Viewer Training in Subjective Assessment

Vittorio Baroncini

A long time ago

When for the very first time I participated in a subjective assessment trial, it was in one of the most famous and referenced test laboratories belonging to a well-known and highly considered European Broadcaster. The impression was a little shocking to me, in that I was sitting there, 3H from the professional grade 1 studio monitor and in a carefully controlled environment, i.e., silent room with low ambient lights and no noise from outside. Everything was perfect other than a small close to irrelevant detail: what were we to do? Then we were ready to go and an old technician (wearing a white smock) came in and read to us in a grave and formal voice a short text saying "This is a visual quality experiment … … thank you for coming and good work". The text read was exactly the one reported in the ITU-R Recommendation BT.500-2. Then he went out, closing the door and the display began to show the video to assess.

In this short paper the importance of good training is stressed as mandatory practice to obtain the best and most stable results possible. Training the subjects immediately before they run a test session has the main goal of letting them understand what they have to look at and how to properly do the scoring. But not only. Participating in a subjective test is not the same as going to a cinema and watching a feature film; you might even be a little upset at having accepted a long and tedious task that will keep you there doing a "stupid task." So viewing subjects must be put in a psychologically favorable disposition. And in this, emotional involvement may certainly help. This can be achieved by explaining the importance of the experiment, to let the subjects feel that they are going to do something special for you and, most of the time, important for the whole scientific community or for the introduction of new services in the digital world.

We were seven people: three seated at 3 times and four at 4 times the height of the screen. During the test, no one was

taking notice of whether we were filling out the scoring sheets properly or commenting on the video on the screen or even joking among ourselves! This was a really shocking experience for me, but you must also consider that at those times the scores collected from 3H and 4H viewers and from all the test sequences were all put together to compute a "grand mean."

Today no one would consider this behavior a "best practice" in subjective assessment. Scores collected from viewers seated at different distances are considered separately, as are scores coming from different video clips. Though this seems quite obvious and easily sharable, not that much has been done so far to harmonize the "instructions to the viewers." All the relevant recommendations suggest to read a text to the viewers that briefly explains what they are to do and how they are to do it. The use of a training session is also recommended, as well as the use of test material that is different from that used for the test.

Now let me disagree with both of the above.

Reading a text certainly has the advantage of providing all the viewers the same information; but it may be that not all the subjects understand the text in the same way, and in any case this tends to result in an aseptic relationship between the test manager and volunteers participating as viewers. As mentioned above, testing is often boring, due to the fact that the same four (may be five) video clips are seen by the viewers so many times and sometimes with very little differences in quality among them. This demands a lot of attention of the viewing subjects, and it almost always causes a lot of frustration as the test sessions progress ("Always the same flowers!" or "Always the same train running beneath a calendar going up and down!").

So it's a "must" that the test manager engage the viewers! But the point is: how?

Certainly paying them is a good start, but it may be not sufficient. In most cases people come to your laboratory without knowing anything about testing, so you must make them feel comfortable about performing the task. But it might be useful to involve them emotionally, telling them that "this is an important experiment, this is the first time such an experiment is being done, and many industries world-wide are waiting for the results," and so on. Motivation works very well for people entering a laboratory for the first time. It happened to me that not a few of my subjects asked for a "participation certificate." So at first talk about the importance and the meaning of the test, and then begin to explain what to watch and what to do with their scoring sheet (or buttons on the screen).

This is another crucial task that, if well done, will allow you to obtain better and more stable results. This is the main role of the practice session (also called training session). You will pick some of the lowest, middle and highest quality video clips of the ones you are using for the test; these video clips will be presented using the same presentation that will be used for the actual test, to simulate the task the subjects will face.

Why do I select video clips that are part of the test set? Because it is important to show to the naïve subjects where it is preferable to look, picture by picture. This allows all the naïve subjects to respond in a more homogeneous way to the stimuli (i.e., to the impairments or improvements in the video clips). I know this is against what is written in most of the traditional literature, but it works! What you have to avoid in the editing of the training session is to show the same video clips in the same order they will be shown at the beginning or during the actual test. Also the training session will consist of not less than five but not more than eight Basic Test Cells.[1]

[1] A Basic Test Cell (BTC) is the sequence of messages and video clips presentations that allows to evaluate a single test condition (also called "test point" - TP)

Before the training session begins, explain the meaning of each grade in the scale selected for your experiment. It is important to provide to them a mental anchor for each grade. As an example for a five level scale ACR test you may explain that 5 is used when no impairment is seen and the video looks perfect, 4 is used when they see or even think they see some artifact, but in any case impairment is difficult to see, 3 is used when some artifact is easily visible, 2 is used when many artifacts are seen and they are clearly visible, and 1 is used when the picture quality is really poor.

During the training session you will stay close to the subjects, verifying that they vote at the appropriate time (not before the video clips are finished and not too late) and helping them to properly understand the meaning of a score; they must not be scared to use the full quality or impairment range available. Furthermore, I strongly recommend that you intervene when you see that a subject scores a perfect image (e.g. a source) with a "3" (or lower rate) or when a very poor quality video clip is scored with a "4" or a "5". You can also provide such comments when the training session is completed, revising the scores they have entered together with the subjects. If you see that one or more subjects made several errors, it is recommended that the training session be played again.

Remember that humans are "really different" from each other, and also what is obvious to you may be hard for people not in your industry to see. I know so many people who told me that after having participated in a test in my laboratory, they were no longer able to fully enjoy a TV program because they were seeing a lot of "impairments" and did not feel as comfortable as before when watching TV!

Last, but not least, the current literature describes how to screen viewing subject for their vision. Well, the training phase allowed me in many cases to screen the subjects for their behavior during the training. Some people who appeared "normal" clearly revealed their psychology when asked to

perform a task that was revealed as too complicate for them to complete. This comes out clearly when you see all the scores flattened to the lower or to the upper grades

Let me conclude by saying that human subjects are one of the main tools a test manager needs to have. You have to select them carefully but mainly you have to train them in the best possible way to avoid getting unstable or even unusable results.

Vittorio Baroncini is a senior researcher at Fondazione Ugo Bordoni a research institute in Rome. He is the responsible for Multimedia and TV quality assessment area. Chair of the MPEG Test Group, Vice-Chair of ITU-R WP6C and co-founder of the Video Quality Expert Group, he is author of many paper to conferences and scientific journals. He is also co-author of two books on MPEG.