Crowdsourcing 2.0: enhancing execution speed and reliability of web-based QoE testing

Bruno Gardlo, Sebastian Egger, Michael Seufert, Raimund Schatz

Angaben zur Veröffentlichung / Publication details:

Gardlo, Bruno, Sebastian Egger, Michael Seufert, and Raimund Schatz. 2014. "Crowdsourcing 2.0: enhancing execution speed and reliability of web-based QoE testing." In *2014 IEEE International Conference on Communications (ICC), 10-14 June 2014, Sydney, Australia*, edited by Farzad Safaei, Leith Campbell, and Der-Jiunn Deng, 1070–75. Piscataway, NJ: IEEE. https://doi.org/10.1109/icc.2014.6883463.

Crowdsourcing 2.0: Enhancing Execution Speed and Reliability of Web-based QoE Testing

Bruno Gardlo, Sebastian Egger
Telecommunications Research Center
Vienna (FTW)
Donau-City-Straße 1
A-1220 Vienna, Austria
{gardlo, egger}@ftw.at

Michael Seufert
University of Würzburg
Institute of Computer Science
Am Hubland
D-97074 Würzburg, Germany
seufert@informatik.uni-wuerzburg.de

Raimund Schatz
Telecommunications Research Center
Vienna (FTW)
Donau-City-Straße 1
A-1220 Vienna, Austria
schatz@ftw.at

***Abstract*—Since its introduction a few years ago, the concept of 'Crowdsourcing' has been heralded as highly attractive alternative approach towards evaluating the Quality of Experience (QoE) of networked multimedia services. The main reason is that, in comparison to traditional laboratory-based subjective quality testing, crowd-based QoE assessment over the Internet promises to be not only much more cost-effective (no lab facilities required, less cost per subject) but also much faster in terms of shorter campaign setup and turnaround times.**

However, the reliability of remote test subjects and consequently, the trustworthiness of study results is still an issue that prevents the widespread adoption of crowd-based QoE testing. Various ideas for improving user rating reliability and test efficiency have been proposed, with the majority of them relying on *a posteriori* analysis of results. However, such methods introduce a major lag that significantly affects efficiency of campaign execution. In this paper we address these shortcomings by introducing *in momento* methods for crowdsourced video QoE assessment which yield improvements of results reliability by factor two and campaign execution efficiency by factor ten. The proposed in momento methods are applicable to existing crowd-based QoE testing approaches and suitable for a variety of service scenarios.

I. Introduction

Traditional quality assurance and optimization in communication networks mainly targeted technical QoS parameters in order to ensure service provisioning at sufficiently high quality levels to the end user. However, over the last decade the focus has shifted from pure QoS centred consideration to a more end-user centric focus on quality termed Quality of Experience (QoE) [1]. For network and service providers this development constitutes a new challenge: for performance improvement, traffic engineering, service management etc. they have to consider QoE as additional source of evaluation and optimization criteria. In order to make these criteria (and thus the end user perspective) applicable, validated QoE scores are required that quantify the impact of influence factors like network QoS and media encoding settings on user-perceived quality.

Such QoE scores - typically in the form of mean opinion scores (MOS) - are obtained via extensive experiments with human subjects conducted in laboratory environments, an established approach known for producing valid and reliable results. The major disadvantage of such lab-based experiments is the fact that they not only require expensive facilities and testing expertise but also incur significant expenses and relatively long campaign setup and turnaround times (typically in the order of weeks). Therefore, lab experiments are not suitable for proof of concept tests or comparisons of different prototype implementations during the development phase. One solution to overcome these constraints is subjective QoE assessment by means of crowdsourcing. In contrast to lab-based evaluation this approach uses web-hosted crowdsourcing platforms such as Amazon Mechanical Turk or Microworkers to have the different quality evaluation tasks executed by remote participants (or 'crowd-workers') via the Internet. The advantages of crowd-based QoE testing can be summarized as: (1) The evaluation tasks require no special lab facilities, (2) they well reflect the real usage scenarios and environments of the test subjects as they are carried out in their typical usage context, (3) they provide rapid results and are thereby capable of getting quick responses to changes in the service setup, and (4) they are considerably cheaper than lab tests (in terms of facilities, test assistants, and user remuneration required). Therefore, QoE crowdsourcing approaches are well suited for comparing different optimization solutions with minimal organizational effort and in a timely manner.

Nevertheless, crowd-based QoE testing also introduces new challenges such as ensuring reliability of remote participants and cheating prevention, challenges that represent major roadblocks to the establishment and widespread adoption of the method. Current crowdsourcing approaches for QoE assessment try to address these issues by introducing reliability screening questions throughout the test which are analyzed *a posteriori* (i.e. after) the crowdsourcing campaign has ended (cf. [2]). This approach leads to quite reliable ratings but due to the strict a posteriori filtering also produces a large amount of unusable ratings by unreliable participants. These unusable ratings incur unnecessary costs and considerably increase execution time of the campaign, thereby offsetting the method's advantage of rapid result acquisition.

In order to overcome these disadvantages, we introduce in this paper a novel methodology for online *in momento* reliability computation and rapid user feedback and results in a minimal number of unreliable ratings and hence reduced campaign execution times. To this end, the remainder of this paper is structured as follows: Section 2 discusses related work on crowdsourced QoE, with a focus on reliability screening mechanisms. Section 3 then discusses two QoE user studies on adaptive video streaming, one using the traditional 'a posteriori', the other using the proposed 'in momento' approach. Section 4 compares the two approaches by analyzing the different study results. Finally, Section 5 derives some conclusions as well as an future outlook from our results.

II. Related Work

A number of crowd-based QoE testing study results and methodologies have been published recently: Keimel *et al.* introduced their improved version of a framework for QoE testing - *QualityCrowd2*, which features a simple scripting language for an easy setup of online tests [3]. Chen *et al.* also discussed audiovisual QoE testing based on crowdsourcing [4]. *CrowdMOS* represents another framework for crowdsourcing studies, introduced by Ribeiro *et al.* [5]. Universal frameworks focused on "crowdtesting" are complemented by several other approaches either specialized on audiovisual quality [6], video streaming services [2] or studies focused on description of the crowd itself [7].

Besides other characteristics, all these studies deal with a common problem, namely the proper screening of the users' reliability and verification of the results. Existing recommendations used for lab-based quality assessment define screening techniques for selection of reliable users [8], however these techniques have been proven as insufficient for online tests scenarios [5], [9]. For that reason, there have been several other techniques proposed for a better selection of trustworthy users. Eickhoff and De Vries [10] designed their task in a way, that user creativity is involved and cheaters are discouraged instead of being detected. Another approach in a similar fashion has been proposed by [11] which designed their tasks such that it is easier to do the task correctly than to cheat, which also helped in gaining more reliable results. However, both of these approaches are not sufficient to reliably detect cheating users. Therefore, Hoßfeld *et al.* introduced several other techniques [2] from psychology, such as consistency questions, content questions, or repeated requests, which can further enhance the reliability of the crowdsourcing results. These approaches together with verification tests [12], [13] have proved that they are viable to increase the efficiency and overall reliability of the crowdsourcing results [14].

However, all the aforementioned works rely on an *a posteriori* results analysis, where the results are examined after the campaign (or a large part of it) is already finished. The major disadvantage of these approaches is the fact that the majority of the gathered (and already paid for) data has to be sorted out due to cheating, bad reliability of the subjects, etc. To address this issue, Gardlo [15] has proposed a major improvement in the reliability screening process by using a two stage design, where in the first stage the crowd is invited to participate in an easy to do, low paid task which serves as a first reliability test as well. Subsequently, the subset of workers which reliably performed the first stage is invited to participate in the second stage, the stage that accommodates the actual QoE assessment. The work analyzed in this paper takes this approach of two stage design, and transforms it to the single stage real-time - i.e. *in momento* - computation of user's reliability. In addition, we enhance this screening approach by providing user feedback that tells the users about their reliability during the QoE assessment. Furthermore, we show that such feedback in conjunction with *in momento* reliability computation can significantly improve the whole crowd-based testing process.

III. Studies, User involvement

The two studies discussed within this paper were both targeted towards the quality evaluation for adaptive video streaming. The main issue with quality evaluation for this service is the large number of possible combinations of video adaptation along the temporal, spatial and image compression dimension with different quality (or resulting bitrate) levels for each dimension. If one wants to additionally consider different content classes it gets clear that this results in a huge number of test conditions to be evaluated throughout one study. In our case this resulted in 85 different video quality adaptation profiles which we wanted to test with three different content classes. Our goal was a yield of 15 ratings per profile and content class which therefore sums up to $N_{Goal} = 85 * 3 * 15 = 3825$ ratings that would be needed to satisfy our goal. Such a number of ratings can obviously not be achieved in a traditional lab setting as this would either need a pretty large number of test users (and therefore high costs) or would face severe fatigue effects of the subjects in case of exhaustive test sessions. In such cases crowdsourcing can serve as an ideal substitute for lab tests as it allows to recruit a large user sample and therefore allows to execute such exhaustive studies in reasonable time and under reasonable economic constraints (cf. [16], [17], [2], [3]).

Therefore, we aimed at executing the study as a traditional crowdsourcing study (Study A) as described in [6], [2], [17] with reliability computation after the campaign has ended and every participant got paid. Within the remainder of this paper we term such studies *a posteriori* reliability computation studies. For such studies the reliability is typically around 33% of the issued ratings. Considering the above postulated $N_{Goal} = 3825$ of reliable ratings this would equal a total number of $N_{total} = 11475$ ratings to be gathered.

In contrast to this *a posteriori* approach we also enhanced the approach of [15] such that it allows for an *in momento* computation of reliability, while the user is still online and proceeding with the assessment (Study B). By doing so we are able to identify reliable participants and allow them to further issue a certain number of ratings. When the online computed reliability score for a subject drops under a certain threshold, the subject is not allowed to further proceed and gets only paid for the issued ratings to this point in time.

Table I gives an overview of the two studies conducted with Study A serving as an example of *a posteriori* reliability computation and Study B serving as example of *in momento* reliability computation. In both studies the participants were asked to rate the video quality of given sequences on an absolute category rating scale (ACR-5). In both studies we used the same technology for video presentation. Users watched the content after the video was fully loaded, hence the QoS related parameters affecting the end-user's perception (jitter, packet loss, network delay, etc.) were eliminated.

	Study A	Study B
Number of ratings	10737	1593
Number of reliable ratings	3483	1377
Reliability [%]	32%	86%
Campaign Duration	6 months	25 days
Payment / Rating	$ 0.2625	$ 0.0834
Reliability Computation	a posteriori	in momento

TABLE I: Overview of the two different crowdsourcing studies and related parameters

A. *Study A: A posteriori approach*

For the execution of Study A an online test framework similar to [2] was implemented, such that each participant watched three different videos and rated the quality afterwards. Each video had a duration of 20s and contained a single one-dimensional video quality adaptation, i.e., after 10s either frame rate, or resolution, or quantization parameter were changed. Additionally conditions with one stalling event, with a change of player size, and reference conditions were tested. Three types of content (action, cartoon, and sport) were used such that in total 85*3=255 conditions had to be evaluated.

The study was available as a micro job on the crowdsourcing platform Microworkers [18]. Every user could participate once and was rewarded with 0.20$ upon completion of the test. As we observed many unreliable participants and the execution time of the campaign was lengthening, we decided to offer the task to specialized worker groups which were offered higher loans (0.30$, later up to 0.50$). First, the task was given to a group consisting of top performers and top earners of Microworkers, i.e., users who have the highest successful job completion rate or have earned the most money. Later, we listed reliable workers on a whitelist by ourselves, and repeatedly hired this whitelist group for participating in the study.

During the six month test phase the test was completed 3579 times. 1161 tests were conducted reliably, such that 3483 ratings of test conditions could be evaluated. To distinguish reliable users and users who did not conduct the test properly, several consistency checks were included. To be considered reliable, a user first had to read the test instructions which also explained a game-like monitor quality pre-test. The clicking behavior during that pre-test immediately revealed whether he read the instructions or not. Then, the user had to watch all videos in their full lengths and answer the content questions correctly. Finally, when answering different questions about a given video, he had to rate the quality consistently, e.g., a user who did not notice any quality change, but also rated that the same video contained a severe quality degradation, was considered inconsistent.

B. *Study B: In momento approach*

The goal of the second study was to verify the performance of an *in momento* reliability analysis approach and how this affects result reliability and execution time. Changes introduced in the *in momento* approach were reflected on several levels of the application design. In particular we aimed at the following goals, which are influencing the overall experience with the test utility for the participants:

- Use implicit reliability measures and mind cumbersome questions that complicate the task completion.
- Cut the overall test time to minimum.
- Camouflage video quality test with screen quality assessment.

To keep the time from entering the application to completing the test as short as possible, we integrated social login possibilities with Facebook, Google+ and Twitter accounts. The advantage of such social networks integration was twofold. First, it helped us to keep track about the user's involvement and his testing history, as the user was uniquely authenticated. Second, with the user's permission to access his profile, we were able to gather demographic data, which could later be used for reliability verification. All this was possible without any requirement to fill questionnaires, and hence cutting the overall testing time.

In order to keep the whole user interface easy to use, clean from any distractions, and straightforward even for less experienced users, we moved all verification methods to the background. Hence the user interface remained simple and users were only required to perform basic actions related to the actual task. In practice this means, that we skipped the majority of the verification questions in favour of interface simplicity. The advantage of these adjustments is that users are neither distracted, nor tired, nor overhauled with too many required interactions.

For the reliability assessment we utilized the two stage reliability framework introduced in [15]. This framework avoids questions targeted towards the content of shown videos as well as repetitive questions for cheating detection. In contrast, for each stage of the test we defined suspicious behaviors and monitored them. If suspicious behavior was detected, the user was assigned reliability penalty points which were used to compute the overall reliability of the respective user.

For the first stage screen quality tests, patterns used for calibration of professional screens in graphic studios [19] were adapted in a way that they reflect the screen setup and watching conditions at the end-user's environment. Pattern (A) displayed equal shapes with different contrast and users had to select all visible shapes, whereas pattern (B) utilized numbers from one to seven in different contrasts which had to be detected by the users. Each questionnaire corresponding to pattern (A) or (B) was designed in a way to better engage user's attention and also contained false answers for better filtering of unreliable users (e.g., if the user answered for pattern (B) option outside of $< 1; 7 >$, he perceived penalty points). Additionally, we used random movement of the testing shapes and numbers to better reflect the dynamic video characteristics and to prevent cheating by sharing of correct locations between participants. We analyzed the clicking behavior of the users, the number of correctly detected shapes, and also the time spent on the page (e.g., a focus time of less than 6sec was considered as a reliability penalty point as the task was not achievable in this short time). For pattern (B) the users had to answer several questions related to the displayed numbers. Example of the first stage screen quality tests with full details about cheat detection is available on GitHub and can be forked [20].

In the second stage related to video quality, we were only focusing on user behavior with respect to the video player, e.g., for each watched video we monitored the playback time, focus

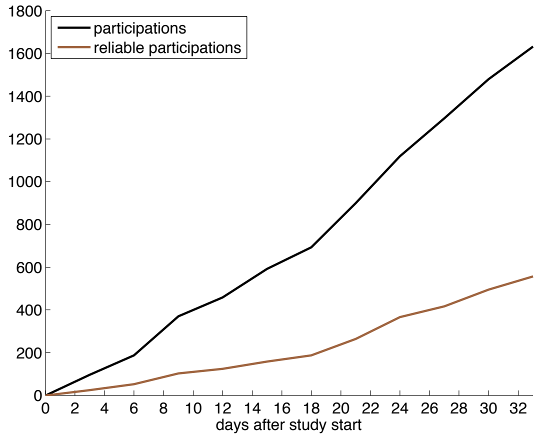

Fig. 1: Study A (a posteriori) participation over time (days).

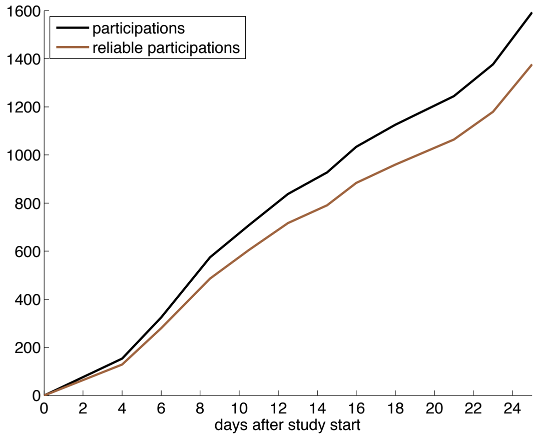

Fig. 2: Study B (in momento) participation over time (days).

time of the web browser, toggling of fullscreen mode, pausing of the video player, and other playback related parameters. We assigned reliability penalty points to the respective user if his behavior was suspicious, e.g., if the video playback was not finished properly.

As some of these aforementioned deviating behaviors or answers directly indicate cheating, and some of them might indicate misinterpretation or unintentional poor performance, we weighted the assigned points accordingly. Suspicious behavior directly indicating cheating were up to 3 penalty points, whereas minor user's mistakes received half point penalty only. After summing up all gathered penalty points we used a hyperbolic tangens function[1] to map the penalty points into reliability percentages and defined a threshold which allowed some minor mistakes, but excluded all users for which the probability of cheating was too high[2]. In terms of the test execution each user was offered to take part in the basic scenario consisting of the screen quality test and an evaluation of one video and was rewarded with 0.10$. If the user's results were reliable, we offered him to rate three more videos for additional financial reward of 0.20$.

IV. Results

The following section describes results from the two crowdsourcing methodologies which were presented above. First, the efficiency of both campaigns and the different approaches within are described. Then, the reliability of the different approaches is compared. Finally, we show how crowd exhaustion effects can be mitigated by the in momento approach, and compare the obtained quality results.

A. *Campaign Efficiency*

To cope with the high number of required ratings, Study A was split into nine identical campaigns to simplify the handling of the campaign(s) on the Microworkers platform. All campaigns utilised the same video quality evaluation task, the only difference were just the users' groups the campaigns were targeted at. Four subsequent campaigns were opened as a basic campaign, where every user was able to participate (Basic). One campaign was only visible to members of Microworkers' top performers and top earners group (Top). Additionally four campaigns were available for a special "whitelist" group (Whitelist). This was a manually created group consisting of selected reliable users, i.e., users who so far always conducted the test properly. Those users who became unreliable were removed from the Whitelist before each start of a new campaign (see Figure 4).

In total 2669 participations resulted from Basic campaigns, 403 participations from the Top campaign, and 497 users from Whitelist campaigns. When all available tasks in a campaign were finished, a new campaign was started. To increase the speed and the efficiency of the study, we also started parallel campaigns: at the beginning with the top performers/top earners group and later with the whitelist group. After 182 days the study was closed. Figure 1 unveils the number of participations over the time. It can be seen that the number of participations as well as the number of reliable participations increase almost linearly. However, the number of reliable participations has a much lower gradient. Due to the linear behavior the extension of a campaign over time does not help to increase its efficiency. Thus, to obtain a higher efficiency, possibilities for increasing reliability have to be identified.

In Figure 2 the corresponding participation plot for Study B is shown. In this study, 1593 ratings were obtained in the basic campaign within 25 days. Again, the linear increases can be seen for both number of participations and number of reliable participations. However, the difference of both gradients is much lower than for Study A which proofs that our in momento approach successfully increased the efficiency of the crowdsourcing campaign.

B. *Comparison of Reliability*

Figure 3 illustrates in detail the reliability for Basic, Top, and Whitelist workers' group of Study A and the in momento campaign of Study B. For Study A, both basic campaigns for all users and top performers/top earners campaign have

[1] A hyperpolic tangens function allows a threshold definition, where reliability significantly drops after reaching certain penalty points.

[2] In our specific scenario we allowed user to get 1 or 2 penalty points out of 22 total points. After receiving more than 2 points, the reliability significantly drops under 91%.

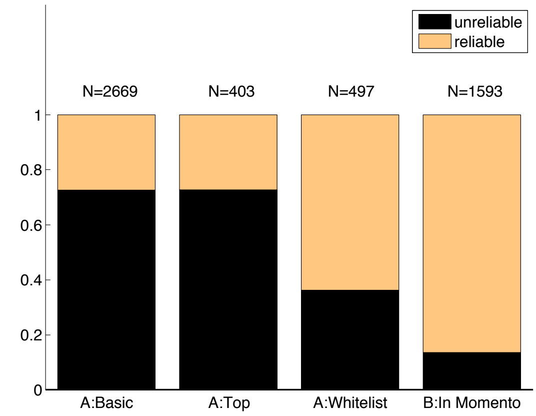

Fig. 3: Share of reliable subjects in tested groups. Basic campaign, top performers/earners campaign, whitelist group (all Study A) vs. in momento campaign (Study B).

a low ratio of reliable participation, which is only slightly above 28%. This implies that Microworkers' recommendation of top users did not provide any increase of reliable users in the study. On the other hand, the whitelist group has a 64% share of reliable participation, showing that our manual effort of filtering reliable users halved the share of unreliable participation. An even better ratio of reliable participation was achieved in Study B with the in momento approach. Study B shows a substantial increase of users' reliability. We were able to collect 1377 reliable ratings out of 1593 participations. This results in a ratio of 86% of reliable participation, which represents an improvement of 58% compared to the reliability achieved overall in Study A.

To put it in a nutshell, we can compare the four approaches by their efficiency with respect to speed and reliability. It can be seen that Basic and Top campaigns are not very reliable. A Whitelist campaign instead is reliable but still very slow. Thus, the best efficiency can be reached by the in momento approach which is reliable and fast.

C. Crowd Exhaustion

Figure 4 shows the performance of the whitelist group in the four subsequent Whitelist campaigns. It is shown that new reliable users were added to the group before each campaign, and unreliable users were removed afterwards. It can be seen that the overall number of participations increases much faster as compared to the number of reliable users listed in our whitelist group. This can be attributed to the still relatively constant ratio of unreliable participation, ranging from 33-39%. This is very surprising, considering the fact, that users who were added to the whitelist group managed to conduct the test properly at least once. Thus, one would assume that these users had understood the test correctly and a repetition should result in another reliable participation. However, this was not the case, instead the results indicate exhaustion of the crowd.

To cope with the problem of crowd exhaustion as witnessed in Study A, we redesigned the application as described in Section III-B. The idea for the *in momento* verification of the reliability, is that some of the users may want to increase their earnings and perform additional tasks, once they had already entered the application. Moreover, users who intentionally only came for one short task assignment, should not be overstressed with other required actions, to keep the involvement high and task cancellation rate very low.

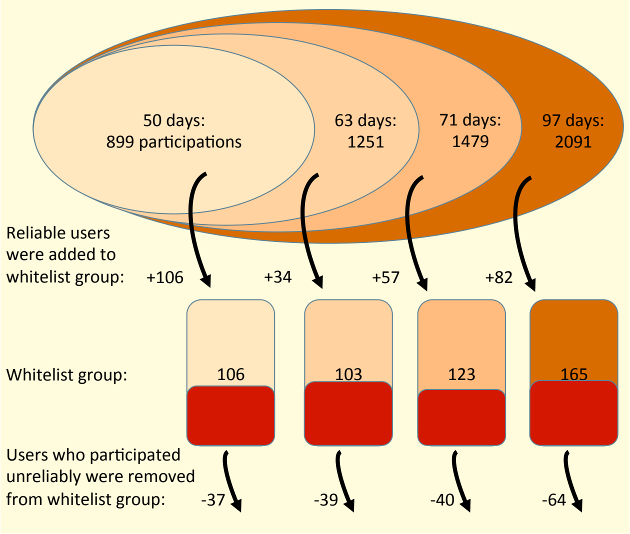

Fig. 4: Whitelist group performance during different iterations.

With all this in mind, we switched from presenting three videos to only presenting one, in order to keep the user's attention. Together with the other changes described in Section III-B, we reduced the overall test session time. The overall testing time in Study A was close to 7 minutes, whereas in Study B the user was able to complete the task in 90 seconds. However, for those reliable users who were interested into earning more money, we offered to continue with an evaluation of three additional videos.

In Study B, in total 673 different users participated, thereof 576 reliable users. 286 of them decided to rate more videos and were paid in the "extra campaign". This represents about 1000 perfect reliable ratings, resulting in an additional increase of reliability. In traditional crowdsourcing campaigns, users are often asked to rate up to 20 video sequences. However, if the user does not want to continue in the assessment after watching a single video file, if he has no option to leave the assessment, his results will be most probably unreliable and useless. Very good perspective about the user's intention represent the fact, that 50% users were not interested in continuing with the campaign, and they decided to get only basic reward of 10$. On the other hand, additional 77 ratings were collected as "volunteer ratings", if users decided to repeat the test, as they were interested to it.

It can bee seen, that the implemented changes had a major impact on the speed of the campaign, and are positively reflected in the users' involvement and reliability. Thus, the exhaustion effects observed in Study A could be mitigated. We found that users are willing to reliably participate in shorter tasks, and intrinsic motivation of the task can further attract additional users.

D. Quality Ratings Results Comparison

For the comparison of the two studies, 12 chosen test conditions, used in both studies, were selected. Thus, the results derived from the video quality ratings which were obtained with the in momento approach could be compared to results from a traditional a posteriori approach. This comparison showed, that the MOS scores for both studies were not statistically significant different. This proves that the in momento approach is comparable to the a posteriori approach in terms of video quality rating results gathered. From this result on can further conclude that switching from three video evaluations per user to only one did not have any negative effect on the reliability of the gained results.

V. Conclusions

In this paper we compared and contrasted different approaches towards the design of applications for QoE crowd-testing. By introducing the concept of *in momento* verification of test participants' reliability, we demonstrated that it is possible to significantly increase the performance of the crowd with careful incentive design, even without repetitive hiring of specialized user groups. Although video quality evaluation is the focus of this paper, the *in momento* approach can also be extended to evaluations of other media files, such as pictures or audio files, as the same reliability principles can be applied for these media experiences as well.

We also found that a repetitive involvement of users improves overall performance only to a certain extent and can cause an exhaustion of the crowd at the risk of declining motivation and poor rating performance. Our proposed *in momento* approach successfully addresses this problem, leading to a doubling of results reliability and a reduction of overall study completion time by a factor of ten. Additionally, it reduces the administrative overhead introduced by traditional *a posteriori* approaches that require extensive data cleaning and group generation with repeated campaign runs. In addition, the proposed rapid feedback component allows for better communication with test participants since any suspicious behavior can be directly communicated to them. It enables users to reflect on their performance and to choose whether to stop or to continue the testing process. Based on these encouraging results we believe that the proposed methods represent an essential step towards making crowd-testing sufficiently mature for widespread adoption by researchers and practitioners alike.

References

[1] P. L. Callet, S. Möller, and A. Perkis (eds), "Qualinet white paper on definitions of quality of experience (2012)," Lausanne, Switzerland, Jun. 2012.

[2] T. Hossfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of youtube qoe via crowdsourcing," in *Multimedia (ISM), 2011 IEEE International Symposium on*, dec. 2011, pp. 494 –499.

[3] C. Keimel, J. Habigt, and K. Diepold, "Challenges in crowd-based video quality assessment," in *Forth International Workshop on Quality of Multimedia Experience (QoMEX 2012)*, Yarra Valey, Australia, Jul. 2012.

[4] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, "A crowdsourceable QoE evaluation framework for multimedia content," in *Proceedings of the 17th ACM international conference on Multimedia*, ser. MM '09. ACM, 2009, pp. 491–500.

[5] F. Ribeiro, D. Florencio, C. Zhang, and M. Seltzer, "Crowdmos: An approach for crowdsourcing mean opinion score studies," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 2416–2419.

[6] B. Gardlo, M. Ries, T. Hoßfeld, and R. Schatz, "Microworkers vs. Facebook: The Impact of Crowdsourcing Platform Choice on Experimental Results," in *QoMEX 2012*, Yarra Valley, Australia, Jul. 2012.

[7] W. Mason and S. Suri, "Conducting behavioral research on Amazon's Mechanical Turk," *Behavior Research Methods*, vol. 44, no. 1, pp. 1–23, Jun. 2011. [Online]. Available: http://dx.doi.org/10.3758/s13428-011-0124-6

[8] I. R. Assembly, "ITU-R BT.500-12 Methodology for the subjective assessment of the quality of television pictures ," 2009.

[9] S.-H. Kim, H. Yun, and J. S. Yi, "How to Filter out Random Clickers in a Crowdsourcing-Based Study? (Research Paper) ," in *BELIV 2012 Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, eattle, WA, USA, Oct. 2012.

[10] C. Eickhoff and A. de Vries, "Increasing cheat robustness of crowdsourcing tasks," *Information Retrieval*, pp. 1–17, 2012.

[11] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with mechanical turk," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '08. New York, NY, USA: ACM, 2008, pp. 453–456. [Online]. Available: http://doi.acm.org/10.1145/1357054.1357127

[12] J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor, "Are your participants gaming the system?: screening mechanical turk workers," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '10. New York, NY, USA: ACM, 2010, pp. 2399–2402. [Online]. Available: http://doi.acm.org/10.1145/1753326.1753688

[13] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," *SIGIR Forum*, vol. 42, no. 2, pp. 9–15, Nov. 2008. [Online]. Available: http://doi.acm.org/10.1145/1480506.1480508

[14] B. Gardlo, M. Ries, and T. Hoßfeld, "Impact of Screening Technique on Crowdsourcing QoE Assessments," in *22nd International Conference Radioelektronika 2012, Special Session on Quality in multimedia systems*, Brno, Czech Republic, Apr. 2012.

[15] B. Gardlo, "Quality of experience evaluation methodology via crowdsourcing," Doctoral Dissertation, Dept. of Telecommunications and Multimedia, University of Zilina, Zilina, Slovakia, September 2012.

[16] M. Hirth, T. Hoßfeld, and P. Tran-Gia, "Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms," *Mathematical and Computer Modelling*, 2012.

[17] ——, "Anatomy of a Crowdsourcing Platform - Using the Example of Microworkers.com," in *Workshop on Future Internet and Next Generation Networks (FINGNet)*, Seoul, Korea, Jun. 2011.

[18] Microworkers. (2013, Feb.). [Online]. Available: http://microworkers.com

[19] Lagom.nl. (2012, April) LCD monitor test images. [Online]. Available: http://www.lagom.nl/lcd-test/

[20] B. Gardlo. (2014, January) Screen quality measurement application. [Online]. Available: https://github.com/St1c/screentest