The Basics of High Fidelity
Part 6: Subjective Testing

In this part we will deal with “the proof of the pudding” which is in the eating, we are going to listen ourselves or ask for the opinion of others.

Let’s start with the question what it is we are going to listen to, what do we want to test? Transparency is again the answer! The perfect sound system is transparent, if there is no audible difference between the idealization and the realization we have reached our HiFi goal. Two fundamental problems arise, how do I get hold of the ideal and if I can, or cannot, hear the difference between the idealization and the realization does the same apply to other listeners? The last problem can be solved pragmatically, we can decide that we will only declare a system to be transparent against a certain ideal if no single person can hear the difference. An extreme hard requirement, especially because subjects can train themselves indefinitely to hear ever smaller differences.

Regarding the first problem, the idealization, there is no general acceptable solution. For simple audio systems that have an electric in- and output and that strive for transparency we can run a transparency test. In this case the input signal is defined as the idealization and the possibly degraded output signal as the realization. In order to judge the signals we we need a loudspeaker or a headphone, which need to be transparent, making the validity of this test limited. In testing electric in-/output systems mostly a headphone is used that allows to have a better controlled reproduction. In order to test for transparency we need to run a double blind test where subjects can switch between the output and input without audible clicks and compare both signals with the input signal (the ideal). For a quality judgment we can also ask the subjects not only to identify the degraded but also give it a rating, mostly using a discrete quality scale. These tests are called Degradation Category Rating experiments [1] [2] and the most widely used rating scale is given in Table 1. If the system under test does not allow to use the transparency paradigm because the idealization is unknown, e.g. for sound enhancers, we can only ask for a personal rating, having no clue against which idealization the quality is judged. These tests are called Absolute Category Rating experiments [1] and the most widely used rating scale is given in Table 1.

Opinion Score	Absolute quality (against an ideal) ACR	Relative quality (against an unknown ideal) DCR
5	Excellent	No audible degradation
4	Good	Audible but not disturbing
3	Fair	Audible, slightly disturbing
2	Poor	Audible, disturbing
1	Bad	Audible, very disturbing

Table 1. Most widely used rating scales, Absolute Category Rating (ACR [1]) and Degradation Category Rating (DCR [1], [2]). If scores are averaged over a large set of subjects the resulting number is called a Mean Opinion Score (MOS). If the MOS score in a DCR experiment is 5.0 the system under test is transparent.

If the audio system under test has no electric coupling points for both in- and output the best way forward is to use the most pure solution of the subjective test problem, the "live versus recorded" transparency test of the complete audio chain; recording, transport/storage and reproduction. After having read parts 1 through 4 we know that there are two "live versus recorded" tests, transparency with respect to the illusion "here and now" and transparency with respect to the illusion "there and then". If we want to create either a "here and now" or a "there and then" illusion with a headphone we need to have access to the so called individualized Head Related Transfer Functions (HRTF's) that describe the sound transformation from the free field to the entrance of our ear. But even if we could measure and implement the complete set of HRTF's of our own ears, head movements will remain a problem, while making the appropriate recording is extremely difficult for both the "here and now" and the "there and then". Also low frequencies are perceived through bone/body transmission, completely ignored in headphone reproduction.

So what remains is a loudspeaker reproduction and we will start with the "here and now" illusion. We start by taking a single loudspeaker that has the same radiation pattern as the source we want to reproduce. Next we make a recording of that source in an anechoic room and reproduce this recording in any room we like and compare the result with the live experience of that source in that room. In the ideal case you can draw a curtain before the loudspeaker and the "live" source and nobody can distinguish the reproduction from the production. Unfortunately many sources have wild radiation patterns making it hard to carry out such a test. However if we take a single source with a smooth radiation pattern, such as a human voice, we can run such a test. The nice thing about the test is that you can run it in any environment, if you run it in a low reverberant room at a close distance the test focuses on the direct field reproduction, if you run the test in a highly reverberant room at a large distance the test focuses on the diffuse field reproduction. This test can be seen as the ultimate HiFi "Turing" test for the "here and now" illusion.

Of course this ultimate test is not very practical and most subjective testing is carried out by using a play back comparison with a headphone to compare a recorded reference with a recorded degraded signal. Judging the degradation with loudspeakers is dangerous for both the "live versus recorded" test as well as the play back comparison test because the reproduction room has a major impact on what differences will be audible. In rooms that have hard surfaces comb filtering becomes a severe problem. It can lead to unexpected results, e.g. when a soft distortion is unmasked by suppression of a loud component in the audio signal due to comb filtering. This makes the statement that an audio compression scheme is transparent dependent on the play back context. It will only be transparent under the same condition as the subjective test was carried out.

Transparency regarding the illusion "there and then" is even more difficult to test. How do we know that the reproduction of a surround recording of a concert hall sounds the same as in the concert hall? It's almost impossible to test and we are more or less forced towards just asking subjects what they "like", the absolute category rating approach. However this is dangerous and can result in making the wrong choices. The best known example in my view was the standardization of the first Japanese digital cellular speech coding standard. In the subjective tests that were carried out in Japan with a number of speech coding schemes a system was preferred that sounded "robot" like. The Japanese preference for high tech tools caused a preference bias. When the system was introduced and people used the system to make calls to their loved ones they were disappointed by the poor speech quality. There is thus no simple answer to the question "what is the best subjective test". We should be aware that the design of the test largely determines the answer and even if the statistics give you a high reliable outcome, it might be exactly the wrong answer. So what shall we do? Well, as the answer is highly dependent on what you want, we will discuss what we really like in Part 7, realizing that it's better to be roughly right then exactly wrong.

[1] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," August 1996.

[2] ITU-R Rec. BS.1116, "Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems," International Telecommunication Union, Geneva, Switzerland (1997 October).

John G. Beerends
Published in Hifi Video Test 4/2008 (in Dutch), translated and updated over the period 2012-2020.