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 need a loudspeaker or a headphone, which needs 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 (ACR) 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 we want to assess the quality of a complete audio chain, recording, transport/storage and reproduction, we have to carry out a “live versus recorded” transparency test in combination with a DCR test. 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 recordings are made in the electric domain, e.g. using a synthesizer or an electric guitar, acoustic references are not available and we con only carry out ACR tests.

If we want to assess the quality of a “there and then” illusion, we need to exclude any degradation caused by the room in which we reproduce the recording, so we have to use a headphone that uses individualized Head Related Transfer Functions (HRTF’s) that describe the sound transformation from the free field to the entrance of our ear. This “live versus recorded” test has to be carried out with recordings that are focused on headphone reproduction and thus on recordings that use an artificial head that matches our head/headphone combination. It requires an individualized set up and can only be carried out in the environment where the recording was made. In this DCR test subjects listen to the recording over their individualized headphone and compare this to the live acoustic event. In this test head movements and the reproduction of low frequencies, which are partly perceived through bone/body transmission, may give rise to non-optimal quality judgements.

If we want to assess the quality of a “here and now” illusion, degradations caused by the room in which we reproduce the recording have to be included, so we have to use a loudspeaker that has the same radiation pattern as the acoustic source we want to reproduce. In this “here and now” approach the sources have to be recorded in an anechoic room and can be compared with the live production of these sources in any reproduction 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 find a loudspeaker set up that allows to match these patterns. 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.

Transparency with loudspeaker reproduction regarding the illusion “there and then” is more difficult to obtain and more difficult to test. How do we know that the reproduction of a concert hall recording in our living room sounds the same as the original performance sounded in the concert hall? We could try to make two recordings of the live acoustic event, one focused on the reproduction over loudspeakers and one focused on the comparison of the live event with the reproduced event using an exact, individualized, binaural recording play back chain. Then we would also have to make a binaural recording of the reproduction over the loudspeakers in our living room and compare this recording with the binaural live recording. Although attempts have been made to carry out such comparisons we will always run into problems, especially when shortcomings in the production are repaired and the reproduced sound may even sound better than the live sound. With loudspeaker reproduction focused on the illusion “there and then” we are

more or less forced towards just asking subjects what they “like”, the absolute category rating approach.

In general the absolute category rating approach is dangerous because subjects will use an unknown ideal in their judgement and such tests may 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. We should thus be aware that the design of the test largely determines the outcome, and that there is no simple answer to the question “what is the best subjective test”. 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.

Part 1: Transparency and Perceptual Measurement Techniques

Part 2: Reproduction Philosophy “Here and Now” versus “There and Then”

Part 3: The Ideal Loudspeaker, Diffuse Field Equalization

Part 4: The Ideal Loudspeaker, Reflections and Resonances

Part 5: Audio Compression

Part 6: Subjective Testing

Part 7: What Do We Really Want

Part 8: Telephony