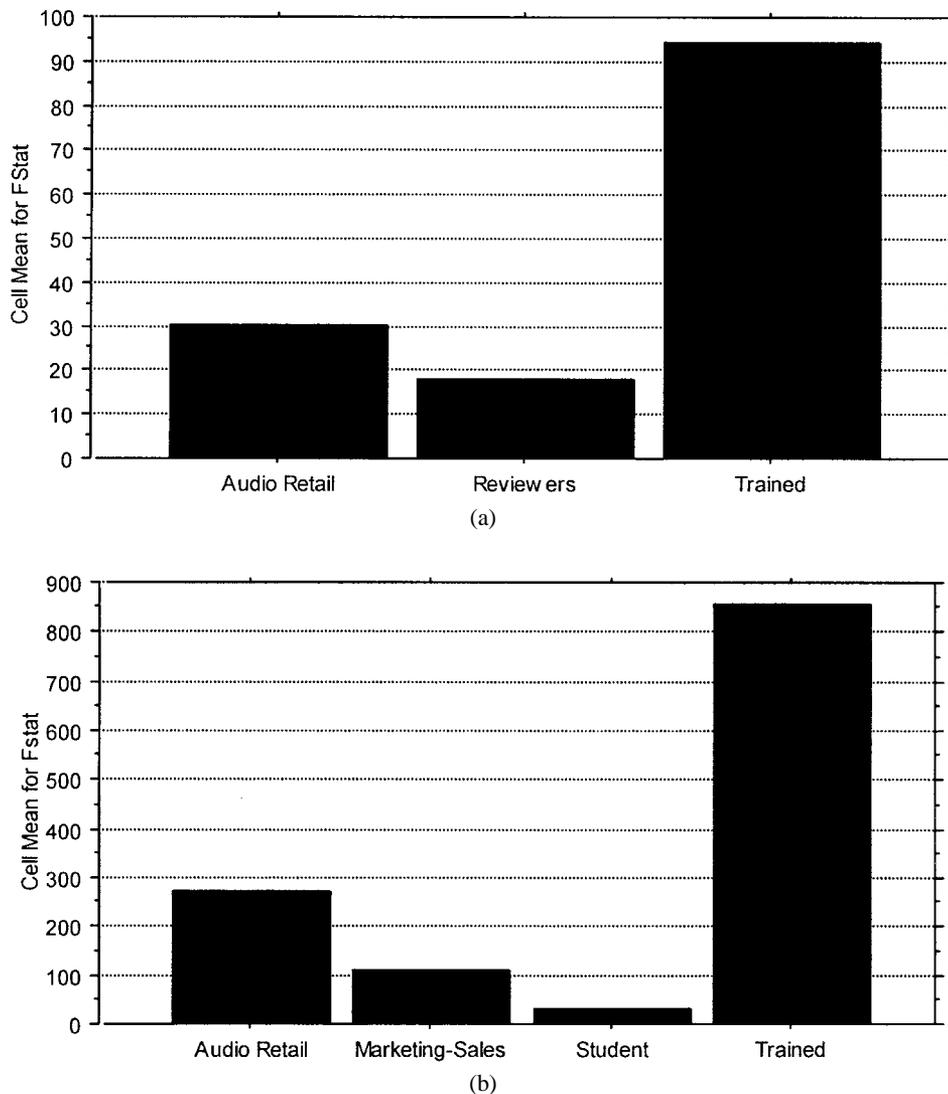


Fig. 8. Mean loudspeaker F statistic F_L as a function of occupation category (a) Four-way test. (b) Three-way test.

which they are rated, most preferred to least preferred.

Loudspeakers P and I are very similar in terms of bass extension and flatness in frequency response that is maintained well off axis. Based on their similarity in measured performance, it is not surprising that the mean difference in the preference ratings was only a 0.34 (a slight preference).

Loudspeaker B was rated third in the four-way test (5.59 preference), 1.92 and 1.58 lower than loudspeakers P and I, respectively. This represents a strong and moderately strong preference. Loudspeaker B has a respectable performance on axis with some gentle undulations in its response. The loudspeaker also has less bass output below 80 Hz compared to loudspeakers P and I. More serious is the rather substantial dip in its sound power response centered at 3 kHz, which is caused by a mismatch in the directivities of midrange and tweeter through their transitional passband regions. Listeners described the subjective effect as a hollow and recessed midrange coloration, which explains, in part, why it scored lower. The qualities of the indirect and reverberant sounds are evidently important since this coloration would not have affected the direct sound heard by the listener. It is important that manufacturers pay attention to these details and have the ability to measure and characterize the complete off-axis performance of loudspeakers accurately.

Now we turn to loudspeaker M, an electrostatic hybrid loudspeaker that received a mean rating of 3.21 in the four-way test, a full 2.38 to 4.3 ratings below the other three loudspeakers. The response curves are not very flat or pretty. There are many visible resonances that are well above threshold [44], [45] and are present in both the on-axis and the off-axis curves. This means that the colorations will be present in both the direct and the reflected sounds in the listening room. The loudspeaker has less bass output below 40 Hz than the other three loudspeakers, and the midrange frequencies are somewhat emphasized. The slope of the sound power curve indicates that the high-frequency output drops dramatically as the listener moves off-axis from the loudspeaker. Listeners who sit off axis will hear a much duller sound than those sitting on axis. Apparently this did not matter in these tests since the listeners who sat on axis (trained listeners) tended to rate the loudspeaker as low or lower than the untrained listeners sitting off axis. The colorations of this loudspeaker were dominant and omnipresent, regardless of where the listener sat in the room.

In conclusion, we can see clear visual correlations between these four sets of measurements and the listeners' preference ratings. The loudspeakers with the flattest, smoothest, and most extended frequency responses received the highest ratings.

4 DISCUSSION

This study reports one of the largest controlled loudspeaker listening tests conducted to date in terms of the sheer number of listeners involved. It is also unique in that most of the listeners (96%) had no formal training and little or no prior experience in controlled tests. One of the most significant findings is that the loudspeaker prefer-

ences of these nominally untrained listeners were very similar to those of the panel of trained listeners. The results may finally validate the use of trained listeners on the basis that their preferences can be extrapolated to a larger population of untrained listeners. The notion that the loudspeaker preferences of trained listeners are somehow biased cannot be used to predict those of reviewers, audio retailers, and the intended (untrained) customer is not supported by scientific data.

The differences between trained and untrained listeners are mostly related to differences in performance. The mean performances of the trained listeners based on loudspeakers F_L values were 3–27 times higher than any of the other four listener occupations measured in this study. Training and experience in controlled tests lead to significant gains in performance so that fewer listeners are required to achieve the same statistical power. The comparatively poorer performance of the students relative to the other three groups of audio professionals suggests that in field job experience can be beneficial to making more reliable judgments of sound quality. This implies that some form of training may be necessary in order to measure statistically significant preferences using more naïve and inexperienced listeners. Fortunately Bech has shown that very little training (four to eight sessions) is required [1].

The trained listeners were also found to use lower preference ratings than the untrained listeners. However, the loudspeaker rank ordering and the relative differences in preference between them were quite similar for both trained and untrained listeners. This means that extrapolations across different listener groups are possible based on the results from trained listeners. Trained listeners were the least forgiving when it came to rating the technically and sonically weakest loudspeaker in the test (for example, loudspeaker M).

The study provides strong validation for the current set of acoustic loudspeaker measurements used to design and test loudspeakers in our organization. There are clear visual correlations between measurements and subjective preference ratings, which supports the earlier findings reported by Toole [20], [21]. While interpreting the loudspeaker measurements still takes some skill and experience, the set of frequency-response curves alone could have largely predicted the outcome of these listening tests. The audio product reviewing industry could do a great service to consumers if they adopted a more meaningful set of technical measurements such as the ones shown here. Unfortunately such measurements are difficult and costly to perform, and beyond the reach of most audio reviewers. In the end it is the listening test that is the final arbiter of performance, and it is here that the reviewers need to spend more time and take greater care. Hopefully doing so will prevent reviewers from recommending two loudspeakers (P and M) as “state-of-the-art” equals when their technical and subjective performances have nothing in common. In retrospect, the only common denominator between these two loudspeakers is price.

It is the author's experience that most of the differences in opinion about the sound quality of audio product(s) in

our industry are confounded by the influence of nuisance factors that have nothing to do with the product itself. These include differences in listening rooms, loudspeaker positions, and personal prejudices (such as price, brand, and reputation) known to strongly influence a person's judgment of sound quality [9]. This study has only reinforced this view. The remarkable consensus in loudspeaker preference among these 268 listeners was only possible because the judgments were all made under controlled double-blind listening conditions.

5 CONCLUSION

The conclusions from this study are summarized in the following.

1) The loudspeaker preferences of trained listeners were generally the same as those measured using a group of nominally untrained listeners composed of audio retailers, marketing and sales people, audio reviewers, and college students.

2) Different groups of listeners use different parts of the preference scale. Trained listeners use the lowest part of the preference scale, indicating they may be more critical and harder to please.

3) Significant differences in performance were measured among the four different occupations. The average F_L values of the trained listeners were 3–27 times higher than those measured by the other groups. The second most discriminating and reliable group of listeners were the audio retailers, followed by the audio reviewers who were about equal to the marketing and sales people. The students had the worst performance, most likely due to their lack of audio experience compared to the other groups.

4) There were clear correlations between listeners' loudspeaker preferences and a set of acoustic anechoic measurements. The most preferred loudspeakers had the smoothest, flattest, and most extended frequency responses maintained uniformly off axis.

5) The rank order of the loudspeaker preferences did not change between the four-way and the three-way tests. However, eliminating loudspeaker I in the three-way test reduced the differences in mean ratings between loudspeakers P, B, and M. The most likely cause is a scaling or context effect related to the number and relative sound quality of the loudspeakers compared in the test.

6) The individual loudspeaker F statistics were on average seven times higher in the three-way test, indicating that rating these three loudspeakers may have been an easier task for the listeners.

6 ACKNOWLEDGMENT

Harman International sponsored this work. The author would like to thank all of the 268 listeners who participated in his study, as well as the engineering interns who helped set up and run many of these tests: Charles Sprinkle, Daniel Faissol, Ara Baghdassarian, and John Jackson. He is also grateful to his wife Valerie, Floyd Toole, and Søren Bech who provided valuable suggestions and corrections to this text.

7 REFERENCES

- [1] S. Bech, "Selection and Training of Subjects for Listening Tests on Sound-Reproducing Equipment," *J. Audio Eng. Soc.*, vol. 40, pp. 590–610 (1992 July/Aug.).
- [2] R. Shively, "Subjective Evaluation of Reproduced Sound in Automotive Spaces," in *Proc. AES 15th Int. Conf. in Audio, Acoustics and Small Spaces* (1998), pp. 109–121.
- [3] S. E. Olive, "A Method for Training Listeners and Selecting Program Material for Listening Tests," presented at the 97th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 42, p. 1058 (1994 Dec.), preprint 3893.
- [4] S. E. Olive, "A Method for Training Listeners: Part II," presented at the 101st Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 44, p. 1160 (1996 Dec.), no preprint.
- [5] S. E. Olive, "A New Listener Training Software Application," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 49, p. 542 (2001 June), preprint 5384.
- [6] T. Neher, F. Rumsey and T. Brookes, "Training of Listeners for the Evaluation of Spatial Sound Reproduction," presented at the 112th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 50, p. 518–519 (2002 June), preprint 5584.
- [7] M. Meilgaard, G. V. Civille, and C. T. Carr, *Sensory Evaluation Techniques*, 2nd Ed. (CRC Press, Boca Raton, FL, 1991).
- [8] F. E. Toole, "Subjective Evaluation: Identifying and Controlling the Variables," presented at the AES 8th Int. Conf.: The Sound of Audio (1990, Apr.), paper 8-013.
- [9] F. E. Toole and S. E. Olive, "Hearing Is Believing vs. Believing Is Hearing: Blind vs. Sighted Listening Tests, and Other Interesting Things," presented at 97th Convention of Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 42, p. 1058 (1994 Dec.), preprint 3894.
- [10] F. E. Toole, "The Acoustics and Psychoacoustics of Loudspeakers and Rooms: The Stereo Past and the Multichannel Future," presented at the 109th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 48, p. 1101 (2000 Nov.), preprint 5201.
- [11] P. Schuck, S. Olive, J. Ryan, F. Toole, S. Sally, M. Bonneville, E. Verreault, and K. Momtahan, "Perception of Reproduced Sound in Rooms: Some Results from the Athena Project," in *Proc. AES 12th Int. Conf.* (1993 June), pp. 49–73.
- [12] S. E. Olive, P. Schuck, S. Sally, M. Bonneville, "The Variability of Loudspeaker Sound Quality among Four Domestic-Sized Rooms," presented at the 99th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 43, pp. 1088–1089 (1995 Dec.), preprint 4092.
- [13] F. E. Toole, "Loudspeakers and Rooms for Stereophonic Sound Reproduction," presented at the AES 8th Int. Conf. The Sound of Audio (1990 Apr.), paper 8-011.
- [14] S. E. Olive, P. L. Schuck, S. L. Sally, and M. E. Bonneville, "The Effects of Loudspeaker Placement on Listener Preference Ratings," *J. Audio Eng. Soc.*, vol. 42, pp. 651–669 (1994 Sept.).

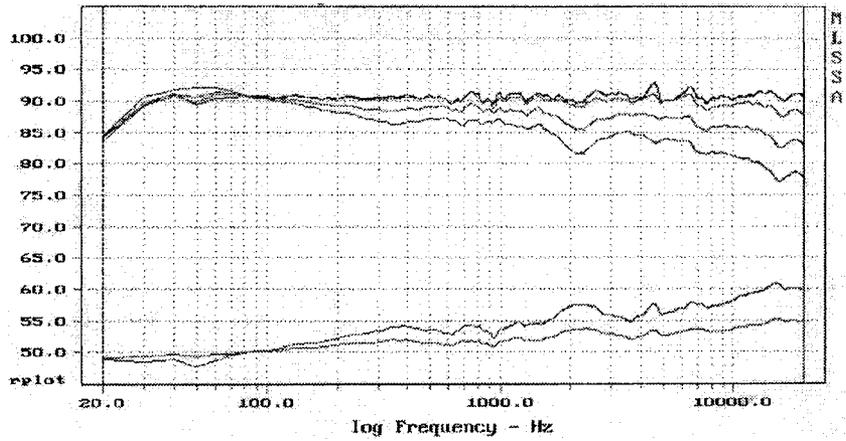
- [15] S. Bech, "Timbral Aspects of Reproduced Sound in Small Rooms. I," *J. Acoust. Soc. Am.*, vol. 97, pp. 1717–1726 (1995).
- [16] R. E. Kirk, "Learning a Major Factor Influencing Preferences for High-Fidelity Reproducing Systems," *J. Acoust. Soc. Am.*, vol. 28, pp. 1113–1116 (1956).
- [17] A. Gabrielsson, U. Rosenburg, and H. Sjogren, "Judgments and Dimension Analysis of Perceived Sound Quality of Sound-Reproducing Systems," *J. Acoust. Soc. Am.*, vol. 55, pp. 854–861 (1974).
- [18] A. Gabrielsson, "Loudspeaker Frequency Response and Perceived Sound Quality," *J. Acoust. Soc. Am.*, vol. 90, pp. 707–719 (1991).
- [19] A. Gabrielsson and B. Lindstrom, "Perceived Sound Quality of High-Fidelity Loudspeakers," *J. Audio Eng. Soc.*, vol. 33 pp. 33–53 (1985 Jan./Feb.).
- [20] F. E. Toole, "Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 1," *J. Audio Eng. Soc.*, vol. 34, pp. 227–235 (1986 Apr.).
- [21] F. E. Toole, "Loudspeaker Measurements and Their Relationship to Listener Preferences: Part 2," *J. Audio Eng. Soc.*, vol. 34, pp. 323–348 (1986 May).
- [22] R. Quesnel, "A Computer-Assisted Method for Training and Researching Timbre Memory Evaluation Skills," PhD Dissertation, Tech. Rep. 1, McGill University, Montreal, P.Q., Canada (2002).
- [23] A. Gabrielsson, "Statistical Treatment of Data for Listening Tests on Sound-Reproducing Treatment," Rep. T. A, Karolinska Institute, Technical Audiology, HH, Stockholm, Sweden (1979).
- [24] A. Devantier, "Characterizing the Amplitude Response of Loudspeaker Systems," presented at the 113th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 50, p. 954 (2002 Nov.), preprint 5638.
- [25] S. E. Olive, B. Castro, and F. E. Toole, "A New Laboratory for Evaluating Multichannel Audio Components and Systems," presented at the 105th Convention of Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 46, pp. 1032–1033 (1998 Nov.), preprint 4842.
- [26] M. R. Jason, "A Real-World Implementation of Current Theory in Loudspeaker Subjective Evaluation," presented at the 90th Convention of Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 39, pp. 385 (1991 May.), preprint 3048.
- [27] M. R. Jason, "Design Considerations for Loudspeaker Preference Experiments," *J. Audio Eng. Soc.*, vol. 40, pp. 979–996 (1992 Dec.).
- [28] S. Clement, L. Demany, and C. Semal, "Memory for Pitch versus Memory for Loudness," *J. Acoust. Soc. Am.*, vol. 106, (1999 Nov.).
- [29] G. E. Starr, and M. A Pitt, "Interference Effects in Short-Term Memory for Timbre," *J. Acoust. Soc. Am.*, vol. 102, pp. 486–494 (1997).
- [30] F. E. Toole, Private conversation (2003).
- [31] A. Parducci, "The Relativism of Absolute Judgment," *Sci. Am.*, vol. 219, pp. 84–90 (1968).
- [32] A. Parducci, "Contextual Effects: A Range-Frequency Analysis," in E. C. Cartrette and M. P. Friedman, Eds. *Handbook of Perception*, vol. 2. (Academic Press, New York, 1974).
- [33] A. Parducci and D. H. Wedell, "The Category Effect with Rating Scales: Number of Categories, Number of Stimuli, and Method of Presentation." *J. Experim. Psychol.*, vol. 12, pp. 496–516 (1986).
- [34] D. H. Wedell and A. Parducci, "The Category Effect in Social Judgment: Experimental Ratings of Happiness," *J. Personal. Soc. Psychol.*, vol. 55, pp. 341–356 (1988).
- [35] D. H. Wedell, A. Parducci, and M. Lane, "Reducing the Dependence of Clinical Judgment on the Immediate Context: Effects of Number of Categories and Type of Anchors," *J. Personal Soc. Psychol.*, vol. 58, pp. 319–329 (1990).
- [36] D. H. Wedell and J. C. Pettibone, "Preference and the Contextual Basis of Ideals in Judgment and Choice," *J. Experim. Psychol: General*, vol. 128, pp. 346–361 (1999).
- [37] E. C. Poulton, "*Bias in Quantifying Judgments*," (Lawrence Erlbaum Assoc., Hove, UK, 1989), 304 pp.
- [38] H. Lawless, "Bias and Context Effects in Ratings," Lecture in Food Science 410: Sensory Evaluation, Cornell University (2002), <http://zingerone.foodsci.cornell.edu/fs410/lectures/context.pdf>.
- [39] R. T. Hurlburt, *Comprehending Behavioral Statistics*, 3rd ed. (Thomas Wadsworth, 2003).
- [40] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Assoc., Hove, UK, 1988).
- [41] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th ed. (Academic Press, New York, 1997).
- [42] A. Gabrielsson, B. Hagerman, T. Bech-Kristensen, and G. Lundberg, "Perceived Sound Quality of Reproductions with Different Frequency Responses and Sound Levels," *J. Acoust. Soc. Am.*, vol. 83, pp. 1359–1366 (1990).
- [43] S. Olive, "Evaluation of Five Commercial Stereo Enhancement 3-D Audio Software Plug-ins," presented at the 110th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 49, p. 543 (2001 June), preprint 5386.
- [44] F. E. Toole and S. E. Olive, "The Modification of Timbre by Resonances: Perception and Measurement," *J. Audio Eng. Soc.*, vol. 36, pp. 122–142 (1988 Mar.).
- [45] S. E. Olive, P. L. Schuck, J. G. Ryan, S. L. Sally, and M. E. Bonneville, "The Detection Thresholds of Resonances at Low Frequencies." *J. Audio Eng. Soc.*, vol. 45, pp. 116–128 (1997 Mar.).

APPENDIX 1 LOUDSPEAKER MEASUREMENTS

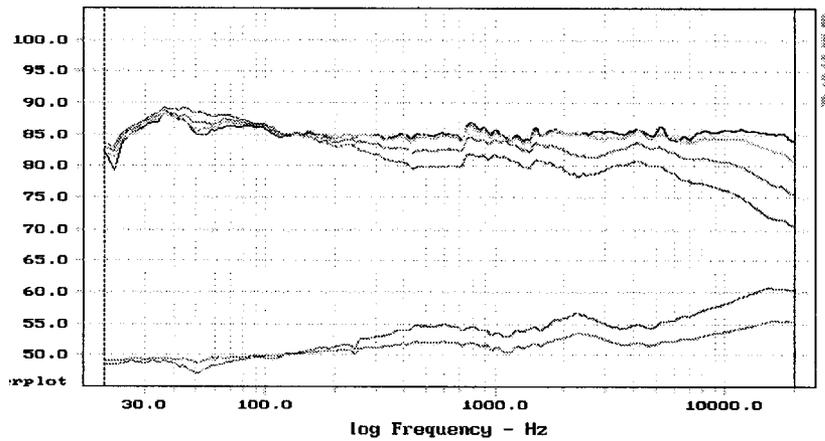
The spatially averaged anechoic measurements of each loudspeaker used in the listening tests in the order in which the loudspeakers were rated, from highest to lowest, are presented in Fig. 9. See section 2.1 for a description of what each curve represents.

APPENDIX 2 INSTRUCTIONS TO LISTENERS

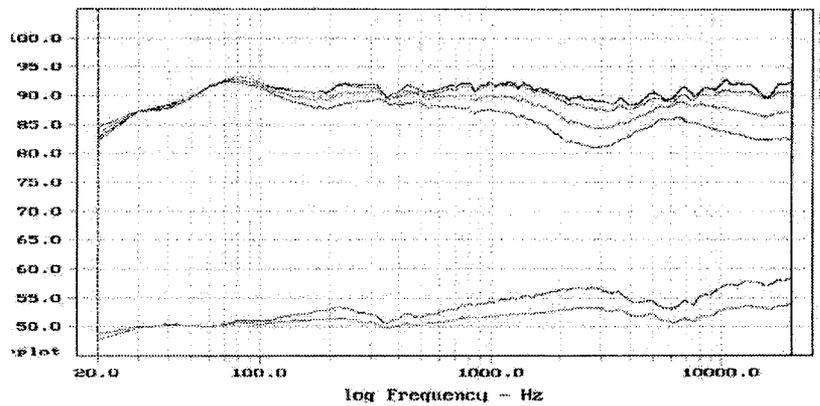
In these tests you will be judging the sound quality of different loudspeakers and rating them according to your personal preference. You MUST enter a rating for each



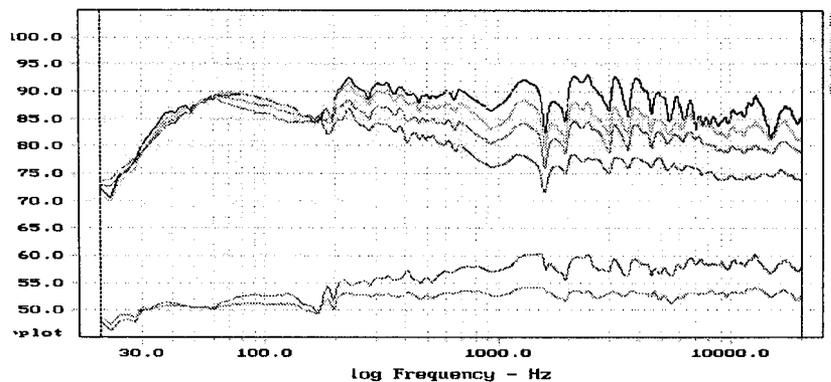
(a)



(b)



(c)

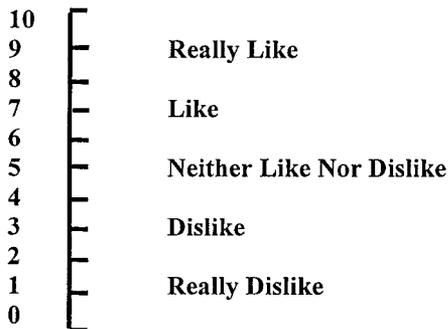


(d)

Fig. 9. Spatially averaged anechoic measurements (a) Loudspeaker P. (b) Loudspeaker I. (c) Loudspeaker B. (d) Loudspeaker M.

loudspeaker in the appropriate box after the program selection has ended. Please enter your ratings using the following preference scale:

Preference Scale



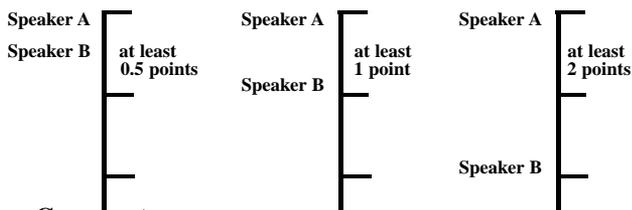
Your ratings can contain up to one decimal place (e.g., 7.3, 2.5).

DO NOT GIVE TIED SCORES IN ANY ROUND.

If you do, the computer will ask you to reenter your ratings.

You should separate your preference ratings among different speakers to reflect your relative preference between two speakers. Use the following guidelines:

Slight Preference Moderate Preference Strong Preference



Comments

Finally, we encourage you to write comments about what you like and dislike about the sound of the speakers you are comparing: what aspects is it about the speaker that makes you prefer it (or not prefer it) over the other speaker(s)?

APPENDIX 3

Tables 4 and 5 are the ANOVA summary table and the Scheffe post-hoc test table for the variable, loudspeaker.

APPENDIX 4

Table 6 lists the 36 different listening groups, showing the number in each group, their occupation category (AR—audio retailer, PR—professional audio reviewer, T—Harman-trained listener, MS—Harman marketing and sales, and S—student), and the dates of the listening tests.

Table 6. Loudspeaker Measurements.

Group	Count	Category	Test Date
Acad1	7	AR	3/11/02
Acad2	6	AR	3/12/02
Acad3	6	AR	3/13/02
Acad4	7	AR	3/14/02
Acad5	4	AR	3/15/02
Acad6	5	AR	4/17/02
Audio reviewers	6	PR	4/16/02
Acad7	8	AR	9/10/02
Acad8	8	AR	9/11/02
HAR	6	T	9/16/02
Acad9	8	AR	9/12/02
Acad10	5	AR	2/11/03
Acad11	6	AR	2/12/03
Acad12	7	AR	2/13/03
Acad13	8	AR	4/23/03
Acad14	5	AR	4/24/03
Subtotal	102		
AC1	6	AR	11/1/01
AC2	3	AR	11/5/01
APP	6	AR	5/1/03
CALP	7	S	11/5/01
HAR	6	T	10/23/02
HCG	9	MS	11/5/01
JBL	12	MS	11/8/01
RD1	7	AR	11/1/01
RD2	7	AR	11/5/01
RD3	5	AR	11/5/01
RD4	6	AR	11/5/01
RD5	5	AR	1/14/02
RD6	23	AR	5/20/02
RD7	22	AR	10/18/02
RD8	13	AR	1/10/03
RD9	5	AR	3/10/03
RD10	6	AR	3/10/03
RD11	6	AR	3/10/03
RD12	5	AR	3/10/03
UC	7	S	2/26/02
Subtotal	166		
Total	268		

THE AUTHOR



Sean E. Olive received a bachelor of music degree from the University of Toronto, Ont., Canada, 1982 and a master's degree in sound recording from McGill University, Montreal, P.Q., in 1986. He is currently pursuing a Ph.D. degree in sound recording at McGill University, investigating the perception of spectral distortion and its effect on listener preference.

From 1986 to 1993 he was a research scientist in the Acoustics and Signal Processing Group at the National Research Council in Ottawa, Ont. There he worked with Dr. Floyd Toole on research related to subjective and objective testing of loudspeakers and microphones, room-adaptive loudspeakers, and the detection of reflections and resonances. Much of this work had been pre-

sented in various AES publications. For two of these papers he received, as a coauthor, AES publication awards in 1990 and 1995 and an AES Fellowship in 1996. Since 1993, he has been the manager of Subjective Evaluation with the R&D group at Harman International in Northridge, CA, where he is responsible for subjective testing of all Harman consumer products and conducting psychoacoustics-related research in sound reproduction.

Mr. Olive is a former chair of the AES Los Angeles Section, a past AES governor, and a current member of two AES technical committees. For the past five years he has taught psychoacoustics and critical listening at the UCLA Extension's recording engineering certificate program.