Improving the Efficiency of QoE Crowdtesting

Ricky K. P. Mok
CAIDA
University of Califorina, San Diego
United States
cskpmok@caida.org

Ginga Kawaguti
NTT Network Technology Laboratories
NTT Corporation
Japan
ginga.kawaguti.nr@hco.ntt.co.jp

Jun Okamoto
NTT Network Technology Laboratories
NTT Corporation
Japan
jun.okamoto.nw@hco.ntt.co.jp

ABSTRACT

Crowdsourced testing is an increasingly popular way to study the quality of experience (QoE) of applications, such as video streaming and web. The diverse nature of the crowd provides a more realistic assessment environment than laboratory-based assessments allow. Because of the short life-span of crowdsourcing tasks, each subject spends a significant fraction of the experiment time just learning how it works. We propose a novel experiment design to conduct a longitudinal crowdsourcing study aimed at improving the efficiency of crowdsourced QoE assessments. On Amazon Mechanical Turk, we found that our design was 20% more cost-effective than crowdsourcing multiple one-off short experiments. Our results showed that subjects had a high level of revisit intent and continuously participated in our experiments. We replicated the video streaming QoE assessments in a traditional laboratory setting. Our study showed similar trends in the relationship between video bitrate and QoE, which confirm findings in prior research.

CCS CONCEPTS

• **Networks** → *Network experimentation*; • **Human-centered computing** → **Field studies**; • **Information systems** → *Multimedia streaming*.

KEYWORDS

QoE assessment; crowdsourcing; network measurements

ACM Reference Format:
Ricky K. P. Mok, Ginga Kawaguti, and Jun Okamoto. 2020. Improving the Efficiency of QoE Crowdtesting. In *1st Workshop on Quality of Experience (QoE) in Visual Multimedia Applications (QoEVMA'20), October 16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3423328.3423499

1 INTRODUCTION

Laboratory-based subjective assessment is the *de facto* standard for assessing quality of experience (QoE). Constrained by cost and time, the scale of assessments is often small. Crowdsourcing-based QoE experiments (QoE crowdtesting [18]) have become popular due to the potential for a large and diverse population of subjects. Researchers publish experiments to public crowdsourcing platforms, such as Amazon Mechanical Turk [4], Clickworker [1], and Microworkers [3]. Users on these platforms participate in experiments remotely over the Internet in exchange for monetary payments.

In traditional QoE crowdtesting design, every subject rates the same number of stimuli and receives a fixed payment. As the length of QoE crowdtesting is often much shorter than laboratory experiments (less than 30 minutes [10, 18] *vs.* more than one hour [31]), a significant portion of the experiment time is spent on delivering instructions and training the subjects to operate the assessment interface. Therefore, subjects can only assess a few stimuli. For video streaming QoE experiments, each assessment can be up to 3 minutes as subjects have to watch the entire videos. Thus, the efficiency of the crowdsourcing campaign can be low. As the number of samples collected from each subject is small, it is hard to examine the intra-rater reliability and mitigate variances introduced by different assessment environments between subjects.

We propose a novel experiment framework to improve the efficiency of crowdsourcing measurements. This framework is generic with respect to the nature of the tasks, but it is most effective for long tasks, particularly QoE assessments. This framework differs from traditional crowdsourcing in that we introduce an *extended study* after the subjects complete the initial one. More specifically, the subjects can revisit the experiment platform to perform more tasks over a period of time. This design has three major advantages. First, the extended study increases the overall efficiency of the experiment campaign, because we do not need to repeat instructions again to returning participants. The second advantage is that, by collecting more ratings from the same subjects, we reduce variances induced by environmental factors across subjects, which leads to more reliable results. Third, the experiment campaign looks more attractive to subjects, as it can lead to a relatively bigger reward than other one-off crowdsourcing tasks.

Apart from the experimental design, an important element of this framework is the subject's engagement with the experiment. The success of this framework relies on trained subjects to revisit the experiment platform to reduce overheads. To this end, we apply gamification, which is defined as *the use of game design elements in non-game contexts* [9], to improve user engagement and experience. Prior research showed that the use of gamification in crowdsourcing tasks can increase the intrinsic motivation of subjects, which can result in higher performance [12] and user activity [16].

QoEVMA'20, October 16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-8158-1/20/10…$15.00
https://doi.org/10.1145/3423328.3423499

We implemented a gamified web-based experiment platform, QUINCE[1] [26], to conduct four types of measurement tasks, including video streaming QoE assessment and network performance measurement. QUINCE employs game elements for three major purposes: an interactive tutorial and interface for providing training to difficult tasks; a user profile system for enhancing user experiences; scores, levels, and badges, to measure subjects progress and provide incentives to participate. We leverage the scoring system to quantify task completion and compute rewards. We group different tasks and employ a cool-down time to throttle the submission rate of tasks according to their difficulty and expected completion time.

We used QUINCE to run two experimental studies on Amazon Mechanical Turk (MTurk) with slightly different parameters to evaluate our framework. We had five major findings:

- More than 70% of subjects enrolled in the extended study. These subjects were highly engaged with the experiment. Half of the them achieved at least 5.5 times more than the minimum required score.
- We received task submissions throughout the entire time period of experiment campaigns. User activities showed a strong diurnal pattern, peaked between 7pm and 11pm. Each subject completed on average 10 tasks per hour.
- Over 98% of enrolled subjects returned to our platform at least once within 24 hours reflecting high intent to revisit.
- We found that applying different cool-down times to different groups of tasks effectively moderated submission rates.
- Our experiment design reduced the cost per QoE rating to USD $0.2-0.32, which was 20-67% lower than when procuring multiple one-off crowdsourced tests.

We further showed that our experiment framework did not degrade the quality of QoE assessments. We used a non-gamified version of QUINCE to perform video streaming QoE assessments in laboratory settings. We found high correlation between assessment results and the ratings collected from the crowdsourcing studies. Our results on the correlation between video bitrates and QoE aligned with prior studies, further validating the reliability of our approach.

We highlight related work in §2. We describe our proposed framework and implementation in §3. We present our experimental results in §4. We conclude the paper in §5, and discuss limitations in §6.

2 RELATED WORK

Researchers have proposed a number of web platforms to facilitate the deployment of QoE crowdtesting. QualityCrowd [19, 20] was a simple web platform for assessing picture or video quality on MTurk. It leveraged Content Delivery Networks (CDNs) to distribute test materials. Rainer et al. [29] proposed a PHP-based platform for a similar purpose. Their platform was configurable and flexible to different assessment methodologies. QoECenter [39] analyzed characteristics of the video source and simulated different video streaming performance scenarios by providing network quality of service parameters. Its front-end could stream videos and collect QoE ratings. A recent work, TheFragebogen [15], facilitated the design of crowdsourcing experiments by simplifying the generation of browser-based questionnaires by providing a number of build-in templates and the capability to capture user behavioral data for screening out potential spammers. These platforms typically employed the traditional *one-off* experiment design, without considering repeated participation.

Gardlo et al. [14] proposed a two-stage approach, called *inmomento*, to improve the reliability of QoE crowdtesting. Their method whitelisted subjects who passed the screening stage, and then invited these subjects to accept video assessment tasks in the second stage. In this case, participation of whitelisted subjects was limited by the number of studies published by the experimenter to the crowdsourcing platforms. In contrast, our proposed design only requires us to set up one crowdsourcing task. Subjects can then participate freely within a period of time.

Instead of using a traditional 5-point Likert scale, paired comparison frameworks showed better reliability in assessing the QoE of video streaming [7, 36] and the web [34]. These frameworks are highly specific to the type of stimulus. Our generic experiment design can adapt these methods to perform subjective assessments.

3 EXPERIMENT DESIGN AND IMPLEMENTATION

Our experiment design comprises two parts—the *initial study* and the *extended study*. Fig. 1 shows the work flow of our experimental design. We offered our experimental study as a regular crowdsourcing task on public crowdsourcing platforms, such as MTurk. The study first presents a task description with information about the nature of tasks, the minimum requirement for the initial study, and details on participating and computing rewards (Step (1)). The subjects then follow a hyperlink at the end of the task description to access QUINCE and start the initial study (Step (2)).

The initial study (Step (3a)–(3c)) is similar to traditional QoE crowdtesting approaches (e.g., [19]). QUINCE first obtains informed consent from subjects and then presents detailed instructions on measurement tasks, followed by a short interactive practice session. When a subject's work (score) meets the minimum requirement, the platform offers the option of enrolling in the extended study (Step 4(a)), and instructs the subject to submit a unique experiment identifier to the crowdsourcing platform (Step (4b)). The enrolled subjects can directly access the platform, and perform any available experiment tasks without repeating the onboarding process (Step (5)). QUINCE automatically generates new measurement tasks for subjects.

After we run the extended study for T days (Step (6)), we close the platform and verify the subject's work (Step (7)). The experimenter uses the identifier submitted by the subject to link between MTurk and the experiment platform. We decide to accept/reject a task submission based on the amount of work submitted by each subject, quantified by subject's score (§4). We pay a fixed amount of reward for the initial study (Step (8a)–(8b)). For the extended study, we pay an additional reward based on the score the subject obtained (Step (8c)). We leverage the bonus payment function in MTurk to pay rewards to subjects.

[1] QUINCE platform is accessible at https://crowdtrace.caida.org.

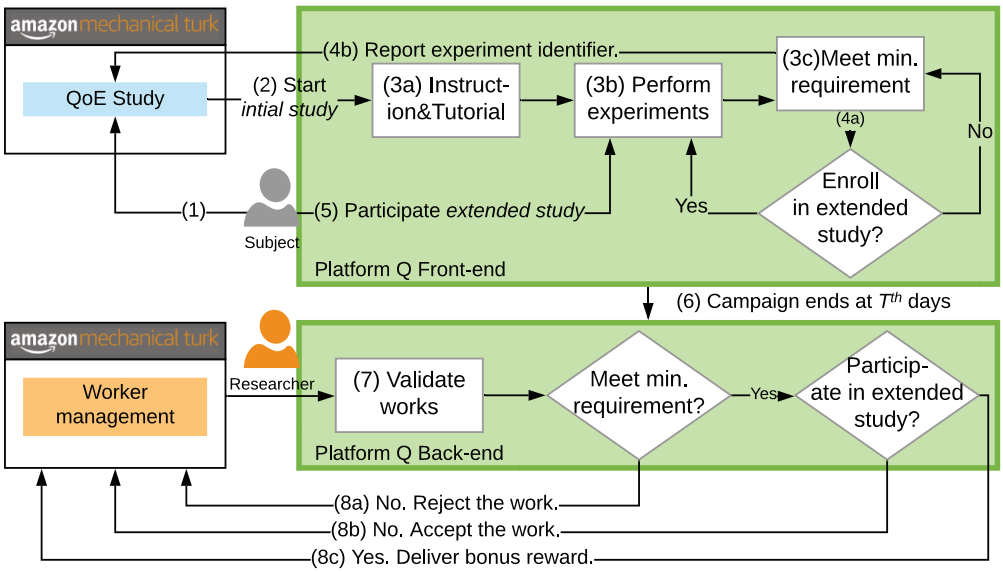


Figure 1: The overall work flow of QUINCE.

3.1 QUINCE- A gamified implementation

We implemented a web platform, named QUINCE, using the Meteor Javascript framework [2] and MongoDB for reactive web design and data storage. We also used the amCharts JavaScript library to render visualization, such as maps and charts. We deployed a video streaming server to support adaptive streaming using Apple's HLS standard.

We employed gamification techniques to increase the motivation of subjects [40], and thus improve the efficiency and accuracy in crowdsourcing [12, 21] and laboratory [35] experiments. We implemented four gamification elements in QUINCE.

(1) *Story/Theme.* Providing a theme enables subjects to experience a vicarious setting [13]. In QUINCE, we presented a simple story in game play: the subject can be a hero who can improve Internet performance. We embedded our experiment tasks into a map-based interface and the 'Missions' tab, and instructed subjects to discover experiment tasks. Subjects could also select an avatar to represent themselves.
(2) *Scores/Points.* A scoring system is one of the most commonly used gamification element in crowdsourcing experiments (e.g., [11]) for boosting motivation and performance of subjects [24]. In our platform, we employed scores to provide feedback on progress and quantify the accomplished work. More importantly, the score earned was proportional to the monetary payment to the subject.
(3) *Levels.* We introduced a "level" system to visualize personal achievements [21] and provide subjects with clear goals and milestones. Subjects could gain "experiences" as they performed more tasks. We designed 10 levels that subjects could achieve in the experiments.
(4) *Badges.* A badge system can increase user activities [16]. In the latest version of QUINCE, we developed three badges that subjects could earn when they 1) proceeded to the next level, 2) completed any task groups 5 times, or 3) visited the platform for 5 days.

Fig. 2 shows the user interface and part of the gamification elements of QUINCE.

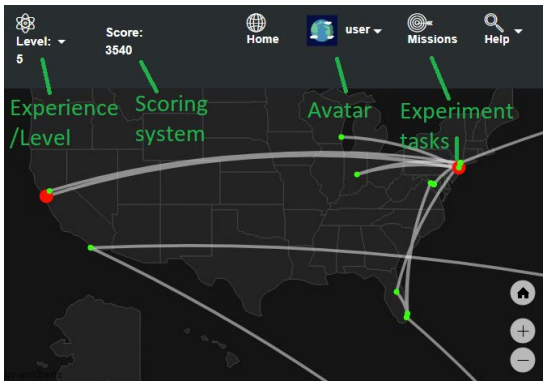


Figure 2: User interface of QUINCE. Gamification features identified in a user of QUINCE.

3.2 Experiment tasks

Unlike existing platforms (e.g., Eyeorg [34], QualityCrowd [19]) that only focus on a few specific experiments, QUINCE can incorporate any browser-based experiment into its interface. We implemented four different, but related types of measurement tasks to study the relationship between QoE of video streaming and topology and performance characteristics of Internet infrastructure.

T1) Network topology measurement. We instructed the subjects to execute the system's built-in traceroute command to measure paths from their computer to IP destinations. Our platform determined the IP destinations based on data from

other network measurement platforms or the hostnames we extracted in previous executions of Task T3.

T2) Network performance measurement. We used web-based speed tests to measure network throughput between a subject's computer and speed test servers across the Internet. We incorporated two tests into QUINCE–M-Lab Network Diagnostic Tool (NDT) [22] (downlink and uplink throughput), and a customized version of fast.com (downlink throughput).

T3) File download. We asked subjects to download dedicated web pages, so we could extract hostnames of CDN caches from the source code of these pages for use as target destinations in subsequent executions of Task T1.

T4) Video streaming QoE assessment. We streamed a short (60-90s) video clip from our own web server using HTTP Adaptive Streaming (HAS) or large-scale video service providers (YouTube and Vimeo). Upon completion of the video playback, we asked subjects to rate their QoE using an Absolute Category Rating (ACR) method (1:Bad–5:Excellent) [27]. We used a customized JavaScript-based video player to insert different impairments, such as rebuffering and switching video quality, to simulate different streaming performance conditions.

Tasks T1-3 collected Internet topology and performance data that help diagnose QoE degradation we observed in QoE assessments (T4), particularly for video streamed directly from providers. The complexity and duration of these tasks varies. For example, the video streaming tasks require at least 1 minute to play the entire video, while subjects can complete the file download task within 10 seconds. Aggressive subjects may choose to perform many easy tasks within a short time. In addition, the long durations of QoE assessments can fatigue subjects and lower the reliability of their assessments [27]. We designed three approaches to regulate task completion rate.

(1) We grouped tasks into *task groups*. Subjects must complete all tasks in a group before they receive more tasks. Table 1 summarizes the composition of tasks in each task group.
(2) We introduced a cool-down period for each task group. Subjects must wait for this period of time before the platform generates new tasks for them. We set the cool-down period according to the nature and requirement of the measurements. For example, we expected the CDN information captured from T3 to be relatively stable and not require high-frequency measurements. Therefore, we assigned a longer cool-down time to those task groups.
(3) We weighted scores that subjects could earn from a task group according to the completion time for a normal subject. The last two columns of Table 1 show the scores that we used in our two experiment studies. We assigned a higher score to task groups containing the video streaming QoE task (T4), due to its longer test duration.

4 EVALUATION

We ran two IRB-approved studies on MTurk in July 2019 (Study A) and December 2019 (Study B). In both studies, we selected workers from the United States with a historical task acceptance rate higher than 95%. Our task requires subjects to earn at least 500 points (*i.e.*, the minimum requirement) in order to receive an initial reward of USD $2. We set the extended study to 7 days (T=7). Upon the completion of the 7-day study, we computed the additional bonus for each subject according to Eqn (1).

$$\text{Bonus (in USD)} = \min\left(\frac{\text{Total score} - 500}{1000}, 48\right) \quad (1)$$

We subtract 500 points from the total score, because that portion of work was paid as the initial reward. Each enrolled subject can receive a maximum bonus of USD $48. The user consents to the minimum requirement and bonus computation information before they start the study. We also set QUINCE to show a pop-up message reminding subjects to submit their experiment identifier in order to receive their initial reward and information regarding the maximum reward when they reach 500 and 50,000 points, respectively.

Table 1: We included different sets of task groups in the two studies. The score assigned to each task group reflects the difficulty of tasks, and the badge awards in study B.

Task Groups	Tasks T1	T2	T3	T4	Cool-down time	Scores (points) in Study A	Study B
G1	✓	✓		✓	3 min	250	200
G2		✓	✓		1 hr	100	50
G3			✓	✓	1 hr	150	N/A
G4	✓		✓	✓	1 hr	N/A	200

We made some changes to QUINCE between study A and B. As mentioned in Table 1, we adjusted the grouping of tasks and score assignments for each group to provide more opportunities for subjects to participate in study B: 4 task groups through the user interface (1×G1, 1×G2, 2×G4), instead of 3 groups in study A (1×G1, 1×G2, 1×G3). We also introduced a badge system in study B. Subjects could receive 10 points for each badge. We examined whether these changes improved subject's performance in §4.1.

Table 2 shows descriptive statistics for the two studies. The subjects were diverse in terms of geographical locations and network providers. Subject approval rates were similar (>85%) across studies. Most rejected subjects did not earn sufficient scores (<500 points). We also identified a few MTurk workers who attempted to cheat by submitting random strings as the experiment identifier. We did not reject those subjects if they satisfied the minimum requirement (500 points). Although some subjects may have provided unreliable QoE ratings if they did not pay full attention during video playbacks, or their computer failed to decode high quality video smoothly. Those subjects contributed to other measurement tasks (T1-T3) that did not rely on their subjective judgment. Over 70% of subjects enrolled in the extended study, indicating intent to continue participation.

4.1 Subject performance

Fig. 3 shows the cumulative distribution function (CDF) of scores earned by all approved subjects. The solid and dotted lines show the CDFs of the scores in study A and B, respectively, with a dashed vertical line at 50,000 points. The minimum requirement was 500 points. We found that {98.6%/93.6%} of the subjects exceeded this requirement in study {A/B}. On the other hand, {1%/8.51%} of the subjects reached 48,000 points, and received the maximum reward.

Table 2: Statistics of the two crowdsourced studies.

	Study A	Study B
Study period	7/29-8/4, 2019	12/13-12/20, 2019
Subjects (Enrolled)	251 (204)	251 (188)
Accepted subjects	219	217
Cities	150	134
ISPs (by AS Numbers)	62	57

Very few subjects conducted further measurement tasks after the pop-up message showed up at 50,000 points. The median score in study A (4,500 points) was higher than in study B (2,750 points), probably due to the lower scores assigned to tasks in study B.

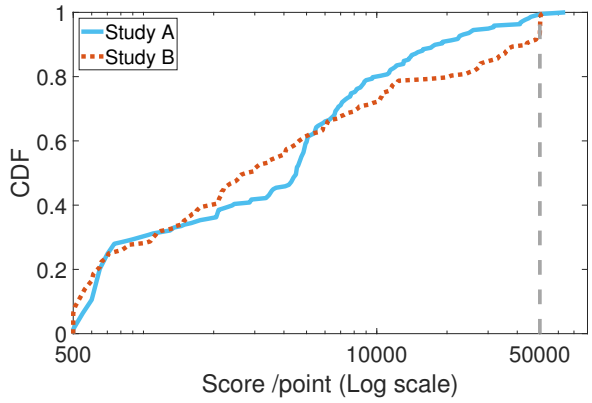


Figure 3: Subjects continued to conduct experiments after meeting the minimum requirement.

Table 3 lists the number of tasks completed by subjects and the experiment expenses. The number of submissions in study B was 32.9%-135% more than in study A, at 4.1% lower cost. Our changes to scoring and the addition of badges further boosted the cost efficiency of study B. In general, our experiment cost was lower than multiple one-off crowdsourced QoE tests. If we conservatively assume subjects spent 50% of their time on the video streaming QoE assessment task (T4), the cost per QoE rating was only USD $0.32 and $0.2 in study A and study B, respectively. On the other hand, a 30-minute crowdsourced test for rating 10 videos could cost USD $4-6 (i.e., $0.4-0.6 per rating) [14, 34]. Therefore, our approach was 20-66.7% more cost effective than traditional approaches.

We analyzed the number of tasks completed by subjects per hour to provide a fine-grain view of their activities. We found a similar pattern for both studies, so we only show the results for study B. The heatmap in Fig. 4a depicts the number of submissions of the four measurement tasks. Each vertical line represents 1 hour; and the color intensity represents the number of submitted tasks during that hour. The solid and dotted curves in Fig. 4b show the total number of task submissions and the number of subjects who completed at least one task within the hour, respectively. We show the subject throughput ($= \frac{\text{Number of submissions}}{\text{Number of subjects}}$) and its linear fit model in solid and dashed line in Fig. 4c, respectively.

Table 3: Subjects submitted thousands of tasks. The changes we made in study B increased its cost efficiency.

Number of submissions	Study A	Study B
T1	4,056	9,531
T2	6,162	8,192
T3	3,712	5,365
T4	4,254	6,429
Experiment cost† (USD)	$2,712	$2,599

† Note: The cost included 40% fee charged by MTurk.

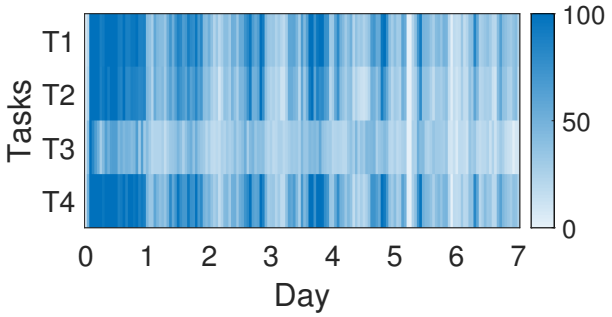
The three subfigures in Fig. 4 are vertically aligned and start at December 13, 2019 00:00am PT. As we published the study to MTurk at 1:35am, we would not expected any submission in the beginning of the figure. In the hour we published the study, subjects performed almost 600 measurement tasks. We received more than 200 submissions per hour from at least 30 subjects throughout the first day. We observed a diurnal pattern, which peaked in the evening time (7pm-11pm). In the peak hours, more than 20 subjects submitted over 200 measurements in an hour. We still received at least 50 measurements per hour from around 10 subjects in the off-peak hours. Collecting more samples during peak hours can help us diagnose QoE degradation due to Internet congestion.

While the total number of submissions largely varied with time-of-day, subject throughput was relatively stable for the first three days. Each subject submitted on average 10 measurements in an hour, except for the first two hours and the last three days of the study. During the first two hours, subject throughput was only 2.75 submission/hour, as subjects went through the tutorial and were unfamiliar with the tasks. After day 4, as subjects were more familiar with measurement tasks, the performance of subjects slightly increased. Engaged subjects submitted more than 20 measurements in late night (1-2am) when the platform had fewer active subjects. Having subjects participating around-the-clock provided us opportunities to investigate the impact of time-of-day on QoE evaluation.

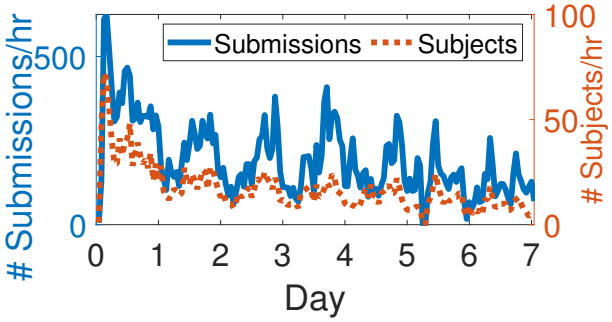
The color of Task 3 (T3) is much lighter than the other tasks in Fig. 4a. This is because T3 was the only task that was not in any task group that had a cool-down time of 3 minutes. The cool-down time mechanism throttled the rate of submission.

We quantified subject efficiency using task completion time (S), which is defined as the difference between the time when subjects clicked on (started) the tasks and the corresponding task submission (completion) time. Because the video clips we used in Task 4 (T4) had different length (60-90s), we subtracted the duration of video clips from the task completion time in T4 to reveal the actual time duration that subjects spent rating their QoE. For other tasks (T1-T3), task completion times included a fixed or short duration for conducting the measurement.

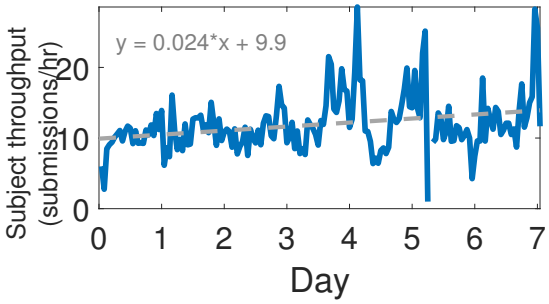
Fig. 5 shows a box-and-whisker plot of task completion times for the four tasks in the extended period ($S_{T1} - S_{T4}$) in the 2nd, 4th, 6th, and 8th boxes, respectively. Further, we presented the task completion times of the subject's first attempt during the initial study ($S^0_{T1} - S^0_{T4}$) in the 1st, 3rd, 5th, and 7th boxes, respectively. After the initial study, subjects were familiar with the tasks, their efficiency increased by reducing median task completion time by



(a) Heatmap of number of submissions for the four tasks. Task T3 showed significantly lower submission throughout the experiment period, because of the longer cool-down period.



(b) Except for the first day of the study, the total hourly number of submissions and the number of active subjects showed a strong diurnal pattern, which had the same peak as Internet usage patterns.



(c) Each subject submitted on average 10 measurements per hour. The linear fit model (dashed line) showed a small positive slope, indicating a constant increasing trend in subject throughput. The model is statistically significant (p <0.001, RMSE=3.89, R^2=0.0823). The intercept and coefficient are also significant (p <0.001).

Figure 4: Hourly performance of subjects in study B.

18.8%-46.1%. Fewer than 2% of measurements took over 120 seconds to finish; these measurements probably reflect loss of the subject's attention.

Take away: Crowdsourced workers can perform longitudinal studies, so long as experimenters provide time and incentives. Longitudinal study lowered overhead and experiment cost. The cool-down period and task group mechanism were effective to control the throughput of subjects. Repeated experiments also constitute

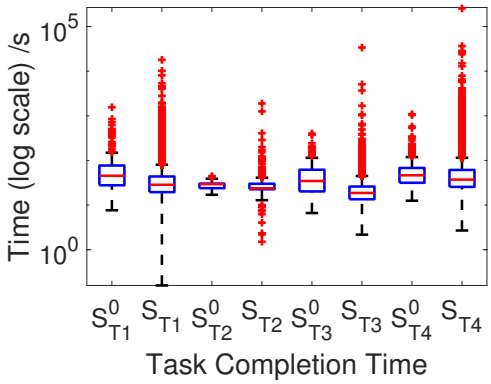


Figure 5: Box-and-whisker plot of task completion times of the first (S^0_T) and subsequent experiments (S_T) of the four measurement tasks. The horizontal red line in each box represents the median value. Subjects generally used less time to complete tasks in extended study, due to training effect. The median task completion time reduced up to 46.1% after subjects' initial trial.

a form of training, which reduces future task completion times, improving efficiency.

4.2 Subject engagement

The key factor to success of our experiment design is that it facilitates and incentivizes repeat participation by the same subjects, which allows longitudinal data collection. Fig. 6 shows the time period between enrolled subjects' consecutive logins to QUINCE. The behavior of the two studies were similar. Interestingly, we found significant fractions of re-login activity (13.9% in study A and 26.1% in study B) happening less than 1 minute after the previous session closed. The reason could be that subjects accidentally closed the browser tab or the connection to our server was unstable. On the other hand, over 98% of subjects in both studies revisited the platform at least once within the next day.

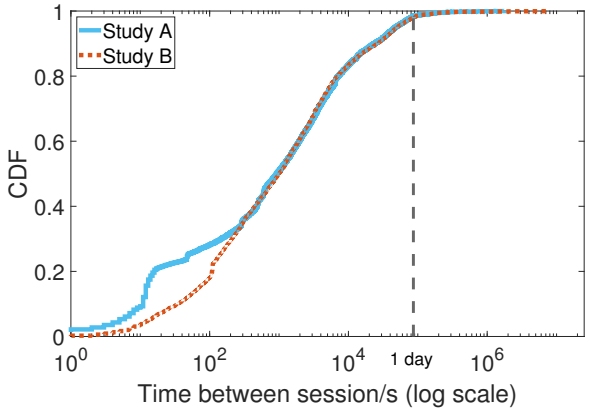


Figure 6: CDFs of times between consecutive logins. Most subjects revisited the platform at least once per day.

Another metric of engagement of subjects is the time they spent on the platform. Fig. 7 shows the distribution of length of all sessions (dotted yellow line), and the length of the first session of subjects who enrolled (dashed blue line) and those who did not enroll (solid green line) in the extended study. Because subjects had to go through a tutorial before performing any tasks, the median length of the first session length was around 30 minutes. We also found that those who enrolled in the extend study generally spent around 6 minutes more in the first session than those who decided not to participate further. Although the length of initial sessions was longer than that suggested for one-off studies [18], the option of participating in the extended study provided sufficient incentive for subjects to spend more time on the initial session. As subjects were familiar with the tasks after the initial study, they could perform tasks more efficiently. Thus, most sessions were significantly shorter than the initial session. The median session length was only 5.5 minutes.

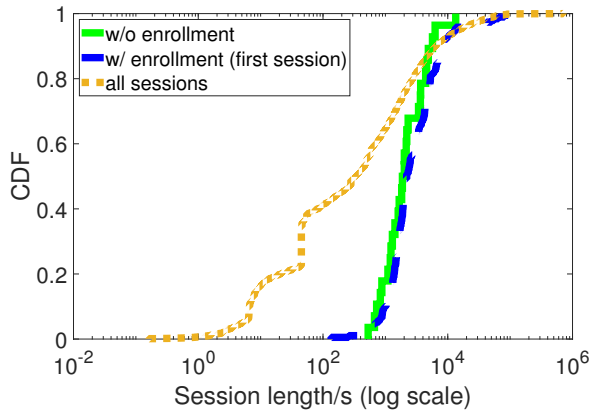


Figure 7: CDFs of the length of login sessions in study B. The first session was often long because of the onboarding process.

Take away: Our experimental design and implementation successfully increased subject engagement. Repeat participants had high familiarity with the experiment tasks, enabling great efficiency in gathering measurements.

4.3 Cross-validation with laboratory-based assessments

For network measurement tasks, we can easily assure the quality of data by enforcing format checks on subject's input. Our platform can immediately detect common errors and provide feedback to subjects to correct their submissions. However, due to the subjective nature of QoE, we cannot easily evaluate the quality of subjective assessments.

We focus on analyzing the data collected from the video streaming QoE task (Task T4) and evaluating the reliability of our experimental design. We simulated 15 streaming conditions: 5 streamed with different resolutions (240p, 360p, 480, 720p, and 1080p) smoothly without any stalling, 3 inserted stalling events, and the rest with video quality changing during playback. We encoded 12 creative common video clips downloaded from YouTube and 2 videos from a commercial video repository into HTTP Live Streaming (HLS) format with 5 quality levels. In each experiment, we applied one of the streaming conditions to one of the randomly selected video clips. Therefore, there were 210 ($= 15 \times 14$) combinations of videos and streaming conditions.

As a source of validation, we conducted laboratory-based subjective assessments using a non-gamified version of QUINCE. We removed all gamification elements and network measurement tasks from the platform. We translated the user interface and assessment forms to Japanese, the native language of the subjects. We invited 16 male and 16 female subjects, aged between 20 and 27, to rate 30 video streams, comprised of 2 videos and 15 streaming conditions we used in the previous crowdsourcing studies. The experiment consisted of three 20-minute sessions. Each subject watched 10 video streams in a session, and rated the QoE using the ACR method (in Japanese) [27] after the completion of each video stream. We presented the videos on 32-in Full HD monitors. The viewing distance was 1-1.5 times of the height of the monitor, mimicking ordinary Internet users watching online videos. We provided 20-minute breaks between sessions to mitigate the effect of fatigue. We also randomized the presentation sequence of video streams to reduce order effects [28].

Although the lab experiment and crowdsourcing study employed the same set of streaming conditions, we cannot directly compare the two sets of Mean Opinion Scores (MOS). We found that some subjects had low network throughput, revealed by the network performance task (T3). The simulated performance in crowdsourcing studies may not have always followed the pre-defined streaming conditions. Therefore, we derived the actual streaming performance from logs gathered by the video player, and leveraged well-established correlations between picture quality, stall events, and the reported QoE ([32]) to cross-validate our results.

We analyzed the QoE of video streaming with five different resolutions (240p, 360p, 480p, 720p, and 1080p) smoothly without any stalling or quality adaptations. Fig. 8 compares the MOSes collected from the laboratory (y-axis) with the crowdsourcing test (x-axis). The raw MOSes (diamond markers) are original ratings we collected in the lab experiment. However, a previous study [6] showed that Japanese subjects tended to rate QoE more conservatively than Western ones. Therefore, we converted the raw MOSes into converted MOSes (circle markers) using Eqn 2, as suggested in [33]. After the conversion, the absolute values between the two sets of data were much closer to each other. We also plotted 95% confidence intervals of the converted MOSes (vertical error bars) and the crowdsourcing MOSes (horizontal error bars). The confidence intervals of crowdsourcing MOSes were smaller than those of laboratory ones, due to the larger sample sizes.

The higher QoE ratings from crowdsourcing subjects than the laboratory ones under the same streaming performance could be due to other uncontrollable factors [5], including environment conditions, equipment, and day of time. Even though we cannot directly compare the absolute value of MOSes, both sets of data showed a similar trend. Subjects in both studies could not distinguish between video resolutions 720p and 1080p. While the laboratory experiment showed insignificant differences in MOS between resolutions 480p and 360p, crowdsourcing subjects rated 360p videos significantly

lower than 480p ones. The QoE of 240p video streams was the lowest in both experiments.

$$Converted\_MOS = 0.8681 * Raw\_MOS + 0.027 \quad (2)$$

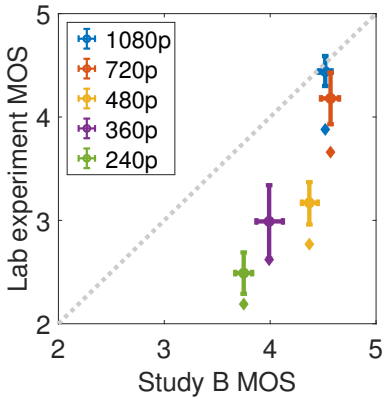


Figure 8: Comparison of MOS measured in the laboratory and QoE crowdtesting. Each color represents MOSes of one video bitrate. The MOS we obtained from both environment followed a similar trend.

To evaluate the reliability of the experiments, we computed Cohen's kappa coefficient [8] to measure agreement between subjects perceiving the same stimuli. Although the mean coefficient of the laboratory experiment (~0.9) was higher than the crowdsourcing one (~0.6), it was at a reasonable level and was similar to the values reported in [18] before filtering outliers. We presented raw MOS data to avoid potential bias between different outlier detection methods, which is a non-goal of this paper. While the sample size in our studies was large, we can further improve the reliability by detecting and removing outliers from the data by applying *a posteriori* methods, such as CrowdMOS [30], HodgeRank [37], and iHT/iLTS/aLTS [38].

Take away: Our results showed that QoE assessment conducted with our new experiment design can achieve reliability level that is similar to traditional crowdsourcing tests, but at a lower cost per rating. We compared the results we gathered in a laboratory experiment, and cross-validated our assessment results using correlation between picture quality and QoE studied in previous research.

5 CONCLUSION

We designed a novel crowdsourcing experiment framework for improving the efficiency of QoE crowdtesting, and implemented the framework as a gamified platform for measuring the QoE of video streaming and Internet performance. We conducted two 7-day studies with Amazon Mechanical Turk. Both studies showed that our experimental design and platform attracted strong subject engagement. Subjects were willing to revisit and perform additional measurement tasks throughout the study durations. Finally, to show the reliability of the QoE rating we collected, we compared the QoE assessment results with a laboratory experiment, and reproduced the well-established correlation between QoE and video bitrates.

6 DISCUSSION

Limitation on task grouping. We bundled tasks to ensure subjects completed tasks that spanned a range of difficulty. A drawback of this approach is that subjects may give up the entire experiment when one hard task blocks them from completing the group. In the case of `QUINCE`, subjects reported that running `traceroute` (T1) can be difficult for novice computer users. We revised our tutorial and instructions in study B based on questions frequently asked by subjects in study A. In study B, we received significantly less feedback about task T1, and productivity increased.

Future work to improve reliability of QoE assessments. Low quality subjects who pay no or little attention to stimuli can induce higher variance in assessment results. We can apply statistical approaches to remove outliers by analyzing distributions of ratings. As future work, we will leverage subjects' activities during the assessments to evaluate their quality. For example, low-quality subjects tend to have short think time and task completion time [17, 23], and straightforward mouse cursor trajectory [25] that minimizes time spent on tasks. These detection models only considered one-off experiments, and might not be applicable to longitudinal experiments. We will investigate potential changes in behavior when subjects revisit our platform for the extended study. The `QUINCE` platform collects metrics (§4) that support such analysis of subject behavior.

ACKNOWLEDGMENTS

We would like to thank UC San Diego undergraduate students (Kenil Vora, Andrew Zhen, Xinpei Tan, Akshit Gupta, Jennifer Chan, Jessica Nguyen, Shania Ie) for their work in implementing the experiment platform. We also thank the anonymous reviewers for valuable comments. This research was supported by NTT Corporation under the project *A reactive crowdsourcing-based QoE monitoring* (20193641) and Nation Science Foundation Award OAC-1724853.

REFERENCES

[1] [n.d.]. Clickworker. https://www.clickworker.com.
[2] [n.d.]. Meteor. https://www.meteor.com.
[3] [n.d.]. Microworkers. https://microworkers.com/.
[4] Amazon. [n.d.]. Mechanical Turk. https://www.mturk.com.
[5] Louis Anegekuh, Lingfen Sun, and Emmanuel Ifeachor. 2014. A Screening Methodology for Crowdsourcing Video QoE Evaluation. In *Proc. IEEE Globalcom.*
[6] Zhenyu Cai, Nobuhiko Kitawaki, Takeshi Yamada, and Shoji Makino. 2010. Comparison of MOS evaluation characteristics for Chinese, Japanese, and English in IP telephony. In *Proc. International Universal Communication Symposium.*
[7] Kuan-Ta Chen, Chen-Chi Wu, Yu-Chun Chang, and Chin-Laung Lei. 2009. A crowdsourceable QoE evaluation framework for multimedia content. In *Proc. ACM Multimedia.*
[8] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.
[9] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From game design elements to gamefulness: defining "gamification". In *Proc. MindTrek.*
[10] Sebastian Egger-Lampl, Judith Redi, Tobias Hoßfeld, Matthias Hirth, Sebastian Möller, Babak Naderi, Christian Keimel, , and Dietmar Saupe. 2017. Crowdsourcing Quality of Experience Experiments. In *Proc. Crowdsourcing and Human-Centred Experiments (Dagstuhl Seminar 15481).*
[11] Carsten Eickhoff, Christopher G. Harris, Padmini Srinivasan, and Arjen P. de Vries. 2012. Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments. In *Proc. ACM SIGIR.*
[12] Oluwaseyi Feyisetan, Elena Simperl, and Max Van Kleek abd Nigel Shadbolt. 2015. Improving paid microtasks through gamification and adaptive furtherance incentives. In *Proc. WWW.*
[13] David R. Flatla, Carl Gutwin, Lennart E. Nacke, Scott Bateman, and Regan L. Mandryk. 2011. Calibration Games: Making Calibration Tasks Enjoyable by Adding Motivating Game Elements. In *Proc. ACM UIST.*

[14] Bruno Gardlo, Sebastian Egger, Michael Seufert, and Raimund Schatz. 2014. Crowdsourcing 2.0: Enhancing Execution Speed and Reliability of Web-based QoE Testing. In *Proc. IEEE ICC*.

[15] Dennis Guse, Henrique R. Orefice, Gabriel Reimers, and Oliver Hohlfeld. 2019. TheFragebogen: A Web Browser-based Questionnaire Framework for Scientific Research. In *QoMEX*.

[16] Juho Hamari. 2017. Do badges increase user activity? A field experiment on the effects of gamification. *Computers in Human Behavior* 71 (2017), 469–478.

[17] Matthias Hirth, Sven Scheuring, Tobias Hoßfeld, Christian Schwartz, and Phuoc Tran-Gia. 2014. Predicting Result Quality in Crowdsourcing Using Application Layer Monitoring. In *Proc. IEEE ICCE*.

[18] Tobias Hoßfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia. 2014. Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing. *IEEE Trans. Multimedia* 16, 2 (2014), 541–558.

[19] Christian Keimel, Julian Habigt, and Clemens Horch. 2012. Video quality evaluation in the cloud. In *Proc. IEEE PV*.

[20] Christian Keimel, Julian Habigt, Clemens Horch, and Klaus Diepold. 2012. QualityCrowd - A framework for crowd-based quality evaluation. In *Proc. IEEE Picture Coding Symposium*.

[21] Tak Yeon Lee, Casey Dugan, Werner Geyer, Tristan Ratchford, Jamie Rasmussen, N. Sadat Shami, and Stela Lupushor. 2013. Experiments on Motivational Feedback for Crowdsourced Workers. In *Proc. ICWSM*.

[22] M-Lab. [n.d.]. NDT (Network Diagnostic Tool). https://www.measurementlab.net/tests/ndt/.

[23] Yoshitaka Matsuda, Yu Suzuki, and Satoshi Nakamura. 2017. A Trade-off between Estimation Accuracy of Worker Quality and Task Complexity. In *Proc. IEEE BIGDATA*.

[24] Elisa D. Mekler, Florian Br uhlmann, Klaus Opwis, and Alexandre N. Tuch. 2013. Disassembling Gamification: The Effects of Points and Meaning on User Motivation and Performance. In *ACM CHI Extended Abstracts*.

[25] Ricky Mok, Rocky Chang, and Weichao Li. 2017. Detecting low-quality workers in QoE crowdtesting: A worker behavior based approach. *IEEE Trans. on Multimedia* 19, 3 (2017), 530–543.

[26] Ricky Mok, Ginga Kawaguti, and kc claffy. 2019. QUINCE: A unified crowdsourcing-based QoE measurement platform. In *Proc. ACM SIGCOMM poster session*.

[27] ITU-T P.913. 2016. *Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment*. ITU-T Recommendation P.913.

[28] Margaret H. Pinson and Stephen Wolf. 2003. Comparing subjective video quality testing methodologies. In *Proc. Visual Communications and Image Processing*.

[29] Benjamin Rainer, Markus Waltl, and Christian Timmerer. 2013. A Web based subjective evaluation platform. In *Proc. QoMEX*.

[30] Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer. 2011. CrowdMOS: An approach for crowdsourcing mean opinion score studies. In *Proc. IEEE ICASSP*.

[31] Raimund Schatz, Sebastian Egger, and Kathrin Masuch. 2012. The Impact of Test Duration on User Fatigue and Reliability of Subjective Quality Ratings. *Journal of the AES* 60, 1/2 (2012), 63–73.

[32] Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia. 2015. A Survey on Quality of Experience of HTTP Adaptive Streaming. *IEEE Commun. Surveys Tuts* 17, 1 (2015), 469–492.

[33] Telecommunication Technology Committee. 2018. TTC JJ-201.01 A Method for Speech Quality Assessment for IP Telephony.

[34] Matteo Varvello, Jeremy Blackburn, David Naylor, and Konstantina Papagiannaki. 2016. EYEORG: A Platform For Crowdsourcing Web Quality Of Experience Measurements. In *Proc. ACM CoNEXT*.

[35] Xiaohui Wang, Dion Hoe-Lian Goh, Ee-Peng Lim, Adrian Wei Liang Vu, and Alton Yeow Kuan Chua. 2017. Examining the effectiveness of gamification in human computation. *International Journal of Human-Computer Interaction* 33, 10 (2017), 813–821.

[36] Chen-Chi Wu, Kuan-Ta Chen, Yu-Chun Chang, and Chin-Laung Lei. 2013. Crowdsourcing Multimedia QoE Evaluation: A Trusted Framework. *IEEE Trans. Multimedia* 15, 5 (2013), 1121–1137.

[37] Qianqian Xu, Tingting Jiang, Yuan Yao, Qingming Huang, Bowei Yan, and Weisi Lin. 2011. Random Partial Paired Comparison for Subjective Video Quality Assessment via HodgeRank. In *ACM MM*.

[38] Qianqian Xu, Ming Yan, Chendi Huang, Jiechao Xiong, Qingming Huang, and Yuan Yao. 2017. Exploring Outliers in Crowdsourced Ranking for QoE. In *ACM MM*.

[39] Lingyan Zhang, Shangguang Wang, Fangchun Yang, and Rong N. Chang. 2017. QoECenter: A Visual Platform for QoE Evaluation of Streaming Video Services. In *Proc. IEEE ICWS*.

[40] Yuxiang Chris Zhao and Qinghua Zhu. 2014. Effects of extrinsic and intrinsic motivation on participation in crowdsourcing contest: A perspective of self-determination theory. *Online Information Review* 38, 7 (2014), 896–917.