Subjective Measurements of Loudspeaker Sound Quality and Listener Performance*

FLOYD E. TOOLE

National Research Council, Ottawa, Ontario K1A 0R6, Canada

With adequate attention to the details of experiment design and the selection of participants, listening tests on loudspeakers yielded sound-quality ratings that were both reliable and repeatable. Certain listeners differed in the consistency of their ratings and in the ratings themselves. These differences correlated with both hearing threshold levels and age. Listeners with near-normal hearing thresholds showed the smallest individual variations and the closest agreement with each other. Sound-quality ratings changed as a function of the hearing threshold level and age of the listener. The amount and direction of the change depended upon the specific products; some products were rated similarly by all listeners, whereas others had properties that caused them to be rated differently. Stereophonic and monophonic tests yielded similar sound-quality ratings for highly rated products, but in stereo, listeners tended to be less consistent and less critical of products with distinctive characteristics. Assessments of stereophonic spatial and image qualities were closely related to sound-quality ratings. The relationship between these results and objective performance data is being pursued.

0 INTRODUCTION

In 1941 Harvey Fletcher presented a classic paper [1] in which he stated that "the properties of the hearing mechanism and the characteristics of the listening location, rather than the properties of the sounds transmitted, will very largely determine the fundamental requirements of the transmission system." In this he sums up much of what has beleaguered audio ever since.

Fletcher's faith in the ear as the final arbiter and as a sensitive and "wonderful instrument for measuring various aspects of sound" [2] has been rewarded in numerous psychoacoustical investigations.

In the domains of professional and consumer audio, however, the idea that listening tests can produce trustworthy results is one that has very limited acceptance. Most people, it seems, believe that we all "hear differently" and that we can have strongly individual preferences not only in music but in sound quality as well because of differences in taste, experience, and hearing ability. There have even been suggestions of geographical influences in sound-quality preference. In addition to these individual biases, the variable influences of the listening room, music selection, price, appearance, advertising, and so on can alter opinions of how products sound. There are enough of these confounding factors that it always seems possible to doubt the real worth of an opinion that is not one's own.

But in spite of the problems, listening tests remain the final arbiters of sound quality from loudspeakers and are the basis for selecting loudspeakers for domestic and professional applications. At the design stage, where calculations and measurements play an important role in the development of products, listening tests provide the confirmation of the design integrity and are essential in the "fine tuning" of the final product.

If our perceptions and preferences are truly individual, it follows that the designer and the user must somehow be "matched" in order for the product to be similarly appreciated. The strength and variety of the individual biases and their distribution in the population would then determine the popularity of the products. One would hope that the majority of people would have similar tastes, otherwise the process of matching products and users would be extremely complicated. At present there is substantial evidence that this untidy scenario is, in fact, commonplace: designers attempt to create products that will have widespread appeal, and individual consumers attempt to maximize their

* Manuscript received 1984 March; revised 1984 September 20.

pleasure. Neither, as it turns out, has a foolproof method for achieving his objectives. The unreliability of typical listening tests simply ensures that the process remains more complicated than it needs to be.

The present work is the result of a concentrated effort to identify and to control the factors contributing to the personal opinions expressed in listening tests. A large-scale series of practical tests involving 42 listeners and 37 loudspeakers and spanning a period of over two years resulted in masses of data, only a portion of which has yet been examined thoroughly. As pointed out in an earlier work [3], if the listening test can be improved to the point of rendering trustworthy subjective data, we will be in a position to assess which technical measures of loudspeaker performance are the most useful predictors of subjective preference. It may then be feasible to prepare a set of rules for the design and evaluation of loudspeakers by purely technical means. Ideally, perfecting the listening test should lead to its obsolescence.

While it is optimistic to think that this will happen easily or soon, there are already indications that some areas of loudspeaker performance are adequately described by measurements. Acknowledging these areas can reduce the complexity of the listening test, allowing it to become more specialized and, therefore, more sensitive. The test is likely also to be simpler and less time-consuming.

Traditionally, scientifically controlled listening tests on loudspeakers have been lengthy, costly, and complicated affairs, so they are rarely done. Anything less could not, however, be trusted to produce statistically reliable data. Consequently it is a further objective of this work to explore the possibility of a simple, relatively fast method by which useful subjective data can be obtained. Only by this means are reliable listening tests likely to be used widely.

1 SUBJECTIVE MEASUREMENTS—DEVELOPING THE TECHNIQUE

1.1 A Brief History

Listening tests are not new. From the earliest days of radio and recordings, competing products have been judged on the basis of their audible performance. Technical measurements are, in fact, the later development.

By the early 1950s there had been a number of serious efforts to exercise some controls on the variables in listening tests and to set objectives by which the performance of loudspeakers could be judged [4]–[7]. LeBel [8] provided a particularly perceptive overview of some contemporary work and argued for a more scientific approach to psychoacoustical studies. Olson [9] was more specific about proper experimental methodology, and even made a case for standardizing the process if and when sufficient knowledge existed. Langford-Smith [10] provided further useful comment and a good bibliography to early work in this field, and Wilson [11] described a technique for jury assessments, using paired comparisons and an analytical listener reporting form, that gave uncommonly consistent results for its time. Regrettably World War II interrupted the proceedings, and further development of the method ceased.

In the succeeding years, work on subjective tests themselves seemed to lose impetus, although statistical analysis made an appearance [12]. Listening tests continued, of course, with the audio press developing its own version, known as the product review. In spite of the large audience and influence that these product assessments had, the tests themselves were usually of the most rudimentary kind. The large variations in opinion resulting from these widely publicized tests simply confused the picture, cultivating a public mistrust in measurements and a reliance on "golden-eared" listeners.

In 1975 Cooke [13] presented a careful analysis of the prevailing practice of listening tests and concluded that a great many yielded results that were so influenced by extraneous factors as to be misleading. He further observed that, at that stage, listening tests were limited in what they could reveal about loudspeaker performance and that, to assess the better loudspeakers properly, listening tests needed to be brought to a new standard of sophistication. This was a view shared by others at that time, and a number of independent efforts were under way to improve the situation.

Perhaps the longest sustained effort in the subjective evaluation of loudspeakers has been that of the BBC Research Department. Much of their work is unpublished, and there are few details about the experimental methods used, but an interesting summary of their findings is given by Harwood [14]. The bulk of the effort of this group seems to have been directed at defining the thresholds of audibility and/or annoyance of various specific technical faults. While the result of this work was evident in some of the loudspeakers produced in the United Kingdom [15], there was unfortunately little public contribution to the developing science of listening tests. Gilford [16] did, however, offer some direction in the design of adequate listening environments.

In Japan there was a sophisticated effort to analyze the results of listening tests, to correlate them with aspects of physical performance, and thus to optimize product design [17]–[19]. In this and work in Scandanavia by Eisler [20], Staffeldt [21], and Gabrielsson and his various coworkers [22] statistical analysis has played a major role. By this means it has been possible to separate the myriad descriptive words and phrases used by listeners to describe perceived qualities of sounds into groups having a common basis. From this have been derived a small number of relatively independent perceptual dimensions that appear to convey the essential descriptions of perceived sound quality. The list as it stands may or may not be complete; it is, however, an important basis for listener analysis of sound quality. The correlations with physical performance were not well developed, although some general

trends were noted by several workers.

There were also smaller but no less serious efforts on the part of several manufacturers and consumer product testing magazines and organizations to bring some order to this confused situation [23]–[26]. The author's own work in this field was also taking shape in the early 1970s, motivated in large part by requests from industry and audio publications in Canada for reliable performance data on loudspeakers and other audio transducers.

A significant event of this period was the decision by the International Electrotechnical Commission (IEC) to develop a standard for "listening tests on loudspeakers." Evidently it was felt that the field could use some leadership from this respected organization. While any simple description of the events is bound to be somewhat unfair, we think it is reasonable to say that the working group labored with an abundance of good intentions and a dearth of useful scientific facts.

Standardization is essential to the smooth operation of international commerce, but a poor standard can sometimes be worse than none at all. There was a feeling within the IEC working group that, at that time, there was not enough knowledge to be able to standardize a specific test procedure. Consequently the document took on a more tutorial tone, and the collection of good advice has been published as a technical report [27] rather than as a full-fledged standard. It is hoped that, with knowledge gained from field use, the original intention of producing a standard can eventually be realized. In the meantime it is clear that the techniques are in need of refinement and scientific documentation.

1.2 Improving the Technique

It is possible, and even relatively easy, to obtain statistically significant results from listening tests, but many of the past efforts have been of limited practical value because the experiments did not adequately control all of the factors bearing on the formation of the listeners' opinions. Most commonly, controls seem to have been relaxed in the acoustical aspects of experiment design. Listening rooms, for example, have in one way or another not been representative of typical listening environments, or have not been specified at all.

In many of the older works, doubt is cast by the state of technology at the time. Microphones and record/reproduction devices performed deficiently in ways that could easily have prejudiced the results of the listening tests. Nevertheless, the major problem was a lack of experimental controls and incomplete descriptions of the experimental procedure in published accounts. LeBel wrote in 1947: "Unfortunately, the basic problem of much audio research is similar in nature. Almost anyone can make a test and get consistent results, but an engineer of long experience may find real difficulty in defining the scope and validity of the result. More scientifically put, consistency is a necessary, but not the sole, requirement for accuracy" [8]. Thirty-seven years later, we seem to suffer from the same problems [3].

The objective of the present work was to examine ways of improving the techniques for obtaining reliable subjective data on loudspeakers. The approach to the improvements was to synthesize, from existing knowledge, the conditions under which optimal listener performance would be likely to occur This is not a straightforward task, as it involves many branches of acoustics, from physical to physiological, as well as electronics and experimental psychology. Success in such a venture is dependent on the thoroughness of the experimenter and the completeness of existing theory.

1.3 Sources of Variability

Reliable measurements of any kind require that the experimenter control the sources of variability to the point where, within an acceptable margin of error, the experiment is reduced to two variables: the independent variable (the input under the experimenter's control) and the dependent variable (the output or desired response). In the present case the loudspeakers are the independent variable and the listener responses are the dependent variable.

In loudspeaker evaluations conducted under normal circumstances the variability in subjective assessments is usually rather large. In part this is undoubtedly due to differences among individual listeners, but it is equally clear that much of the fluctuation in opinion is caused by what could be called the "nuisance variables." Many of these are well known, but others may not yet have been identified.

The nuisance variables associated with the physical environment appear to be as follows:

Listening room
Loudspeaker position
Listener position
Relative loudness (of compared sounds)
Absolute loudness (of all sounds)
Program material
Electronic imperfections
Stereo (peculiar technical problems).

The nuisance variables associated with the listeners themselves tend to fall into intellectual, psychological, or physiological categories and appear to be as follows:

Knowledge of the products
Familiarity with the program
Familiarity with the room
Familiarity with the task
Judgment ability or aptitude
Hearing ability (physical impairment)
Relevant accumulated experience
Listener interaction and group pressure
Stereo (conflicts between spatial and sound-quality aspects of reproduction).

Nuisance variables are also to be found in the experimental procedure and the manner of recording and scaling listener responses, as follows:

Identification of the perceptual dimensions
Scaling of the perceptual dimensions

Anchoring or normalization of individual scales
Effects of context and contrast
Effects of sequence and memory
Experimenter bias.

1.4 Controlling the Variables

Some of these variables are well known and have been the object of specific research. Some are likely to have a greater influence on subjective opinion than others. Still others, perhaps, have not yet been properly identified or acknowledged. In the interim it is necessary to make some decisions that may be more arbitrary than scientific in order simply to make a useful start.

The importance of these decisions is amplified by the nature of the tests. If, for example, listeners were required merely to respond in a relative sense by indicating the presence and direction of a preference, it might be adequate merely to ensure that the nuisance variables be kept constant. When absolute responses are involved, on the other hand, the conditions created by some of the nuisance variables can introduce constant errors in the results. Each decision is an opportunity for the experimenter to bias the test results.

1.4.1 Controlling the Technical and Environmental Variables

It is common knowledge that the listening room is a major factor in determining certain aspects of the sound of loudspeakers. Møller [28], for example, illustrates well the influence of the room and of loudspeaker placement on listener preferences. Olson [9] and others have emphasized the importance of listening in an environment appropriate to the product: a loudspeaker intended for the domestic market should be auditioned in a "typical" living room. Meeting this requirement is not entirely straightforward, since many domestic listening rooms (perhaps even a majority) have acoustic flaws that strongly characterize their sound. Still, there are common features, and it seems reasonable to specify a room with the appropriate volume and proportions in which the disposition of sound scattering and absorbing furnishings is typically domestic. In constructing a specific room it is possible, of course, to meet the overall requirements while avoiding many of the undesirable characteristics of rooms built without benefit of acoustical design.

The room used here was described in a recent paper [3] and is the prototype for the recommended room specified in the current IEC publication [27]. In brief, the room is 6.7 × 4.1 × 2.8 m (22 × 13.5 × 9.2 ft) in size, with a reverberation time of 0.34 ± 0.08 s from 250 Hz to 4 kHz, rising to 0.8 s at 40 Hz and falling to 0.2 s at 10 kHz.

It is important to note that the IEC publication also permits rooms that fall into rather liberal dimensional ranges. With care it is possible to design other rooms with proportions that ensure a good frequency distribution of the important room modes. (It is also possible to construct problem rooms.)

The important room modes are those that are most actively involved in the acoustical coupling of the loudspeakers to the listeners. Conventional formulas for the dimensional ratios of listening rooms place equal importance on all calculable room modes, an unrealistic requirement [3], [16]. Furthermore, in practical rooms the theoretical predictions of resonance frequencies and standing-wave patterns are compromised by the fact that the room boundaries are not perfectly flat and sound reflecting. Sound-absorbing surfaces and walls introduce phase shifts in the reflected sounds, an effect that sometimes can be used to advantage by positioning low-frequency membrane absorbers to move modal nulls away from listener head locations. The fine-tuning of a listening room is essentially empirical since the arrangement of loudspeakers and listeners is one that, in stereo, is severely restricted.

Fig. 1 shows the room arrangements used in these experiments. The loudspeakers were elevated, if necessary, to place the mid- and high-frequency drivers at the listeners' ear height when seated. Listeners were seated in low-backed chairs at locations determined by careful acoustical measurements [3, Appendix].

Ensuring equal loudness among the sounds being compared is of fundamental importance. Illényi and Korpassy [29], for example, found that listener ratings of studio monitor loudspeakers correlated well with their relative loudness.

In the present experiments, power amplifier gain was automatically adjusted to compensate for differences in loudspeaker sensitivity. The preset adjustments were made using A-weighted sound-level measurements when the test loudspeakers were fed with pink noise. A Brüel and Kjaer 4134 microphone, pointed toward the ceiling, was placed in the middle of the listener area at ear height. Listeners were later asked to report on loudness as a part of the listening questionnaire so that, in the event that the measurements of sound level failed to provide equalization of loudness, a subjectively satisfactory balance could be achieved. In practice there have been problems only with loudspeakers exhibiting very uneven frequency responses, where the apparent loudness depended upon the spectral content of the program material.

The sound level for program playback would, on the face of it, seem to be an important experiment parameter, given the sound-level dependence of so many aspects of the hearing system. Nevertheless Staffeldt [21] was unable to find any significant change in listener judgments with program levels ranging over 25 dB. Gabrielsson and Sjögren [30], on the other hand, did identify significant interactions between subjective ratings and sound level, with the strength and direction of the interaction depending upon the individual, the loudspeaker, and the program. Judging from the published measurements, it would appear that the loudspeakers tested by Staffeldt were rather poor by today's standards, and because of large measured differences, they may have been relatively easy to identify at any sound level. The same could be said of some loud-

speakers used by Gabrielsson and Sjögren. The conflicting experiences may be entirely the result of different perception thresholds for different technical flaws. For example, problems of directivity and resonance are likely to be detectable over a large range of sound levels, while response to spectral balance would predictably change because of the frequency and amplitude dependence of loudness.

There seems to be no perfect solution, but Gabrielsson and Sjögren conclude that program levels ranging from the original "true-to-nature" to 10–15 dB lower, and deliberately varied, are a practical compromise. The listeners in the present tests, all experienced in audio, tended to select levels that fell within a very narrow range, tending toward "true-to-nature" reproduction. Once established, the levels were maintained throughout the tests to avoid further complications due to this variable.

Choosing program material represents one of the most obvious opportunities for prejudicing the results of listening tests; virtually all experimenters have commented on the sensitivity of listener opinions to program selections. One of the most serious and most common sources of bias is the quality of the originating recording studio, a point made well enough by Walker in 1953 [6] and Somerville in 1954 [7], who described some early BBC experiences. There has been some progress since then, but without industry standards for monitoring, commercial recordings still can exhibit substantial variations in overall frequency response [3], [31]. The solution in the present experiments has been to use only carefully selected commercial recordings (thereby, of course, allowing experimenters to bias the results by their own taste and judgment), or to use recordings of traceable origin, some specially made, in which monitoring was done at realistic sound levels and the studio equipment was of high-fidelity caliber. Program material for the monophonic tests was selected for minimal interference effects in the stereo-to-mono conversion. Electronic imperfections should clearly be minimized, since there are no well-defined limits to the detectability of the various technical flaws.

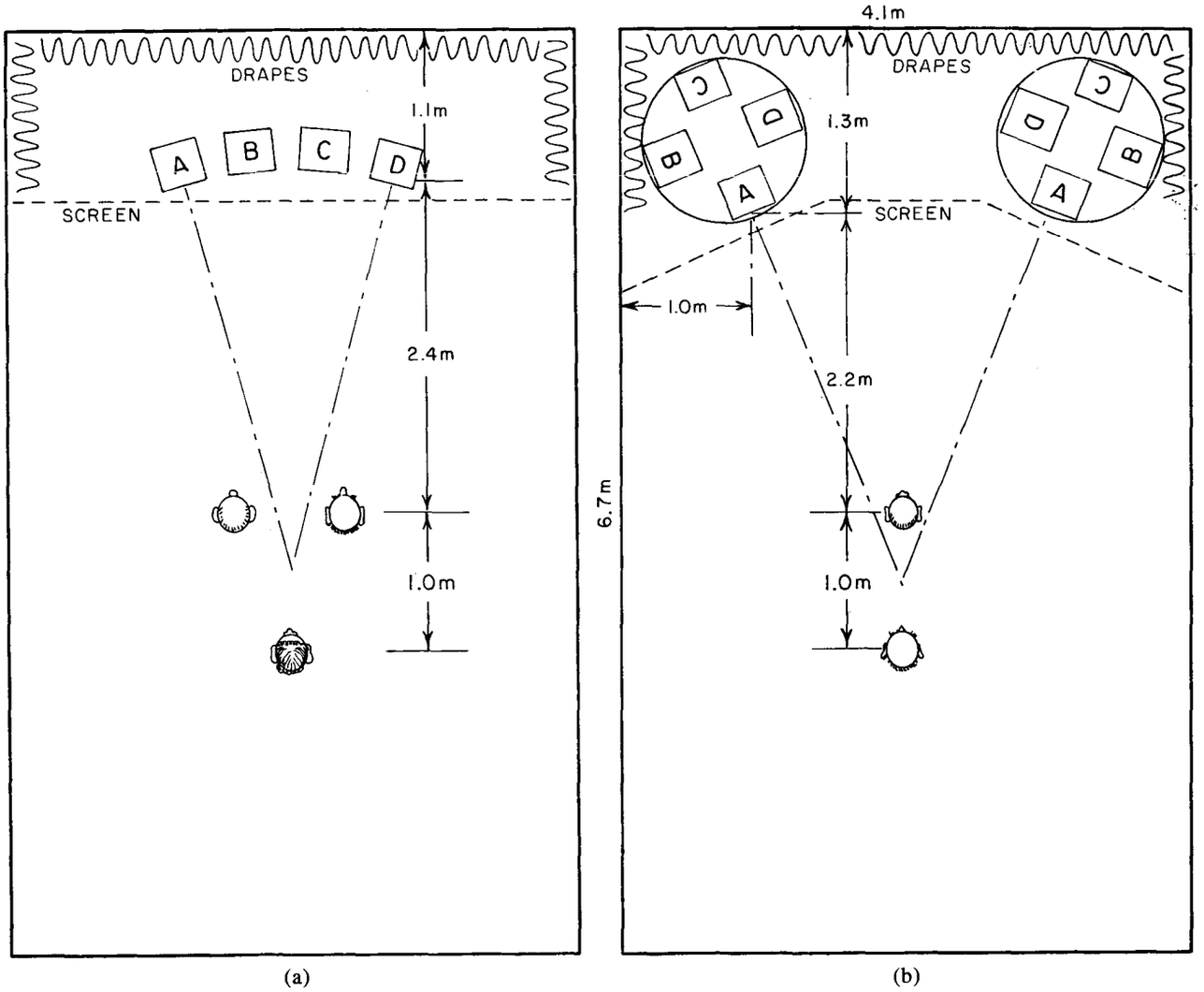


Fig. 1. Listening room arrangements for (a) monophonic and (b) stereophonic loudspeaker comparisons. In the stereophonic comparisons both listeners are within 5° of the loudspeaker axes. In the monophonic comparisons all listeners are within 9° of the loudspeaker axes. Front-row chairs are slightly lower than the rear chair so that the listener in the rear has an unobstructed acoustical view of the loudspeakers. The floor is carpeted; the ceiling is hard; the side walls between the listeners and the loudspeakers are hard and flat; the remaining walls are covered with sound absorbing and scattering objects.

Whether one employs stereophonic or monophonic listening is dependent on the importance attached to stereo itself. In the present experiments, the fundamental problem being addressed was transduction accuracy, so it seemed reasonable at the outset to avoid the acoustical interference occurring at the listeners' ears due to the excessive number of signal components in stereo presentations [3]. The complete evaluation did include stereophonic listening, but in the present learning situation it was decided to make that a matter for separate study.

1.4.2 Controlling the Listener Variables

Listeners who are aware of the identities of the products under test can hardly be considered impartial in their assessments. In these experiments, such partiality was eliminated by placing the loudspeakers behind an acoustically transparent but visually opaque screen. Furthermore, in most of these tests, the listeners were not even aware of the sample population from which specific test objects were drawn.

A lack of familiarity with the program material, the room, and the rather unusual (and fairly demanding) task naturally creates difficulties for new listeners. Experience is the only solution to this problem. In the early years of these experiments the program tape was quite short, and it was the practice to allow new listeners one or more rehearsal rounds before beginning the experiment proper. In the present tests the program tape lasted approximately 30 min and contained considerable redundant material, so most listeners were able to respond with some confidence within this time. Consequently results were accumulated from the first listening session; subsequent data processing may reveal the influences, if any, of experience.

One might naturally assume that if there are parallels with other acquired skills, there should be evidence of individual differences in aptitude or innate ability. Some listeners, by virtue of age, heredity, disease, traumatic injury, or noise exposure, may also have peripheral hearing apparatus that does not perform normally. And then there is experience: the sum total of critical listening practice, memories of reference sounds, expectations of reproduced sounds, sensitivities to various audible defects, and so on [32]. On the face of it, individuals would indeed appear to be distinctive, but perhaps this is not so much a variable to be controlled as one to be studied.

Fortunately various experiments have contributed to our perspective on the selection of listeners. Among recent work, Gabrielsson et al. and Killion and Tillman have both observed the superiority of listeners with critical listening experience. Gabrielsson [22], [30] found that listeners experienced in hi-fi listening exhibited higher reliability and a better ability to differentiate between products than did listeners without this background, even when the latter had considerable experience as concertgoers and musicians. Killion and Tillman [33] came to very much the same conclusion, stating that "where population sampling is not a major concern, one trained subject appears to be worth as many as eight untrained subjects."

In the present experiments, listeners covered a broad range of age and, therefore, experience. Many were musicians, all were interested in music, and all were either professionally involved with audio or were seriously involved as audio hobbyists. It was a carefully selected population in that all listeners brought to the listening tests a similar seriousness of purpose.

Whatever population of listeners is chosen, communication among them during and between listening sessions must be controlled rigorously to avoid group biases [34]. Even with precautions it is not uncommon for some groups to vote seemingly as a unit. Single listeners are, for this reason, the ideal choice, but usually impractical. A workable compromise is to divide the listeners into several small groups that have no opportunity to interact.

1.4.3 Controlling the Experiment Variables

Some of the most difficult problems in experiments of this kind are the identification of the parameters to be judged, the provision of simple and relevant scales for quantifying the judgments, and finding legitimate ways of comparing the assessments of individuals who may have attached different meanings to the descriptions of the perceptual dimensions or the response scale or both. These problems clearly require study before reliable decisions can be made.

Considerable study has already been devoted to the perceptual dimensions of sound-quality assessments [4], [20], [22], [35], [36], and there is a developing optimism that we can indeed quantify several relatively independent aspects of reproduced sounds. The present work draws from the research by Gabrielsson and his colleagues in the preparation of the response-reporting forms for sound-quality assessments.

It is well known that opinions about a particular loudspeaker can be influenced by the other loudspeakers that are available for comparison. The degree of influence will depend on whether the comparisons occur in rapid succession, as in the traditional A–B comparisons, or after intervals of time, as in so-called single-stimulus tests. Even the order of listening to different products can affect the judgments. All of these factors need to be taken into account in designing the experiment. Finally experimenters should not themselves be able to impose their personal biases—conscious or subconscious—on their listeners. All of the experiments in the present work were at least single-blind, and about 60% were double-blind.

1.5 Choosing an Experimental Method

Fundamental to this enquiry is the notion that it is not only possible to elicit from listeners responses of "better than" or "worse than," but also quantitative measures of *how much* better or worse. Simple rank ordering does not reveal the subjective distances between sensations. The measurement of subjective difference is, however, by no means easy.

1.5.1 A Discussion of the Options

The considerable published literature on the subject of the measurability of subjective events includes descriptions of several experimental methods that are capable, with appropriate data processing, of providing insight into the mathematical relationship between the magnitudes of the stimulus and the response [37]–[39]. Such techniques may be useful ultimately, but at the present stage there are fundamental uncertainties about the precision of sound-quality evaluations.

Consistency in the listener responses, though necessary, is not sufficient. It must also be shown that the responses are related in an orderly and logical manner to the physical continuum: the measured performance of the objects under test. Consistent judgments must be seen to fit within a logical framework related to physical reality so that the subjective and objective results may be said to express the same things in different ways. This, also, is not a simple problem, since the objective measurements themselves are a matter of question and debate. In fact, obtaining a numerical subjective rating for a product may well be easier than devising a comparable single-number objective rating.

Among the methods commonly employed for subjective evaluations are those that involve listening to and rating sounds one at a time (the single-stimulus method), and those that involve listening to a pair of different sounds before the ratings are required (the paired-comparison method). The latter includes the popular A/B or A/B/X techniques [40], [41]. Less popular in scientific experiments, but perhaps the most common method in real-life listening experiences, is what could be called the multiple-comparison method. All over the world there are dealers' showrooms, manufacturers' demonstration rooms, and even some recording control rooms in which any of several types of loudspeakers can be auditioned at the push of a button. Various manufacturers supply gain-compensating switching units for just this purpose. Such listening comparisons rarely take into account even the most obvious nuisance variables, and so the results are usually shrouded in doubt. Indeed there appear to be occasions when the nuisance variables are deliberately manipulated to bias the results.

Nevertheless, with appropriate controls, there may be advantages to an adaptation of this technique. Unlike experiments in which a judgment is called for at each sound presentation, everyday judgments often represent the result of an accumulated impression (an "integration") of a set of sound presentations. Once an integrated impression is formed, the individual rating appears to depend upon its place in a collection of similar ratings [42]. The restricted context of isolated paired comparisons could be a source of error, and this argues for at least several such comparisons within an experiment. Others [14], [43] have found more reasons to criticize restricted A/B tests.

In the multiple-comparison method it is acknowledged that all judgments are essentially relative: the rating of a stimulus depends upon the other stimuli with which it is presented. With the multidimensional sounds from loudspeakers, the effects of context are certain to be complex. Experience shows that in selecting an experimental set of loudspeakers the experimenter affects the subjective ratings of all of them. For example, the subjective rating of a mediocre product presented in the company of inferior products may well be higher than that given to the same product when it is compared to a group of superior ones.

The adaptation level theory proposed by Helson [44] holds that the comparative basis for category judgments is a weighted mean of the relevant perceptual values. Judgment, in other words, is relative, moving in opposite directions from a middle point, which is the perceptual "center of gravity" of the experimental range of perceived values, called the adaptation level. The theory may not be complete, but it is consistent with much common experience and considerable scientific data. It may be important, therefore, to ensure that, in any experiment, listeners be exposed to a more or less standard range of sound qualities, from good to bad, so as to establish a stable context for judgments.

Judgments are also influenced by memory. Sounds will tend to be compared with the perceptual memory of previously presented stimuli, for instance, or with some remembered response criterion, both of which are susceptible to bias and variability effects [45] Thus the order of presentation of the test sounds affects the ratings. Randomization is the usual solution.

Presenting sequences of stimuli for comparison raises the inevitable questions of how long the stimuli should be, and how long the interval should be between comparisons. To neither question is there a perfectly satisfactory answer. Prolonged exposure to one sound stimulus allows the listener to adapt to certain parameters of that sound. In the extreme, it becomes the basis for future comparisons, perhaps causing erroneous judgments [32], [46]. Short exposures presume that listeners can detect, identify, and quantify all of the relevant perceptual dimensions in the allotted time.

Instantaneous comparisons have long been viewed as a superior means of making useful evaluations, particularly when the differences are small. Scientific evidence indicates so far that when comparisons are delayed there is a tendency for the variability of judgments to increase and for discrimination to be reduced [46], [47]. Therefore the long-term "at home" listening exposure to one product at a time is a method that is plagued by nuisance variables and does not, consequently, lend itself to scientific evaluations. Adequate controls would be possible, but extremely difficult to maintain. The method that is psychologically the most persuasive, and certainly the most satisfying to the listener, may also be the most susceptible to bias and error.

In summary, it seems clear that relatively rapid comparisons are advantageous for maximum discrimination and minimum variability in the judgments. It is almost axiomatic that if comparisons are to be made among a

group of loudspeakers, the sequence of presentations should be balanced and randomized—all loudspeakers should be auditioned an equal number of times in sequences following all other products, and in an order that is randomized so as not to be predictable by the listeners. This places a practical upper limit on the number of products that can be included in an experimental set.

Also, since the particular selection of products included in the experiment establishes, to some extent, the norm by which all the products are rated, it may be advisable to ensure that listeners are exposed to a somewhat standardized range of sound qualities. A practical method has been to add to the test population some loudspeakers that are known from experience to represent useful "anchor" points on the subjective rating scales [3]. For example, a group of "good" test products may need some "poor" anchors, and vice versa.

Anchor products, therefore, need not be the current "best," or "reference," products preferred by some product reviewers. They can stand anywhere on the quality scale. The choice of anchors should, above all, be based on their previous performance in technical and listening tests; they should also be physically stable and reliable, since over months or years of listening they will be assumed to be the same; and they should have no strong idiosyncrasies, such as extremes of directionality or colorations that will cause them to be easily identified by listeners. This last requirement disqualifies some products with low ratings, since they might allow judgments associated with a recognizable product rather than developed from independent assessments of sound quality.

1.5.2 Scaling the Listener Responses

There are several formal methods of measuring subjective responses. Of these, perhaps the most direct is "magnitude estimation," in which subjects are required to estimate the strength of an event as a proportion of its original or reference intensity. With those aspects of sound that can be measured in amounts, for instance, such as loudness (the overall amount of the signal) or "brightness" (the relative amount of a portion of the signal), the stimulus allows the magnitude estimation to be tested by addition: when more of the same kind of stimulus is added, the measured magnitude increases. Unfortunately the mathematical relationship that links the listener's estimate of magnitude with the physical measure of the stimulus is not linear, but rather a power or log function.

Other aspects of loudspeaker performance, however, appear to operate on the principle of substitution rather than addition. A new stimulus is substituted for an old, but in a different location on the continuum or scale. The substitutive continua tend to be linearly related [39], and fall naturally to the use of partition scaling, where the subject locates a response on a numerical or otherwise partitioned scale that extends over some portion of the perceptual continuum. The subjective aspects of sound *quality* (as opposed to sound *quantity*) would appear, on the face of it, to fall into this category.

Selecting a standardized format for recording listener opinions about loudspeaker sound quality may seem to be a hopeless quest. The language of critical listeners tends to be closer to the language of poetry than of scientific measurement. Fortunately many words and expressions are simply different ways of saying the same thing. The work of Gabrielsson [35] reveals that underlying a multitude of descriptors there are only a small number of substantially independent perceptual phenomena.

In the present experiments, listener response reporting forms were developed (Figs. 2 and 3). For sound-quality assessments Gabrielsson's eight perceptual dimensions and two overall ratings—pleasantness and fidelity—were used. So as not to restrict the language or content of listener responses, a comments column was provided, and it is gratifying that it was used regularly. In fact, as a result of monitoring these comments, it may be possible to expand or modify the repertoire of dimensions used in the rigid format. Assessments of spatial quality were required in some of the tests, and these were made on a number of scales (Fig. 3) that seemed to embrace most listener comments in a series of pilot tests. The selection of these perceptual dimensions was not as scientifically rigorous as the selection of the sound-quality dimensions. However, there were very few additional comments from listeners in the succeeding experiments, indicating that the questionnaire allowed listeners to describe adequately their impressions of stereo "imaging."

The two overall ratings, pleasantness and fidelity, are in different ways a synthesis of the perceptual dimensions in sound quality. The fidelity rating is intended to reflect the extent to which the reproduced sound resembles an ideal. With some music and voice the ideal may be a recollection of live sound; with other material the ideal must be what listeners imagine to be the intended sound. Pleasantness is self-explanatory. The use of this rating is an attempt to achieve a less technical assessment that might be a parallel to what some people call "musicality." In any event, it is a rating that some listeners find less intimidating than "fidelity," and others regard as a necessary adjunct.

In general, however, the fidelity rating is regarded as the single number that sums up a listener's opinion of the sound heard. Forcing the listener to respond on a variety of other, more analytical scales is useful for gaining further insight into the perceptual process and, eventually, for diagnosing the virtues and problems of specific products. Whether they are immediately useful or not, the analytical ratings are important as a mnemonic device, and force listeners to evaluate several important aspects of the sound before arriving at an overall assessment. Inexperienced listeners, especially, may tend to concentrate only on the most obvious features.

The scale chosen for the present experiments is one

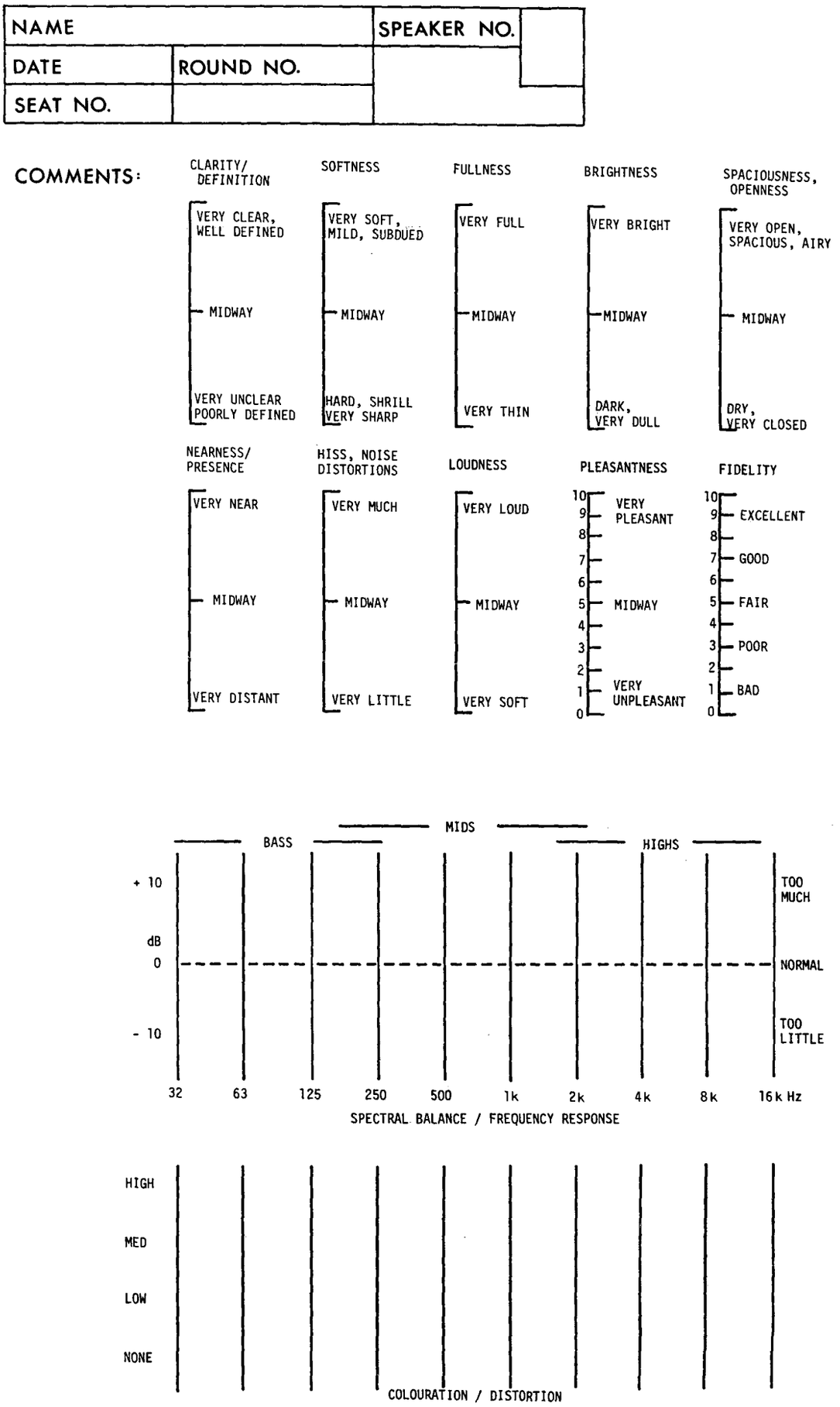
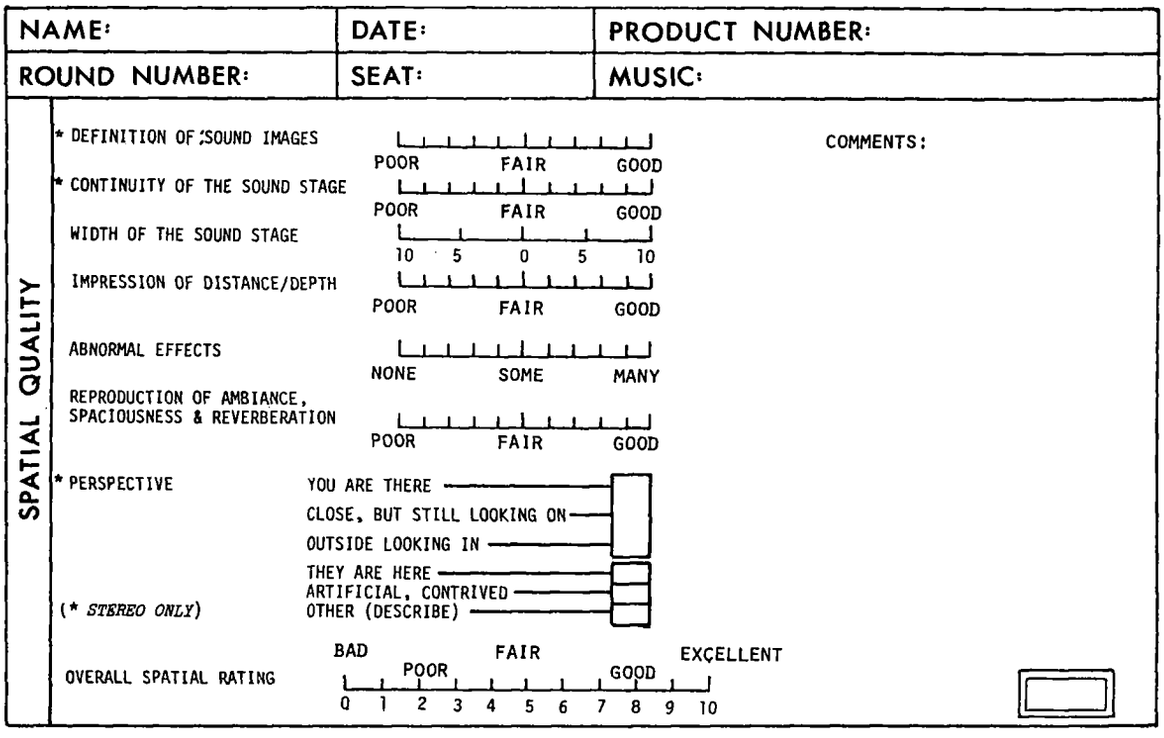


Fig. 2. Questionnaire used by listeners for evaluation of sound quality. See Appendix for instructions to listeners.



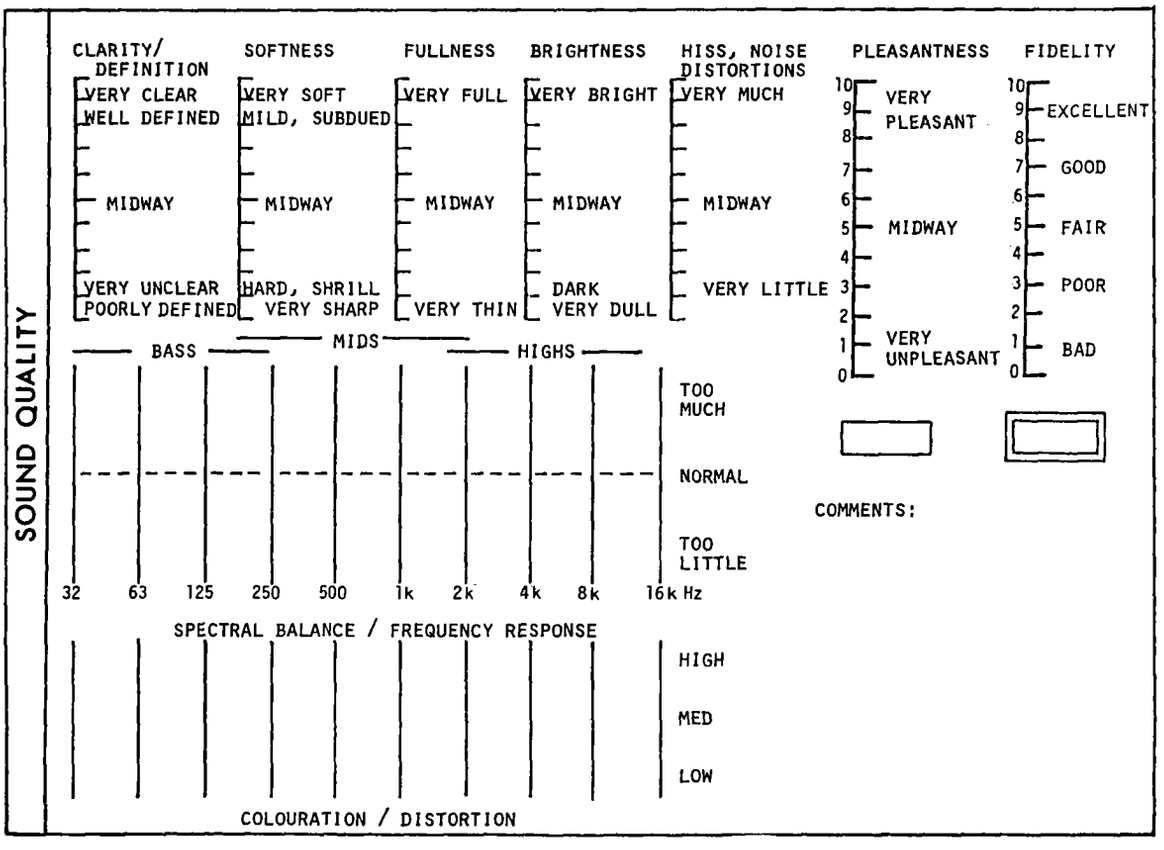


Fig. 3. Questionnaire used by listeners for evaluation of sound and spatial quality. See Appendix for instructions to listeners.

that we have used for several years, which is also the scale suggested in the IEC [27]. Ratings of pleasantness and fidelity are made on parallel numerical scales. Beside the scales are verbal indications of the meanings to be attached to the numbers. In these experiments listeners were instructed to regard 10 on the fidelity scale as a reproduction that is perfectly faithful to the ideal, no improvement being possible. The number 0, on the other hand, denotes a reproduction that has no similarity to the ideal—a worse reproduction cannot be imagined.

To further "anchor" the meaning of the fidelity rating numbers in the mind of the listener, it was suggested that a telephone might be rated 2.0 and a typical portable radio 4.0 on this cale. (Gabrielsson and Lindström [48] recently tested the suggestion and found that in reality a radio was rated between 2.0 and 3.0 and the telephone between 0.2 and 0.7.) It is normally expected that ratings of 0 and 10 will not be used; in the range between, listeners may report the subjective ratings with one decimal place.

A significant problem with experiments of this kind is the extent to which listeners' opinions are influenced by the program material. It is almost self-evident that transducer problems will be revealed only when the test signal possesses appropriate spectral or temporal characteristics; an oboe is not likely to reveal low-bass problems, chamber music is not likely to stress the upper regions of dynamic range, and so on. The variations and interactions are numerous, and it is a problem made more complex by the fact that the measuring device (the listener) is itself susceptible to different interactions with the test signal. Auditory masking is sufficiently strong, for instance, that the music itself reduces the audibility of some kinds of distortion [49]. Distortions that are inaudible for some orchestrations may suddenly be revealed in another selection of instruments or sound levels. Similarly, individuals differ in their hearing sensitivity, particularly to high-frequency sounds.

Various experimenters [21], [22], [3] have noted the interactions between listener ratings and program selections. In the present experiments the decision was made to avoid the complication of requiring individual assessments for each selection of music, as it rapidly multiplies the amount of data to be processed and reveals information of value only if it is possible to relate the assessments to technical features of the specific program selections. Instead, listener reactions to individual music selections produced "scatter diagrams" on the perceptual dimension scales. The overall ratings of fidelity and pleasantness represented the individually weighted and integrated assessment of all perceptual dimensions for all musical selections.

1.5.3 Technical Ratings

To some listeners the qualitative descriptions are merely abstract ways of describing aspects of performance that are more precisely (and perhaps more directly) described by technical terms. To persons who associate what they hear with a resonant boost at 3 kHz, it seems imprecise to say that the sound appears "hard," "a little bright," and has some "presence." For such listeners, forms are provided (Figs. 2 and 3) on which to draw frequency response and distortion curves. The results often tend to be somewhat impressionistic, but a few listeners are remarkably astute as judges of technical performance.

1.5.4 Experimental Apparatus

Apart from the listening room itself, the only specialized apparatus was a gain-compensating selector switch that allowed for equal-loudness reproduction through any one of four loudspeakers (or stereo pairs). Gain adjustment was by potentiometers placed ahead of the power amplifier, and all switch contacts were by redundant relay contacts. An illuminated display identified by number the loudspeaker being used.

Program material was derived from either high-quality commercial phonograph recordings or analog or digital master recordings. In the early experiments reported here, multiple replay was achieved by transferring the program excerpts to analog tape (Revox A-700), although the majority employed a digital (PCM) tape recorder (Technics SV-P100). Disk transfer was with Shure V-15V and Technics EPC-P205C Mk3 cartridges in a Technics SL-1000 Mk II turntable/tone-arm/base combination, fitted with an Oracle mat. Bryston 1B preamplifiers and 4B power amplifiers completed the complement of recording and playback apparatus.

The monophonic and stereo/mono series I tests used material excerpted from the following commercial disks, with the exception of about 12 min of the Canadian Broadcasting Corporation (CBC) series I test tape, which was excerpted from CBC analog master tapes

Bach: *Brandenburg Concertos*; English Chamber Orchestra, Raymond Leppard (Philips 6747 166)
Handel: *Messiah*; The Academy of Ancient Music, Christopher Hogwood (L'Oiseau Lyre D189D3)
Orff: *Carmina Burana*; London Symphony Orchestra, André Previn (EMI, Mobile Fidelity MFSL 1-506)
Mussorgsky-Ravel: *Pictures at an Exhibition*, Chicago Symphony, Solti (London LDR-10040)
Laudate (choral collection) (Proprius PROP 7800)
Jazz at the Pawnshop (Proprius PROP 7778)
Peoria Jazz Band (Opus 3 79-00)
Harry James (Sheffield Lab-6)
Christopher Cross (Warner Bros. QBS 3383)
Pink Floyd: *The Wall* (CBS 36183)
Offenbach: *Rock Bottom* (Spectra Scene SS1702)

The program for CBC series I was recorded on the Revox A-700 at 380 mm/s without noise reduction. Subsequent test programs were recorded on the Technics SV-P100 PCM digital audio recorder.

In stereo/mono series II tests, the musical program consisted of transfers made from analog and digital

master tapes of concert hall and studio recordings of known origin.

The air-conduction hearing threshold levels of the listeners were measured with a Madsen model OB40 audiometer, calibrated to ISO-1964 hearing threshold levels.

2 THE EXPERIMENTS—TESTING THE TESTS

The very large number of experimental variables precluded a randomized examination of all their influences. It seemed probable at the outset, however, that one of the most influential factors would be the acoustical coupling of the loudspeaker to the listener through the listening room. Accordingly, the initial series of tests were monophonic, thereby avoiding the positional restrictions of stereo listening.

For all the experiments, listeners were instructed to avoid exposure to high-level sounds for at least 12 hours prior to the first session. Upon arrival on the first day, listeners' air-conduction hearing threshold levels were measured for both ears, over the frequency range of 125–8000 Hz. (Standard audiometric measurements lose precision rapidly above 8 kHz.)

Standardized instructions were presented either verbally or in written form (Appendix). Questions were answered and listeners were seated in the numbered, specified locations. In successive rounds, listeners moved in rotation from seat to seat to average out some of the residual room-position effects.

Loudspeakers were picked in sets of four from the test group. The selection was ordered only to the extent necessary to ensure that each product was auditioned the appropriate number of times.

2.1 Monophonic Listening Tests

2.1.1 Procedure

In the monophonic tests a "round" consisted of auditioning four different loudspeakers positioned in a row behind a screen [Fig. 1(a)]. An operator sat at the back of the room and switched manually from loudspeaker to loudspeaker in quasi-random sequence throughout the entire program tape. Switching took place at intervals of between 5 and 15 s, depending mainly on the nature of the music. As much as possible, the switching occurred between repetitive passages, in sympathy with the music. Operators were used in rotation so that no one person's habits would prejudice the results. Most of the time, the operator was not aware of the products in the round he or she was switching. Listener requests to do their own switching were denied unless there was only a single listener. The switching sequence was found to be a major source of nonverbal communication in groups.

Listeners correlated their impressions with the large display numbers. Between rounds, the loudspeakers' locations were changed, or the products themselves were exchanged, according to a prearranged sequence; the relation between the display number and the product was altered; the loudness levels were equalized; and the system gain was set to provide the same playback sound level at each session. The rest period for listeners was at least 30 min, during which time they were cautioned not to discuss the tests. The listeners were not told which products were under evaluation until the entire series of tests was completed. There were no more than six listening sessions per day.

Tests of this kind have been conducted in this laboratory for several years; most, however, involved only a few loudspeakers and listeners. The first experiments reported here were the result of a large-scale investigation of monitor loudspeakers conducted in collaboration with the Canadian Broadcasting Corporation/Radio-Canada, a large nationwide network of AM, FM, and television stations. Loudspeakers were needed in three power output classes to equip large popular and classical music studios, modest station control rooms, and portable recording and remote-broadcast units. The decision to meet these diverse needs with loudspeakers meeting sound-quality standards that were not only high but similar meant that the project was particularly well adapted to the procedures described here. The scale of the project meant that a large body of coherent experimental data would result [50].

Other data came from experiments involving loudspeakers and listeners selected to test specific hypotheses. The details of these will be discussed in the presentation of the results.

2.1.2 CBC Series I—A Large Test with Some Real-Life Problems

CBC series I involved 16 people who auditioned 16 loudspeakers. The listeners were all highly successful professional audio people with pressing responsibilities. As a consequence, the tests were disrupted by people arriving late, departing in the middle of the series, and being forced to cancel the appointment entirely. Such are the problems of real life.

To accommodate the enforced time constraints, it was necessary to place five listeners in a group rather than the recommended maximum of three. The tests were organized in three groups, the sixteenth listener exchanging places with an early departure to complete the series. Each group stayed two days, participating in five rounds one day and six the next.

That the listeners would, in their turn, occupy inferior listener positions was a problem. There was also a problem with the music program: compared to high-quality disk transfers, some of the broadcast tapes were noisy, of audibly restricted bandwidth, and distorted. In addition, several of the listeners had significant hearing loss—in two cases sufficiently serious to place them in the impaired category.

As a result, the precision of series I was so compromised that it was regarded from the outset as a "coarse filter," capable of identifying the poorest products but incapable of resolving the fine differences among the best. Nevertheless, as it turned out, series I produced an important set of data.

2.1.3 CBC Series II

Based on the group fidelity ratings, but considering also the relative prices, seven loudspeakers were selected for the final test. Listeners for this series were selected for their youth, hearing performance, and experience. Some could be classified as "golden ears" in audiophile terms.

The 12 listeners participated in groups no larger than three, using excerpts from commercial records transferred to digital tape as program material. Each listener made three or four assessments of each loudspeaker over a period of two days.

2.1.4 Audiophile Series I

These tests were conducted for a consumer audio publication and involved seven review products and one anchor product from the previous CBC series. All four listeners were audiophiles and audio writers, two of whom had experience in loudspeaker design. All had experienced the test procedure on several previous occasions, and all had essentially normal hearing. Each listener made four assessments of each loudspeaker over a period of two days.

2.1.5 Audiophile Series II

To expand the data base for statistical analysis and to resolve a matter of scaling and variation in fidelity ratings, four loudspeakers were selected from the audiophile series I group, and 12 new listeners participated alone and in groups of two or three. There was in this series a deliberate attempt to expand the age and hearing loss range compared to the previous two, but ironically, it proved difficult to find people with significant hearing loss. Each listener made four assessments of each product within the same day. All listeners were serious audiophiles.

2.2 Results and Discussion

Although listeners reported on all categories in the response form (Fig. 2), the analysis of the complete data is not relevant to the immediate purposes of this paper. The following discussion therefore pertains only to the final overall rating: the fidelity rating. Loudspeakers are identified by letter codes.

2.2.1 Normalization of the Data

In subjective measurements, it is perhaps inevitable that individual test subjects—in this case listeners—will adopt slightly different scale references in reporting on the same phenomena, however carefully the design of the experiment seeks to impose standardization. In order that individual responses may be directly compared to one another, and that group data be compiled, a process of "normalization" is necessary, which adjusts the individual results to conform to an overall average scale.

Fig. 4, for example, shows the raw results of CBC series II tests. The amount of scatter in the individual listener ratings does not inspire confidence in the closely spaced group mean ratings. However, inspection of the data reveals that some individuals, such as "open triangle" and "circled asterisk," are simply responding on a lower portion of the fidelity scale than, for example, "black dot" or "asterisk." There is a different fixed bias for individual sets of data that appear to be independent of the relative ratings of the products.

If we assume that all listeners rate the products in approximately the same manner; we can measure the different scaling biases by determining the "center of gravity," or mean value, of each person's judgments on the complete set of loudspeakers tested. The scaling bias can then be removed by adjusting each listener's ratings by the appropriate constant amount to bring all the listeners' experiment means to the same number. In terms of the relative ratings of the products, this target number, or norm, is immaterial. However, to retain a semblance of absolute meaning in the ratings, the norm was here established to be the group mean—the mean of all listeners' ratings on all products in this experiment.

The assumed linearity of the fidelity scale (Sec. 1.3.2) permits the use of a variety of important statistical parameters such as the arithmetic mean; the product-moment correlations, and so on. In the present case it means that the constant corrections are additive rather than multiplicative [37]. Normalization, therefore, is a simple process of adding to, or subtracting from, each listener's ratings the amount necessary to bring

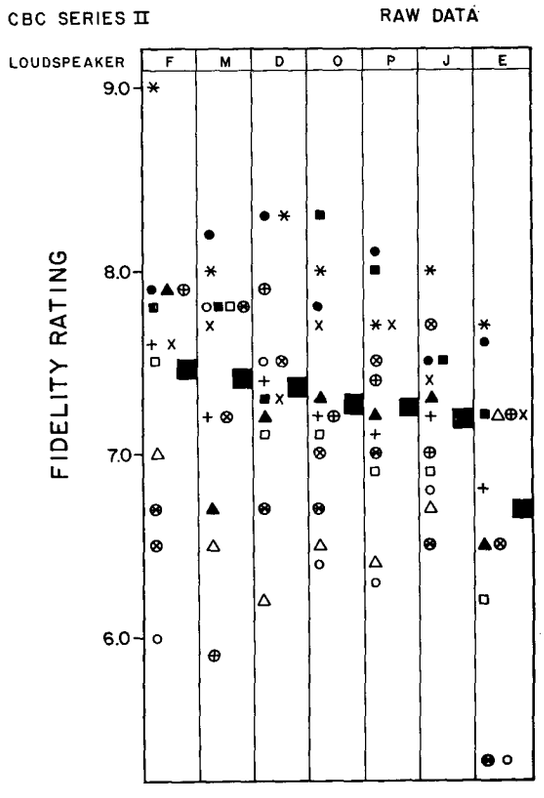


Fig. 4. Means of fidelity ratings for individual listeners (small symbols) and group of 12 listeners (large squares) for loudspeakers in CBC series II. Data are shown in raw form. Loudspeakers are identified by code letters.

the experiment mean for the individual listener into agreement with the overall experiment mean.

Following this procedure, we can transform the raw data for CBC series II in Fig. 4 into the normalized data presented in Fig. 5. It should be noted that the individual listener's relative ratings of the products are not changed, nor are the relative or absolute group ratings altered. The reduction in scatter is, however, substantial.

Immediately evident in the normalized data are the similarities and the occasional differences in the individual listener ratings of the products. The grouping of the ratings for loudspeaker O is particularly striking, where eleven of twelve listeners rated the product within 0.3 fidelity scale units. Even in the other product ratings there is good evidence of a central tendency among the individual listener ratings. Notable in another respect is the rather broad spread of ratings for loudspeaker M, where it would seem that some listeners differed in their tastes, or in their ability to hear certain attributes of this product.

In general, however, these 12 listeners rated six of the seven loudspeakers in very much the same class, and with near unanimity placed loudspeaker E in a lower category.

2.2.2 Comparing the Ratings: A New Variable Is Revealed

The importance of any experimental result is enhanced if it can be repeated. In this case the second experiment has already been done, within the context of CBC series I. Fig. 6 shows the averaged ratings for the 16 listeners in series I for all 16 loudspeakers in the test. The data from CBC series II are shown for comparison, but another normalization was necessary. As pointed out in Sec. 1.3.1, there is a tendency for ratings to be influenced by the range and distribution of ratings that occur within the experimental set of loudspeakers. In this case, the seven products in CBC series II, judged in isolation, were assessed on average to be 0.3 scale units lower than when judged as part of a larger group containing products with substantially lower ratings. Given the opportunity to compare good products with inferior ones directly, listeners rated the good products higher than when the good products were compared with one another in isolation. Consequently for this comparison, the ratings of series II have been elevated by 0.3 scale units.

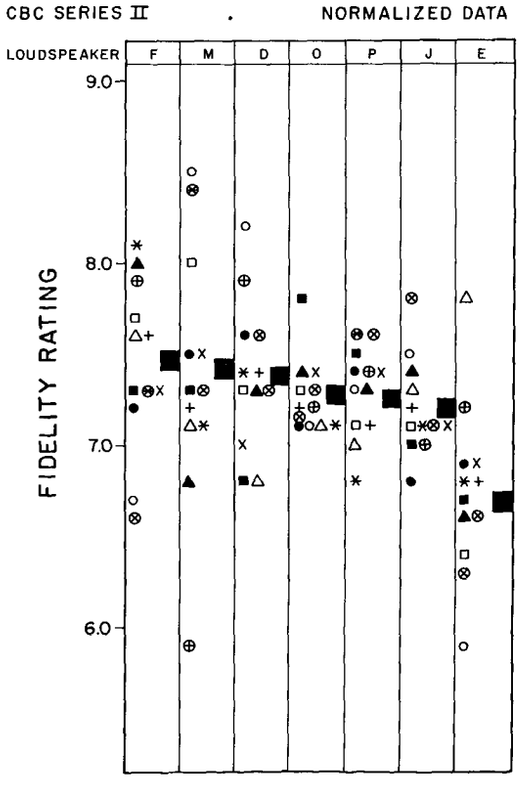


Fig. 5. As Fig. 4, after normalization of data.

The comparison between series I and series II is satisfying in that at least the data are similarly grouped; but there are some substantial shifts and reversals of ratings. In looking for logical explanations for these differences, it became apparent that listeners with the most consistent responses tended to agree most closely with each other, whereas listeners exhibiting large variations in repeated product ratings tended to produce nonconforming averaged ratings.

All listeners in CBC series II exhibited "mean standard deviations" (a measure of lack of consistency) of less than 1.0 fidelity scale unit. Grouping CBC series I listeners according to this criterion revealed that the inconsistencies noted earlier (F and O in Fig. 6) were clearly attributable to those listeners with high judgment variability and that, in isolation, the seven listeners with low variability showed good agreement with the 12 listeners in CBC series II.

That individual listeners should not perform identically comes as no surprise. Neither is it surprising that groups of individuals can share opinions about sound quality. In the past such trends in opinion have largely been attributed to taste or conditioning. What *is* new is the appearance of a relationship between opinion itself and the stability of the opinion when expressed repeatedly.

It would be especially useful if a common factor among listeners exhibiting the same kind of performance could be identified. Not anticipating rapid success, but pursuing one obvious course, a careful examination was made of the listeners' audiometric performances.

Combining the audiograms of listeners in the two categories of judgment variability produced the results of Fig. 7. From this it is fairly evident that the high-variability listeners tend to have less sensitive ears (a higher hearing level; 0.0 dB represents the statistical norm for healthy ears) than the more consistent judges. At the same time, some of the consistent listeners in CBC series I had substantially reduced sensitivity at high frequencies. The one feature of hearing threshold performance that relates most directly with judgment variability is the hearing level at middle and lower frequencies.

To test the hypothesis, the data were simplified by averaging the hearing threshold levels above and below 1000 Hz, and these new data were plotted against the corresponding mean standard deviations for the individual listeners (Fig. 8). The correlation coefficients and "best-fit" straight lines confirm that the more reliable indicator of listener variability is the hearing level at frequencies *below* 1000 Hz.

That the hearing level should be a factor in subjective judgments of sound quality was not entirely unexpected. That the strong association should be with the hearing level at the lower frequencies is surprising, especially since the relationship is strongly developed over the range of hearing level less than 20 dB, a range that in audiometric terms is regarded as representing acceptably normal hearing [51]. Perhaps the possession of hearing that is adequate for speech communication, the conventional criterion of normality in hearing, is insufficient for the especially critical task of judging sound quality.

Generally speaking, hearing loss at low frequencies is accompanied by at least the same loss at higher frequencies, although this is not invariably the case. In particular, if the hearing loss is purely conductive (that is, excess attenuation in the outer and middle ear), the high frequencies may or may not be affected. Positively identifying conductive hearing loss as a significant determinant of performance in listening tests is not possible without more comprehensive audiometric tests, but hearing loss at low frequencies is a strong indicator [51].

The two listeners with low-frequency hearing level in the vicinity of 30 dB are obviously functioning with a handicap. It is interesting that their performances are somewhat better than would be predicted by a best-fit line plotted through the remaining data points. It may be significant that these are among the most experienced listeners in the group, and they take pride in their judgment ability. Perhaps constant practice has allowed them to overcome their handicaps partially.

One common contributor to conductive hearing loss is age, and it is reassuring (only in a scientific sense) to see in Fig. 9 a moderate positive correlation between judgment variability and the age of the listener.

2.2.3 A Measure of the Error Due to Nuisance Variables

Fig. 10 shows the mean standard deviation of fidelity ratings plotted against the low-frequency hearing level

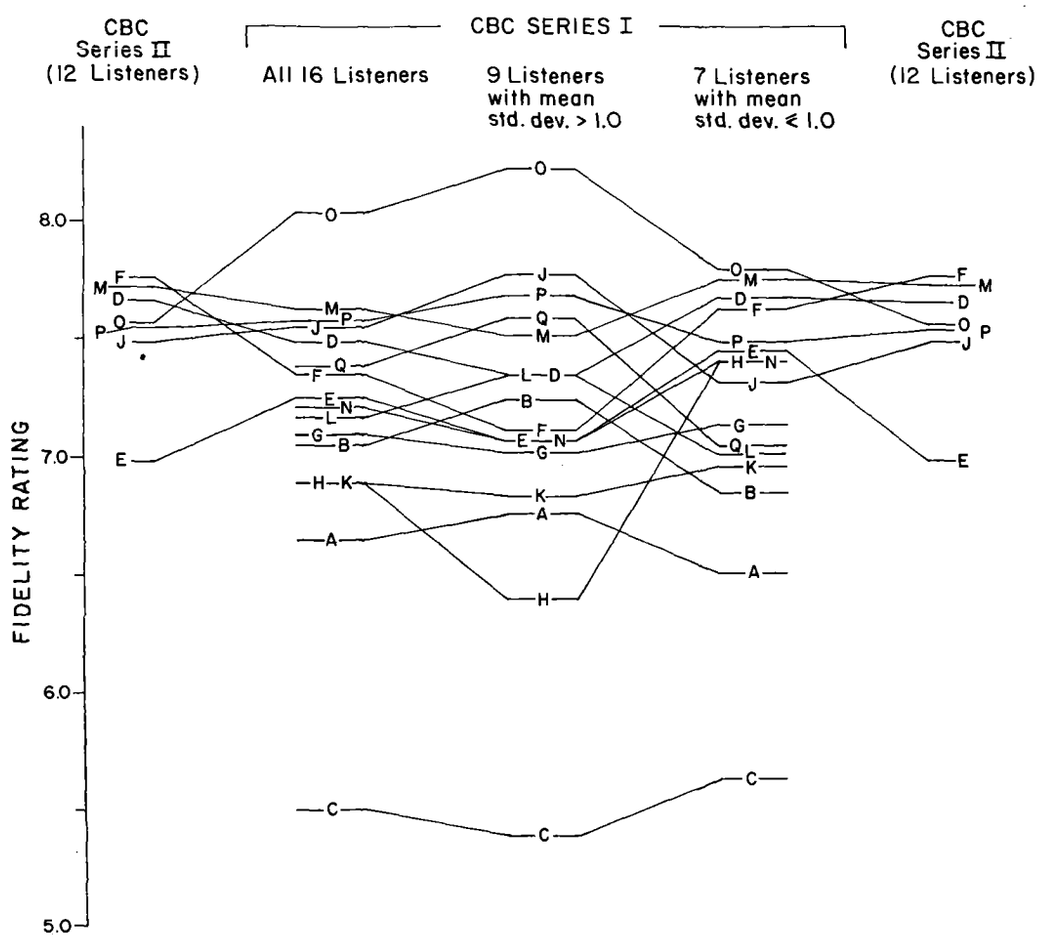


Fig. 6. Comparison of fidelity ratings for listeners in CBC series I and II experiments. Letter symbols represent group mean fidelity ratings for appropriate loudspeakers. In this presentation the CBC series II ratings have been elevated by 0.3 scale units to normalize the ratings of this set of products with the same set of products in series I; see text for an explanation. For convenience the series II data are shown on both sides of the display.

data for the listeners of CBC series I and II. Because of the careful selection of the listeners in series II, the data points are so tightly grouped that a trend can barely be discerned and the correlation coefficient is rather low (0.28). The slope of the linear regression is nevertheless very close to that seen in Fig. 6, though lower on the graph.

Recalling that the precision of CBC series I was compromised by reduced control of several nuisance

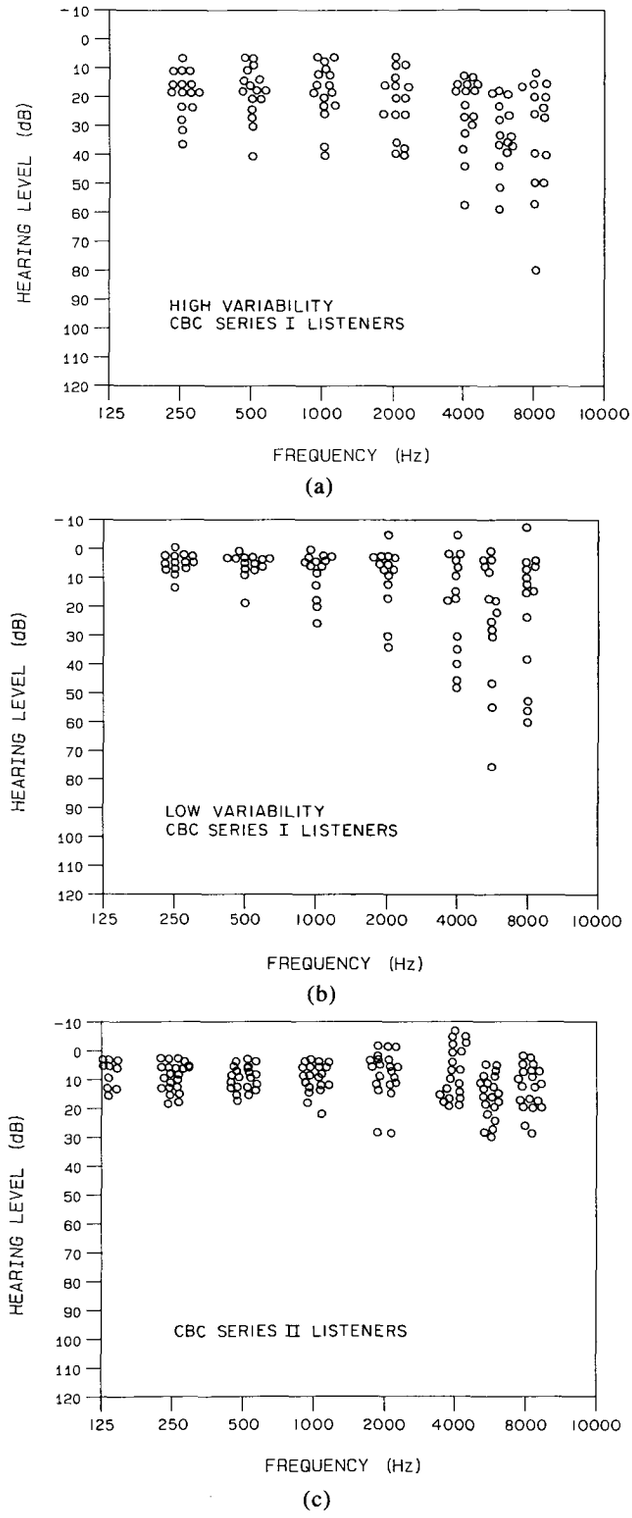


Fig. 7. Hearing threshold measurements for both ears of listeners participating in CBC experiments. (a) Series I listeners with mean standard deviations greater than 1 scale unit. (b) Series I listeners with mean standard deviations less than 1 scale unit. (c) Series II listeners (all listeners had mean standard deviations less than 1 scale unit).

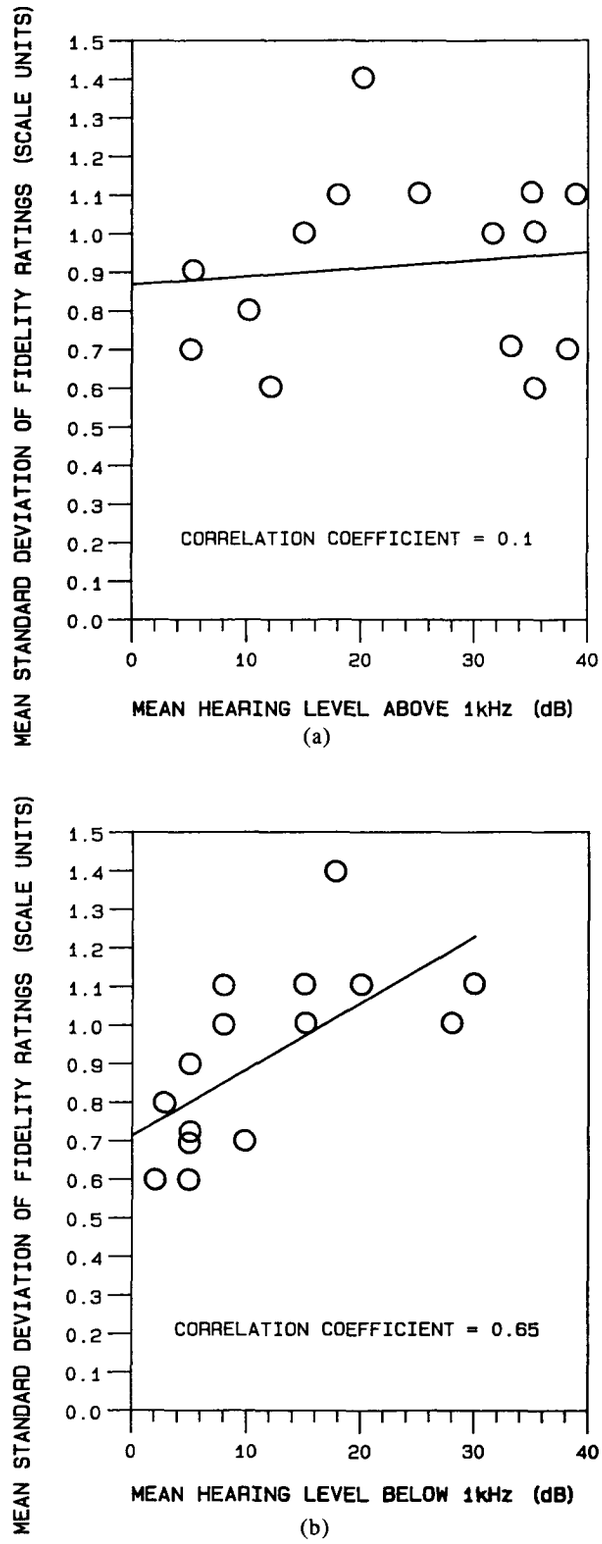


Fig. 8. Relationship between judgment variability (mean standard deviation of fidelity ratings) and hearing threshold level averaged over frequencies (a) above and (b) below 1 kHz. The best-fit straight lines (linear regressions) through the data are shown, as are the correlation coefficients. The data are from CBC series I.

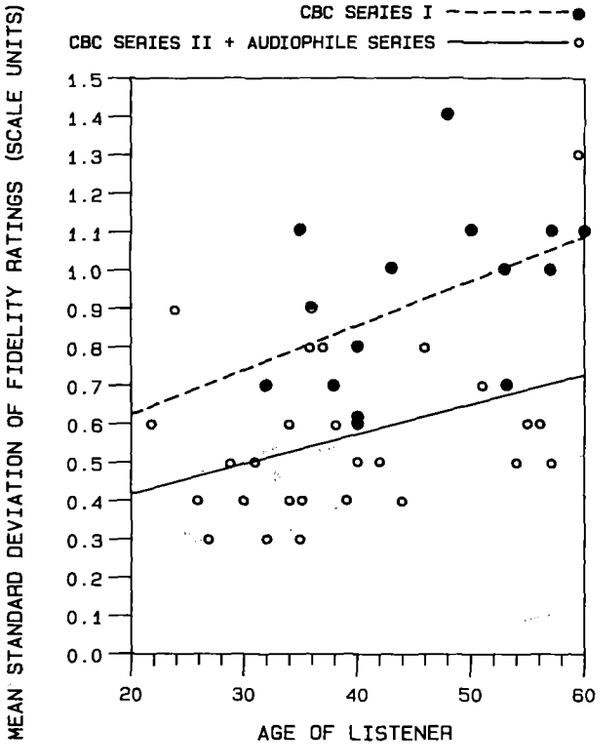


Fig. 9. Relationship between judgment variability (mean standard deviation of fidelity ratings) and age of listeners. The best-fit straight lines are shown for data from CBC series I and also for data from CBC series II and audiophile series combined. The correlation coefficient is 0.46 for the former and 0.36 for the latter.

variables (Sec. 2.1.2), it is interesting to see what appears to be a measure of the amount of error contributed by the reduction in experimental controls. Comparing the two sets of data in Fig. 10, the grouping of listener performances is sufficiently tight in both cases that the vertical difference between the distributions is very clear. Using the linear approximations as guides and extrapolating backward to the "perfect" listener at zero hearing level it can be seen that the mean standard deviations of listener fidelity ratings were reduced by about 0.35 scale units, or by about a factor of 2 between the two series.

It would be satisfying if the reduction in judgment variability could be attributed positively to the improved experimental method. However, there is another possible explanation: listeners in series I worked with ratings that covered much more of the fidelity scale than was occupied in series II (Fig. 6); the reduced range of responses could have led to a proportionately smaller range of judgment errors. Clearly another experiment is needed to resolve the uncertainty.

The two audiophile series experiments provided the resolution to this problem and served also as a test of the test. In these experiments, with the exception of one loudspeaker, the products were different, the listeners were different, and the program music was different. However, the room and the experimental controls were as closely as possible the same as those exercised in CBC series II.

The detailed results are discussed later; for now it is sufficient to show the data of Fig. 11, which reveals

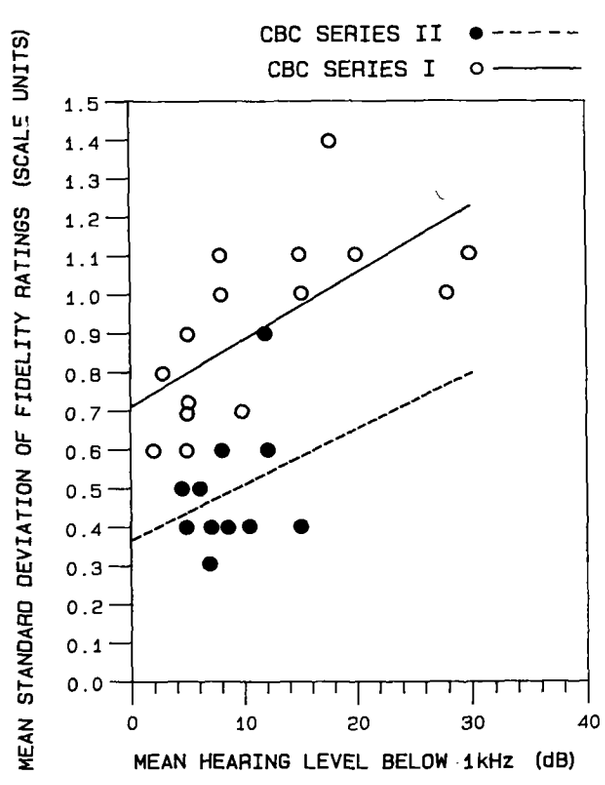


Fig. 10. Comparison of judgment variability data for listeners in the two CBC listening tests.

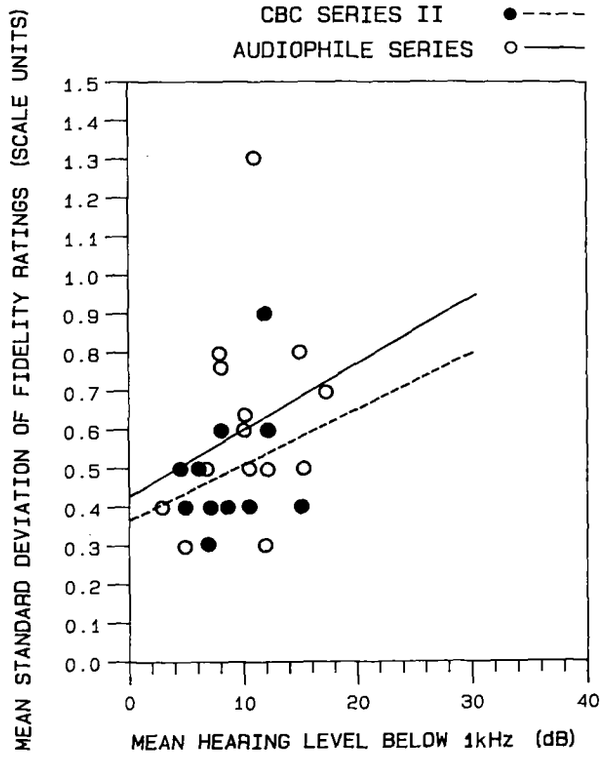


Fig. 11. Comparison of judgment variability data for listeners in the similarly well controlled tests CBC series II and audiophile series I and II.

that 28 listeners in three separate experiments exhibited essentially the same performance in terms of judgment variability and hearing level. Since the range of the fidelity ratings in the audiophile experiments was as large as that in CBC series I, we may conclude that the factor-of-2 improvement in judgment consistency in the CBC series II was most probably due to the improved control of nuisance variables.

Given the fact that the CBC series I was, by normal standards, a well-controlled experiment, there seems to be considerable potential for erroneous judgments in essentially uncontrolled everyday listening experiences, particularly those in which repeated assessments are not called for, and listeners are not screened.

Audiophile series I (Fig. 12) is an example of an optimized experiment, with four experienced listeners with near-normal hearing. The results are unambiguous. Individually the listeners exhibited mean standard deviations that averaged 0.55 fidelity scale units, and across the group, the agreement was of a similar order. Loudspeaker D also appeared in the CBC tests (Fig. 6), and it is interesting to look at the similarity of the assessments. Comparing the results of the seven low-variability listeners in CBC series I, all 12 listeners in CBC series II, and the present four listeners, the group normalized fidelity ratings for this loudspeaker were 7.67, 7.67, and 7.62, respectively.

Results of this caliber are impressive, and in the present example they provided a strong basis for consumer product reviews in which the publication could justify expressing somewhat more than the usual amount of candor.

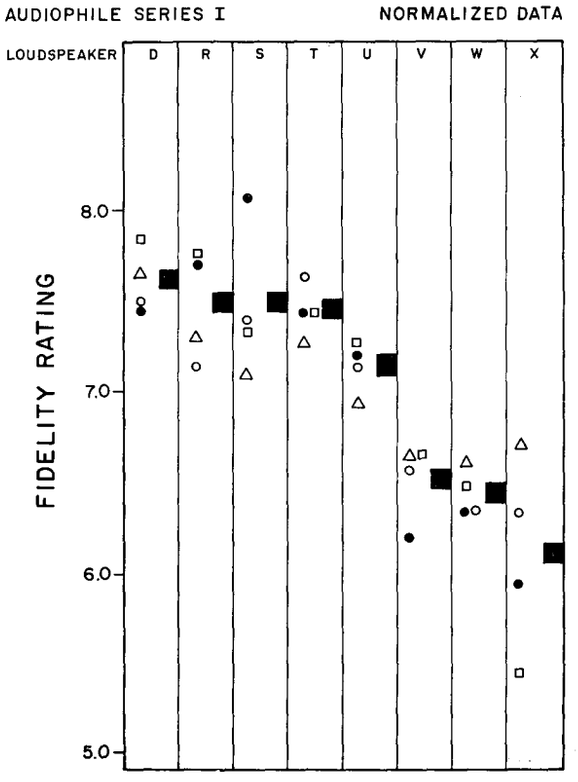


Fig. 12. Mean fidelity ratings for individual listeners (small symbols) and group of four listeners (large squares).

2.2.4 Judgment Bias in Individual Listeners

Evidence of bias has already been seen in listeners grouped according to judgment variability (Fig. 6), but it is clearly a matter worthy of more extensive examination.

Fig. 13 displays individual fidelity ratings on four of the products in CBC series I. The ratings have been displayed on the horizontal axis according to the mean hearing level, measured at frequencies below 1 kHz, exhibited by the 16 listeners in the test. Best-fitting straight lines were computed to fit the data. However, it appeared from visual inspection that the two listeners with about 30-dB hearing level did not always conform to the same trends exhibited by listeners with 20-dB hearing level or less. Consequently a second set of best-fitting lines were calculated to fit these data.

It is evident that as the hearing level increases, the fidelity ratings of some loudspeakers rise and others fall. Fig. 14 displays the best-fitting lines computed for all 16 loudspeakers in CBC series I, showing a wide diversity of trends.

As a further examination of this phenomenon, and as a test of earlier observations, the data were reorganized to show the fidelity rating as a function of listener hearing level below 1 kHz, listener judgment variability, and listener age; Fig. 15 shows the results for loudspeaker H. The resulting trends and correlations are in agreement with previous findings: the fidelity rating can change as a function of low-frequency hearing level, judgment variability, or age, and the trends in each case are similar, suggesting that those three parameters are correlated with each other—a conclusion reached earlier.

It may not be entirely facetious to suggest that the best loudspeakers are those that improve upon—or at least maintain—their high ratings with increasing listener age, reduced hearing sensitivity, and failing discrimination.

In audiophile series II, loudspeakers D, U, V, and X were selected from the previous series to provide test objects whose ratings were evenly spaced over a wide range of the fidelity scale. As a test of the earlier findings, a new group of 12 listeners was chosen, covering a wide range of age and hearing performance. In Fig. 16 the resulting data have been pooled with those from series I, and the listeners were grouped according to judgment variability. Again listeners with high variability exhibited bias in their fidelity ratings. In the case of loudspeaker D the bias was downward; the same bias is evident for this product in Fig. 6 and, with hearing level as the parameter, in Fig. 10. Loudspeaker U showed a similar downward shift in ratings by listeners with high variability.

Loudspeakers V and X, on the other hand, showed fairly strong reverse trends, with a considerable amount of skew in the distribution of ratings. Assessments of loudspeaker X, in particular, covered a very large range of fidelity ratings. The grouping of listeners into two classes of judgment variability is, however, not as re-

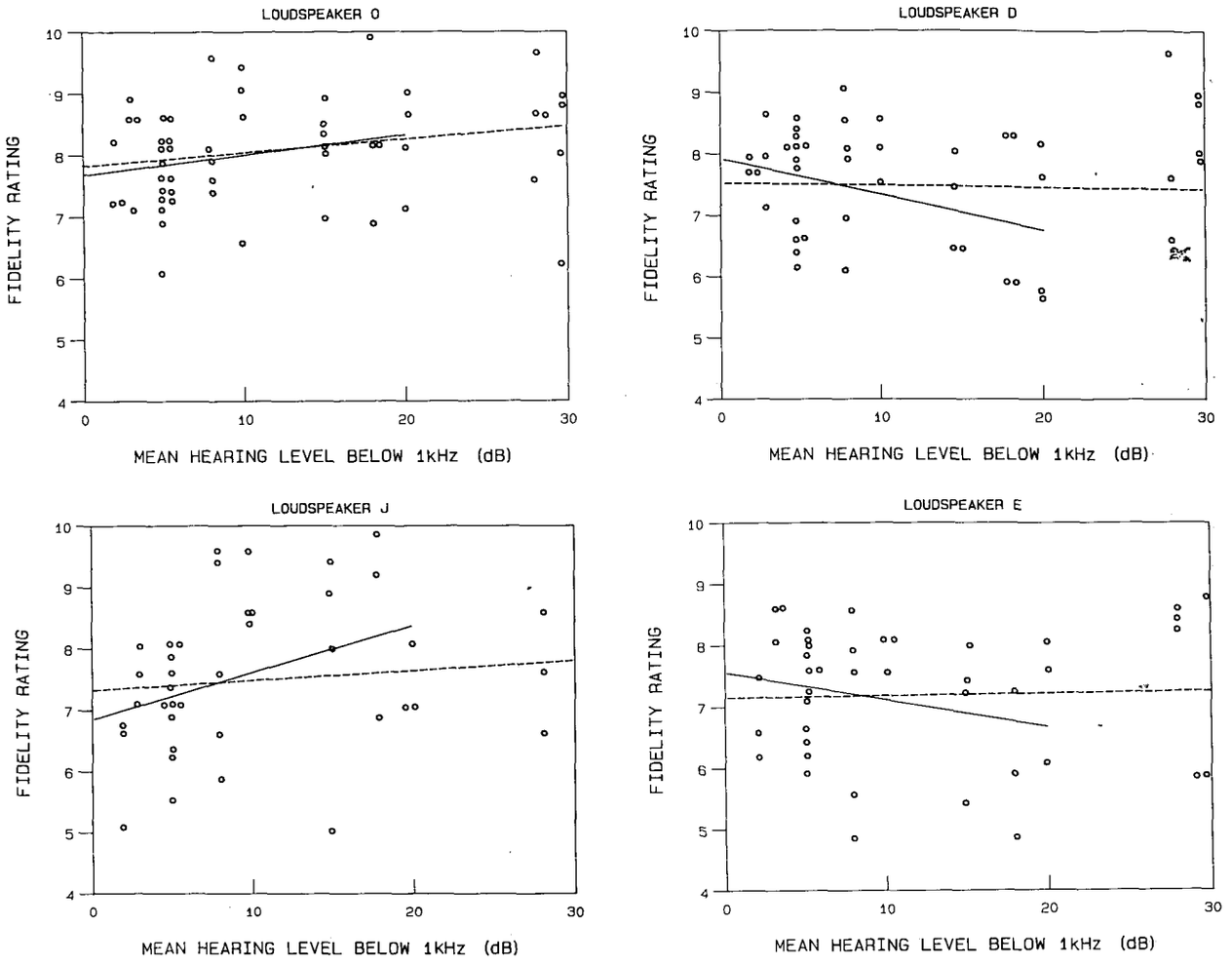


Fig. 13. Relationship between fidelity ratings and hearing threshold levels of the listeners. The vertical axis displays the individual fidelity ratings accumulated throughout the CBC series I experiments, organized along the horizontal axis according to the listeners' mean hearing levels at frequencies below 1 kHz. The broken line is the best-fit straight line through the complete set of data; the solid line fits the data excluding the two listeners with the highest hearing levels.

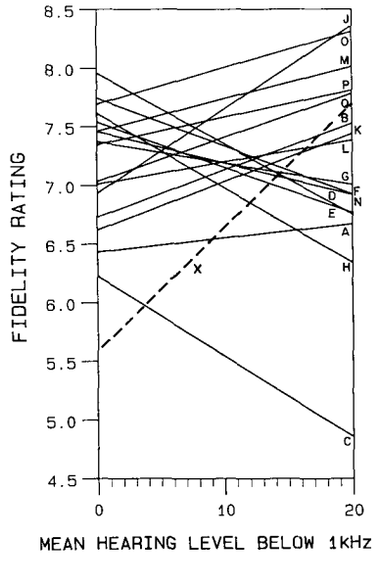


Fig. 14. Superimposition of best-fitting lines, of the kind shown in Fig. 13, for all 16 loudspeakers in the CBC tests. The broken line pertains to loudspeaker X from audiophile series II, which is discussed in Sec. 2.2.4.

vealing as a detailed look at the relationship between fidelity rating and hearing threshold level. The result of such an analysis for loudspeaker X, added to Fig. 14, shows that listeners with widely different hearing levels can have clearly opposing opinions of this particular product.

2.2.5 *Scale Factor as a Source of Judgment Variability*

When considering the performance of groups of listeners it is occasionally evident that some of the apparent differences in opinion are really different expressions of the same opinion. For example, the listener represented by the open triangles in Fig. 12 has taken a very conservative view of rating differences. From the top to the bottom rating in his collection of data the range was merely 1.05 scale units. In contrast, the listener identified by the open squares used a range of 2.4 scale units. Another listener, identified in Fig. 16 by the horizontal lines through his points, used a range of about 3.3 scale units for a selection of the same products.

These differences in scaling are reflections of different slopes in the functions relating the stimulus magnitude to the magnitude of the listener response. Since the stimulus in these experiments (the fidelity of sound produced by the loudspeakers) is not quanitifed, this factor is more than usually difficult to manage. This is another matter requiring further study, and possibly normalization.

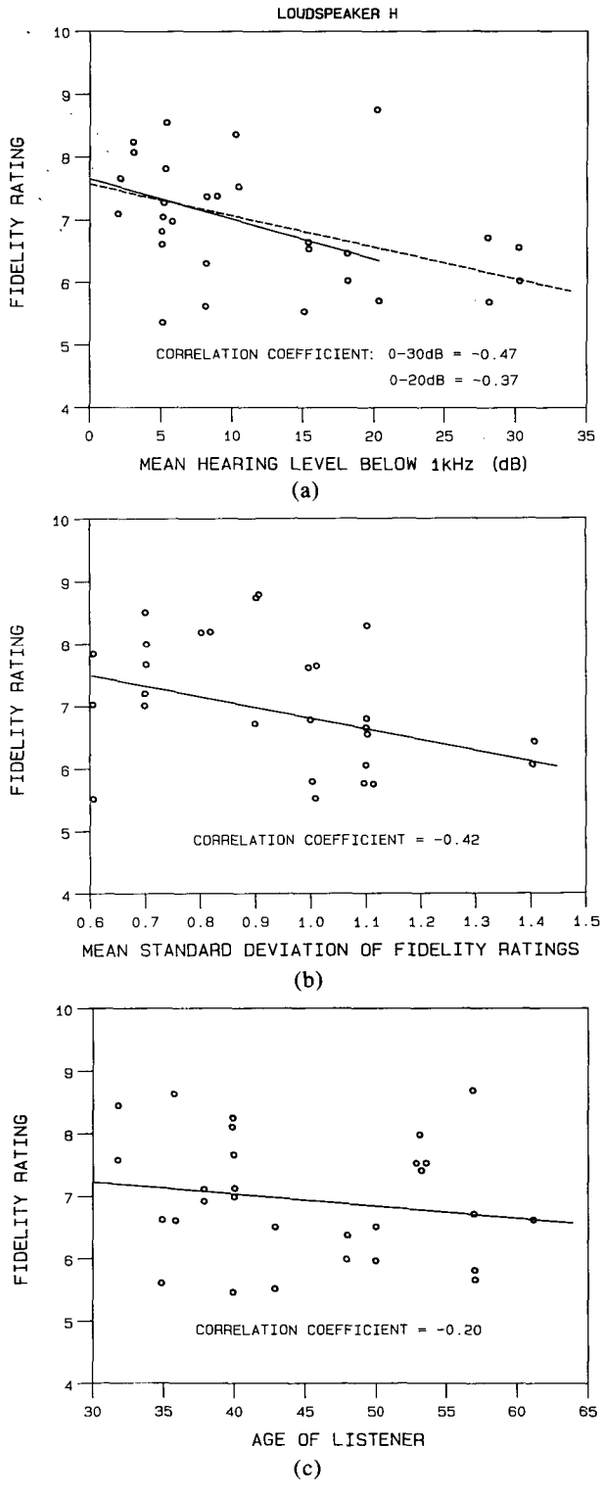


Fig. 15. Relationship between fidelity ratings and (a) hearing threshold level, (b) judgment variability, and (c) age for one loudspeaker.

2.2.6 *Distribution of Fidelity Ratings and Application of Statistical Procedures*

In experimental work, repeated measurements of the same quantity will vary as a result of the cumulative effect of errors from several sources. Often the errors will vary in such a way that the distribution of measurements will conform to a statistically normal distribution. Where this is so, many statistical procedures are available to the experimenter wishing to probe the causes underlying the variations in the results and the significance of the differences in specific data. With the numerous sources of variation in experiments of the present kind it seems reasonable to expect that they will interact randomly and that the assumption of normality will be justified. On the other hand, if there are strong biases or nonlinearities in scaling, such an assumption may be unwarranted.

Testing the normality of a distribution of data requires confirmation that the four fundamental properties of a normal distribution are met. First there should be evidence of a central tendency: the data should tend to cluster around the mean of the distribution. Second, the distribution should be symmetrical above and below the mean. Third, the sizes of the individual data samples should be unrestricted in either direction. Fourth, the mean, the median, and the mode of the distribution should all have the same value.

Using the experimental data for loudspeaker D, the most thoroughly tested of all the products in these experiments, and selecting the most homogeneous body

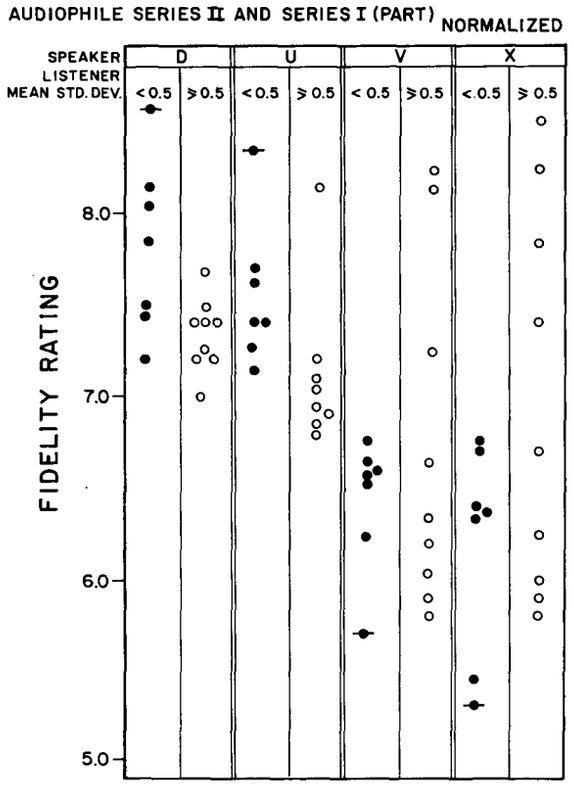


Fig. 16. Mean fidelity ratings on four loudspeakers by 16 listeners who have been grouped according to the variability in their judgments.

of data, from CBC series II and audiophile series I and II, Fig. 17 was prepared to test the statistics. With 104 individual fidelity ratings, each the result of 30 min of comparison listening in a similar experimental context, there should be sufficient data to see if, to a first approximation, the assumption of normal distribution is justified.

Apart from the finite response scale of 0 to 10, which does not fully meet the third requirement, there would appear to be a *prima facie* case for accepting the notion that the underlying data distribution is normal. Fig. 17(a) shows a theoretical normal distribution curve beside the real data, indicating the shape the data distribution might take if given enough samples.

Nevertheless, within the total sample shown in Fig. 17(a) there are subgroups of data. For example, Fig. 17(b) and (c) shows data grouped by listeners' judgment variability. Again we see evidence of the reduced fidelity rating given to this loudspeaker by listeners with higher variability. Further, there is confirmation of the important observation that listeners with the lowest individual variability agree most closely with each other and show the lowest group variability. Listeners with high individual variability may, as in this and many other cases, show quite low group variability, but with a bias.

Unfortunately this fairly tidy pattern could be upset by those listener/product combinations that result in the heavily skewed data obtained in audiophile series II (Fig. 16). Such data could quite easily be multimodal (exhibiting more than one central tendency), depending upon the selection of the listeners, a fact that makes the uncritical application of statistical processing methods somewhat hazardous.

2.2.7 Statistical Significance of the Results

With the present understanding of the data it is possible to apply some statistical tests selectively. Avoiding the obviously nonnormally distributed data produced by listeners with high variability in audiophile series II, an analysis of variance was performed on the data of the four experiments [52]. In all cases the results were significant at the $p \leq 0.001$ level, which means that within each series there is at least one pair of group fidelity ratings that are sufficiently different that the difference would occur by chance less than once in a thousand times.

The significance of a difference between any two specific products depends on the variance in judgments on those particular loudspeakers, and other factors. However, it is evident from the data of CBC series II, and other experiments using listeners with near-normal

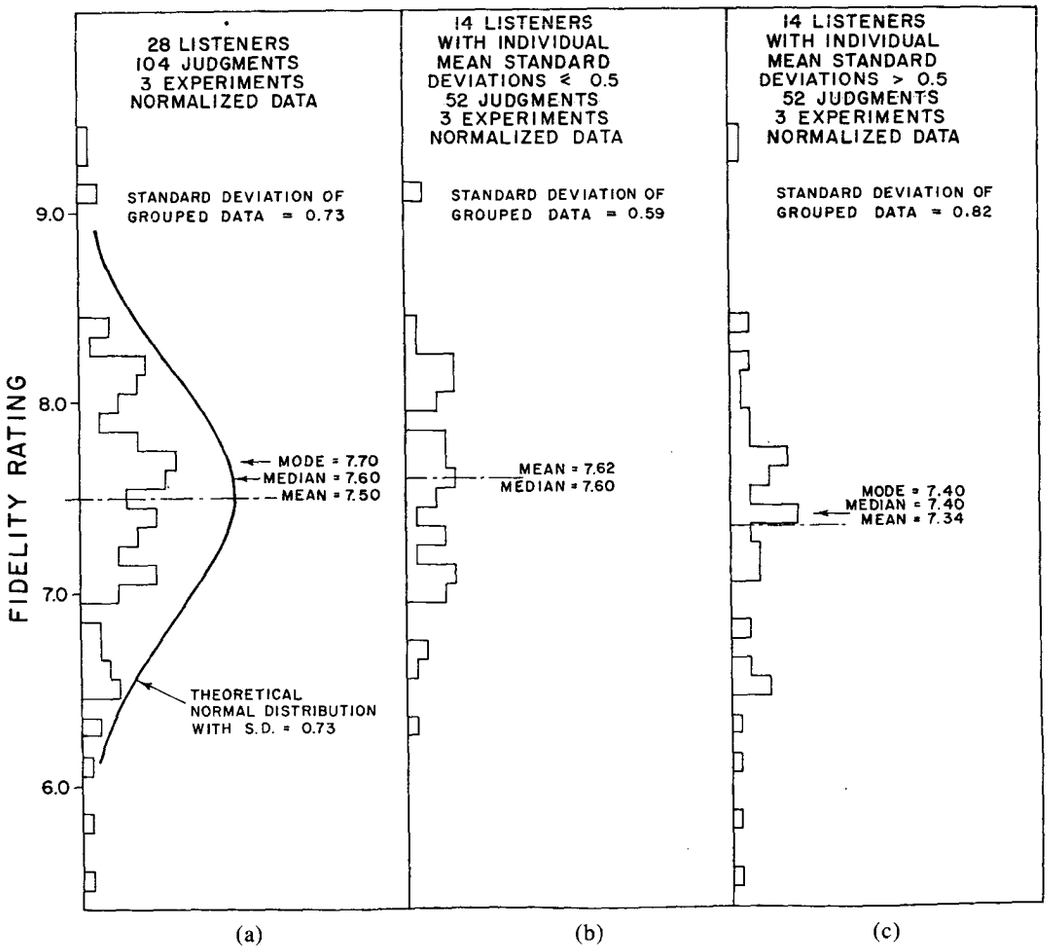


Fig. 17. Distribution of all fidelity ratings of loudspeaker D accumulated in the tests of CBC series II and both audiophile tests. The data were normalized.

hearing levels, that group ratings differing by 0.5 to 0.8 of a fidelity scale unit can be significant at the $p \leq 0.001$ level. This resolution would appear to be adequate for most purposes.

The process of establishing the precise levels of statistical significance for all observed differences is unfortunately not straightforward. In contrast with the paired-comparison or single-stimulus methods, the present method does not provide data in an ideal form for analysis. In this case the relative ease and rather natural manner of gathering the data must be balanced against the problems of analysis and providing a theoretical background [37].

In any event, unless the experimental controls are impeccable and the selection of listeners is less than arbitrary, results that are statistically significant may well be incorrect or inappropriate.

3 STEREOPHONIC VERSUS MONOPHONIC LISTENING

3.1 Effect on Fidelity Ratings

In the stereo comparisons of stereo/mono series I, four loudspeakers, adjusted to equal height, were placed on the turntables and rotated into the same positions for listening [Fig. 1(b)]. This positional substitution method is cumbersome, but it removes an important nuisance variable. Trials conducted with side-by-side comparisons of stereo pairs proved to be unsatisfactory, as differences created by the shifting stereo "stage" were sometimes greater than the real differences between loudspeakers. Monophonic comparisons were made using the method described earlier [Fig. 1(a)].

Eight 3-min musical excerpts from the commercial recordings listed in Sec. 1.3.5 were transferred to the PCM tape recorder, which had a convenient search–rewind–play feature for multiple repeats of program selections. In the stereo tests the sound was muted during the few seconds it took to rotate the next test objects into position. The program selections were repeated as necessary to ensure a uniform number and duration of exposures for each loudspeaker. The randomized loudspeaker presentations occurred at 5–15-s intervals.

Listeners were interrogated only on aspects of sound quality and completed one questionnaire (Fig. 2) for each loudspeaker during the 30–40-min mixed program. The listeners were all audio professionals or audiophiles; eight began with the stereo test and seven began with the mono test.

Fig. 18 shows the normalized results. The small symbols represent the mean ratings for individual listeners; the large block represents the group mean.

Loudspeakers P, D, and CC were rated very similarly in both tests, but listeners reacted quite differently to loudspeaker DD (selected on the basis of its poor technical performance). The stereophonic presentation was clearly flattering to this product; however, the variation in listener ratings was also substantially greater (a standard deviation of 0.82 for DD compared to 0.47–0.50 for P, D, and CC).

The similarity of the ratings for P, D, and CC indicates that the room arrangements and sound reflection and diffraction by adjacent comparison loudspeakers in the monophonic tests were not important factors in the tests. This potential problem [53] may have been neutralized by the practice of exchanging loudspeaker positions, wherein each product takes its turn in each location and the spacing between the products is varied.

3.2 Effect on Judgment Variability

Viewed overall, the mean standard deviation for all listeners was 0.48 in the monophonic tests and 0.63 in the stereophonic tests, indicating a loss of precision in stereophonic assessments.

Looking at it in more detail, the relationship between judgment variability and low-frequency hearing level is shown in Fig. 19. In the monophonic test the pattern is very similar to those seen earlier (Fig. 11) and the correlation coefficient is slightly higher (0.54 versus 0.25–0.28). In stereo the slope of the relationship is considerably steeper and the correlation coefficient a confidence-inspiring 0.70. Listeners with the nearest to normal hearing levels performed with similarly low variability in both stereo and mono tests; listeners in

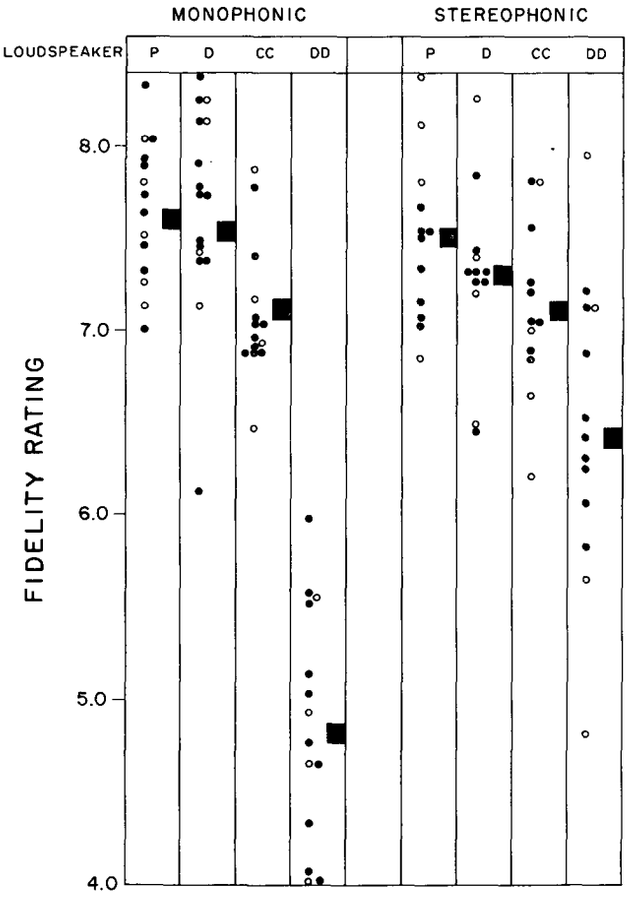


Fig. 18. Stereo/mono series I mean fidelity ratings for individual listeners (small circles) and group of listeners (large squares) for four loudspeakers assessed in separate monophonic and stereophonic listening tests. The results for listeners with high judgment variability are indicated by open circles.

general, however, exhibited greater variation in the stereo tests.

The conclusion seems to be that stereophonic listening increases the variability in individual judgments and places even more severe constraints on the selection of listeners in order to maintain low judgment variability.

3.3 Effect on Judgment Bias

In Fig. 18 the mean ratings by listeners with higher than average judgment variability are identified by open circles. There are no clear indications in these limited data of any trends in the assessments of these individuals, compared to the other listeners. In the stereophonic tests, however, there was a tendency for the ratings by these listeners to be on the extremes of the rating distributions, indicating a nonconformity, but not a consistent one.

3.4 Relationship between Fidelity Ratings and Spatial Quality

Thus far the tests have avoided a direct examination of listener impressions of the spatial qualities, or "imaging," as the popular jargon would have it. Stereo/mono series II addressed this matter. Listeners completed the questionnaire shown in Fig. 3 in repeated assessments of three loudspeakers selected for particular technical features. Two products, E and BB, have been

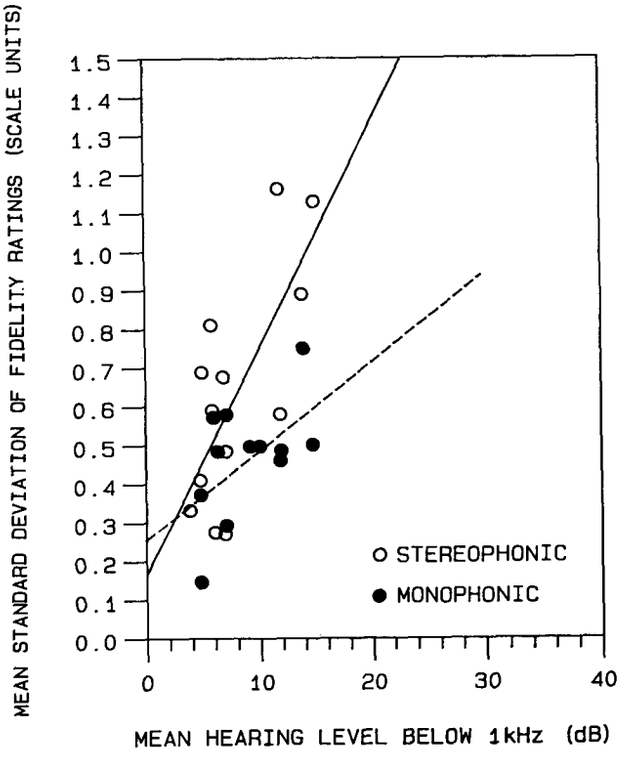


Fig. 19. Relationship between judgment variability and hearing threshold level for the same listeners assessing the same loudspeakers in interlaced stereo and mono tests. The correlation coefficient between stereo data and hearing level was 0.70; for mono data it was 0.54.

praised by the audio press for both high sound quality and excellent spatial reproduction. The monophonic evaluations were performed using the loudspeakers on the left-channel turntable, so that in this experiment loudspeaker position is not a variable. The test procedure was also the same in both stereo and mono modes. There were ten listeners, all audiophiles with essentially normal hearing.

The musical program consisted of transfers made from analog and digital master tapes of concert hall and studio recordings of known origin. The first selection was of a mixed choir recorded in a concert hall using a distant multimicrophone technique. The second selection was a chamber ensemble performance, incorporating strings, bass, and percussion, recorded using a Blumlein coincident microphone pair in a concert hall setting. A jazz selection followed (bass, piano, guitar, and percussion), a multimicrophone studio recording. The final selection was a multimicrophone studio recording of popular music that was given the full treatment of signal processing for special spatial and spectral effects.

The master-tape-quality recordings were regarded as being particularly important in the stereophonic experiments to avoid the distractions of surface noise and tape hiss that interfere with localization judgments in high-quality reproduction systems. It is possible, for example, for background hiss to be associated with ambiance.

The results (Fig. 20) were similar in some respects to those of the previous experiment. In sound quality, the highly rated loudspeakers scored similarly in both stereo and mono tests. The bottom-ranked product (BB) received higher fidelity ratings in the stereo tests.

Asked to assess the spatial qualities of the monophonic reproductions, listeners readily rated the "width" and "depth" of the sound images, along with abnormal effects and spaciousness. Overall spatial ratings paralleled the sound-quality assessments, but all scores were lower.

In the stereophonic tests the spatial ratings were even closer to the sound-quality ratings. The differences were not significant in fact. Plotting the stereophonic fidelity ratings against the corresponding spatial-quality ratings (Fig. 21) shows a strong relationship (correlation coefficient 0.7) between the two ratings. In other words, the fidelity rating is a good prediction of spatial quality, and vice versa. The best-fit straight line indicates a slight tendency for the fidelity ratings to be higher than the corresponding spatial-quality ratings, but the data are so closely clustered that this trend may not be real. As it stands, however, the implication is that even with no merit to the spatial quality there would still be some sound quality—a notion that seems reasonable.

In the event that the overall spatial rating masks some interesting compensating factors, the distributions of the individual listener ratings for the spatial dimensions are shown in Fig. 22. Each histogram includes about 40 separate ratings. The vertical line represents the mean rating.

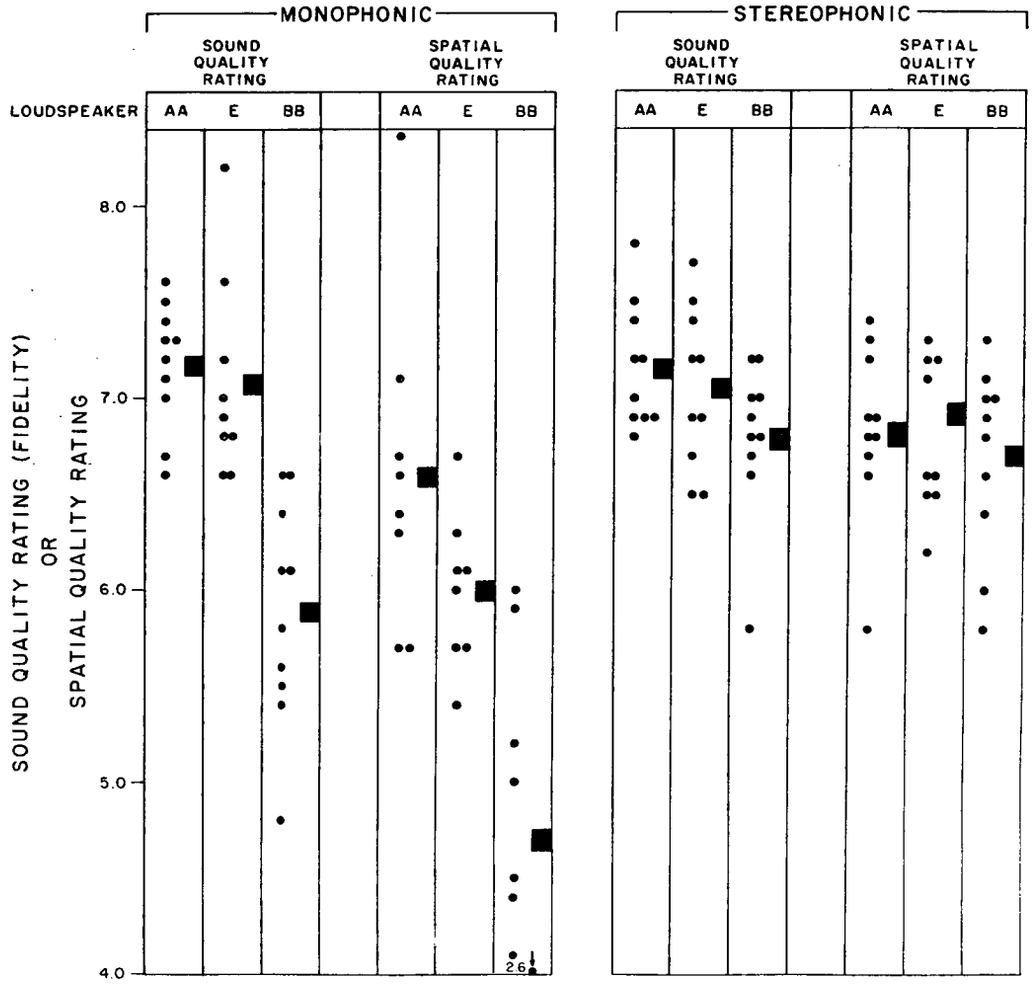


Fig. 20. Stereo/mono series II mean fidelity ratings for individual listeners (small circles) and group of listeners (large squares).

In terms of the definition of the spatial images and the continuity of the sound stage, all three loudspeakers were about equally rated. In the width of the sound stage, the loudspeakers received similar mean scores, but it is interesting to look at the distributions of the individual ratings. The loudspeakers most frequently credited with the widest sound stage were loudspeakers E and AA (12 and 10 ratings, respectively, in categories 9 and 10, as opposed to only 5 ratings for loudspeaker BB). This agrees with the observations of relative monophonic source size.

Of the remaining spatial ratings, only the abnormal spatial effects seem to be discriminatory, and again it is loudspeaker BB that is singled out as producing somewhat more of these effects than the others. From listener comments it appears that the distinctive abnormality consisted of the illusion of sounds originating close to the listener's head or center images far forward of the remainder of the sound field.

Since the kind of music and the method of recording are both determinants of stereophonic imaging and spatial illusions, it is interesting to look at the ratings as a function of music and/or recording method. Fig. 23 shows the overall spatial quality ratings for the four musical selections used here. There appears to be no clear indication of preference in the reproduction of the two concert hall recordings, whether they were recorded with multiple microphones (choral) or a Blumlein coincident microphone pair (chamber). Loudspeaker AA was less satisfactory in its rendering

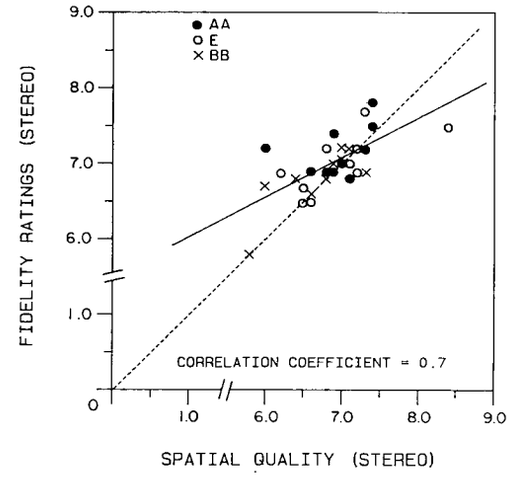


Fig. 21. Relationship between sound quality (fidelity) ratings and spatial quality ratings in stereophonic listening for the three loudspeakers in stereo/mono series II.

of the multimicrophone studio recording of a jazz combo. An examination of the kind shown in Fig. 22 using this music selection reveals that loudspeaker AA received the lowest scores in all categories of analysis, with the largest disparities occurring in the categories of "depth of sound stage" and "impression of ambiance and reverberation."

The multimicrophone studio recording of popular music incorporated several essentially monophonic sound components appearing variously in the left, right, and both (center image) channels. With this program selection loudspeaker BB was given an inferior overall spatial rating. Analyzing the ratings as done in Fig. 22 reveals that for this particular music selection, loudspeaker BB received the lowest scores in all categories of analysis, with the largest disparities occurring in "depth of sound stage," "abnormal effects," and "impression of ambiance and reverberation." Returning to the spatial data obtained during the monophonic listening sessions, a similar analysis by categories reveals precisely the same overall result. The criticisms of loudspeaker BB in the popular music portion of the stereo test may be associated with the difficulties that this loudspeaker has in reproducing satisfactory spatial illusions from essentially monophonic components in the stereophonic mix.

3.5 Mono versus Stereo Ratings—Why the Difference?

That the highly rated loudspeakers retained their high positions in both modes of listening encourages faith in the results. Nevertheless, some loudspeakers that were poorly regarded in monophonic assessments received much higher ratings in stereophonic tests.

Some of this may simply be related to the second sound source and the lateral dispersion of real and apparent sound sources that results. In the prestereophonic era, various methods using reflected and diffracted sound created the impression of a larger sound source. In

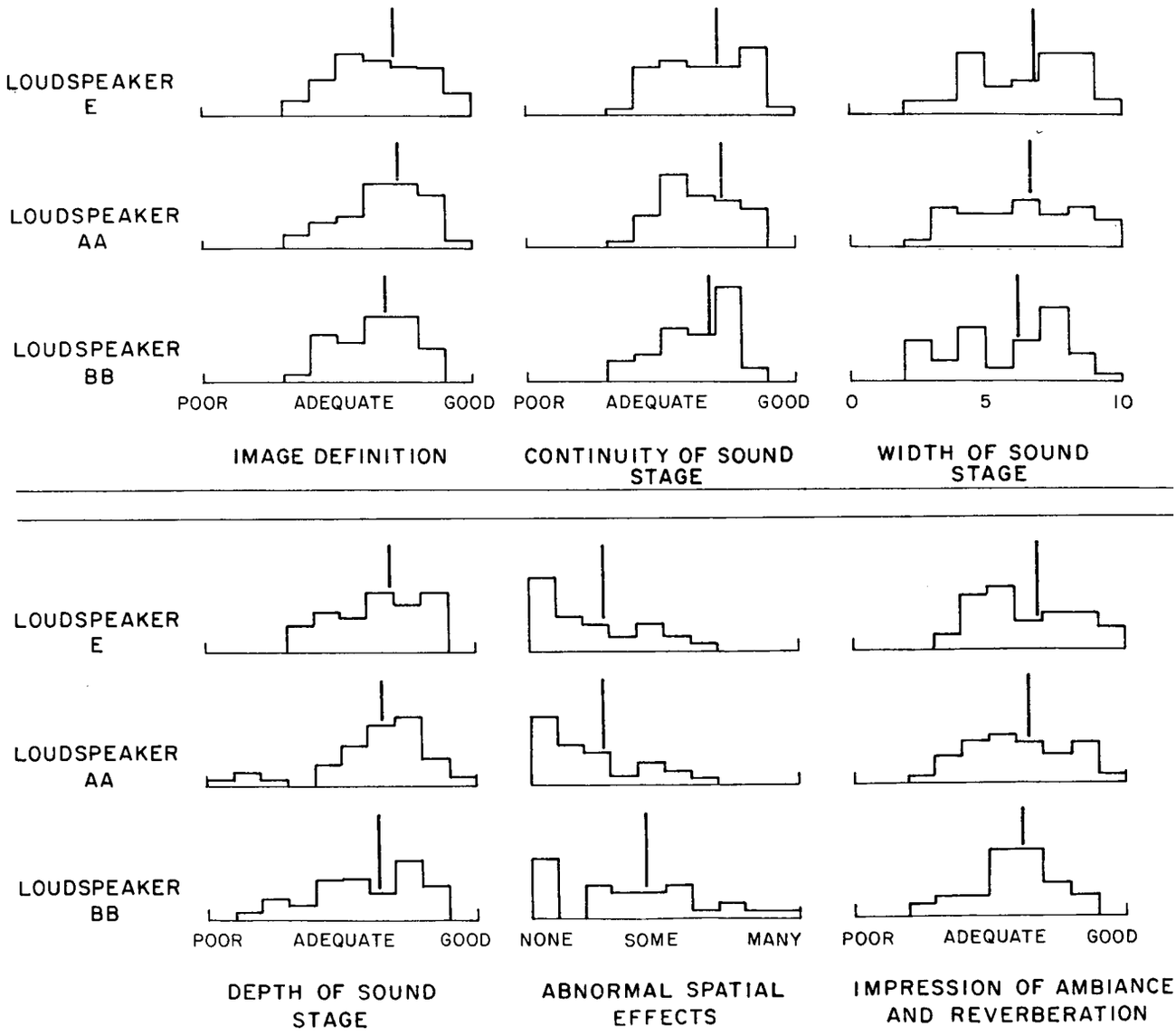


Fig. 22. Histograms showing cumulative listener ratings of various spatial dimensions in stereo/mono series II. The vertical line above each histogram is the mean response.

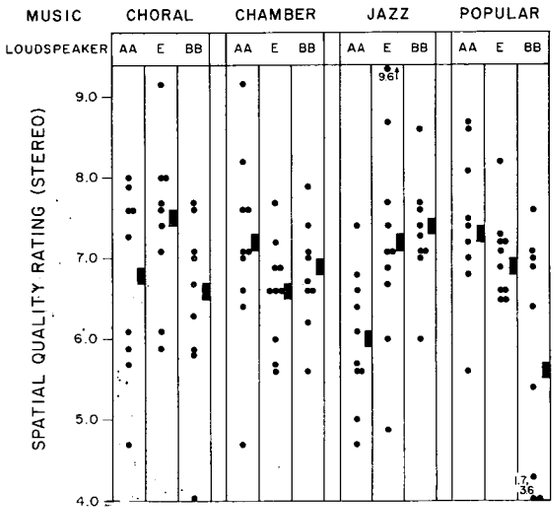


Fig. 23. Overall spatial ratings of loudspeakers in stereo portion of stereo/mono series II for each of the musical selections.

1951 Canby [54] commented that such techniques result in a "far greater naturalness than a point-source speaker can ever give. Secondary (expected) results are a larger tolerance towards poor reproduction of various sorts. . . ." Since an illusion of spaciousness is appropriate to most music, the first statement is understandable. The second comment is interesting, particularly the "expectation."

In 1961 Moir [55] observed that stereophonic reproduction could result in an apparent reduction in certain forms of distortion, a point that was reaffirmed by Dougherty in 1973 [56].

Part of the improvement from stereophonic presentations derives from the sharing of the signal between two channels. The components in each have less to do, and some problems, such as intermodulation distortion, may be less aggravated. However, there are reasons to believe that this mechanistic explanation is incomplete.

The binaural hearing process is well adapted to dealing with spatially complicated sounds. The "cocktail-party effect" is the obvious example of spatial discrimination, wherein certain sounds are attended to and others are perceptually suppressed. Without this ability aural perception would be severely hampered.

By observing that in stereo the instruments of the orchestra, and also the distortion products, are distributed in space, it becomes immediately clear why at least some of those unwanted sounds are less objectionable. Background noises, including electronic and tape noise, are usually uncorrelated in the two channels and present the listener with a large diffuse noise "image" resembling, in some ways, well-recorded reverberation. Distortions originating in one loudspeaker may not be proportionally matched in the other because of differences in the signals (or the loudspeakers) and would, therefore, be localized differently. In monophonic presentations everything is superimposed at the loudspeaker and, logically, would be more objectionable.

4 SOUND-QUALITY ASSESSMENTS AND HEARING PERFORMANCE

It is no surprise to find that listeners with reduced hearing sensitivity perform distinctively. Compared to listeners with normal hearing they simply hear less, and less well. The literature is sprinkled with evidence of deteriorated or altered perceptual performance in people with subnormal hearing. Naturally the emphasis has been on diagnostic and remedial techniques for the hearing impaired.

In the present tests it was assumed at the outset that listeners with severe hearing impairment would not perform normally. What was not expected was that there would be well-defined trends in variability and bias within the range of hearing levels conventionally regarded as normal. Clearly the assessment of sound quality is a very demanding task.

Existing data on perceptual deterioration generally exclude listeners with normal hearing, but it may be reasonable to extrapolate backward from evidence gathered using listeners with severe impairments. Assuming this, reduced hearing sensitivity could result in alterations in the temporal integration of short-duration sounds [57]. In other words, the detection and probable loudness of short sounds may be out of proportion with sustained sounds, resulting in a signal-dependent dynamic-range distortion.

The ability to localize sounds would also be affected [58], implying a reduced ability to differentiate sounds in space and, perhaps, to discriminate binaurally against unwanted sounds and reverberation.

The important relationships between frequency, sound pressure, and loudness [59] are likely to be altered by abnormal hearing. Assuming that these relationships are either defined at birth or established in the early years, it means that changes in the peripheral hearing mechanism occurring later in life may not be perfectly compensated, or even compensatable, by learning [60]. Such a view embraces the observation that, thus far, listeners with nearest to zero hearing level all exhibited low variability, but a few listeners with high hearing levels have shown lower variability than the group. That these listeners all had a long-term involvement with sound-quality assessment suggests that they may have been able to compensate partially for their "handicaps."

To go much further than this would be to venture into areas of speculation. It is sufficient for now that there are plausible explanations for the observed effect.

The present tests were scheduled carefully to avoid listener fatigue or, worse, temporary hearing loss. In many real-life situations, however, sound quality is assessed by listeners whose ears are not operating with maximum efficiency. In recording studio control rooms, for instance, sound levels are often high enough to produce at least temporary hearing loss. It seems highly probable that the all-important decisions made under these conditions are prejudiced by the temporary condition of the listener's auditory system.

A test of the effect was conducted by the Australian Broadcasting Corporation [61], with the expected results that 25-min exposures to 100-dB(A) popular music produced measurable temporary threshold shifts and related changes in spectral balance in recordings made before and after the exposures.

4.1 Selecting Listeners—Whom Do You Trust?

It is an inevitable question, and the answer is not likely to be universally applauded, but—whose opinion is most worthy?

From the evidence accumulated thus far it would appear that listeners with the smallest variations in their judgments are the ones to trust. A good indication of this is a near-zero hearing threshold level at frequencies below 1000 Hz. Although the correlation between judgment variability and hearing level was good, and the results were significant in the present tests, it is important to remember that these were all experienced listeners. Other studies [22], [30], [33] suggest that less select listeners should be screened for obvious hearing disabilities and then evaluated by their performance in the listening test itself. The important experience, in this context, seems to involve the analysis of sound quality specifically. Listeners with extensive experience as musicians, sound recording engineers, and producers were indistinguishable from serious audiophiles in their assessments.

With continued experience in the present tests, listeners developed quite stable rating scales such that individual ratings are not so biased by the ratings of other products in the group. For example, one experienced listener rated loudspeaker D 25 times in eight different experiments conducted over an 18-month period. Without normalization, his fidelity rating judgments averaged 7.7, with a standard deviation of 0.42 [compare with Fig. 17 for 28 listeners (normalized) and the same product]. With other products also, this listener's ratings followed the group means. This, it would seem, is a listener to trust. Yet, in his ratings of loudspeaker D there are two at 8.5 and one at 6.7, which proves that even with 30 min of listening under excellent conditions, mistakes are made. Independent repetitions are essential.

5 SUMMARY AND CONCLUSIONS

With careful preparation it is possible to conduct listening tests on loudspeakers yielding results worthy of being called "subjective measurements." Meticulous attention to the acoustical, psychological, and experimental variables is rewarded by subjective ratings that are reliable and, as will be shown in a future publication, logically related to certain aspects of measured performance. That these assertions are not supported by much common audio experience relates, it is believed, to the lack of necessary controls in conventional listening situations. Even a modest relaxing of controls can allow experimental errors and biases to mask real differences between products.

Central to the "calibration" of the measuring system is the selection of listeners. Individual listeners can exhibit different amounts of variability in repeated judgments, and at the same time they can produce different averaged ratings. These differences in performance are not random, but are related to the hearing threshold levels of the listeners. The selection of the participants in the test will therefore to some extent determine the result itself and the confidence that can be placed in it.

The sensitivity of the sound-quality ratings to different listeners depends upon some property of the loudspeakers. Certain products can be rated similarly by all listeners; others can elicit strongly different opinions from listeners with different hearing threshold levels. From a commercial point of view, the most viable loudspeaker designs are likely to be those with the widest acceptance by listeners. At present it would appear that some designs place the designer and the satisfied customers in a specific minority of the population.

That listener judgments could be influenced by hearing sensitivity was not unexpected. That the correlations were well developed within the conventional "normal" range of hearing threshold levels was not anticipated. Clearly, hearing criteria based on speech intelligibility are not sufficiently rigid for sound-quality assessments. Other factors such as age and listening experience are also involved, but there is little doubt that hearing performance is a major factor in these demanding tests. Experienced listeners with the nearest to normal hearing threshold levels individually exhibited the most consistent judgments and, collectively, showed the closest agreement with each other.

Identifying these people in the context of the established, heavily controlled experiments was straightforward. Identifying them in more conventional circumstances is another matter. The hearing threshold level is an indicator, but because of the other factors, the probability of error is likely to be high. Unfortunately the listeners likely to express aberrant opinions do not otherwise distinguish themselves; they can be as talented, knowledgeable, and articulate as the listeners who seem to "speak for the masses."

Listening in stereo produced sound-quality ratings very close to those achieved in monophonic comparisons, with one significant exception: products with apparently obvious flaws in monophonic listening received substantially higher sound-quality ratings in stereophonic presentations. In aspects of spatial reproduction, it was found that several important spatial dimensions were just as clearly revealed in monophonic tests. The dimensions added by stereophonic presentations seemed to be less dependent on the loudspeakers themselves than, perhaps, the program material. Accurate sound reproduction and good spatial representations appear to go hand-in-hand; a good loudspeaker, used in pairs, becomes a good stereo loudspeaker.

For critical loudspeaker evaluations it is probably important to examine the performance in both ster-

eophonic and monophonic tests. To omit the monophonic assessments would seem to be equivalent to leaving the most sensitive measuring instrument on the shelf.

In general the results of these investigations support the practice of subjective evaluations, but only under carefully controlled circumstances. Listeners are willing and capable measuring instruments, but they often yield data that are dominated by factors other than the one under test. The casual expression of opinions in conventional listening tests are really measurements that are performed without a calibrated instrument, without standardized physical conditions, and without a stable measuring scale. By comparison it is reasonable to suggest that careful interpretations of the appropriate technical measurements may well be more reliable indicators of loudspeaker performance. The next paper in this series addresses this subject specifically.

6 ACKNOWLEDGMENT

The author wishes to thank René St. Denis for his valuable assistance throughout years of these tests, David Bennett and Daniel St. Georges for their helpful collaboration in the CBC tests, Wieslaw Woszczyk and McGill University for excellent master recordings, and David Kelln for his transcriptions and experimental assistance. Of course, without the many listeners who generously volunteered their time and energy to these experiments none of this would have been possible. To them my sincere thanks.

7 REFERENCES

[1] H. Fletcher, "Hearing the Determining Factor for High-Fidelity Transmission," Bell System Telephone Tech. Monograph B-1351 (1941); reprinted in *Audio*, vol. 42, pp. 24, 26, 49 (1958 July); pp. 45–52, 63 (1958 Aug.); pp. 34–36, 53 (1958 Sept.).

[2] H. Fletcher, "The Ear as a Measuring Instrument," *J. Audio Eng. Soc.*, vol. 17, pp. 532–534 (1969 Oct.).

[3] F. E. Toole, "Listening Tests—Turning Opinion into Fact," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 30, pp. 431–445 (1982 June).

[4] V. Salmon, "Imagery for Describing Reproduced Sound," *Audio Eng.*, vol. 34, pp. 14–15, 28 (1950 Aug.); pp. 14, 29, 30, 32–35 (1950 Sept.).

[5] F. H. Brittain, "Loudspeakers: Relations Between Subjective and Objective Tests," *J. Brit. Inst. Radio Eng.*, vol. 13, pp. 105–109 (1953 Feb.).

[6] P. J. Walker, "The Loudspeaker in the Home," *J. Brit. Inst. Radio Eng.*, vol. 13, pp. 377–380 (1953 July).

[7] T. Somerville, "The Establishment of Quality Standards by Subjective Assessment," *Acustica*, vol. 4, pp. 48–50 (1954).

[8] C. J. LeBel, "Psycho-Acoustical Aspects of Listener Preference Tests," *Audio Eng.*, vol. 31, pp. 9–12, 44–48 (1947 Aug.).

[9] H. F. Olson, "Subjective Loudspeaker Testing," *IRE Trans. Audio*, vol. 1, pp. 7–9 (1953 Sept.-Oct.).

[10] F. Langford-Smith, *Radiotron Designers Handbook*, 4th ed. (Wireless Press, Sydney, 1953).

[11] P. Wilson, "A Repeatable Technique for Listening Tests," *J. Audio Eng. Soc.*, vol. 15, pp. 73–75 (1967 Jan.).

[12] F. Olson and K. Schjonneberg, "Listening Test Methods and Evaluation," *J. Audio Eng. Soc.*, vol. 9, pp. 29–36 (1961 Jan.).

[13] R. E. Cooke, "Loudspeaker Listening Tests—Useful or Misleading?" *Studio Sound Broadcast Eng.* (1975 Dec.).

[14] H. D. Harwood, "Some Factors in Loudspeaker Quality," *Wireless World*, vol. 82, pp. 45–54 (1976 May).

[15] C. D. Mathers, "Design of the High-Level Studio Monitoring Loudspeaker Type LS58," British Broadcasting Corp. Research Dept. Rep. 1979/22 (1979).

[16] C. L. S. Gilford, "The Acoustic Design of Talk Studios and Listening Rooms," *J. Audio Eng. Soc.*, vol. 27, pp. 17–31 (1979 Jan./Feb.).

[17] R. Miyagawa, T. Nakayama, and T. Miura, "Design of Reproduced Sound Quality by ESP Method," *Proc. 6th ICA*, A-5-14, pp. 129–132 (1968).

[18] Matsushita Electric Industrial Co., Ltd., "Psycho-Acoustical Measuring System, PACMS," *Volt*, pp. 38–44 (1973 May); pp. 46–52 (1973 June).

[19] M. Kommarmura, K. Tsuruta, and M. Yoshida, "Correlation Between Subjective and Objective Data for Loudspeakers," *J. Acoust. Soc. Jpn.*, vol. 33, pp. 103–115 (1977 Mar.).

[20] H. Eisler, "Measurement of Perceived Acoustic Quality of Sound Reproducing Systems by Means of Factor Analysis," *J. Acoust. Soc. Am.*, vol. 39, pp. 484–492 (1966).

[21] H. Staffeldt, "Correlation Between Subjective and Objective Data for Quality Loudspeakers," *J. Audio Eng. Soc.*, vol. 22, pp. 402–415 (1974 July/Aug.).

[22] A. Gabrielsson and H. Sjögren, "Perceived Sound Quality of Sound Reproducing Systems," *J. Acoust. Soc. Am.*, vol. 65, pp. 1919–1933 (1979 Apr.).

[23] E. M. Villchur, "A Method of Testing Loudspeakers with Random Noise Input," *J. Audio Eng. Soc.*, vol. 10, pp. 306–309 (1962 Oct.).

[24] J. R. Kissinger, "The Development of the Simulated Live vs Recorded Test into a Design Tool," presented at 35th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 17, p. 86 (1969 Jan.), preprint 609.

[25] U. Rosenberg, "Loudspeaker Measurement and Consumer Information," Statens Provningsanstalt/rapport, C-Ra-244/P66-313, 11486 Stockholm (1973).

[26] L. Seligson, "How Consumers Union Tests Speakers," *Stereo Rev.*, vol. 33, pp. 62–66 (1969 Feb.).

[27] "Listening Tests on Loudspeakers," International Electrotechnical Commission, Publ. 268-13: Sound System Equipment, pt. 13, in press.

[28] H. Møller, "Relevant Loudspeaker Tests," Brüel and Kjaer Application Note 15-067; also presented at 47th Convention of the Audio Engineering Society, "Relevant Hi-Fi Tests at Home," *J. Audio Eng. Soc. (Abstracts)*, vol. 22, p. 272 (1974 May).

[29] A. Illényi and P. Korpassy, "Correlation Between Loudness and Quality of Stereophonic Loudspeakers," *Acustica*, vol. 49, pp. 334–336 (1981 Dec.).

[30] A. Gabrielsson and Haken Sjögren, "Preferred

Listening Levels and Perceived Sound Quality at Different Sound Levels in High Fidelity Sound Reproduction," Karolinska Institutet, Rep. TA 82, 100 44 Stockholm 70 (1976 Mar.).

[31] S. E. Borja, "How to Fool the Ear and Make Bad Recordings," *J. Audio Eng. Soc. (Communications)*, vol. 25, pp. 482–490 (1977 July/Aug.).

[32] R. E. Kirk, "Learning, a Major Factor Influencing Preferences for High-Fidelity Systems," *J. Audio Eng. Soc.*, vol. 5, pp. 238–241 (1957 Oct.); H. E. Riorden, "Comments," *J. Audio Eng. Soc. (Letters)*, vol. 8, p. 269 (1960 Oct.) and author's reply, *ibid.*, p. 269.

[33] M. C. Killion and T. W. Tillman, "Evaluation of High Fidelity Hearing Aids," *J. Speech Hearing Res.*, vol. 25, pp. 15–25 (1982 Mar.).

[34] S. E. Asch, "Effects of Group Pressure upon the Modification and Distortion of Judgments," from *Readings in Social Psychology*, rev. ed. (Henry Holt and Company, New York, 1952).

[35] A. Gabrielsson, "Dimension Analysis of Perceived Sound Quality of Sound Reproducing Systems," *Scand. J. Psychol.*, vol. 20, pp. 159–169 (1979).

[36] A. Gabrielsson and B. Lindström, "Scaling of Perceptual Dimensions in Sound Reproduction," Tech. Audiology Rep. 102, Karolinska Institutet, Stockholm (1981 Oct.).

[37] F. N. Jones, "Overview of Psychophysical Scaling Methods," in E. C. Carterette and M. P. Friedman, Eds., *Handbook of Perception*, vol. 2, chap. 10 (Academic Press, New York, 1974).

[38] J. P. Guilford, *Psychometric Methods* (McGraw-Hill, New York, 1954).

[39] S. S. Stevens, *Psychophysics* (Wiley, New York, 1975).

[40] S. P. Lipshitz and J. Vanderkooy, "The Great Debate: Subjective Evaluation," *J. Audio Eng. Soc.*, vol. 29, pp. 482–491 (1981 July/Aug.).

[41] D. Clark, "High-Resolution Subjective Testing Using a Double-Blind Comparator," *J. Audio Eng. Soc. (Engineering Reports)*, vol. 30, pp. 330–338 (1982 May).

[42] A. Paraducci, "Contextual Effects: A Range-Frequency Analysis," in E. C. Carterette and M. P. Friedman, Eds., *Handbook of Perception*, vol. II, chap. 5 (Academic Press, New York, 1974).

[43] R. N. Marsh, "Double-Blind Listening Tests," *Wireless World (Letters)*, vol. 88, p. 63 (1982 Oct.).

[44] H. Helson, *Adaptation Level Theory* (Harper & Row, New York, 1964).

[45] A. Sandusky, "Memory Processes and Judgment," in E. C. Carterette and M. P. Friedman, Eds., *Handbook of Perception*, vol. 2, chap. 3 (Academic Press, New York, 1974).

[46] D. Shanefield, "The Great Ego Crunchers: Equalized, Double-Blind Tests," *High Fidelity*, vol. 45, pp. 57–61 (1980 Mar.).

[47] D. Bindra, J. A. Williams and J. S. Wise, "Judgments of Sameness and Difference: Experiments in Decision Time," *Science*, vol. 150, pp. 1625–1627 (1965 Dec.).

[48] A. Gabrielsson and B. Lindström, "Perceived Sound Quality of High-Fidelity Loudspeakers," presented at the 74th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 31, p. 972 (1983 Dec.), preprint 2010.

[49] G. Slot, *Audio Quality* (N.V. Philips' Gloeilamenfabrieken, Eindhoven, The Netherlands, and Illiffe, London, 1964).

[50] D. A. Bennett and F. E. Toole, "Choosing Monitor Loudspeakers for a National Broadcasting Network," presented at the 72nd Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 30, p. 944 (1982 Dec.), preprint 1906.

[51] J. Sataloff, *Hearing Loss* (Lippincott, Philadelphia, 1966).

[52] J. L. Bruning and B. L. Klintz, *Computational Handbook of Statistics* (Scott, Foresman, Glenview, IL, 1968).

[53] J. Moir and E. Hands, "The Unwanted Speaker Effect," *HiFi News Record Rev.*, vol. 27, pp. 40–41, 45 (1982 Oct.).

[54] E. T. Canby, "The Other Side of the Wall," *Audio Eng.*, vol. 35, pp. 24, 40 (1951 Jan.).

[55] J. Moir, *High Quality Sound Reproduction*, 2nd ed. (Chapman and Hall, London, 1961).

[56] E. H. Dougherty, "Stereophony and the Musician," *BBC Eng.*, pp. 3–6 (1973 May).

[57] C. B. Pedersen, "Brief-Tone Audiometry," *Scand. Audiol.*, vol. 5, pp. 27–33 (1976).

[58] R. Hausler, S. Colburn, and E. Marr, "Sound Localization in Subjects with Impaired Hearing," *Acta-Oto-Laryngologica*, Suppl. 400 (1983).

[59] F. E. Toole, "Loudness, Applications and Implications to Audio," *dB—Sound Eng. Mag.*, vol. 7, pp. 27–30 (1973 May); pp. 25–28 (1973 June).

[60] P. Plath, "Signal Perception in Noise-Induced Hearing Loss," *Acustica*, vol. 29, pp. 47–52 (1973).

[61] D. H. Woolford, "An Aspect of Aural Perception Related to Loudspeaker Monitoring," *14th Nat. Convention Dig., Inst. Radio Electron. Eng.* (Australia), pp. 226–227 (1973).

APPENDIX

INSTRUCTIONS TO LISTENERS

In these experiments we are comparing various sound recording and reproducing methods and devices. You, the listeners, are the measuring instruments. You are required to listen carefully and analytically, and to report what you hear. In order to simplify the processing of the data you are required to answer several specific questions about the quality of the reproduced sound. However, it is important that you not restrict your criticism to just those aspects that are questioned. Please use the "Comments" sections to express any other opinions that are appropriate. We are still learning what questions should be asked.

The following definitions will help explain what is meant by some of the responses. Feel free to ask questions if anything is unclear.

At the beginning of each listening session fill in the blanks at the top of each page. You will be told how many pages to prepare and the round number.

Each sheet contains your responses to one of the test sounds that is identified by the lighted number in front

of you. Each sound will be repeated several times for each piece of music so that you can get a good sense of comparison for all of the questions that must be answered. At first this may be difficult but practice will improve your speed and ability.

Avoid communication of your feelings by sounds or gestures. Operate independently. Do not discuss the test results during the rest periods. You will be told the results when they have been processed.

Thank you.

Definitions: Spatial Quality

Definition of the sound images—Refers to the extent that different sources of sound are spatially separated and positionally defined. Images should not move as the pitch of the music rises and falls. The size of the image should be appropriate to the source of the sound.

Continuity of the sound stage—Is the display of sound images continuous, left to right, or are there illogical groupings of images, with large gaps in between? Is the reverberation uniformly displayed or is it concentrated in strange places?

Width of the sound stage—Refers to the left–right display of sound images. The response scale represents the one in front of you in this room. Mark on it the left and right limits or boundaries of the sounds you hear. Do not include vague reverberant sounds, only those of the orchestra.

Impression of distance or depth—Should be judged on the basis of a satisfactory impression of instruments at various distances. An unsatisfactory reproduction would have all of the instruments at one distance (two-dimensional), or some of them too close or too far, and so on.

Abnormal effects—Refer to spatial sensations that do not occur in common experience. For example, it is possible for some sounds to appear to stretch between you and the screen, perhaps even some of the sounds will appear inside your head. Other sounds may appear to have no location, when you know the instrument should be precisely localized.

Perspective—Refers to your general impressions of the experience. A good reproduction of a good recording with natural room or hall acoustics should suggest that "you are there" at the performance, complete with a sense of the enveloping ambient sound. A less perfect reproduction could separate you from the performance, giving the impression that you are "close, but still looking on." In a still worse reproduction it may seem that you are listening through an opening between the loudspeakers. It is as though you were "outside looking in"—there is no impression of being within the ambient sound. Other recordings may appear to transport the musicians to the listening room, "they are here." The ambiance is that of the listening room, and the instruments sound close. Still other recordings are created as abstract special effects, with no attempt to simulate a realistic experience.

Definitions: Sound Quality

Clarity/definition—Refers to the ability to hear and distinguish different instruments and voices within complex orchestrations. The individual notes should also be distinguishable, with well-defined attacks, not diffuse or muddled.

Softness—Refers to the quality of high-frequency sounds. These should be smoothly natural, neither overly subdued and mild nor excessively hard, shrill, strident, or sharp.

Fullness—Refers to the quantity of low-frequency sounds and their balance with respect to the middle- and high-frequency sounds. Good sound should be neither too full nor too thin.

Brightness—Refers to the balance of the high-frequency sounds with respect to the middle- and low-frequency sounds. Good sound should be neither too bright nor too dull.

Pleasantness—Is an overall rating that concentrates on the pleasantness or lack of aggravations and annoyances in the reproduced sound.

Fidelity—Is the overall rating that describes how closely the reproduced sound approaches your impression or recollection of the original or "perfect" sound. This is the one rating that sums up the previous analytical sound-quality ratings. You *must* give a fidelity rating, it is the single-number indication of your opinion. Please report this score as a number (use one decimal if you wish) in the box provided. The number 10 represents perfection. A telephone might score between 0 and 1, and a small portable radio might score 2 or 3.

THE AUTHOR

Floyd E. Toole was born in Moncton, New Brunswick, in 1938. He received a B.Sc. degree in electrical engineering in 1960 from the University of New Brunswick and a Ph.D. and D.I.C. in electrical engineering in 1965 from the Imperial College of Science and Technology, University of London, England. Since then he has been with the acoustics section, division of physics, National Research Council, Ottawa.

Dr. Toole's early research was concerned with sound localization and the mechanisms of binaural hearing. After an interval of activity in the measurement and

control of noise, including organizational and standards-writing work with the Canadian Standards Association, he returned to audio. In recent years he has been involved with loudspeakers, rooms, and listening tests. A routine program of measurements and listening tests is regularly used by loudspeaker manufacturers, acoustical consultants, and audio publications for purposes that range from product design to product reviewing. A parallel, research-oriented effort is aimed at improving the precision and utility of measurements and listening tests. Part of this energy has been put into working groups of the International Electrotechnical Commission where he is active in standards writing for loudspeaker measurements, listening tests, headphones, and amplifiers. In the field of professional audio, Dr. Toole has designed recording studios, control-room monitor loudspeakers and sound-reinforcement systems for large multipurpose concert halls and theaters.

Dr. Toole is a member of the Audio Engineering Society, the Acoustical Society of America, and the Canadian Acoustical Association.