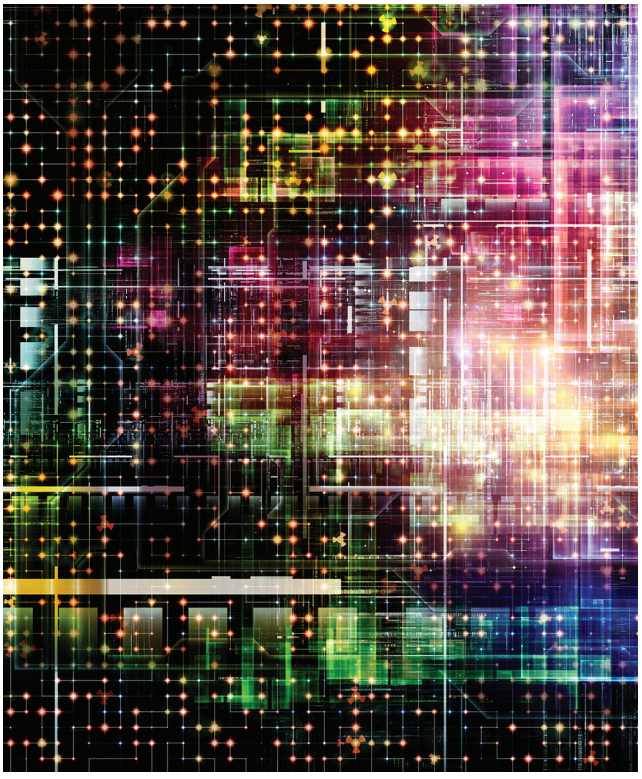

# Crowdsourcing Quality-of-Experience Assessments

**Tobias Hossfeld,** *Julius-Maximilians-Universität Würzburg*

**Christian Keimel,** *Technische Universität München*

**Christian Timmerer,** *Alpen-Adria-Universität Klagenfurt*

**Crowdsourced quality-of-experience (QoE) assessments are more cost-effective and flexible than traditional in-lab evaluations but require careful test design, innovative incentive mechanisms, and technical expertise to address various implementation challenges.**

The quality of experience (QoE) of an application or service can be defined in many ways. For example, the ITU Telecommunication Standardization Sector (ITU-T) defines QoE as "overall acceptability," while the European Network on Quality of Experience in Multimedia Systems and Services (Qualinet) equates it with the "degree of delight or annoyance of the user." Regardless of its precise definition, QoE is generally acknowledged to be an evolution of quality of service (QoS), with QoS metrics regarded as objective and QoE metrics as subjective.[1]

Although QoS-to-QoE mappings exist and some provide meaningful guidance, the most common means to determine QoE is through user evaluations in which subjects evaluate a given stimuli—such as content, an application, or a service—within a controlled laboratory environment. Depending on the setup, such user studies can be costly in terms of preparation and execution time as well as human resources, but they generally yield reliable results.

In recent years, crowdsourcing has gained momentum in various application domains,[2,3] including as a cost-effective alternative to in-lab QoE evaluations. Instead of conducting the study in a controlled environment, researchers use special platforms (Amazon Mechanical Turk, Microworkers), social networks (Facebook, Twitter, LinkedIn), or email campaigns to recruit subjects, who participate in the study via the Web from their home or office or on the go.

Figure 1 highlights some of the many advantages of crowdsourced QoE evaluations. These include the rapid availability of numerous users—a campaign typically takes only minutes or hours—from diverse backgrounds; natural test environments; heterogeneous client devices and software; various network access technologies (wired, Wi-Fi, 3G/4G, and so on) in worldwide locations; the ability to


Published by the IEEE Computer Society

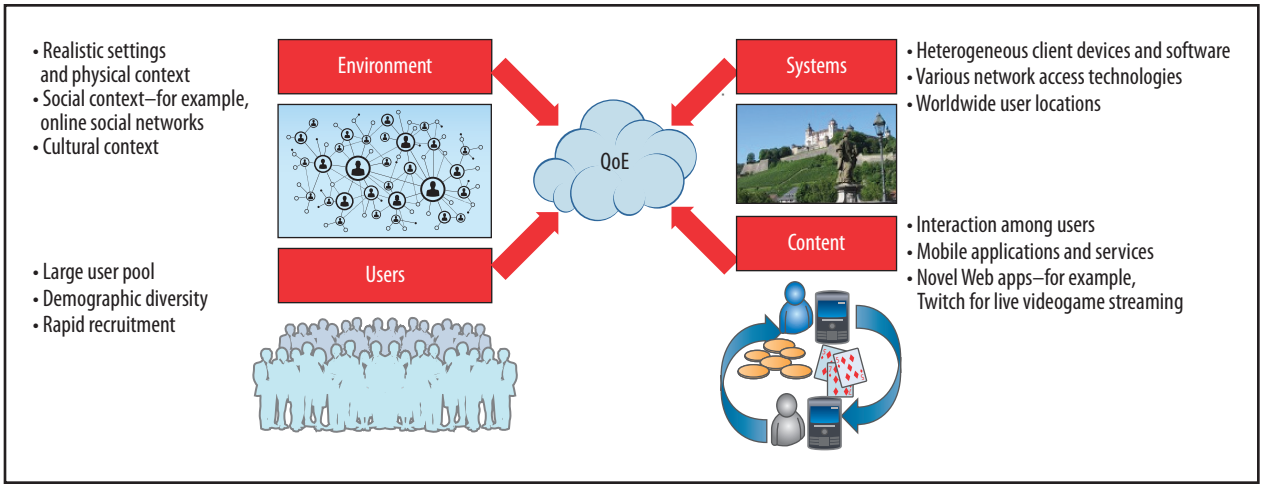


**Figure 1. Advantages of crowdsourced quality-of-experience (QoE) assessments.**

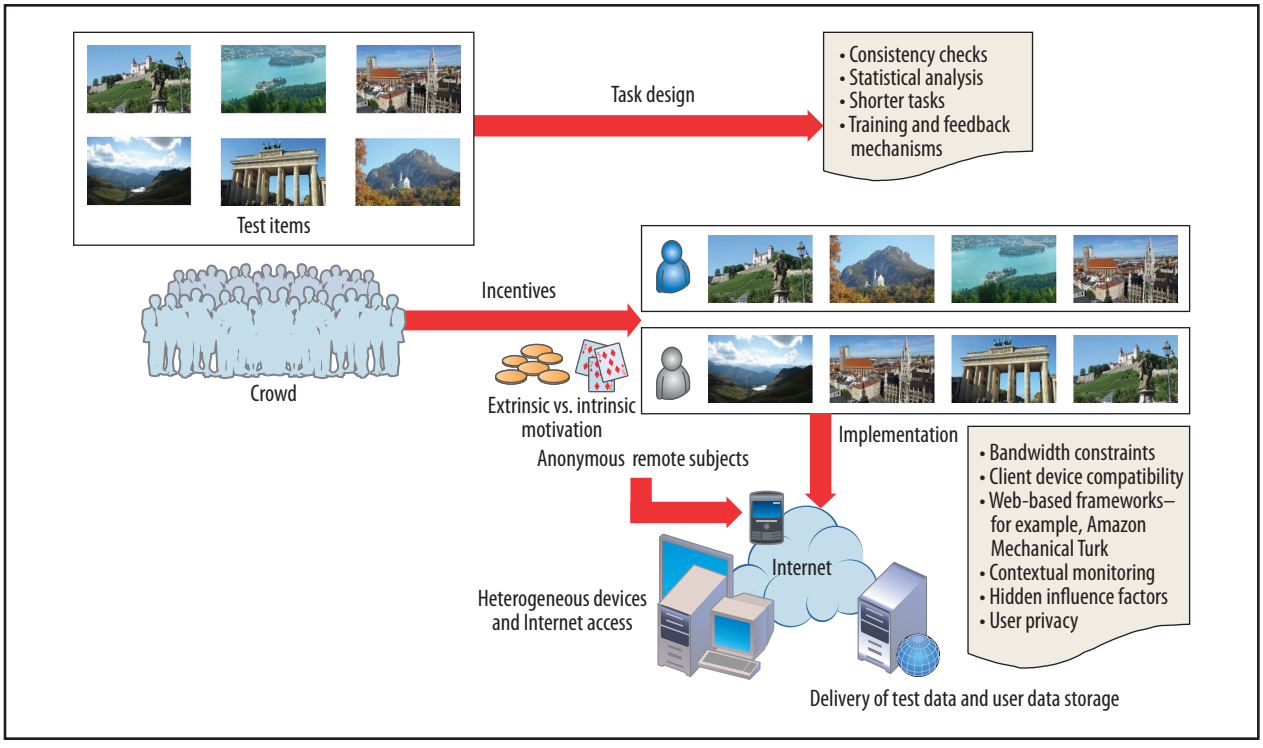


**Figure 2. Challenges of crowdsourced QoE assessments.**

evaluate user interaction and mobile apps and services under realistic conditions; and the avoidance of costs associated with experimental facilities, in-lab personnel, and traditional participant recruitment schemes (including, for example, payment and reimbursement).

On the other hand, user responses in crowdsourced QoE studies are generally considered less reliable due to the lack of controlled conditions. In addition, contextual tools must be designed to address the many hidden influence factors.

Here, we discuss the challenges of crowdsourced QoE evaluations with respect to test design, participant incentives, and implementation, which Figure 2 summarizes. We also describe new opportunities for this approach enabled by emerging technologies.

## TEST DESIGN

Crowdsourced QoE evaluations require a different conceptual approach than in-lab studies due to participants' remoteness and anonymity, the need for shorter tasks,

and the absence of onsite personnel to provide instructions or answer questions.[4]

## Testing methodology

As participants are typically remote, anonymous users conduct the test with their own equipment at a place and time of their preference, so researchers must employ rigorous consistency checks and statistical analysis methods to ensure reliable results. For example, results obtained from a QoE assessment of videos compressed with MPEG-AVC/H.264 at varying bitrates were completely different for in-lab and crowd-based subjects.[4] Possible explanations for this discrepancy are heterogeneous hardware among subjects or improper training. Researchers must extract such contextual factors in crowdsourced settings and analyze their hidden influence on user ratings.

In addition, crowdsourced QoE tests typically consist of shorter tasks, on the order of minutes, than comparable in-lab tests. Thus, traditional in-lab QoE tasks recommended by standardization bodies such as the ITU-T can't be directly migrated to crowd-based environments but instead must be split into multiple smaller tasks. However, test items still need to capture the entire quality range. The situation is complicated by the fact that in most cases not all subjects assess all test items, resulting in an imbalance in the total number of ratings per item.

Researchers are exploring ways to address these issues to ensure greater reliability in crowdsourced QoE assessments.[5] Recent YouTube studies using absolute category rating (ACR) scales yielded similar results for crowd- and lab-based participants. However, a verification methodology is still needed because subjects can have difficulty judging ACR scores consistently or could give false ratings by not paying attention to the scoring procedure.

Pairwise comparison of QoE evaluations is a promising solution: subjects need only provide comparative judgments.[6] The large size of the user pool in crowdsourced QoE evaluations makes it possible to cope with the quadratic growth of comparisons arising from the number of test items. Still, better statistical methods for analyzing pairwise-comparison results, especially during runtime, are needed and are active areas of research.

## Training and feedback

The lack of onsite monitors and reliance on a Web interface in crowdsourced QoE evaluations necessitates proper training and feedback mechanisms. Instructions to subjects must be clear and include descriptions of the nature and purpose of the test, what to evaluate, and how to rate the quality of or compare items.

Proper task designs and statistical methods can help prevent or at least detect subjects' misunderstanding of a task, uncertainty about rating scales, sloppy execution, or fatigue. However, direct feedback between test supervisors and participants can be even more effective: non-real-time feedback can include comments, contact forms, or forums, while real-time feedback—which is practical only for short tests given that users can conduct the test whenever they want—can include chat and social networking apps. As users are free to decide which tasks to conduct and in what order, incentives might need to be incorporated into the test to ensure compliance with certain guidelines.

# PARTICIPANT INCENTIVES

Successful crowdsourcing for QoE evaluations depends on the ability to incentivize users to participate. Motivation can be extrinsic or intrinsic.

## Extrinsic motivation

In commercial crowdsourcing platforms, extrinsic motivation is typically the key driver. The user doesn't participate to gain satisfaction from conducting the task—that is, the actual QoE assessment—but by receiving some reward, such as money, credit, or points. Increasing extrinsic motivation leads to faster task and study completion,[7,8] which seems desirable, but rapid task and study completion also result in smaller demographic diversity—for example, users might all be in the same time zone. Moreover, increasing monetary rewards or other payments won't necessarily lead to better data quality, as financially motivated users often complete studies sloppily.[8] Careful statistical analysis is thus clearly required to avoid poor-quality results.

Data quality can also be improved by letting users decide whether to stop or to continue QoE tasks.[9] Those who desire to increase their earnings could perform additional tasks, but only if a reliability threshold is exceeded; users who only want to participate for a short time could leave the study earlier. Nevertheless, this approach requires automated reliability mechanisms and advanced statistical output analysis of user ratings.

Even reliable extrinsically motivated users are affected by the payoff of a task, representing a contextual influence factor on QoE assessments. A recent study shows that the payment level influences the ACR score but not qualitative factors.[8] Researchers need to better understand this effect to develop appropriate rating normalization schemes.

## Intrinsic motivation

Users can also have intrinsic motives to participate in QoE evaluations. These motives can be altruistic—for example, to help advance scientific research or support a particular community—or selfish, such as a desire for entertainment.[10] Studies indicate that, in general, an increase in intrinsic motivation

leads to higher data quality.[7]

*Gamification* involves designing assessment tasks that users carry out as a side effect of playing online games, especially interactive ones.[11] Recent studies indicate that gamification reduces false ratings by a factor of five and that innovative, creative tasks are less likely to invite cheating.[12] However, as gamification is strongly task related, there are no general guidelines on how to design an interesting game for QoE assessment.

## IMPLEMENTATION

Although in principle crowdsourcing could be used for any type of QoE assessment, there are practical limitations on the potential scope, mainly due to bandwidth constraints and the inability of subjects' (consumer) devices to present certain stimuli. For example, it's currently infeasible to perform crowd-based ultra HDTV or high-dynamic-range imaging QoE studies. Moreover, some crowdsourcing service providers specify in their policies that participants can't be required to download and install software. Therefore, the complete QoE must be conducted via a standard Web browser without task-specific plug-ins.

Instead of implementing each crowdsourced QoE assessment from scratch, researchers can use existing Web-based frameworks.[9] These enable researchers to focus on task design as well as lower the hurdle to crowd-based solutions by abstracting practical deployment issues and providing basic reliability and monitoring functionalities.

Generic crowdsourcing platforms such as Amazon Mechanical Turk or Microworkers are generally preferable to specialized platforms or aggregators.[5] They provide more flexibility in test and task design; access to a huge, globally distributed crowd of potential test subjects; and filter or qualification mechanisms to select specific users based on, say, their location. Depending on the assessment goal, social networks like Facebook can also be used, but they often present implementation limitations and the subject pool might not be as diverse.

The variability of test participants and environments makes crowdsourcing attractive for QoE evaluations, but it also necessitates contextual monitoring and checking for potential hidden influence factors. To ensure that subjects perform evaluation tasks as designed, test administrators must be able to verify that environmental conditions—for example, background illumination in visual QoE assessments—are correct and subjects' devices comply with test parameters. Similarly, researchers need to capture latent factors such as subjects' expectations, demographic characteristics, and possible impairments that could influence result reliability.[5]

Researchers could leverage information in subjects' social media profiles to gain insight into hidden influence factors without the need to use additional questionnaires, as well as choose test subjects even more selectively. To address privacy concerns, however, they should inform subjects about the use and purpose of such data.

## NEW RESEARCH OPPORTUNITIES

As current technologies mature and new ones emerge, crowdsourcing will enable novel opportunities for QoE assessments.

Some QoE studies that require special hardware like eye trackers could be realized in crowdsourcing settings via built-in cameras in laptops, tablets, and smartphones. In addition, such devices include other sensors that could capture additional contextual information—for example, the user's precise location, environmental conditions such as light and noise levels, and device orientation. Devices' increasingly powerful processing capabilities also make it possible to extract discrete features from audiovisual data in real time for later analysis.

Furthermore, the pervasiveness of mobile devices and anytime, anywhere Internet connectivity enable more extensive field-based QoE trials. These would provide more realistic insights into users' behavior by allowing them to interact with applications as well as other participants in a wider variety of natural everyday settings and situations. To facilitate training, test administrators could use online chat or even voice/video communication to provide assistance during the assessment comparable to that of in-lab moderators.

Along with these new opportunities, however, come a host of technical challenges as well as privacy-related issues, as capturing additional contextual information could significantly reduce participants' anonymity.

Crowdsourcing is becoming increasingly popular as a QoE assessment tool. It's more cost-effective and flexible than conducting in-lab evaluations, and can quickly provide a large and diverse pool of subjects who

interact with applications and services as well as with one another in realistic everyday settings. Emerging communication technologies will enable exciting new research opportunities. At the same time, this approach requires careful test design, innovative incentive mechanisms, and technical expertise to address various implementation challenges.

## References

1. P. Le Callet, S. Möller and A. Perkis, eds., *Qualinet White Paper on Definitions of Quality of Experience*, v1.2, European Network on Quality of Experience in Multimedia Systems and Services, Mar. 2013; www.qualinet.eu/images/stories/QoE_whitepaper_v1.2.pdf.
2. M. Masli, "Crowdsourcing Maps," *Computer*, vol. 44, no. 11, 2011, pp. 90–93.
3. J. Riedl and E. Riedl, "Crowdsourcing Medical Research," *Computer*, vol. 46, no. 1, 2013, pp. 89–92.
4. T. Hossfeld and C. Keimel, "Crowdsourcing in QoE Evaluation," *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller and A. Raake, eds., Springer, 2014, pp. 315–327.
5. T. Hossfeld et al., "Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing," *IEEE Trans. Multimedia*, vol. 16, no. 2, 2014, pp. 541–558.
6. K.-T. Chen et al., "A Crowdsourceable QoE Evaluation Framework for Multimedia Content," *Proc. 17th ACM Int'l Conf. Multimedia* (MM 09), 2009, pp. 491–500.
7. J. Rogstadius et al., "An Assessment of Intrinsic and Extrinsic Motivation on Task Performance in Crowdsourcing Markets," *Proc. 5th Int'l Conf. Weblogs and Social Media* (ICWSM 11), 2011; www.mediateam.oulu.fi/publications/pdf/1405.pdf.
8. M. Varela et al., "Increasing Payments in Crowdsourcing: Don't Look a Gift Horse in the Mouth!," *Proc. 4th Int'l Workshop Perceptual Quality of Systems* (PQS 13), 2013; https://pqs.ftw.at/workshop-program/papers/s14_Varela.pdf.
9. T. Hossfeld et al., "Survey of Web-Based Crowdsourcing Frameworks for Subjective Quality Assessment," *Proc. 16th Int'l Workshop Multimedia Signal Processing* (MMSP 14), 2014; http://infoscience.epfl.ch/record/199537/files/MMSP2014_CS.pdf.
10. A.D. Shaw, J.J. Horton, and D.L. Chen, "Designing Incentives for Inexpert Human Raters," *Proc. 2011 ACM Conf. Computer Supported Cooperative Work* (CSCW 11), 2011, pp. 275–284.
11. C. Timmerer and B. Rainer, "The Social Multimedia Experience," *Computer*, vol. 47, no. 3, 2014, pp. 67–69.
12. C. Eickhoff et al., "Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments," *Proc. 35th Int'l ACM SIGIR Conf. Research and Development in Information Retrieval* (SIGIR 12), 2012, pp. 871–880.

***Tobias Hossfeld*** *heads the Future Internet Applications and Overlays group in the Department of Communication Networks at Julius-Maximilians-Universität Würzburg, Germany. Contact him at hossfeld@informatik.uni-wuerzburg.de.*

***Christian Keimel*** *is a researcher at Technische Universität München, Germany. Contact him at christian.keimel@tum.de.*

***Christian Timmerer,*** *Social Computing column editor, is an associate professor at Alpen-Adria-Universität Klagenfurt, Austria. Contact him at christian.timmerer@itec.aau.at.*