On Training the Crowd for Subjective Quality Studies

Tobias Hossfeld

Experiences: Crowd vs. Lab Tests

Crowdsourcing enables new possibilities for quality evaluation by conducting subjective studies with the crowd of Internet users. The advantages are the large number of test subjects, fast turn-around times, and low reimbursement costs of the participants. Further, crowdsourcing allows easily addressing additional features like diverse populations or real-life environments.

Moving the evaluation task into the Internet, however, generates additional challenges and differences from lab studies in conceptual, technical and motivational areas (Hossfeld & Keimel, 2014). Due to the remoteness of the test participants, reliability of test results requires advanced test design including consistency checks, content questions, etc. as well as statistical analysis methods such as outlier detection, as not all test conditions will be typically assessed by all subjects in crowdsourcing. Hossfeld et al. (2014) provided best practices for test design and analyzed statistical methods that lead to similar subjective results for crowdsourcing and laboratory studies, e.g. for initial delays and stalling of online video streaming (Hossfeld et al., 2012). Nevertheless, quality tests of videos compressed with H.264/AVC at different bitrates and transmission errors differed absolutely for lab and crowd studies (Hoßfeld & Keimel, 2014). The reasons for the difference may be hidden influence factors in crowdsourcing due to heterogeneous hardware like subjects' screens or improper training sessions.

Training Sessions in Crowdsourcing

The conceptual differences arise mainly from the fact that crowdsourcing tasks are usually much shorter (5-15 min.) than comparable laboratory tests and due to the lack of a test moderator. The user is guided via the web interface through the tests, including an explanation about the test itself, what to evaluate and how to express the opinion. The training of subjects is mostly conducted by means of qualification tests. Nevertheless, in case of any problems with understanding the test, uncertainty about rating scales, sloppy execution of the test, or fatigue of the test user, appropriate mechanisms or statistical methods have to be applied. Therefore it is more difficult to ensure subjects have fully understood the training, in particular as no direct feedback between supervisors and subjects is possible. Due to the short task duration in crowdsourcing, demo trials to familiarize the subject with the test structure and practice trials not included in the analysis significantly decrease the efficiency of a test and increase the costs. Hossfeld and Keimel (2014) show that without any worker training and reliability questions the results are significantly worse than with lab or advanced crowdsourcing designs. Training phases must be included in the task design!

Integrate a Feedback Channel

In general, all questions from the subjects should be answered. A feedback channel can be implemented, e.g., via comments, a contact form or forums. For allowing direct feedback, a communication chat (e.g., via social network apps) is possible, but only for small or short tests, as crowdsourcing users conduct the test whenever they want until the number of required subjects is reached.

As a side effect, this helps to increase the reputation of the test administrator, as participants tend to gather in virtual communities and share their experiences with certain tests and tasks.

Two-stage Test Design

Hoßfeld et al. (2014) propose a general recommendation for crowdsourcing quality tests, the two-stage test design. The first stage includes a simple and easy to do task which tests the reliability of users, gathers a huge subject pool, gathers (demographic) information about the users, is very short (less than 1 min.) and low paid. Also, the training session including demo and practice trials is performed in this stage. This creates a pseudo reliable, trained group of users, who will be later invited to the actual quality test, which presents the second stage, as illustrated in Figure 1. In our experiments, creating this pseudo-reliable panel increases the overall efficiency by more than 60 % in terms of costs and reliable results, which is the major argument for introducing the first stage. Nevertheless, reliability mechanisms in the second stage and post-screening are required to ensure a reliable data set. This design only works with same pool of participants gathered in first stage. Hence, a series of tests should be done in a reasonable time frame, otherwise the training session may be useless.

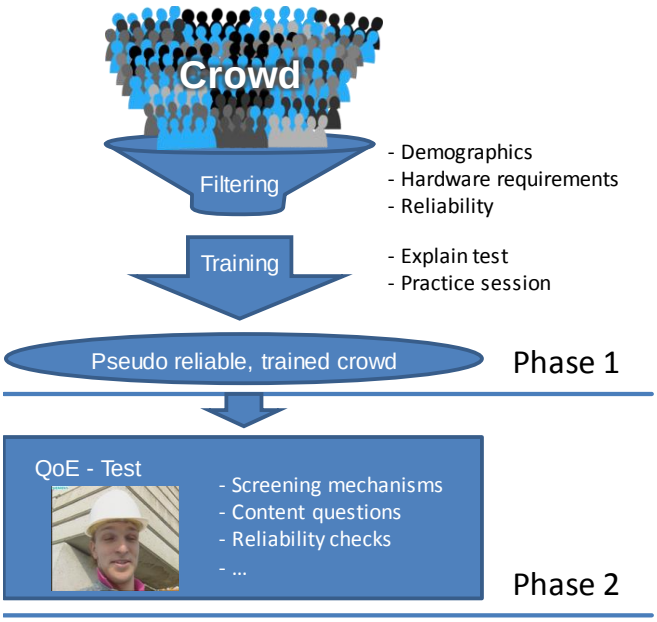

Figure 1. Two-stage design for crowdsourcing subjective studies.

In Momento Methods

Another possibility to cope with efficiency and costs of training sessions compared to actual quality tests in crowdsourcing is introduced by Gardlo et al. (2014). The basic idea of the *in momento* approach is that users are shown an *in momento* verification of their reliability and that users decide whether to stop or to continue the test, but only if a reliability

threshold is exceeded. Users who want to increase their earnings are allowed to perform additional tasks, while users who intentionally only came for one short task assignment should not be overstressed. Users may also be allowed to continue a test session after a certain time, but an upper limit needs to be specified so as not to lose the effect of the training session. As a result of their approach, the performance of the crowd in their study was significantly increased with lower overall costs and more reliable results. Nevertheless, this approach requires automated reliability mechanisms and advanced statistical output analysis of the user ratings which are even more complex than for the two-stage design.

References

Hossfeld, T., Egger, S., Schatz, R., Fiedler, M., Masuch, K., & Lorentzen, C. (2012, July). Initial delay vs. interruptions: between the devil and the deep blue sea. Proceedings of Fourth International Workshop on Quality of Multimedia Experience (QoMEX 2012).

Hossfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., & Tran-Gia, P. (2014). Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing. *IEEE Transactions on Multimedia*, 16(2), 1-18. doi: 10.1109/TMM.2013.2291663.

Hossfeld, T., & Keimel, C. (2014). Crowdsourcing in QoE Evaluation. In S. Möller & A. Raake (Eds.), *Quality of Experience: Advanced Concepts, Applications and Methods* (pp. 323-336). Springer: T-Labs Series in Telecommunication Services, ISBN 978-3-319-02680-0.

Gardlo, B., Egger, S., Seufert, M., & Schatz, R. (2014, June). Crowdsourcing 2.0: Enhancing Execution Speed and Reliability of Web-based QoE Testing. Proceedings of the IEEE International Conference on Communications (ICC 2014).

Tobias Hossfeld is heading the FIA research group "Future Internet Applications & Overlays" at the Chair of Communication Networks, University of Würzburg. He finished his PhD in 2009 and his professorial thesis (habilitation) "Modeling and Analysis of Internet Applications and Services" in 2013. He has been visiting senior researcher at FTW in Vienna with a focus on Quality of Experience research. He has published more than 100 research papers in major conferences and journals and received the Fred W. Ellersick Prize 2013 (IEEE Communications Society) for one of his articles on QoE.