



One Shot Crowdttesting: Approaching the Extremes of Crowdsourced Subjective Quality Testing

Michael Seufert¹, Tobias Hofffeld²

¹Institute of Computer Science, University of Würzburg, Germany

²Chair of Modeling of Adaptive Systems, University of Duisburg-Essen, Germany

seufert@informatik.uni-wuerzburg.de, tobias.hossfeld@uni-due.de

Abstract

Crowdsourcing studies for subjective quality testing have become a particularly useful tool for Quality of Experience researchers. Typically, crowdsourcing studies are conducted by many unsupervised workers, which rate the perceived quality of several test conditions during one session (mixed within-subject test design). However, those studies often show to be very sensitive, for example, to test instructions, design, and filtering of unreliable participants. Moreover, the exposure of several test conditions to single workers potentially leads to an implicit training and anchoring of ratings. Therefore, this work investigates the extreme case of presenting only a single test condition to each worker (completely between-subjects test design). The results are compared to a typical crowdsourcing study design with multiple test conditions to discuss training effects in crowdsourcing studies. Thus, this work investigates if it is possible to use a simple “one shot” design with only one rating of a large number of workers instead of sophisticated (mixed or within-subject) test designs in crowdsourcing.

Index Terms: subjective quality testing; crowdsourcing; Quality of Experience; training effect; anchoring effect; study design

1. Introduction

Quality of Experience (QoE) is the subjective perception of the quality of a service as a whole. Both network and service providers are putting increasing emphasis on QoE and want to deliver the best service to the end user to achieve a high customer satisfaction. To identify the influence of certain parameters and to quantify their impact on the mean opinion scores (MOS), extensive subjective studies have to be conducted to obtain a QoE model. Crowdsourcing, i.e., the outsourcing of small tasks to a large crowd, has been successfully used for subjective QoE assessments (e.g., [1, 2]). Crowdsourcing studies show several advantages compared to classical laboratory studies, for example, in terms of price, speed, and the more realistic setting of service consumption (especially, context and system factors). Nevertheless, the design of a crowdsourcing QoE study is crucial to avoid typical pitfalls from the heterogeneous environment and unsupervised nature of the study conduction [3].

Typically, there is a trade-off between time consumption, number of ratings, cost, and data quality (e.g., [4, 5]). Thus, the participants of crowdsourced subjective quality tests are exposed to several test conditions to balance the cost and time consumption of such studies and reach a high data quality. However, even short tests potentially introduce implicit training and anchoring of the test participants, as, during the perception and rating of a subsequent test conditions, they can always rely on the past test conditions. The training effect names the phe-

nomenon that the workers learn from the test instructions and/or the presentation of the test conditions what influence factors to focus on and to account for during the ratings. The anchoring effect, on the other hand, stems from previous experience of test conditions with worse, similar, or better quality, which allows participants to anchor the current test condition based on the previous ratings. The goal of this study is to investigate these effects by comparing a typical crowdsourced quality study (mixed within-subject test design) to an extreme approach in which every participant is exposed to only one test condition (completely between-subject test design). The rationale here is that if a large number of users rate a condition, we will end up with the true MOS. Thus, the (mixed) within-subject test design would not be required. The extreme “one shot” design is nowadays easily possible with crowdsourcing, but it has to be investigated if it results in different QoE models than existing sophisticated test designs. Eventually, the question shall be raised if training and anchoring is desired in subjective quality testing and what impact it has on resulting QoE models.

Therefore, this paper is structured as follows. In Section 3, the methodology of this work is described and results are presented in Section 4. Section 5 discusses the results of this study and gives an outlook to future work.

2. Background and Related Work

2.1. Test Design

When conducting subjective studies, different test designs are possible (e.g., [6]). A within-subject (or repeated measures) test design typically exposes all conditions of the independent variable to the participant. To avoid effects of the order of stimuli, the test conditions are typically presented in a randomized order. The advantage of within-subject design is that the individual differences between the participants can be controlled, which is especially beneficial in a crowdsourcing environment (e.g., reliability of workers). However, if the number of independent variables and/or the number of conditions per variable is high, long studies result, which may introduce training effects and fatigue [7]. [8] transferred within-subject experiments from the lab to the Amazon MTurk platform. They noticed a close match of absolute results even when studies were repeated multiple times online, such that they concluded that also between-subject designs could be conducted robustly with MTurk.

As crowdsourcing tasks should be designed to be short, current subjective QoE studies are mostly mixed within-subject studies, in which many participants are recruited and each participant is exposed to a subset of stimuli. Thus, a high number of participants is needed to cover all conditions with a sufficient number of ratings. As eventually all ratings for a condition are

combined, the advantages of within-subject test designs to control individual differences are lost. Then, also a between-subject test design (e.g., “one shot”) should be possible.

In a between-subject test design, the test participants are divided into groups, and different stimuli (i.e., test conditions) are exposed to each group. This results in shorter studies, but individual differences between the test participants may influence the results. In case of a factorial test design, separate groups of subjects are required for each combination of the different values of the independent variables. Among others, [9] found that inconsistency occurred occasionally to most online participants, and they concluded that a within-subject design is essential for empirical crowdsourcing studies. Nevertheless, monitoring of task execution allows to immediately detect inconsistencies and filter out such participants (e.g., [10]).

2.2. Effects due to Training Phase

Training the test participants is a concept, which is widely used in laboratory studies. It can also improve the quality for crowdsourcing tasks, which require a high objective accuracy (e.g., visual object detection [11]). For QoE-related crowdsourcing studies, significantly different results were observed without any worker training and additional quality assurance mechanisms, and training was found to increase the similarity of results between laboratory and crowdsourcing (e.g., [12]). The question remains whether training in perceptual quality studies only helps the workers to understand the rating task, or whether it also implicitly suggests the notion of perception/quality of the researchers themselves (self-fulfilling prophecy), and thus, leads to biased results.

Possible reasons for the changed perception after training phases include well-known cognitive biases like anchoring or serial position effects. The anchoring heuristic [13] refers to the tendency of people relying heavily on the first piece of information offered (anchor) and using additional information for incremental adjustments to reach their estimate. In case of a quality study, thus, the training of a participant how to rate (e.g., training with good/bad quality condition, pointing participants’ focus to distortions) could act as anchor for the subsequent conditions and ratings. In a similar fashion, the serial position effects [14, 15], i.e., the tendency of recalling items best, which were presented first (primacy effect) or last (recency effect), also might influence perceptual quality studies. Thereby, participants could implicitly compare the currently presented stimulus with this first or last test condition they were exposed to.

[16] observed for web QoE that a person’s current experience of service quality is shaped by past experiences (referred to as ‘memory effect’). In particular, in addition to the current QoS level, the user experienced quality of the last downloaded web page has to be taken into account. For video QoE of HTTP adaptive streaming, [17] did not find ‘recency time effects’ considering the time how long high quality is played out after the last quality switch.

3. Methodology

A subjective quality test for investigating the influence of the bitrate on the perceived quality of H.264 video sequences was conducted via crowdsourcing (cf. [18]). Five source sequences in 1080p at 25 fps were used, which cover a wide variety of characteristics. The source video sequences were downsampled using *ffmpeg* tool to standard resolution (576p) and cut to a length of 10s to meet the possibly low Internet connections of

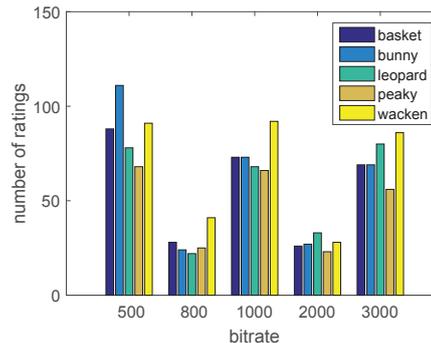


Figure 1: Number of ratings per test condition

the crowd workers and were encoded using the *x264* implementation. Five different test conditions were prepared for the study (500, 800, 1000, 2000, 3000 kbps).

We used an established online test framework (e.g., [1, 10]), which follows the best practices described in [3]. The study was advertised on Microworkers with a compensation of 0.30\$. Every worker had to watch five videos with different content each in randomized order, and rate the quality afterwards on a 5-point ACR scale. To avoid network influences during the playback (e.g., initial delay, stalling), all videos were downloaded to the local browser cache before the playback. The specific bitrate of a video was chosen with equal probability from the five test conditions. As too few reliable ratings remained for analyzing the anchoring and recency effects (see Sections 4.1 and 4.2), we re-run the campaign only with the 500, 1000, and 3000 kbps conditions to gather enough ratings for these analyses. After the typical filtering (i.e., consistency questions, improper test execution, faulty test presentations), 1445 ratings were obtained. Figure 1 shows the histogram of the final number of ratings per test condition.

4. Numerical Results

In this section, we first identify the presence of anchoring and serial order effects in our crowdsourcing study. Then, we compare the typical study design with five conditions to the “one shot” design.

4.1. Anchoring and Primacy Effect

In the conducted crowdsourcing study, no explicit training phase was included. Therefore, the anchoring of the participants can only happen with the first presented test condition, and thus, this effect coincides with the primacy effect. Note that the participants watch five videos with different content each, so the anchoring happens with a different content than the ratings used for this analysis, which is ignored here due to the randomization of the content order. Figure 2 shows the average ratings of participants for one content, which were exposed to a low quality (500kbps, green bars) first or a high quality (3000kbps, yellow bars) first, respectively. The mean opinion scores (MOS) and their 95% confidence intervals are shown for each test condition and are compared to all ratings (blue bars). The sets, which investigate the anchoring, contain approximately a fifth of all ratings. We chose the *basket* content, as it had the most ratings for the anchoring sets. As they were rated less frequently, for 800 and 2000 kbps, only few (2-5) ratings remained for the analy-

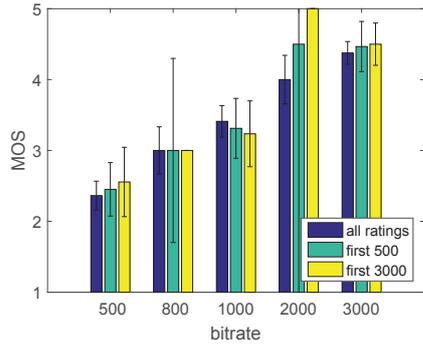


Figure 2: Anchoring effect for content *basket*. Comparison of ratings when low (500kbps) or high (3000kbps) quality condition was rated first.

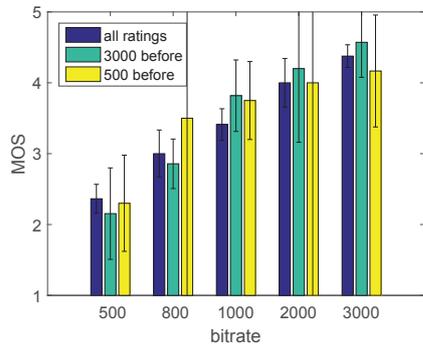


Figure 3: Recency effect for content *basket*. Rating results when low (500kbps) or high (3000kbps) quality condition was rated before.

sis, which explains the large and zero size confidence intervals, respectively. For the other, more frequently tested conditions, more (> 15) ratings could be used. In this case, the MOS differences are quite small and the confidence interval overlap for both anchoring on worst and best quality. Thus, we cannot see an anchoring effect in the QoE results of our crowdsourcing study.

4.2. Recency Effect

Figure 3 investigates the recency effect. Therefore, all ratings, which were given for the next video after the the participants watched and rated a video with the highest quality (i.e., 3000 kbps) and the lowest quality (i.e., 500 kbps), were collected. In the figure, these sets are compared to all ratings (blue bars) for the content *basket*. The x-axis depicts the different test conditions and the y-axis shows the MOS and the 95% confidence intervals. Again, few (2-7) ratings remained for 800 and 2000kbps, which explains the very large confidence intervals for these conditions. It can be seen that the exposure to the highest quality test condition (green bars) causes lower ratings for bad conditions and higher ratings for good conditions. For the ratings after the worst quality (yellow bars), the opposite behavior is visible. Medium quality conditions seem to be rated better in this case. Still all confidence intervals overlap, which shows that there are no statistical differences, and thus, the effects might be negligible.

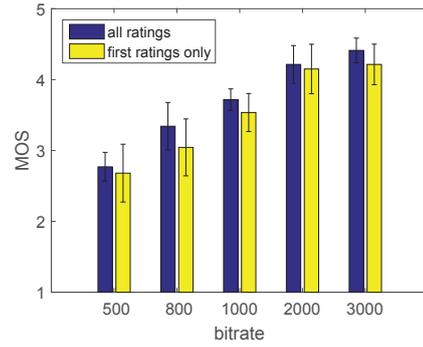


Figure 4: Comparison of traditional crowdsourcing study with “one shot” test design for content *wacken*.

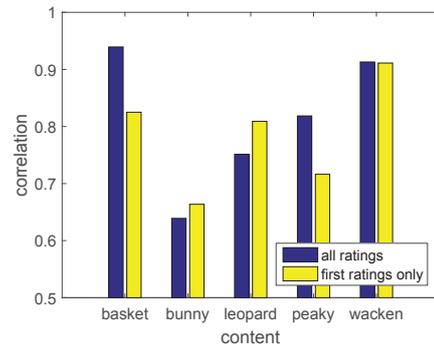


Figure 5: Pearson correlation coefficients between MOS and bitrate for all contents.

4.3. ‘One Shot’ vs. All Ratings

To compare the typical study design and the “one shot” design, for each video content, the ratings are rearranged into two sets. The first set contains all ratings given about the five different test conditions. The second set contains only those ratings, which were given as the first rating of a participant. Thus, in the second set, all ratings were given without any experience from previous test conditions, as it would be the case if only one test condition was presented to each worker. Thus, it can be assumed that these ratings are free from any training or anchoring effects. Obviously, the number of ratings in the second set is only one fifth of the number of ratings in the first set.

Figure 4 directly compares the two sets for the content *wacken*, which has the most ratings overall. First of all, it can be seen that the video bitrate is positively correlated to the MOS, i.e., a higher bitrate results in a higher MOS. For all ratings (blue bars), the selected test conditions show a nice distribution over the rating scale with almost constant MOS differences between adjacent test conditions. When comparing the set of all ratings to the set of the first ratings (yellow bars), lower MOS values for the set of first ratings are clearly visible although the confidence intervals overlap. Nevertheless, the test conditions show a very similar distribution over the MOS scale. This means that although the absolute numbers are slightly different, a one shot test design results in a similar QoE model in our study.

Generalizing this result to all contents, Figure 5 shows the Pearson correlation coefficient between the MOS values and the bitrate of the test condition for all ratings (blue) and the first rat-

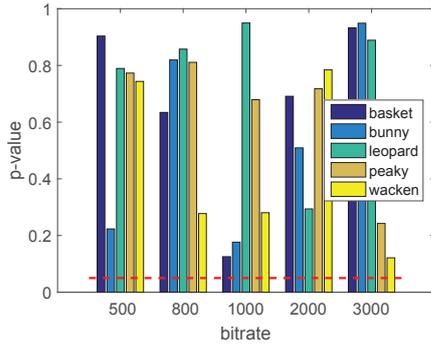


Figure 6: p -values of Kruskal-Wallis test per content and per test condition.

ings only (yellow). The results for *wacken* presented in Figure 4 both reach a high correlation of 0.9. Also for the other contents the correlations are similar for both all ratings and first ratings only. It can be seen that all correlation coefficients are strongly positive (larger than 0.6) and that the difference between the correlation coefficients for traditional (all ratings) and the “one shot” (first ratings only) crowdsourcing are smaller than 0.1.

To assess the difference between the two approaches statistically, a Kruskal-Wallis test, i.e., a non-parametric test for comparing distributions, was conducted per content and per test condition. The resulting p -values for the null hypothesis that the two sets of ratings (all ratings, first ratings only) come from the same distribution are depicted in Figure 6. The x-axis shows the different test conditions, while the five contents are represented by differently colored bars. It can be seen that all p -values are larger than 0.05 (dashed red line), thus, the null hypothesis cannot be rejected. This means, there are no statistical significant differences for the two sets.

To sum up, in the conducted crowdsourcing study, no significant anchoring and serial order effects were visible. Furthermore, no difference could be observed between the traditional approach of a mixed within-subject study and a between-subject test design. This confirms the assumption that a “one shot” crowdsourcing approach feasible and is able to provide similar results than a traditional crowdsourcing study.

5. Discussion and Outlook

In this work, we investigated anchoring and serial order effects in our conducted crowdsourcing study and compared the MOS results with a “one shot” crowdsourcing approach. The results suggest that, when training effects (anchoring or serial order effects) are negligible, an extreme “one shot” design, in which every worker is exposed to only one test condition, seems to be applicable. However, it remains an open issue what happens when training effects are present or even desired.

Unfortunately, the applied methodology of this paper could not be transferred to other crowdsourcing and laboratory studies on subjective quality, e.g., [19], mainly due to missing rating timestamps. Moreover, due to the small number of participants in long within-subject laboratory studies, a random assignment of test conditions leads to very few first ratings per condition and large confidence intervals, which make a “one shot” investigation unfeasible. Furthermore, if training is needed, a “one shot” design is not economic because of the time (i.e., costs) needed for training the participant for a single rating.

To better understand the cognitive biases in crowdsourcing and laboratory studies, dedicated experiments have to be conducted, which investigate the influence of anchoring, primacy, recency, etc. on the perception and the resulting ratings. Also the impact of training phases has to be quantitatively assessed and the bias of QoE results has to be discussed. Ideally, future subjective quality studies, which need a training phase, should report that the training of participants did not bias their perception. Therefore, appropriate methodologies and measures need to be defined.

In future work, we will not only continue the investigation of the “one shot” test design, but also tackle the opposite extreme of long crowdsourcing studies, in which participants rate a large number of test conditions, i.e., we will investigate if crowdsourcing is applicable for a completely within-subject test design. Several related works exist also from different applications of crowdsourcing, e.g., [20, 21, 22], but dedicated studies for crowdsourced QoE testing are missing. It is expected that many effects (e.g., anchoring, training, fatigue, boredom, distraction) are clearly visible in such studies and possibly influence the resulting QoE models.

6. Acknowledgements

This work is funded by Deutsche Forschungsgemeinschaft (DFG) under grants HO 4770/2-1, TR257/38-1: “*Design and evaluation of new mechanisms for crowdsourcing as emerging paradigm for the organization of work in the Internet*” .

7. References

- [1] T. Hößfeld, R. Schatz, M. Seufert, M. Hirth, T. Zinner, and P. Tran-Gia, "Quantification of YouTube QoE via Crowdsourcing," in *Proceedings of the International Workshop on Multimedia Quality of Experience - Modeling, Evaluation, and Directions (MQoE)*, Dana Point, CA, USA, 2011.
- [2] J. A. Redi, T. Hößfeld, P. Korshunov, F. Mazza, I. Pova, and C. Keimel, "Crowdsourcing-based Multimedia Subjective Evaluations: A Case Study on Image Recognizability and Aesthetic Appeal," in *Proceedings of the 2nd International Workshop on Crowdsourcing for Multimedia (CrowdMM)*, Barcelona, Spain, 2013.
- [3] T. Hößfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force Crowdsourcing," 2014, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003).
- [4] J. Rogstadius, V. Kostakos, A. Kittur, B. Smus, J. Laredo, and M. Vukovic, "An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 2011.
- [5] A. Mao, E. Kamar, and E. Horvitz, "Why Stop Now? Predicting Worker Engagement in Online Crowdsourcing," in *Proceedings of the 1st AAAI Conference on Human Computation and Crowdsourcing*, Palm Springs, CA, USA, 2013.
- [6] G. Charness, U. Gneezy, and M. A. Kuhn, "Experimental Methods: Between-subject and Within-subject Design," *Journal of Economic Behavior & Organization*, vol. 81, no. 1, pp. 1–8, 2012.
- [7] R. Schatz, S. Egger, and K. Masuch, "The Impact of Test Duration on User Fatigue and Reliability of Subjective Quality Ratings," *Journal of the Audio Engineering Society*, vol. 60, no. 1/2, pp. 63–73, 2012.
- [8] S. Komarov, K. Reinecke, and K. Z. Gajos, "Crowdsourcing Performance Evaluations of User Interfaces," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris, France, 2013.
- [9] A. Abdul-Rahman, K. J. Proctor, B. Duffy, and M. Chen, "Repeated Measures Design in Crowdsourcing-based Experiments for Visualization," in *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, Paris, France, 2014, pp. 95–102.
- [10] B. Gardlo, S. Egger, M. Seufert, and R. Schatz, "Crowdsourcing 2.0: Enhancing Execution Speed and Reliability of Web-based QoE Testing," in *Proceedings of the International Conference on Communications (ICC)*, Sydney, Australia, 2014.
- [11] H. Su, J. Deng, and L. Fei-Fei, "Crowdsourcing Annotations for Visual Object Detection," in *Proceedings of the Workshops at the 26th Conference on Artificial Intelligence (AAAI)*, Toronto, ON, Canada, 2012.
- [12] T. Hößfeld and C. Keimel, "Crowdsourcing in QoE Evaluation," in *Quality of Experience*. Springer, 2014, pp. 315–327.
- [13] A. Tversky and D. Kahneman, "Judgment Under Uncertainty: Heuristics and Biases," in *Utility, Probability, and Human Decision Making*. Springer, 1975, pp. 141–162.
- [14] J. Deese and R. A. Kaufman, "Serial Effects in Recall of Unorganized and Sequentially Organized Verbal Material," *Journal of Experimental Psychology*, vol. 54, no. 3, pp. 180–187, 1957.
- [15] B. B. Murdock Jr., "The Serial Position Effect of Free Recall," *Journal of Experimental Psychology*, vol. 64, no. 5, pp. 482–488, 1962.
- [16] T. Hößfeld, S. Biedermann, R. Schatz, A. Platzer, S. Egger, and M. Fiedler, "The Memory Effect and its Implications on Web QoE Modeling," in *Proceedings of the 23rd International Teletraffic Congress (ITC)*, 2011, pp. 103–110.
- [17] T. Hößfeld, M. Seufert, C. Sieber, and T. Zinner, "Assessing Effect Sizes of Influence Factors Towards a QoE Model for HTTP Adaptive Streaming," in *Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX)*, Singapore, 2014.
- [18] M. Seufert, O. Zach, T. Hößfeld, M. Slanina, and P. Tran-Gia, "Impact of Test Condition Selection in Adaptive Crowdsourcing Studies on Subjective Quality," in *Proceedings of the 8th International Conference on Quality of Multimedia Experience (QoMEX)*, Lisbon, Portugal, 2016.
- [19] K. Fliegel, "QUALINET Multimedia Databases v5. 5," 2014. [Online]. Available: <http://dbq.multimediatech.cz/>
- [20] D. R. Saunders, P. J. Bex, and R. L. Woods, "Crowdsourcing a Normative Natural Language Dataset: A Comparison of Amazon Mechanical Turk and In-lab Data Collection," *Journal of Medical Internet Research*, vol. 15, no. 5, p. e100, 2013.
- [21] M.-H. Pan and H.-C. Wang, "Enhancing Machine Translation with Crowdsourced Keyword Highlighting," in *Proceedings of the 5th ACM International Conference on Collaboration Across Boundaries: Culture, Distance & Technology (CABS)*, Kyoto, Japan, 2014, pp. 99–102.
- [22] J. Cheng, J. Teevan, and M. S. Bernstein, "Measuring Crowdsourcing Effort with Error-Time Curves," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, Seoul, Republic of Korea, 2015, pp. 1365–1374.