Do Scale-Design and Training Matter for Video QoE Assessments through Crowdsourcing?

Bruno Gardlo
The Telecommunications Research Center Vienna
A-1220, Vienna, Austria
gardlo@ftw.at

Sebastian Egger
Austrian Institute of Technology
A-1110, Vienna, Austria
sebastian.egger@ait.ac.at

Tobias Hossfeld
University of Duisburg-Essen
Essen, Germany
tobias.hossfeld@uni-due.de

ABSTRACT

Crowdsourcing (CS) has evolved into a mature assessment methodology for subjective experiments in diverse scientific fields and in particular for QoE assessment. However, the results acquired for absolute category rating (ACR) scales through CS are often not fully comparable to QoE assessments done in laboratory environments. A possible reason for such differences may be the scale usage heterogeneity problem caused by deviant scale usage of the crowd workers. In this paper, we study different implementations of (quality) rating scales (in terms of design and number of answer categories) in order to identify if certain scales can help to overcome scale usage problems in crowdsourcing. Additionally, training of subjects is well known to enhance result quality for laboratory ACR evaluations. Hence, we analyzed the appropriateness of training conditions to overcome scale usage problems across different samples in crowdsourcing. As major results, we found that filtering of user ratings and different scale designs are not sufficient to overcome scale usage heterogeneity, but training sessions despite their additional costs, enhance result quality in CS and properly counterfeit the identified scale usage heterogeneity problems.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Evaluation

Keywords

Crowdsourcing; QoE; Reliability; Methodology; Scales

1. INTRODUCTION

Video delivery over the Internet accounts for a large share of nowadays traffic. Recently, the delivery method has seen a change from constant quality streaming to HTTP adaptive streaming with variable video quality. As current video quality prediction models do not properly consider varying video quality, the interest in QoE prediction models for adaptive video quality streams has gained momentum. For the development of these prediction models, subjective data for training algorithms and models is needed. However, gathering the necessary subjective data is a Sisyphean task, as quality adaptation can be achieved on several dimensions such as spatial, temporal and image compression. In addition, a multitude of different quality profiles of different lengths and different complexity appears in real world measurements [22]. Altogether this results in a huge quantity of different video quality profiles that have to be subjectively evaluated.

Crowdsourcing is an alternative approach to traditional laboratory tests for conducting subjective quality testing. However, the implementation of known test methodologies and setups from lab tests in the CS domain is not straight forward due to the Internet-based environment and remote test participants. As a result additional challenges and differences in the conceptual, technical and motivational areas emerge [8]. To address these challenges, [9] provides best practices for quality assessment with crowdsourcing, addressing the design, implementation and reliability assessment for successful CS quality testing. However, several of thess practices are either targeted towards well perceivable impairments or qualities (stall events, attractiveness of images) or they propose paired testing on discriminative rating scales (DCR). The complication with video quality under spatial, temporal and image compression impairments is the fact, that their perception and respective rating on an ACR scale is slightly more complicated than e.g. for stall events. Whereas the latter one is a clearly visible impairment, the aforementioned ones are more subtle on a perceptual level. Hence, video QoE assessment in CS environments is particularly challenging and leads to disparities between lab gathered results and CS gathered results (for identical video quality settings) as described in [2, 15]. Such scale heterogeneity issues are well known [20] and can be based on several issues such as scale design, cultural backgrounds, language differences in labels etc. Hence, in this paper we set out to identify if these problems can be overcome by scale design and certain training sessions. In particular, the following hypotheses are investigated by means of crowdsourcing experiments.

Hypothesis 0 *The scale usage problem still occurs in crowdsourcing after filtering unreliable user ratings.*

Hypothesis 1 *The scale design overcomes the scale usage problems in crowdsourcing.*

Hypothesis 2 *Proper training overcomes the scale usage problems in crowdsourcing.*

CrowdMM'15, October 30, 2015, Brisbane, Australia.
 ISBN 978-1-4503-3746-5/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2810188.2810193.

The remainder of this paper is structured as follows. Section 2 provides background on crowdsourcing for quality assessment and emphasizes some differences between lab and CS tests. In particular, related work and results from literature are revisited that indicate the need for investigations on scale design in CS tests and best practices compiled out of the results. Furthermore, lab based guidelines for subject training are reviewed. Section 3 introduces the study and scale design of our CS experiments to investigate the impact of different rating scales and training sessions on performance and efficiency. The numerical results are analyzed in Section 4. First, the applied methodology for outlier detection and the efficiency of the scales are described, before the user ratings for various test conditions are compared. Finally, Section 5 concludes this work and provides practical guidelines on the scale design for researchers conducting crowdsourcing experiments.

2. BACKGROUND AND MOTIVATION

For assessing QoE of a large variety of different test cases, crowdsourcing has recently gained considerable attention [1, 3, 9, 18, 26] as a fast and economical alternative to lab based QoE assessments. In that respect it is applied to several quality evaluation tasks such as in e.g., image quality [18], audio quality [1] and video quality [3]. But CS introduces conceptual differences compared to lab tests [8, 15] and CS methods still face certain challenges [9]: scale usage problems (defined in Section 2.1; discussed in Section 2.2). data quality and reliability of user ratings (Section 2.3), motivation of the crowd and its influence on the results (Section 2.4), unsupervised user-training and lack of test moderator (Section 2.5). Those issues makes it difficult to compare CS results to respective lab tests.

2.1 Definition of the Scale Usage Problem

In this paper, we define the *scale usage problem* as follows in order to investigate the hypothesis **H0**, **H1**, **H2**.

Definition 1 If more than 10 %[1] of users use less than 75 % of the rating scale, then the scale usage problem occurs.

In case of a 5-point scale, this means that more than 10 % of the subjects use 3 or less contiguous items. Note that we do not differentiate if users only use the upper part, the lower part, or the middle part of the scale. The used range Δ_i of user i is defined as the difference between the maximum rating M_i and the minimum rating m_i of that user: $\Delta_i = M_i - m_i$. This definition makes only sense, if each user of course experiences the entire quality range during the test. In that case, it is expected that the user is able to discriminate the different quality settings.

We assume that proper training of the subjects before the actual test and evaluation phase will empower the user to discriminate the qualities and to rate them adequately on the used rating scale. However, in crowdsourcing subjective studies are typically only in the order of a few minutes [9]. Therefore, a training of a few minutes decreases the overall efficiency of CS and increases the costs which are directly related to the task completion time. Crowdsourcing platforms like MicroWorkers check the payment per worker depending on the expected task completion time to ensure a minimum wage. Let us assume that a user evaluates $N = 5$ videos which requires time $t_1 = 40\,\mathrm{s}$ per video (for loading the video, watching the video, evaluating the video and also answering content questions to filter unreliable users [9]). If the training consumes time $t_0 = 100\,\mathrm{s}$, then 30 % of the payments are dedicated to training. It is tempting to overcome the scale usage problem by other means like scale design.

[1]The 10 % value was arbitrarily chosen by analyzing the results of the lab study. However, the results in Figure 3 are unambiguous and other values lead to the same conclusions.

2.2 Related Work on Rating Scales

In terms of subject training and scale usage [1, 26] have proposed paired testing as a solution for eliminating offsets between different CS campaigns and laboratory tests. Although this off course minimizes offsets between different test campaigns, it only provides relative ratings instead of absolute category ratings. This is useful for comparing different implementations of algorithms or codecs but provides less insight in the actually perceived quality of the customer.

Therefore, industry is rather interested in absolute category ratings (ACR) as they compare well to several other customer satisfaction measures that are typically used to assess product offerings, as well as questions about various aspects of the customer's interaction with the company [20]. A major drawback of such scales is that their usage often varies between different users. Furthermore, users tend to avoid both ends of the scale, thus the votes tend to saturate before reaching the end points as shown in [2, 15]. Additionally, language and cultural differences regarding the 'distance' between scale labels for a given ITU scale as reported in [14, 24] make it difficult to compare results across cultural or international boundaries. These different usage patterns have been termed *scale usage heterogeneity problem* by [20] and introduce biases to many of the standard analyses conducted with subjective rating data. Such problems impair comparability between lab and CS results severely as depicted in Figure 1 with MOS values ranging between 2.28 and 4.25 – in contrast to the lab results where the full scale from 1 to 5 is utilized. Although these data shows a strong correlation between the MOS values (with a Pearson correlation coefficient of 0.91), the mean absolute error and the maximum absolute error are 0.62 and 1.56, respectively. This clearly depicts *the problem of scale usage heterogeneity in crowdsourcing.*

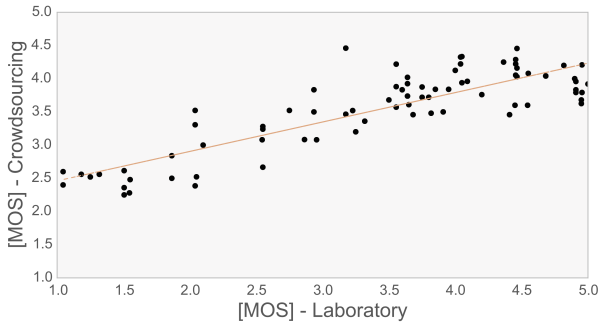

Figure 1: Difference between mean opinion scores (MOS) for laboratory and crowdsourcing results of identical test cases, taken from [2, 15].

Another issue within crowdsourcing is the fact that the language (usually english) the CS workers receive instructions in, and the scale descriptors are labeled in, is (often) different from their native language. Hence, different scale designs can influence the scale usage and the resulting mean opinion scores. Therefore, *the unambiguous design of rating scales* is essential for acquiring proper results from CS campaigns. However, as best design practices on scale designs to be used in CS experiments are missing, one of the aims of this paper is therefore a comparison of different scale designs and their ability to reduce the magnitude of scale usage heterogeneity problems. By doing so we want to identify a scale design that increases the information content of data acquired in crowdsourcing campaigns. Our working assumption is that the implementation of the rating scale in terms of design and number of points influences the rating scale usage problem which is related to **H1**.

2.3 Data Quality and Filtering

Unreliable user ratings in crowdsourcing may be caused by several factors, like problems in executing the test due to hardware or software issues, a wrong understanding of the test, language problems, or sloppy execution of the test. Therefore, it is evident to filter out unreliable users and their ratings as suggested in [3, 9, 17]. The scale usage problem was already visible in the results in [8], but the scale usage problem was not addressed and discussed directly, although a difference of about 1 point on a 6-point scale was existing between lab and CS results at the lower and the upper edge of the scale. The conclusions in [8] showed only that unreliable users have a severe impact on the results which requires screening mechanisms and reliability checks. Reliability checks can be included in the test design [9] (e.g. content questions, gold data, consistency checks). In such cases during the test feedback is given to the subjects about low data quality in order to avoid misunderstandings and motivate users to conduct tests seriously [3]. Furthermore, data from the reliability checks is then used for statistical outlier detection and screening [7].

In contrast, we will investigate the hypothesis **H0** whether the scale usage problem still occurs in crowdsourcing after filtering unreliable user ratings.

2.4 Bias in User Ratings and Normalization

However, still absolute differences between lab and CS evaluations were observed in [8] which could be tackled by typical normalization procedures like z-Scores. The resulting curves of the Z-scores showed no significant differences between lab and CS results. As reason for those absolute differences, the motivation and incentives of the subjects or context of the test settings were mentioned but again the rating scale usage problem was not addressed.

In [23], a bias in the user ratings towards the top-end of the scale was observed which was even stronger in case of higher rewards for the subjects to conduct the test. Normalized z-scores showed however no differences anymore. Those results demonstrate that the scale usage problem in CS may touch different dimensions beyond uncertainty of users how to rate [10] and requires proper means to overcome this issue. In [17], anchors were successfully used to overcome the scale usage problem in CS. To this end, a subset of stimuli was evaluated in the lab and then 5 test conditions corresponding to the 0th, 25th, 50th, 75th, and 100th percentiles of the distribution of quality scores were determined and used throughout the CS test for scale anchoring purposes. However, in [17] a training session was included which we would like to overcome due to cost efficiency in CS.

We investigate the scale design in the hypothesis **H1** whether proper scale design without training of workers is sufficient to overcome the scale usage problem in CS.

2.5 Influence of Training

In lab QoE assessments these effects are typically countered with extensive trainings of the subjects before the actual tests take place as recommended e.g. in [13] and [12]. In CS such exhaustive trainings can not be implemented, as CS tasks are typically much short than lab tests in order not to lose crowd workers attention and ensure reliable results [9]. Thus, no or shorter training sessions are used in CS. However, due to the lack of a test supervisor in CS, an explanation about the test itself, what to evaluate and how to express the opinion are even more important. Thus, training seems to be mandatory to avoid any misunderstandings in executing the test and to make the subject aware of the rating scale [6]. Without any training of the test subjects in CS and reliability checks the obtained quality assessment results are significantly worse than with lab or advanced crowdsourcing designs [6, 8].

We additionally investigate hypothesis **H2**: whether proper training is able to overcomee the scale usage problems in crowdsourcing.

3. SCALE AND STUDY DESIGN

3.1 Implementation of the Rating Scales

Although rating scales in QoE are standardised in terms of number of scale labels, neither their appearance nor their modality (discrete or continuous) is (cf. [12, 13]). However, this is also true for other disciplines such as user experience research or psychology [5, 19], where only the number of items is given, whereas design and modality are not given. To identify differences between scales, the authors in [11, 16] compared scales with different numbers of descriptors as well as discrete and continuous scales. Their results showed no evidence that continuous scales provide more discrimination power or better accuracy than the discrete category scales.

Therefore, we decided to limit the *scale designs* to be compared to discrete scales with five or nine clickable scale items according to the absolute category rating (ACR) scales described in [12, 13]. For the design of the scales as depicted in Figure 2 we used the red-green colour scheme as well known semaphore from traffic control and other commonly used semaphores for satisfaction ratings in the Internet as stars and thumbs-up and thumbs-down. For Scale 1 and 6 we also included non-clickable anchor points at the end of the scale as also used in [12]. Additionally, we decided to include a vertical (Scale 4) and horizontal (Scale 5) ACR-5 scale without any design. Scales were implemented in a way to ensure 100% compliance with the current browsers.

3.2 Lab and Crowdsourcing Experiments

For proper comparing the different scale designs and the resulting rating quality we used videos of different qualities as identical stimulus for each scale. As *video content* for the study we used a 20 second sports clip in 720p, encoded with the x264 software encoder into five different

Table 1: Description of settings in the experiments.

id	setting	scales	training	reliable participants
lab	laboratory	1	yes	39
CS1	crowdsourcing	1–7	no	476
CS-T	crowdsourcing	1,2,6	yes	138

quality levels by utilizing 2-pass constant bitrate encoding at {250,500,800,1250,2400} kbps, resulting in five different stimuli for the study. The content was displayed fullscreen on each subject's screen throughout the study. A description of the different experiments and the number of reliable participants can be found in Table 1.

For the detection of reliable subjects and rating score outliers for each bitrate condition, we used a three step approach. The first step for subject reliability detection was the online reliability check described in [3]. As a second step we used the β_2 screening method described in [13]. In addition, we applied an outlier detection as described in [4], which eliminates individual ratings on condition level instead of subject level as before.

4. RESULTS SCALE COMPARISON

The results of the crowdsourcing experiments are analyzed next. Section 4.1 investigates the scale usage problem after filtering and outlier detection (**H0**). Section 4.2 analyses the rating scale design (**H1**), while Section 4.3 quantifies the influence of training un scale usage (**H2**).

4.1 Filtering and Outlier Detection

In terms of scale efficiency, Table 2 shows the relative number of detected outliers for each scale. One can clearly

Figure 2: Different scale designs as used in this study. The scale designs are available under Creative Commons Attribution 3.0 Austria License at https://github.com/St1c/ratings.

Table 2: Reduction of confidence intervals for each scale and quality level (QL) after outlier removal for CS w/o training.

Scale	Outlier Ratio	QL 1	QL 2	QL 3	QL 4	QL 5
Scale 1	6.05%	0.164	0.000	0.032	0.048	0.090
Scale 2	10.82%	0.178	0.058	0.016	0.121	0.141
Scale 3	8.44%	0.168	0.096	0.023	0.049	0.083
Scale 4	10.31%	0.129	0.027	0.034	0.114	0.054
Scale 5	11.24%	0.168	0.070	0.056	0.060	0.089
Scale 6	5.96%	0.051	0.006	0.055	0.090	0.074
Scale 7	9.35%	0.181	0.043	0.061	0.086	0.131

see that Scale 1 and 6 produce a lower number of detected outliers and are more efficient than the other scales used.

Next, confidence intervals for each scale and quality level were computed with and without filtering. As a result of the outlier detection, confidence intervals decreased and the reduction of the confidence intervals is described in Table 2. It can be seen that the reduction of confidence intervals is largest around the end point of the scales, hence successfully counteracting the scale usage heterogeneity issue. However, Fig. 3 depicts the scale usage for the different scales without training according to **Definition 1**. More than 30 % of the users only use 50 % of the scale or less. Thus, the filtering approach has only partially solved the scale usage heterogeneity issue on the lower part of the scale, but the problem is existing in CS confirming **H0**. Filtering is recommended for data processing in future CS campaigns to ensure high data quality, but does not solve the scale usage problem.

4.2 Analysis of Rating Scale Design

For further analysis we removed now the outliers from the dataset. The MOS results for each bitrate and each scale, after above described outlier detection and following filtering, are depicted in Figure 4. The trend of the scores for the different scales is pretty similar across the different video quality levels. The majority of the scales does not provide statistical significant different mean opinion scores. However, there are two exemptions: 1) Scale 6 provides fairly high scores that are statistically different from three other scales for the 250 kbit/s condition[2] and 2) Scale 4 provides high scores that are significantly different to three other scales for the 500 kbit/s condition[3].

Furthermore, we wanted to identify if the larger number of choices of the ACR-9 scale results in a more even distribution of ratings compared to the ACR-5 scale and thereby reduces standard deviations for the ACR-9 scale as described in [25]. Distribution of relative number of ratings per scale shows that the scale items between the scale descriptors of the ACR-9 scale are utilized significantly less then the labeled items. Also in terms of standard deviation (STD) for each quality level the difference between the scales can be neglected (ACR-5 avg. STD = 0.78, ACR-9 avg. STD = 0.83) and is not improved.

From the above presented scales one can conclude that, in terms of mean opinion scores obtained, the different scales provided no significantly different results (with two exemp-

[2]ANOVA Results: Scale 3: H = 6.531, ρ = 0.012; Scale 4: H = 8.050, ρ = 0.006; Scale 5: H = 10.856, ρ = 0.001;

[3]ANOVA Results: Scale 2: H = 6.857, ρ = 0.010; Scale 3: H = 11.618, ρ = 0.001); Scale 5: H = 15.001, ρ = 0.000;

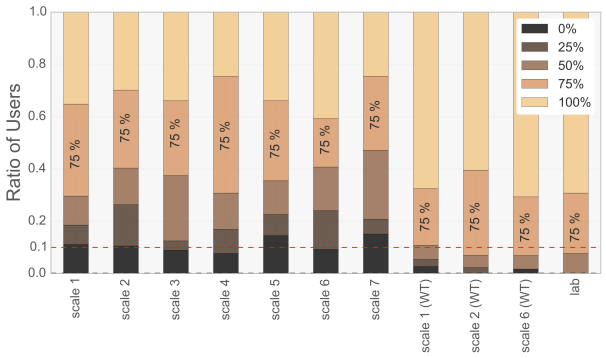

Figure 3: Ratio of users with a certain relative scale usage $^{\Delta_i}/_K$ for the test settings. The used range Δ_i of user i is defined as the difference between the maximum rating M_i and the minimum rating m_i of that user: $\Delta_i = M_i - m_i$. The absolute value is normalized by the number K of rating items. The red line indicates the threshold value $\Theta = 0.1$ according to Definition 1. If more than Θ users use less than 75 % of the rating scale, then the scale usage problem is defined to exist – which is the case for scale 1 – 7 w/o training.

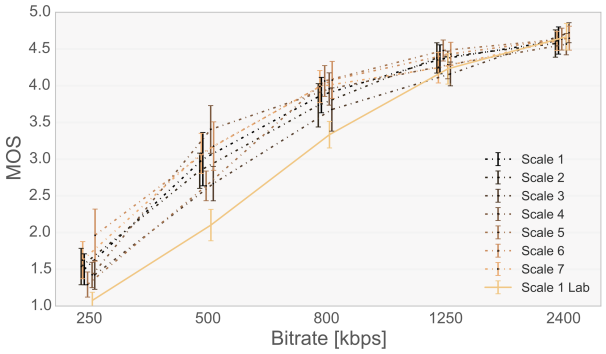

Figure 4: MOS and confidence intervals for each scale and each respective bitrate condition without training in comparison to the lab results.

tions) for identical quality levels (aside from the two exemptions at 250 kbit/s and 500kbit/s) after reliability and outlier detection. However, in terms of outliers that have to be removed there exist differences. Scale 1 and Scale 6 perform clearly better and are hence more efficient than the other scales. A positive side effect of the outlier detection across the majority of the scales is the reduction of confidence intervals towards the scale end points. We want to note that the MOS results of the 9-point scales are not statistically different across all quality levels, but their results tend to be more positive, especially for Scale 6 towards the end of the scale, which indicates the existence of the scale usage heterogeneity issue for this scale. In terms of a (theoretically) more even rating distribution across scales with more items (9-point) we could not prove that in our data. Our results showed that the subjects were mainly using the labeled scale items. From these results, we conclude that Scale 1 is the best choice upon the range of tested scale designs. It is 1) easy to understand and use 2) the lower

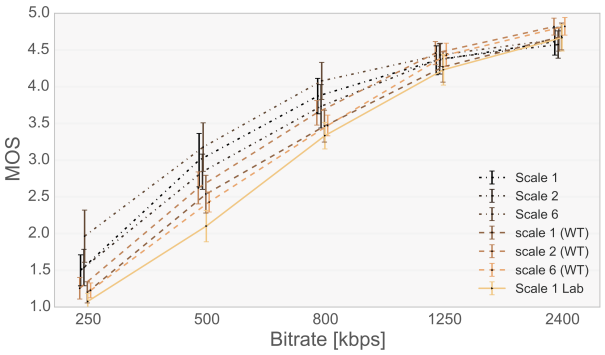

Figure 5: MOS and confidence intervals for selected scale with and without training session.

number of scale items helps also to enhance usability of the scale (cf. [21]), which is key in CS and 3) the graphical representation with the traffic light colour semaphore also contributes to enhanced usability. However, hypothesis **H1** needs to be rejected and scale design without training is not sufficient to overcome scale usage heterogeneity.

4.3 Influence of Training on Scale Usage

To test the performance of the Scale 1 we conducted laboratory assessment with 39 users. Lab environment was adjusted to match similar conditions to those faced by the users on crowdsourcing platforms[4]. The testing session contained the same content as used in the crowdsourcing study, however this time a preceding training session describing 3 videos of very bad, average and very good quality level was included. Training sequences were selected from the same sport video clip source, however the selected scene was different from the one used in the actual test. It should be also pointed out, that in the laboratory assessment, we did not have to remove any outliers. Although Scale 1 proved to be the right choice upon the range of tested scale designs, the MOS ratings from the CS campaign didn't match those collected in the laboratory assessment. Figure 4 shows consistent differences between lab and crowdsourcing results mainly for the quality levels QL 2–3. The same obersavation is valid for all tested scales.

To verify the impact of the training session on the scale usage, we selected 3 representative scale designs - Scale 1 complemented by Scale 2 as a representative for not very efficient design (high number of unreliable answers - Table 2), and Scale 6 as the representative for the ACR9 scale.

We take a look again at Figure 3 which describes the scale range usage by users. Keeping in mind **Definition 1** we can conclude that the introduction of the training phase significantly improves results from crowdsourcing campaigns. For all three tested scales we can see similar scale usage distributions as the distribution achieved in the lab study, i.e. more than 90% users use at least 75% of the scale range.

MOS and confidence intervals for selected scales with (WT) and without training session are depicted in Figure 5. Again, the training session helped to close down the gap between lab and crowdsourcing experiments - with one exception: QL 2

[4]Users used browser video player on an average laptop in a living room environment; Users' age: <30 years (19), 30-44 years (17), 45+ years (9); 19 women, 20 men

(500 kbps) shows statistically significant difference between those two environments, consistent for all selected scale designs.[5] With all the aforementioned results we conclude that hypothesis **H2** cannot be rejected. Thus, proper training is sufficient to overcome scale usage heterogeneity - independent of the scale used.

5. CONCLUSION

While crowdsourcing is a powerful tool for researchers to conduct quality assessments in an easy and fast way, its usage is also ambiguous as the effect of CS on the user ratings is not fully understood by the community yet. In particular, we investigate in this paper whether the scale usage problem exists in CS and whether it can be overcome by an appropriate scale usage design. Changing the scale design avoids additional costs to conduct the subjective study, in contrast to training sessions of participants which prolong the experiment duration.

In our CS experiments, seven different scale designs with and without training were investigated and compared to a lab study. In the CS tests without training, we clearly identified the scale usage problem in CS, although unreliable user ratings were already filtered out. However, our results show that various scale implementations do not lead to significantly different results and are not suitable to overcome the scale usage problem in CS. From a practical point of view, the 5-point ACR scale with the proposed traffic light semaphore design in conjunction with the outlier detection performs most efficient in terms of outliers – and hence costs for the experiments.

On the contrary, we could prove that training sessions are successful to enhance result quality in CS and to properly counterfeit scale usage heterogeneity problems. We conclude that training sessions, despite their additional costs, must be included in CS campaigns in order to ensure heterogeneous scale usage.

6. REFERENCES

[1] Chen, K.-T., Wu, C.-C., Chang, Y.-C., and Lei, C.-L. A crowdsourceable qoe evaluation framework for multimedia content. In *Proc. of the 17th ACM Multimedia* (New York, NY, USA, 2009), MM '09, ACM, pp. 491–500.

[2] Gardlo, B. *Quality of Experience Evaluation Methodology via Crowdsourcing*. PhD thesis, University of Zilina, 2012.

[3] Gardlo, B., Egger, S., and Seufert, M. Crowdsourcing 2.0: Enhancing Execution Speed and Reliability of Web-based QoE Testing. In *Proc. IEEE ICC, Sydney, Australia* (June 2014).

[4] Grubbs, F. E. Procedures for detecting outlying observations in samples. *Technometrics 11*, 1 (1969), 1–21.

[5] Hassenzahl, M., Diefenbach, S., and Göritz, A. Needs, affect, and interactive products - Facets of user experience . *Interacting with Computers 22*, 5 (2010), 353 – 362. Modelling user experience - An agenda for research and practice.

[6] Hossfeld, T. On Training the Crowd for Subjective Quality Studies. *VQEG eLetter 1* (Mar. 2014).

[7] Hossfeld, T., Hirth, M., Redi, J., Mazza, F., Korshunov, P., Naderi, B., Seufert, M., Gardlo, B., Egger, S., and Keimel, C. Best practices and recommendations for crowdsourced qoe-lessons learned from the qualinet task force" crowdsourcing".

[5] ANOVA results: Scale 1: (H = 6.414, ρ= 0.013), Scale 2: (H = 10.972, ρ= 0.001), Scale 6: (H = 6.986, ρ= 0.010)

[8] Hossfeld, T., and Keimel, C. Crowdsourcing in QoE Evaluation. In *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller and A. Raake, Eds. Springer: T-Labs Series in Telecommunication Services, ISBN 978-3-319-02680-0„ Mar. 2014.

[9] Hossfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., and Tran-Gia, P. Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing. *Multimedia, IEEE Transactions on 16*, 2 (Feb 2014), 541–558.

[10] Hossfeld, T., Schatz, R., and Egger, S. Sos: The mos is not enough! In *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on* (2011), IEEE, pp. 131–136.

[11] Huynh-Thu, Q., Garcia, M., Speranza, F., Corriveau, P., and Raake, A. Study of rating scales for subjective quality assessment of High-Definition video. *Broadcasting, IEEE Transactions on 57*, 1 (2011), 1–14.

[12] International Telecommunication Union. Subjective video quality assessment methods for multimedia applications. *ITU-T Recommendation P.910* (April 2008).

[13] International Telecommunication Union. Methodology for the Subjective Assessment of the Quality of Television Pictures. *ITU-R Recommendation BT.500-13* (Jan. 2012).

[14] Jones, B. L., and McManus, P. R. Graphic scaling of qualitative terms. *SMPTE journal 95*, 11 (1986), 1166–1171.

[15] Keimel, C., Habigt, J., and Diepold, K. Challenges in crowd-based video quality assessment. In *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop on* (2012), IEEE, pp. 13–18.

[16] McKelvie, S. J. Graphic rating scales: How many categories? *British Journal of Psychology 69*, 2 (1978), 185–202.

[17] Redi, J., and Povoa, I. Crowdsourcing for rating image aesthetic appeal: Better a paid or a volunteer crowd? In *2014 International ACM Workshop on Crowdsourcing for Multimedia* (2014), ACM, pp. 25–30.

[18] Redi, J. A., Hossfeld, T., Korshunov, P., Mazza, F., Povoa, I., and Keimel, C. Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal. In *2nd ACM international workshop on Crowdsourcing for multimedia* (2013), ACM, pp. 29–34.

[19] Robertson, J. Likert-type scales, statistical methods, and effect sizes. *Commun. ACM 55*, 5 (May 2012), 6–7.

[20] Rossi, P. E., Gilula, Z., and Allenby, G. M. Overcoming scale usage heterogeneity. *Journal of the American Statistical Association 96*, 453 (2001), 20–31.

[21] Seow, S. C. Information theoretic models of hci: A comparison of the hick-hyman law and fitts' law. *Hum.-Comput. Interact. 20*, 3 (Sept. 2005), 315–352.

[22] Seufert, M., Slanina, M., and Egger, S. "To Pool or not to Pool": A Comparison of Temporal Pooling Methods for HTTP Adaptive Video Streaming. In *Proc. of the QoMEX , Klagenfurt, Austria* (July 2013), IEEE.

[23] Varela, M., Mäki, T., Skorin-Kapov, L., and Hossfeld, T. Increasing payments in crowdsourcing: don't look a gift horse in the mouth. In *4th international workshop on perceptual quality of systems (PQS 2013). Vienna, Austria* (2013).

[24] Virtanen, M., Gleiss, N., and Goldstein, M. On the use of evaluative category scales in telecommunications. In *Human Factors in Telecommunications* (1995).

[25] Winkler, S. On the properties of subjective ratings in video quality experiments. In *Quality of Multimedia Experience (QoMEX 2009)* (July 2009), pp. 139–144.

[26] Wu, C.-C., Chen, K.-T., Chang, Y.-C., and Lei, C.-L. Crowdsourcing multimedia qoe evaluation: A trusted framework. *Multimedia, IEEE Transactions on 15*, 5 (Aug 2013), 1121–1137.