The gullible listener

John Gardiner

The pratfalls of subjective reviewing, reviewed.

EVERY CHRISTMAS the Royal Institution presents a series of scientific lectures for young people. In England, these are televised every evening during the first week of the New Year and they make for compulsive viewing. Last year we had Professor Carl Sagan talking about planets, and the discussion about space exploration and our knowledge of the solar system was fascinating. However there was much of incidental interest in these lectures for anyone involved in scientific investigation and objective assessment through observation.

Examples were given of the many pitfalls awaiting the observer who deduces a theory from too few facts and then uses the new theory to 'prove' something else! Now this sounds like familiar territory to anyone who has been following the audio scene for the last few years. Perhaps we can learn something from Professor Sagan.

One of his examples concerned Percival Lowell, a respected astronomer in his time, who made a study of Mars, and from his observations he drew a map. Now we *know* that there are no canals on Mars because the Viking probe photographed the surface in great detail. Yet for years, based on Lowell's observations, it was popularly believed that Mars had canals because his telescope observations had 'proved' it. In fact, it was the brain of the willing believer which saw the canals, and where visual information was missing, it was the brain which filled out the detail. A series of blobs became a straight line and a straight line became a canal. But a canal is a man-made waterway. *Ergo* there must be, or have been life on Mars! All this is now discredited but it was once a very plausible theory.

Observation and emotion

If the eye is so easily deceived, what of the ear? The key message from Professor Sagan was that we should mistrust all observations where judgement can be obscured by emotion. That, I think, is a very powerful message for reviewers of all persuasions, and it is particularly relevant to the testing of equipment and the reviewing of records. All of us must have been guilty at some time or other of liking or disliking something for the wrong reasons. That is why I like to see emotional (i.e. subjective) tests supported where possible by unambiguous objective test data. If there is no apparent correlation between the two, then it would seem probable that one or the other is incorrect. Either the wrong tests have been applied or the listening panel has been misled by some unrecognised factor.

Audio is, after all, not a Black Art, but an artistic science and what can be observed can be explained. That is not the same as saying we are able to explain it: sometimes we cannot – immediately. This is why there are areas of disagreement between reviewers; some will devise a theory to fit the known facts, others will wait for research to uncover further facts. There is no reason to doubt the sincerity of reviewers as observers but if a number of us diverge widely in our findings, we cannot all be right. We can, however, all be wrong, which is a sobering thought and should make us choose our words with care!

There are many theories about at the moment on which I have reservations: I don't dispute that certain phenomena are *possible* but I do require proof that they exist, and even more that they are *significant*. I believe there is a great danger today of chasing the ghosts which haunt the fringes of high fidelity instead of concentrating on more material and fundamental issues.

Tests and theories

An example is the controversy which has raged recently about the characteristic sound of some equipment, and the theories which the debate has produced (Ref. 1). If there is fundamental disagreement between two groups of observers (whether they be reviewers, engineers, musicians, manufacturers or hi-fi enthusiasts) when assessing the relative merits of certain pieces of equipment, this would indicate that it is probably not the equipment which should be criticised but the method of testing it.

It also seems to me that if there is a conflict between objectively derived data and that derived by subjective observation, we should not too hastily dismiss the objective data. Most reviewers have far more experience in measurement techniques than they have of interpreting the results of subjective evaluation. This is not in any way to decry listening tests; they are an essential part of reviewing. But it is important that the results should be statistically and intelligently analysed, so that we don't have to resort to extravagant phrasing to disguise a paucity of information.

Some of the BBC's Research Department reports are well worth studying in this connection as they often give details about their tests and their analysis of the results.

A-B testing

One of the first decisions a listening panel has to make is whether to do an A-B type of test or whether to consider each piece of equipment separately and so make an absolute judgement without reference to other equipment. If an A-B test is chosen the next question is whether the two items to be compared should be fed from synchronous music sources, so that when a changeover occurs there is no break in the signal. Or whether there should be a small time lag so that when switching from A to B it is possible to hear a repeat of the last few bars heard on A. In my experience most panels choose the synchronous approach, yet in many ways it is more logical to have a time lag.

The next point about an A-B test is that it is only possible to compare two sources at a time. Therefore, if several ▶

JOHN GARDINER is an independent technical writer and audio equipment tester/reviewer from Woking in Surrey, UK. He spent a number of years in the BBC Technical Operations Department in both Radio and TV studios and is currently with the Decca Recording Company running a technical liaison department. He is the author of two books on Tape recording and won the 1977 BASF Audio Writer of the Year award. Incidentally, the Concise Oxford defines 'pratfalls' as: fall on buttocks; humiliating failure.

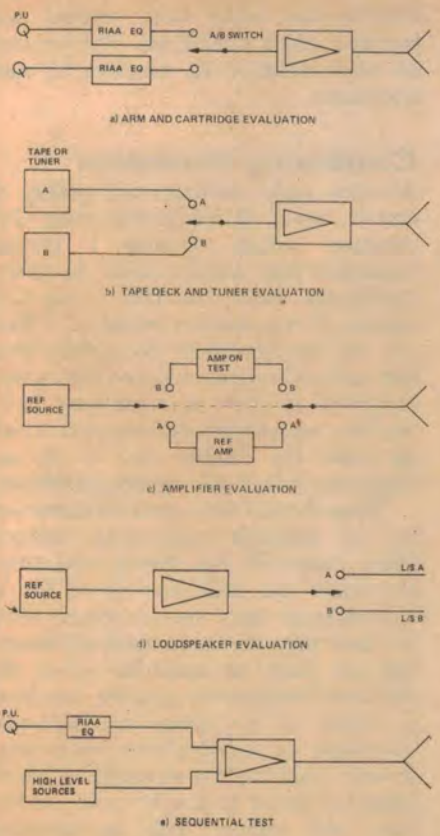

Figure 1. **Alternative listening tests.**

units are to be tested there is the choice either of selecting one unit and using this as a reference, or selecting a well known proprietory unit against which all the others can be compared. The latter approach would seem to have more to commend it because the reader will probably have some knowledge of the proprietory unit and thus he will have some yardstick against which to relate the panel's judgement to his own experience. The difficulty is, of course, that the reader's judgement is coloured by his own room acoustics and therefore the yardstick will only help some listeners.

What is important about this type of test is that verbal comments are kept to a minimum and that the panel do not compare notes during the marking. One effective method is a deviation scale with a central O representing the reference source and a scale extending either side from, say, -5 to +5. A series of judgements can then be recorded quickly with a minimum of effort. Time should be allowed between tests for the marking to take place as it is difficult to write and to listen critically at the same time. The duration of the listening session is of great importance as it is fairly widely accepted that 30 minutes of concentrated listening is the maximum that should be expected from an observer before his findings become unreliable. This assumes that we are looking for subtle differences in quality rather than major defects that are readily detectable by more crude methods.

There are also certain psychological problems associated with this type of listening test. For instance there is a tendency to prefer the second of two sources in an A-B test and so it is necessary for the reference channel to be changed, and for the panel's marking to be cross checked against the changes. It follows that the observers should not know at any time which equipment they are listening to as this can easily influence their judgement.

Ambiguity

In a series of experiments I was recently involved with, to evaluate new methods of recording master tapes, it was found that tests had to be carefully programmed if ambiguous results were to be avoided. One problem is that the brain automatically regards any test of this sort as a challenge. This means that although on one level it is trying to make an impartial judgement, there is another part of the brain looking for clues in the test material. A slight amount of tape hiss here, a drop out there, perhaps a particular pattern of clicks on one channel; any of these things may be latched on to and subconsciously used as a clue to determine which is the reference source. The brain is very good at this sort of thing and it is very hard to guard against it.

Another factor is the relative response and relative loudness of the systems being compared. A significant number of listeners will prefer the louder source regardless of other factors and a discrepancy of as little as 1 dB can give misleading results. Similarly, a variation in amplitude/frequency response in the mid-frequency band of 1 dB will lead to erroneous observations. Thus if a ½ dB rise in one source corresponds with a ½ dB dip in the other, we must look very carefully at the test results.

It has also been established that a large number of individual observations should be tabulated and statistically analysed. The analysis can then be given a 'confidence level' which indicates the chances of the result being accidental (Ref. 2). Ideally, the same test should be given to more than one panel. Most panellists regard it as a defeat if they can hear no difference between two sources and it is advisable to do some dummy runs to make them aware of this. It is surprising how many people if given a non-operational 'A-B switch' will prefer B to A, although no change of source has taken place.

In the space available it is not possible to discuss A-B testing in great detail but I think the above points are sufficient to show that an A-B test must be painstakingly set-up, and that it is very easy for erroneous data to be collected. This is not to say that such a test is no good, merely to put in a word of caution. I believe that a reviewer has a responsibility, where possible, to give some objective support for his findings. Frequently this proves impossible, particularly with transducers.

It follows, I think, that if a simple A-B test is difficult to set-up, the problems of setting up similar tests to evaluate, say, six systems are immense and should not be minimised. In fact the question that should be asked is, whether the detection of subtle and elusive differences between one piece of equipment and another are significant to the reader. There is perhaps a danger of a reviewer writing for himself or his colleagues, rather than for the readers who pay his fees!

Sequential testing

The alternative to the A-B type of test is even more frightening in the margin for error it offers. For want of a better expression we will call this the sequential test, and throughout this discussion we are, of course, concerned with marginal differences in performance. We assume that anyone with more than a passing interest in high fidelity will be able to identify major shortcomings, regardless of the test method used.

Now there are two basic faults which are common to very many comparative reviews and it is difficult to eliminate them. Firstly, there is the time it takes to remove one system and set up another so that it can be assessed under identical conditions. This means that a multiple listening test must be prolonged, with breaks in concentration whilst the equipment is re-aligned. Protracted and frequently interrupted listening sessions do not usually make for accurate observation. Secondly, it frequently happens that when one unit is substituted for another, an unwanted variable is introduced. One example is a multiple loudspeaker test: if the units have different sensitivities, the gain of the reference drive amplifier will have to be changed in order to produce the same sound pressure level, and this may have an unsuspected influence on the panel's judgement. Conversely, with amplifier tests it is feasible that some amplifiers will interface better with the chosen loudspeaker than others and again give misleading results. ▶

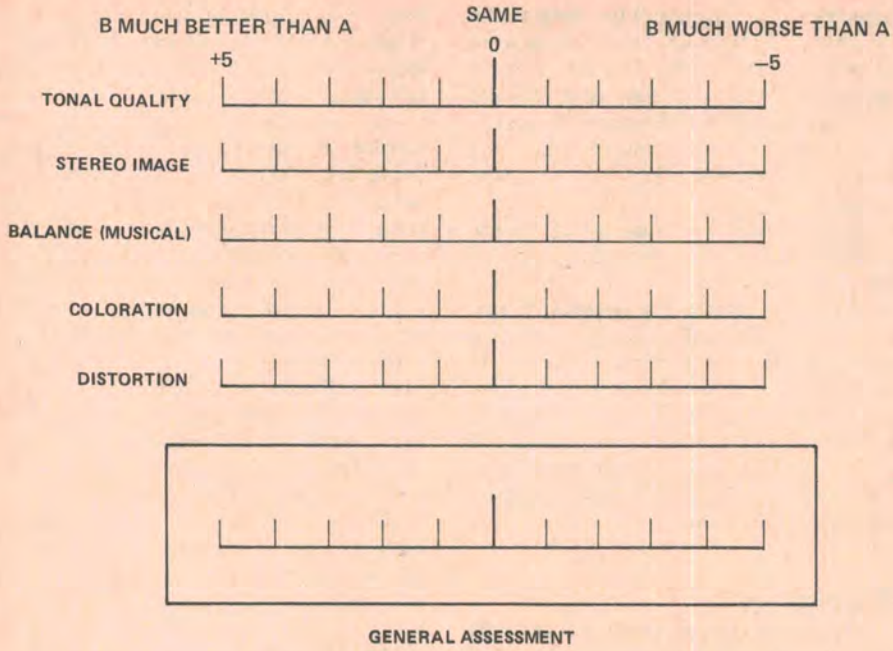

Figure 2. **Sample of listening panel questionnaire suitable for statistical analysis.**

In any form of scientific experiment it is standard practice to have a 'control' sample against which to assess the effect of a test. The idea is to keep the control in the same environment as the samples and then to examine the effect of varying just one condition on each of the samples. At the end of the experiment if a sample differs from the control it may reasonably be assumed that the change was caused by the variable factor. If more than one variable is introduced the experiment is of doubtful value. The implication of this is obvious: any listening tests in which more than one condition is varied at any time, is suspect.

There is another snag to sequential testing and it is that the ear no longer has a reference. It has to make an absolute judgement of sound quality, and the ear, like the eye, can be misled. Everyone is familiar with optical illusions and there is no denying that they exist. It would be surprising if the ear were not fallible in a similar way; and if it is to be deprived of a reference against which to test its judgement, it is likely to suffer from cumulative delusions. It is self-evident, I think, that the more combinations of equipment which are used in a listening test, the greater are the chances of an erroneous conclusion.

The phase debate

There has been correspondence in the technical press recently on matters directly related to listening tests (Ref. 3). It is perhaps worth summarising some of these arguments, as they highlight the difficulties facing the listener who is trying to make a valid judgement. For instance, there is the matter of phase. We are not concerned with inter-channel phase differences which are readily audible, but with absolute phase. What happens if we reverse the connections to both loudspeakers in a given system? A number of distinguished scholars have applied themselves to this one – and arrived at opposite conclusions! However there are some grounds for believing that changing the overall phase of a system may subtly affect the quality of reproduction.

The argument goes something like this: musical instruments often produce asymmetric waveforms and at the time of recording, a positive going wave produces a forward movement of the loudspeaker cone. Any correctly set-up reproducing system will be arranged so that in-phase signals cause identical movements from both loudspeakers. Hitherto, no one has seriously bothered to ensure that the polarity of the reproducing loudspeaker is the same as for the original recording. Therefore, what was originally a forward movement of the loudspeaker cone could become a backward one and vice versa. Hence the compressions and rarefactions of the sound wave produced will be in the opposite sense to that of the live sound.

The ear itself is an asymmetric detector and the suggestion is that it could be sensitive to such an absolute phase reversal. In case there is anything in this theory the inevitable conclusion must be that all listening tests should be carried out under both normal and reversed phase conditions. The amount of work entailed is, needless to say, enormous.

Continuing discussion

Another topic concerns the quality of certain types of connecting wire. One reviewer whose opinion I respect maintains that certain cables do have a significant effect, but that a top class system is required to reveal it. I have not yet experimented along these lines and so have an open mind on the matter. However, if there is something in it, we have yet another variable with which to tease the fallible ear. Is all our equipment identically wired, gentlemen?

Then there is the matter of distortion on the available programme sources. Peter Baxandall has shown that some amplifiers are near perfect so far as conventional tests are concerned (Ref. 4). How then do we perform a listening test on such an amplifier when the available programme sources can have *in excess* of ten times the harmonic distortion of the amplifier under test? It well could be that an amplifier which auditions badly is simply more perfect than the others and is revealing deficiencies elsewhere in the system.

In this article we have deliberately assumed the mantle of the Devil's Advocate and looked at the black side of subjective testing. Obviously tests which are conscientiously performed frequently give sensible and acceptable results. But conflicts do arise and when there is dispute common sense frequently goes on holiday. Some manufacturers are now refusing to submit any equipment for review by any magazine.

In this way everyone suffers.

What is required is some standardisation of listening tests so that there is better correlation between various reports whilst leaving room for individual preferences. We must train our ears and we must beware of feeding them with ambiguous data and expecting them still to give accurate assessments. We should also understand that the ear, like the eye, is fallible and that because someone else claims to hear something: "It Ain't *Necessarily* So." ●

REFERENCES:

1. 'Positive Feedback' and correspondence, Hi-Fi News. Nov. '77 to Feb '78 Letters: Wireless World. Jan and Feb. '78
2. 'Facts from Figures' by M.J. Moroney (Pelican). Pages 246 to 270.
3. Letter: Wireless World, J. Moir. Jan '78.
4. 'Audible Amplifier Distortion'. Peter Baxandall. Wireless World. Nov '77.