International Telecommunication Union

ITU-T Technical Report

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(10 MAY 2018)

PSTR-CROWDS
Subjective evaluation of media quality using a crowdsourcing approach

Summary

This technical report introduces the basic concepts of crowdsourcing, its application in quality assessment studies, and describes general principles of a subjective methodology for assessing media quality using micro-task crowdsourcing. Micro-task crowdsourcing offers fast, low cost, and scalable approaches by outsourcing tasks to a large number of participants. In addition, crowdsourcing also provides a large diverse pool of participants, and a real-life environment for quality assessment of multimedia services and applications. Nevertheless, crowdsourcing methods cannot be understood as direct implementations of laboratory testing methods in an Internet-based environment, due to factors they inherit from the nature of crowdsourcing. Therefore, crowdsourcing experiments should be designed differently. The proposed principles for crowdsourcing subjective assessment methods enable experimenters to collect a large number of media quality ratings (video, image, speech, and audio-visual) in a short period of time from a diverse population of participants and in realistic environments.

NOTE: Reference to commercial offerings is made solely for informative purposes and does not imply any sort of endorsement or approval of the specific services.

Keywords

Absolute Category Rating, crowdsourcing, crowdtesting, experimental design, media quality assessment, micro-task crowdsourcing, subjective evaluation

Change Log

This document contains Version 1 of the ITU-T Technical Report on "*Subjective evaluation of media quality using a crowdsourcing approach*" approved at the ITU-T Study Group 12 meeting held in Geneva, 1-10 May 2018.

Editor: Sebastian Möller
TU Berlin
Germany

Tel: +49 30 8353 58465
Fax: +49 30 8353 58409
Email: sebastian.moeller@tu-berlin.de

CONTENTS

Page

List of Tables

Page

NO TABLE OF FIGURES ENTRIES FOUND.

List of Figures

Page

Technical Report ITU-T PSTR-CROWDS

Technical Report ITU-T
Subjective evaluation of media quality using a crowdsourcing approach

Summary

This technical report introduces the basic concepts of crowdsourcing, its application in quality assessment studies, and describes general principles of a subjective methodology for assessing media quality using micro-task crowdsourcing. Micro-task crowdsourcing offers fast, low cost, and scalable approaches by outsourcing tasks to a large number of participants. In addition, crowdsourcing also provides a large diverse pool of participants, and a real-life environment for quality assessment of multimedia services and applications. Nevertheless, crowdsourcing methods cannot be understood as direct implementations of laboratory testing methods in an Internet-based environment, due to factors they inherit from the nature of crowdsourcing. Therefore, crowdsourcing experiments should be designed differently. The proposed principles for crowdsourcing subjective assessment methods enable experimenters to collect a large number of media quality ratings (video, image, speech, and audio-visual) in a short period of time from a diverse population of participants and in realistic environments.

1 Scope

Media quality assessments (MQA) rely fundamentally on subjective studies in order to capture humans' perception of stimuli quality. Different Recommendations were developed explaining the subjective testing methods in the laboratory like ITU-T P.800 [32] for speech, ITU-T P.910 [34] for video and ITU-T P.911 [35] for audiovisual services.

This technical report briefly introduces the basic concepts of crowdsourcing, its application in subjective quality assessment, and describes general principles of subjective test methods and experimental designs for assessing media quality using crowdsourcing. Crowdsourcing taps into the collective intelligence of the public at large to complete media quality assessment tasks that would normally either be performed by media quality experts themselves, and/or be carried out by a limited number of selected participants in a laboratory setting, based on the corresponding standardized methods.

Standardized laboratory experiments are executed to gather reliable and accurate perceived media quality ratings from a number of recruited participants in controlled experimental setups. The crowdsourcing approach enables experimenters to access a larger pools of participants in order to collect participants' feedback on topics presented to them via innovative experimental design approaches that improve the quality of the data collected.

Limitations: While this technical report provides a general guidance on testing methods for media quality assessments using a crowdsourcing approach, the methods addressed were tested and confirmed for passive image, speech and video quality testing with crowdsourcing only. No guidance can be given for interactive media services so far. Recommendations on implementations of these guidelines for specific media are currently under study by ITU-T Study Group 12.

2 References

[1] Archambault, D., Purchase, H. and Hoßfeld, T. 2017. *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments: Dagstuhl Seminar 15481, Dagstuhl Castle, Germany, November 22–27, 2015, Revised Contributions*. Springer.

[2] Cooke, M., Barker, J., Lecumberri, G. and Wasilewski, K. 2013. Crowdsourcing in speech perception. *Crowdsourcing for speech processing: Applications to data collection, transcription and assessment*. (2013), 137–172.

[3] Dai, P., Rzeszotarski, J., Paritosh, P. and Chi, E.H. 2015. And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions. *Proc. of 18th ACM CSCW* (2015).

[4] Doan, A., Ramakrishnan, R. and Halevy, A.Y. 2011. Crowdsourcing systems on the world-wide web. *Communications of the ACM*. 54, 4 (2011), 86–96.

[5] Recommendation ITU-T P.808 2018. Subjective evaluation of speech quality with a crowdsourcing approach.

[6] Egger-Lampl, S., Redi, J., Hoßfeld, T., Hirth, M., Möller, S., Naderi, B., Keimel, C. and Saupe, D. 2017. Crowdsourcing Quality of Experience Experiments. *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer. 154–190.

[7] Estellés-Arolas, E. and González-Ladrón-De-Guevara, F. 2012. Towards an integrated crowdsourcing definition. *Journal of Information science*. 38, 2 (2012), 189–200.

[8] Gadiraju, U., Kawase, R., Dietze, S. and Demartini, G. 2015. Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. *Proceedings of CHI* (2015).

[9] Gardlo, B., Egger, S., Seufert, M. and Schatz, R. 2014. Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing. *In Communications (ICC), 2014 IEEE International Conference*. (2014), 1070–1075.

[10] Gardlo, B., Ries, M. and Hoßfeld, T. 2012. Impact of screening technique on crowdsourcing QoE assessments. *Radioelektronika (RADIOELEKTRONIKA), 2012 22nd International Conference, IEEE*. (Apr. 2012), 1–4.

[11] Gardlo, B., Ries, M., Hoßfeld, T. and Schatz, R. 2012. Microworkers vs. facebook: The impact of crowdsourcing platform choice on experimental results. *Quality of Multimedia Experience (QoMEX), 2012 Fourth International Workshop*. (Jul. 2012), 33–36.

[12] Geiger, D., Seedorf, S., Schulze, T., Nickerson, R.C. and Schader, M. 2011. Managing the Crowd: Towards a Taxonomy of Crowdsourcing Processes. *AMCIS* (2011).

[13] Hirth, M., Hoßfeld, T. and Tran-Gia, P. 2011. Anatomy of a crowdsourcing platform-using the example of microworkers. com. *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on* (2011), 322–329.

[14] Hoßfeld, T., Hirth, M., Korshunov, P., Hanhart, P., Gardlo, B., Keimel, C. and Timmerer, C. 2014. Survey of Web-based crowdsourcing frameworks for subjective quality assessment. *In Multimedia Signal Processing (MMSP), 2014 IEEE 16th International Workshop*. (2014), 1–6.

[15] Hoßfeld, T., Hirth, M., Redi, J., Mazza, F., Korshunov, P., Naderi, B., Seufert, M., Gardlo, B., Egger, S. and Keimel, C. 2014. Best Practices and Recommendations for Crowdsourced QoE-Lessons learned from the Qualinet Task Force" Crowdsourcing". (2014).

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

[17] Hoßfeld, T., Seufert, M., Hirth, M., Zinner, T., Tran-Gia, P. and Schatz, R. 2011. Quantification of YouTube QoE via crowdsourcing. *InMultimedia (ISM), 2011 IEEE International Symposium*. (2011), 494–499.

[18] Hoßsfeld, T. and Keimel, C. 2014. Crowdsourcing in QoE Evaluation. *Quality of Experience*. Springer. 315–327.

[19] Howe, J. 2006. The rise of crowdsourcing. *Wired magazine*. 14, 6 (2006), 1–4.

[20] ITU-T Handbook 2011. Practical Procedures for Subjective Testing.

[21] Keimel, C., Habigt, J., Horch, C. and Diepold, K. 2012. Qualitycrowd–a framework for crowd-based quality evaluation. *Picture Coding Symposium (PCS), 2012* (2012), 245–248.

[22] Kittur, A., Nickerson, J.V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M. and Horton, J. 2013. The future of crowd work. (2013), 1301.

[23] Malone, T.W., Laubacher, R. and Dellarocas, C. 2009. Harnessing crowds: Mapping the genome of collective intelligence. *Massachusetts Institute of Technology*. (2009).

[24] Martin, D., Carpendale, S., Gupta, N., Hoßfeld, T., Naderi, B., Redi, J., Siahaan, E. and Wechsung, I. 2017. Understanding The Crowd: ethical and practical matters in the academic use of crowdsourcing. *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer International Publishing. 27–69.

[25] Naderi, B. 2018. *Motivation of Workers on Microtask Crowdsourcing Platforms*. Springer International Publishing.

[26] Naderi, B. 2018. Who are the Crowdworkers? *Motivation of Workers on Microtask Crowdsourcing Platforms*. Springer. 17–27.

[27] Naderi, B., Möller, S. and Mittag, G. 2018. Speech Quality Assessment in Crowdsourcing: Influence of Environmental Noise. *44. Deutsche Jahrestagung für Akustik (DAGA)* (2018).

[28] Naderi, B., Polzehl, T., Beyer, A., Pilz, tibor and Möller, S. 2014. Crowdee: Mobile Crowdsourcing Micro-task Platform - for Celebrating the Diversity of Languages. *Proc. 15th Ann. Conf. of the Int. Speech Communication Assoc. (Interspeech 2014)* (Singapore, Sep. 2014).

[29] Naderi, B., Polzehl, T., Wechsung, I., Köster, F. and Möller, S. 2015. Effect of Trapping Questions on the Reliability of Speech Quality Judgments in a Crowdsourcing Paradigm. *16th Ann. Conf. of the Int. Speech Comm. Assoc. (Interspeech 2015). ISCA* (2015), 2799–2803.

[30] Naderi, B., Wechsung, I. and Möller, S. 2015. Effect of Being Observed on the Reliability of Responses in Crowdsourcing Micro-task Platforms. *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on* (2015), 1–2.

[31] Polzehl, T., Naderi, B., Köster, F. and Möller, S. Robustness in Speech Quality Assessment and Temporal Training Expiry in Mobile Crowdsourcing Environments. *16th Ann. Conf. of the Int. Speech Comm. Assoc. (Interspeech 2015). ISCA* 2794–2798.

[32] Recommendation ITU-T P.800 1996. Methods for subjective determination of transmission quality.

[33] Recommendation ITU-T P.800.2 2016. Mean opinion score interpretation and reporting.

[34] Recommendation ITU-T P.910 2008. Subjective video quality assessment methods for multimedia applications.

[35] Recommendation ITU-T P.911 1998. Subjective audiovisual quality assessment methods for multimedia applications.

[36] Recommendation ITU-T P.912 2016. Subjective video quality assessment methods for recognition tasks.

[37] Recommendation ITU-T P.1401 2012. Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models.

[38] Reips, U.-D. 2002. Standards for Internet-based experimenting. *Experimental psychology*. 49, 4 (2002), 243.

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

[40] Tabachnick, B.G. and Fidell, L.S. 2012. *Using Multivariate Statistics*. Pearson Education.

[41] Zhao, Y. and Zhu, Q. 2014. Evaluation on crowdsourcing research: Current status and future direction. *Information Systems Frontiers*. 16, 3 (2014), 417–434.

3 Terms and definitions

3.1 Terms defined elsewhere

This Technical Report uses the following terms defined elsewhere:

3.1.1 crowdsourcing [36]: Obtaining the needed service by a large group of people, most probably an on-line community.

3.1.2 crowdworker [5]: Person performing a crowdsourcing task.

3.1.3 job [5]: A template for tasks including questions and all the information necessary for a crowdworker to accept and complete that task. A task is an instantiation of a job for a particular crowdworker. An experiment may contain one or more jobs.

3.1.4 job provider [5]: Person or entity who creates a job in a micro-task crowdsourcing platform, also known as requester.

3.1.5 micro-task crowdsourcing [5]: Crowdsourcing simple and small tasks in an open call, to a large and undefined crowd which are usually reimbursed by a monetary reward per each piece of work they perform.

3.1.6 micro-task crowdsourcing platform [5]: A platform which manages the relationship between crowdworkers and job providers including maintaining a dedicated panel of crowdworkers and providing required infrastructure like creating jobs, poll of tasks for crowdworkers, and payment mechanisms.

3.1.7 test [36]: Subjective assessments in a crowdsourcing environment.

NOTE – 3.1.7 follows terminology presented in [16].

3.1.8 task [36]: Set of actions that a crowdworker needs to perform to complete a subscribed part of the test.

NOTE – 3.1.8 follows terminology presented in [16].

3.1.9 question [36]: A single event that requires an answer for a crowdworker. A task contains many questions.

3.1.10 vote [33]: A subject's response to a question in a rating scale for an individual test sample or interaction.

3.2 Terms defined here

This Technical Report defines the following terms:

None.

4 Abbreviations

ACR	Absolute Category Rating
API	Application Programming Interface
CCR	Comparison Category Rating
CI	Confidence Interval
DCR	Degradation Category Rating
IQA	Image Quality Assessment
MOS	Mean Opinion Score
MQA	Media Quality Assessment
SQA	Speech Quality Assessment
QoE	Quality of Experience
UI	User Interface
URL	Uniform Resource Locator
VQA	Video Quality Assessment

5 Crowdsourcing[1]

5.1 Introduction

In 2006, the term crowdsourcing was introduced by combining crowd and outsourcing. It refers to ”the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call” [13, 19]. In principle, crowdsourcing utilizes the potential ability of a large group of people, who are connected through the Internet, to carry out a certain task [12]. During the time, numerous communities have been developed on the Internet, under the umbrella of crowdsourcing, to solve a broad range of problems by applying different collaboration methods [4]. Some examples from different domains include Wikipedia, Linux, Stack Overflow, and Amazon Mechanical Turk[2] (MTurk).

Besides the application domain, crowdsourcing systems can be classify by considering who are the people composing that community, what they do, why and how they do it [4, 7, 23, 41]. An important aspect is whether crowd members are unpaid volunteers or paid in exchange of their service. The commercial crowdsourcing platforms with paid members expanded in various dimensions [13]. Typically, these platforms act as a mediator and provide tasks given by employers to the crowd either in the form of contest or open call [13]. A particular case is the micro-task crowdsourcing, which provide a broad range of small tasks in an open call, to a large and undefined crowd who are usually reimbursed by a small monetary reward per each piece of work they perform.

[1] This chapter is based on [25].
[2] https://mturk.com

5.2 Micro-task Crowdsourcing

Micro-task crowdsourcing provides a remarkable opportunity by offering a highly scale able, on demand and low-cost pool of geographically distributed workforce for completing complex tasks that can be divided into a set of short and simple online tasks like annotations and data collection [14, 21, 22, 28]. Typically, *crowdworkers* are paid on a piecework basis. *Tasks* are short, often formatted in a form-like webpage and are as simple as a computer-literate worker should be able to perform one in couple of minutes. Tasks are offered using a crowdsourcing *platform* which manages the relationship between crowdworkers and *job provider*. A job provider creates a *job* in the crowdsourcing platform which is a collection of tasks that can be done by one or many crowdworkers in exchange for monetary rewards.

5.2.1 Platforms

The platforms are maintaining a dedicated crowd of workers and providing required infrastructure like mechanism for creating jobs, pool of tasks for crowdworkers, payment mechanisms, and in some cases additional services like quality control or worker selection mechanisms [14]. Typically, the crowdsourcing platform provides a web access with responsive web design (e.g. MTurk, Microworkers[3], and CrowdFlower[4]) and in some cases a native mobile application (e.g. Crowdee[5]) for crowdworkers.

A simplified crowdworking process (excluding the payments) is illustrated in Figure 1. First, the job provider creates a job on the platform which may contain external content (like speech material) to be assessed by the workers (process A in Figure 1). Typically, platforms provide different templates for jobs (e.g. survey job, moderation of an image), and additionally permit job providers to create a custom designed job using raw HTML code. A job acts as a customized template. The job provider also specifies properties of the job such as reward per assignment, and the number of repetitions requested per task, and worker's requirements such as location of crowdworker. During the task creation process, one or more tasks will be created for the job depending on the requested repetitions and the number of block of dynamic contents. For instance, a survey job with 100 repetitions leads to 100 tasks and a speech-quality assessment (SQA) job with 20 block of stimuli and 24 repetitions will leads to 480 tasks (a block typically contains 10 or more stimuli). After creating tasks, they will be added to the pool of available microtasks and be ready to be performed by the crowdworkers. Furthermore, the platform holds corresponding funds (i.e. the sum of rewards expected to be paid to the crowdworkers, and platforms fee) for liability.

Meanwhile, the crowdworkers browse the list of available jobs in the platform. By selecting a particular job, a crowdworker requests a task from that job to be reserved and assigned to him/her (process B in Figure 1). After performing the task, the crowdworker will submit the corresponding response. The submitted responses are stored in the platform's repository of answers.

Last but not least, job providers can inquire the list of submitted responses (process C in Figure 1). Typically, the job provider has a limited time to validate the submitted responses and accept or reject them. Otherwise, the platform will accept the submitted answer automatically. By accepting the answer, the corresponding crowdworker will be paid, and by rejecting it, the crowdworker gets a notification and the corresponding task will return to the pool of available tasks to be taken by other workers. Platforms may offer additional features, including different job templates, pre-checking of the job by experts (e.g. CrowdFlower), rewards estimation (e.g. Microworkers), and built-in quality control (e.g. CrowdFlower). Typically, platforms provide an API connection for the job providers to

[3] https://microworkers.com
[4] https://www.crowdflower.com
[5] https://crowdee.de

automate the process of creating a job, appending tasks, inquiring answers and approving or rejecting them (which usually provides advanced features not available using GUI).

In addition, platforms use different terminologies for crowdworkers, tasks and job providers. A crowdworker is called a "Worker" in most of the platforms but a "Contributor" in CrowdFlower. The job is called a "Project" in MTurk, and "Campaign" in Microworkers. A task (an instance of a job) is called a "HIT" (Human Intelligence Task) or an "Assignment" in MTurk. The job providers, are called "Requesters" in MTurk, "Employers" in Microworkers and "Customers" in CrowdFlower.

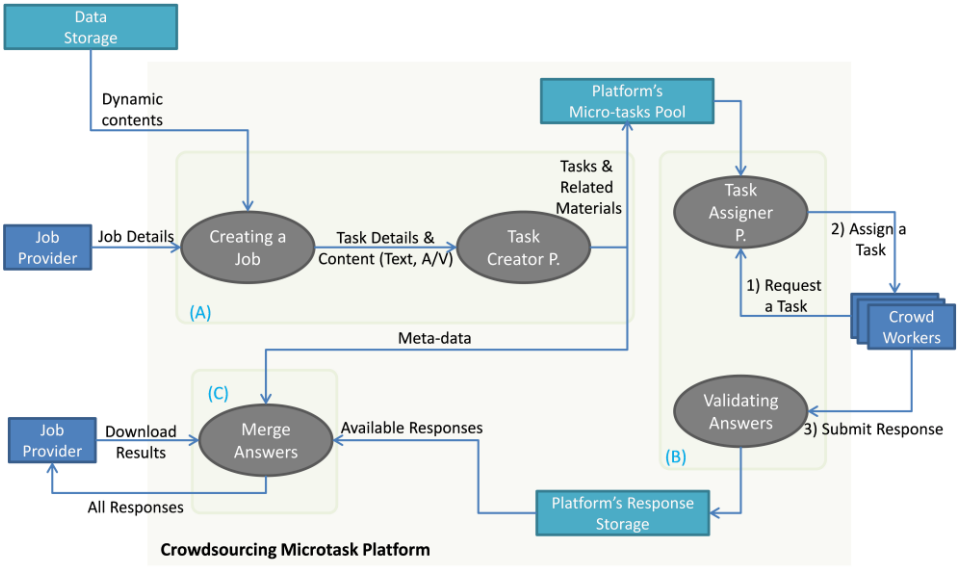

Figure 1: General workflow of the crowdsourcing micro-task platforms, after [25].

5.2.2 Crowdworkers

Result of studies in 2016 shows that the demographics, socioeconomic status, and crowdworking conditions of the workers differ considerably between and within platforms[6] [26]. For items like household size, educational level, and yearly income, the Indian MTurk workers share similarities with the worker from developing countries of other platforms. Workers from the Western countries shared similarities together, irrespective of the platform they use for crowdworking. As a result, job providers should consider the within and between platform differences of crowdworkers when deciding on the platform to use and selecting their participants. In case that their research results may be influenced by demographics and socioeconomic status of participants, adequate sampling procedures should be considered to avoid potential biases in the results.

In addition, workers mostly work at home (>80%) and secondly from their office despite the platform they use. For platforms that provide native smartphone application a considerable group of people also work in their way. The crowdworkers mostly used their desktop computer or laptop for performing their tasks. Workers from developing countries work more (>15%) using their phones than their colleagues from Western countries (for detailed analyses see [24]).

[6] Note that the crowdworker population changes rapidly during time.

6 Quality Assessment using Crowdsourcing

6.1 Introduction

Crowdsourcing as an emerging paradigm in the Internet offers new possibilities for QoE research by providing a global pool of participants for shifting QoE experiments into the Internet. The potential benefits of crowdsourced QoE studies are the investigation a) of participant influence factors due to the diverse population and heterogeneity of users, b) of realistic environment influence factors due to the real-life environment of the subjects, and c) of reduced costs and turnaround times. However, crowdsourcing invokes a variety of challenges due to the uncontrolled setting, e.g. a potential unreliability of participants, ill-designed jobs in absence of an experiment moderator, and the statistical analysis of data obtained in that way.

6.2 Differences to laboratory experiments

Applying crowdsourcing to quality tests brings new challenges which are inherent to its web-based experimentation origin. Here, the experimenter is looking for a fast response from a very large numbers of subjects in a real-world situation and accepting to have less control over the experiment in comparison to the laboratory [2]. In the following, the main known differences between laboratory and crowdsourcing tests are highlighted.

Conceptual Differences: In crowdsourcing, the test duration is shorter (5-15 min [14]) than in the laboratory which normally leads to a mixed within-between-participants study design (e.g. in a speech quality rating task, each task contains only a subset of the entire dataset). Workers can easily drop out from a job consisting of several parts and just rate a subset of samples [18, 39]. Therefore, a robust statistical analysis is necessary [39]. In the absence of a test moderator, a training session with feedback to the participant and the absence of fatigue during the experiment cannot be ensured.

Participants and their Motivation: In crowdsourcing, participants are more diverse than in the laboratory. The selection of a homogenous group of participants is important for some tasks. Relevant factors include gender, age. Other factors can include language history, normality of hearing, and linguistic background [2] depending upon the media being tested. For example, it is crucial to measure the participants' hearing ability (audiograms) remotely for speech and audio quality assessment tasks, as many participants do not realize their mild or even moderate hearing loss [2]. In the absence of a test moderator, the selection of participants is commonly based on their self-reported data, which depends on the participants' trustworthiness. In addition, the reliability of the collected responses in a crowdsourcing test may suffer from some participants who do not work as instructed, or who do not focus enough on the test and provide arbitrary responses in order to minimize effort [30]. The source of the workers' motivation and the task design have an effect on the reliability of outcomes [10, 16, 30]. Comparing outcomes from volunteer workers (recruited through Facebook) and paid crowdworkers showed that the volunteers are less likely to finish the task, but when they complete it, their results are more reliable than the results from the paid workers [11]. A two-stage experimental design for screening out untrustworthy data is recommended [2, 16] and examined in the case of video quality assessment [16]. In addition, including reliability check methods in crowdsourcing studies is highly recommended [15]. Gold standard questions (also known as trapping questions, honeypots) can not only be used in post-hoc reliability check analysis; properly designed and positioned questions may encourage workers to provide more consistent responses [30]. Comparing MOS ratings of a laboratory and a crowdsourcing test employing different types of gold standard questions, showed that the highest correlation (and saturation with less number of ratings from crowdworkers) is achieved when gold standard stimuli with encouraging messages were used [29].

Technical and Environmental Factors: Crowdworkers are participating in studies using their own hardware which, in contrast to hardware usually provided in laboratory tests, can widely vary, is uncalibrated and not high-end. In a listening experiment, it is important that stimuli are transmitted without distortions (other than the ones to be judged) and their reproduction is not compromised by poor-quality sound cards or headphones. In IQA, device form factor, operating system information, web browser, and screen size are factors that highly affect the data collected from a crowdsourcing test. In SQA tests using mobile crowdsourcing, it was reported that narrow-band (NB) speech files tend to be rated with a lower quality in the mobile crowdsourcing study in comparison to the laboratory study [29, 31]. Depending on the crowdsourcing platform or framework used for conducting the study, different audio playback solutions are available. Forcing workers to install a separate software package is not recommended [14]. In addition, the acoustic environment (noise, reverberation) and the environmental context (other applications used in parallel, messages, calls) can influence the quality of the responses directly, or indirectly by affecting attention of crowdworkers. It has been shown that surrounding noise has influence on speech quality rating and type of headphone used (in-ear, open/close back) can reduce that effect [27].

All of mentioned environmental aspects can be considered as hidden influence factors in the crowdsourcing setting, but they also lead to a more realistic consideration of real-life influences.

6.3 Influence factors

Following factors should be consider when designing a crowdsourcing test.

6.3.1 Task influence factors

Here, the question is how crowdsourcing tasks should be designed to achieve a high reliability and validity of the results, and how the reliability can be measured and improved. As abovementioned, crowdsourcing tests' duration is shorter and with absence of a moderator (see discussion above). Task design influencing factors like dynamically imposing retraining of crowdworkers based on temporal timeouts [31] bonuses for participating in full study [39], consistency check and specific trapping questions [29] showed to enhance the reliability of results in the crowdsourcing study.

6.3.2 Participant influence factors

The impact of participant abilities (such as viewing and hearing characteristics) and motivators (e.g. financial) on the outcome of crowdsourcing quality experiments should be considered and methods for controlling them need to be developed. Crowdworkers differ in a number of characteristics from participants in laboratory experiments. As a result from the shorter crowdsourcing task length compared to the laboratory, a higher number of different crowdworkers is required to collect the same number of ratings. As a result of the highly diverse pool of participants, variable and diverse user ratings are collected. In addition, crowdsourcing platforms reach out to specific population groups. As crowdsourcing platforms are online platforms, only computer-literate persons will participate in the tasks. Also, due to the prevailing financial motivator, the participating group will also show certain income characteristics which might differ from the target population of service users. See [24] for details on demographics of crowdworkers in different platforms.

6.3.3 System influence factors

With the increasing popularity of different mobile and web crowdsourcing micro-task platforms, it is important to study the effect of platforms' characteristics, and the user device, on the crowdtesting results. It is necessary to find out how important those factors are and how to monitor or control them. Crowdsourcing experiments are conducted on participants' devices, in different environments (wherever they are in case of mobile crowdsourcing), and in the absence of a test moderator. Devices may differ in terms of hardware (e.g. soundcard, connected headphone, volume

settings) and software (e.g. OS). The bandwidth of the Internet connection may vary and participants may use their devices in different ways (e.g. monaural/binaural listening, portrait or landscape viewing). However, it is a possibility either to detect the device type, or to ask users about their used hardware and settings. Further, users may be motivated to follow a specific usage instruction (e.g. conduct the test in a quiet place). So far, the impact of these factors on the outcome of crowdsourcing-based quality assessment tasks is not clear.

6.3.4 Environment influence factors

Since crowdtesting is conducted in the user's natural environment with little to no control over the parameters that could affect the test, it is important to consider the environment's influence factors on the perceived quality of media test. In crowdsourcing experiments, the physical environment where the experiment is conducted differs between participants. To help mitigate the differences, experimenters can instruct crowdworkers to do their work in a particular environment (which will however not be equivalent to a laboratory environment with respect to ambient noise, reverberation, etc.). Following by monitoring the environmental conditions by asking the crowdworkers, or by analyzing incoming sensory signals such as microphone signal, movement and position sensor signals, events logged on the smartphone, etc.). The monitoring may also clarify whether the crowdworker has been distracted by parallel activities or events, either involuntarily or voluntarily.

6.4 Frameworks

One of the following approaches should be adapted when implementing the experiment depending on the purpose and requirements of the test:

- Using in-built functionalities of the host crowdsourcing platform
- Using the crowdsourcing platform for recruiting the crowdworkers, and conduct the study in a separate infrastructure

In case of using a separate infrastructure, it is recommended to use a web-base framework to ease moderating the experiment.

The basic functionality of a framework includes:

- The creation of the test (by supporting common testing methodologies like ACR, DCR, PC),
- The execution of the test (by supporting training, task design, task order, screening), and
- The storage and access to the result data.

Some frameworks provide advance functionalities like online reliability check (e.g. in-momento crowdsourcing [9]) or scripts for analyzing the collected data (e.g. crowdMOS [39]). A detailed comparison between available frameworks can be found in [6].

7 Transferring Lab experiments to the Crowd

The crowdsourcing tests cannot be considered as direct implementations of laboratory testing methods in an Internet-based environment due to the abovementioned fundamental differences. In this section, principles are provided which guide an experimenter through transferring a passive MQA laboratory experiment to the crowdsourcing experiment.

7.1 Database structure

It should be noted that the equipment used by the crowdworkers in the experiment cannot be assumed as known, high-end and identical for each crowdworker. Meanwhile, the test materials will be transferred through an Internet connection to the crowdworker's device. Those should be considered during feasibility study of applying crowdsourcing test.

Besides that, there is no difference between the preparation of source materials for laboratory tests and crowdsourcing test. Therefore, source recordings and selection of circuit conditions should be prepared as specified in the corresponding Recommendation considering the variability in the test equipment and feasibility of transferring the test martial via the Internet.

7.2 Experiment design

It is recommended that the test is limited in size by the maximum length of session possible for a given media without fatigue, distraction and possibility of losing the collected ratings. Because of the nature of a crowdsourcing study, a typical crowdsourcing micro-task takes couple of minutes to complete. Therefore, when the dataset contains large number of test stimuli, it is recommended to split an experiment session to a chain of tasks in the crowdsourcing test (i.e. rating job). It is recommended that each task duration be not more than a couple of minutes. It is possible, that crowdworkers may perform a number of but not all of tasks available for them in a test which can lead to an error variance caused by individual differences. Therefore, one of the following approaches should be adapted depending on the database structure:

- Applying a Balanced Blocks experimental design as described in [20]. As a result, crowdworkers should be assigned into groups such that the entire corpus of speech materials is rated by the workers as a whole, but each group rates only a subset of that corpus. For a SQA test, considering the corpus contains *t* talkers, each spoken *s* samples, and *N* degradation conditions then usually *s* stimulus sets can be created each containing $t \times N$ stimuli (i.e. *t* stimuli per each condition). As a result each stimulus set should be evaluated in one task of rating job by a listening panel i.e. a crowdworker should be able to take only one task from the rating job.
- Applying a mixed subject study design and motivating crowdworkers to perform multiple tasks from the rating job. Later, consider individual differences during statistical analyses (i.e. mixed models considering individual differences as random effect). Assuming the database contains *s* stimuli, and each task of rating job will include *k* stimuli, then $\left\lceil \frac{s}{k} \right\rceil$ stimulus sets should be created by randomly selecting stimuli from the dataset (either on demand or precomputed). It is recommended to give an extra reward (bonus) to the crowdworkers who perform a sequence of tasks from the rating job ideally evaluating 50% or more of the entire dataset to reduce the error variance associated with individual differences.

NOTE – A short duration of crowdsourcing test session lessens the opportunities for workers being exhausted and technical failures such as internet connectivity. It is recommended that the number of stimuli in a testnot exceed 50-70 overall when performing an IQA, and 20 for SQA.

7.3 Validity and reliability of collected data

The inadequacy of collected data through micro-task crowdsourcing was addressed numerously. Different reasons were given [29, 30, 38]:

- Task related: ill-designed jobs like given instruction may be misunderstood,
- Worker related: participants who do not focus enough, share their attention with a parallel activity and do not work as instructed or try to maximize their monetary benefit with minimum effort,
- Environment related: workers may be interrupted or environmental factors interact with the condition under study (e.g. environmental noise in SQA)
- System related: workers may use inadequate equipment.

To reduce their impact, different quality control approaches were proposed such as using gold standards, majority voting, and behavioral logging to evaluate reliability of the collected data in

post-processing [3, 8]. In the context of QoE, researchers have used additional methods like content questions and consistency tests [17, 18].

Quality control approach must be applied during test preparation, execution, and data screening. In each section corresponding details are given.

7.4 Test Procedure

7.4.1 Preparation

Experimenters must be aware of the following practices that help strengthen their test design, yielding more reliable data collected from crowdworkers in a crowdsourcing study:

1. The use of clear instructions. Unlike in a laboratory setting, an experimenter does not directly interact with crowdworker. As a result, instructions given to participants must be specific and unambiguous, leaving no chance for misinterpretation.
2. Allow crowdworkers to communicate with experimenters to discover any encountered problems. This holds great importance as it provides the experimenter with valuable information during the analysis. A crowdworker could contact the experimenters to inform them of any technical difficulties faced while performing the test. For example, in an IQA, a worker could inform the experimenter that the images are loading slowly, causing the worker to give low ratings or a version of web-browser does not perform as expected. Methods of communication can include but are not limited to email, blogs, community forums, etc.
3. Consider appropriate design and parameters (i.e. duration, reward, etc.) of each task. Previous studies showed that interestingness, profitability and expected workload of a crowdsourcing task influence workers' decision on taking part on it, and perceived interestingness and difficulty of a task influence performance of workers on that [25]. Therefore it is recommended to consider following aspects:
    a) Overall experiment enjoyability
    b) Task duration and complexity
    c) User Interface (UI) logistics (e.g. language, UI design, page complexity, input methods, etc.).
    d) Compensation relative to the test duration and complexity
    e) Test clarity and the user's ability to understand the task
4. It is recommended to use a multi-step crowdsourcing test design. As a result the workflow contains three jobs: Qualification Job, Training Job and Rating Job (cf. Figure 2). Each job contains an instruction followed by a list of questions. A question might be static (e.g. 'In what year were you born?') or includes dynamic part(s) (e.g. a URL pointing to a stimulus which should be evaluated by the worker in that question). The qualification job should be used to test workers ability, environment and equipment adequateness for participating in the study. The training job should be used for anchoring and prepare workers to perform the rating job. Within a task from the rating job, workers assess stimuli set associated to that task. Based on the experiment design a worker may be able to perform one or more task from the rating job. It is recommended to consider re-training the workers following a certain period. Detailed recommendation on each of those jobs are given in the following sections.

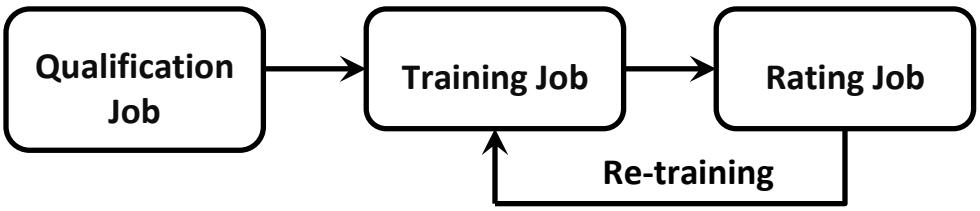

Figure 2: Workflow of crowdsourcing-based MQA

It is recommended that experimenters clearly state the following information in instruction to better inform workers of the task at hand:

1. Eligibility requirements and exclusion criteria
2. Estimated time it takes to complete a task (estimated from pilot studies)
3. Expectation from participating in a test such as the type of responses that may result in rejected work, or bonuses.
4. The identity of your research group. While remaining anonymity may help protect the reputation of the research lab, stating affiliations helps to build trust with the crowdsourcing community.

7.4.1.1 Qualification job

Within this job the purpose of the study should be explained to the crowdworkers, and checked if they are eligible to participate in the study considering their demographic data, abilities, environment, and system requirements. Other evaluations may be considered depending to the aim of the study. It is recommended to use the platform's in-built functionalities to make this job just accessible to the crowdworkers who have performed very well in other jobs. Statistics like the *Task approval rate* (e.g. 98% or more) in combination with a sufficiently high *number of approved task* are recommended. It should be noted that those filters do not guarantee that selected crowdworkers will perform well in the following jobs; therefore the experimenter must use their own qualification and gold standard questions (see clause 7.3.1.3) to check the reliability of submitted responses.

Based on the response to this qualification job, a randomly selected group of crowdworkers (who satisfied the prerequisites) should be invited to participate in the experiment i.e. getting access to the training job. The experimenter should consider inviting three to five times of the number of listeners which is expected to evaluate each stimulus.

NOTE 1 – This job should be performed by large number of crowdworkers to be able to screen for a target group of participants. Therefore, it should be short and sufficiently paid.

NOTE 2 – Depend on the employed crowdsourcing platform, advance information about crowdworkers might be available in form of profile properties. As a result, experimenter may be able to grant access to this job just to workers with specific profile properties (e.g. location, demographic information, and habits). Typically, platforms do not directly share workers' profile information with job providers.

NOTE 3 – Depend on the media under study, the experimenter must carefully create the screening test to assess participant ability (e.g. language test, and hearing screening test for speech, and color blindness for image), environment (e.g. recording 10 seconds of environmental noise for speech) and equipment (e.g. asking about type of headphone for speech or check the display resolution for image) adequateness to partake in the study based on the goals of experiment.

NOTE 4 – Experimenter must compensate workers for participating in this job when they perform their task as explained despite their eligibility to participate in the next step.

7.4.1.2 Training job

Within this job, test instructions should be given to the participants followed by a preliminary list of stimuli (i.e. samples). Participants should give their opinions about the quality of sample on a given scale. Based on the media under study standard scale given in corresponding Recommendation must be used. No suggestion should be made to the crowdworkers that the samples include the best or worst condition in the range to be covered. However, in the selection of samples attention should be applied to approximately cover the range from worst to best quality to be expected in the test. Samples should be selected from the original dataset to be representative of the contents represented in the whole set. For example, when running an IQA, if photos in the overall stimuli set show balls or cars, ensure that the training image set include photos of both balls and cars. This should be applied to other features of the stimuli (e.g. lighting of image or gender of speaker for speech).The order of presentation of stimuli should be randomized, and each crowdworker should receive the same stimuli for training.

By submitting a response to this job, a temporary access to the rating job should be granted to the crowdworker, i.e. assigning a qualification to that crowdworker. As long as the qualification is valid, the crowdworker can perform tasks from the rating job. Ideally, the access should expire within 60 minutes after granting, which requires the crowdworker to perform the training job again after expiration [31]. For SQA it is recommended that the access should not last for more than 24 hours.

NOTE 1 – Ratings for the training set should be discarded from the analysis.

NOTE 2 – Based on the experiment design, the experimenter may consider to combine training and rating job.

7.4.1.3 Rating job

Within this job, first an instruction should be given to worker, following by necessary environmental and system check. Finally, the selected set of stimuli should be presented to the crowdworker. They should give their opinion using a standard scale that must be selected following corresponding Recommendation based on media under study. For each stimulus one question should be added to this job. The rating job should contain validity check methods (including gold standard question) designed following the requirements of clause 7.4.1.4. The experiment should include appropriate number of gold standard questions depend on the number of stimuli in the set and the media under study.

The order of presentation of stimuli in the set should be randomized on runtime for each crowdworker. The experimenter should assign bonuses to crowdworkers who evaluate 50% or more stimuli in the dataset. It is recommended to force the system to download the entire set of stimuli under test in the rating job before the crowdworker can start to rate them, in order to avoid any delay in the rating procedure which might affect the rating.

It is recommended to collect more votes per stimulus than the equivalent laboratory based experiment. The number of votes per stimulus should be increased when the number of stimuli in the set presented in one rating job decreases.

NOTE 1 – The job provider may consider monitoring the crowdworker’s behavior during the test, including focus time of the browser tab and completion time for each question.

NOTE 2 – The job provider should warn the crowdworkers in advance that this job needs to download some materials which are free of charge but the downloading file size can lead to some network usage cost.

7.4.1.4 Validity check

Experimenters must incorporate different validity check methods in their jobs ensuring the worker's full attentiveness to the task, and adequateness of environment and system used by the worker for the given task. The validity check for the environment and system should be designed based on the type of media under study.

It is recommended to use Gold Standard Questions to check attention of worker when performing their task. The gold standard question (also known as trapping question) is a question that the experimenter knows its answer. Crowdworkers shall be able to give a correct answer easily when they completely and consciously follow the test instruction. Designing of gold standard questions depends on the type of stimuli under study. It is recommended (when possible) that the gold standard questions fulfill the following conditions:

- It should not be easily recognizable unless the crowdworker follow the procedure of the test (no visual and contextual differences with other questions in the rating job)
- The effort of concealing the cheating would be as high as the effort of providing reliable answers
- It makes crowdworkers aware of the importance of their work, in order to motivate

A gold standard question may be either 1) an open-ended question or 2) worker being asked to select a specific answer from a multiple-choice question (including the opinion scale). Some examples of those questions include, but is not limited to the following:

- Captchas or computation of simple text equations (recommended for survey and qualification job). For example, "two plus 3=?", "Which of these countries contains a major city called Cairo? (Brazil, Canada, Egypt, Japan)".
- Consistency tests where experimenters ask a question at the beginning of the test, then follow-up with a question that relates to (or exact duplicate of) the first question to ensure validity (recommended for survey and qualification job). For example, experimenters can ask "In which country do you live?" followed later in the test by the question "In which continent do you live?"
- Questions about stimuli content that was displayed last to the worker. These questions could be formed around the content type, or stimuli impairments (recommended for IQA and VQA). For example, experimenters can ask the multiple choice question of "Which animal did you see in the video? (Lion, Bird, Rabbit, Fish)" after showing a video clip showing one of these animals. Another example, if a video is stalling, the question asked could be: "Did you notice any stops to the video you just watched? (Yes, No)". Note, however that such questions can only be used to check for obvious impairments in this case and not to collect a quality rating as these are subjective in nature. In addition, workers should not be able to return to the previous page and check the stimuli again.
- Workers are explicitly asked to select a specific item from scale (recommended for SQA). For example, a voice message should be recorded asking worker "Please select the answer *X* to confirm your attention now." where *X* can be any item from the opinion scale (e.g. X = Poor, or Fair in the ACR test). Five variations of this message (one per each opinion scale item) should be created. A variation of the recorded message should be appended to the first couple of seconds from randomly selected stimuli from dataset to create the *trapping stimuli set*. One or more trapping stimuli should be randomly added to the stimuli set as a normal rating question.

In addition, experimenters can monitor workers' general interactions with tasks (e.g. time it takes to answer questions). For example, to increase the number of valid results in a VQA, a warning message can be displayed to the worker, such as "you did not watch more than 70% of the video" and then allow the worker to watch the video again or continue the test. Once workers are aware of

this control mechanism, the percentage of completely watched videos will increase, and thus the number reliable answers will increase. Experimenter may consider using other user interaction parameters like reasonable rating time (not too fast, not too slow) and user focus based on browser use (user opened another tab). More on reliability scores can be found in [16].

NOTE – Experimenters must not add too many validity check questions, as otherwise the assessment task will become too lengthy and rather cumbersome to complete.

7.4.2 Execution

Similar to regulations of handling human subjects in a laboratory study, experimenters must be knowledgeable of geographical and institutional regulation. It is recommended to:

1. Obtain informed consent from participants before beginning to engage with a test
2. Maintain data confidentiality and privacy of participants
3. Explicitly state participants' right to opt out of the study at any time

In crowdsourcing, an experimenter must provide information that describes the study, the potential risks and benefits, and compensation to the crowdsourcing platform. Consent is obtained actively by requiring the worker to agree to the terms and conditions of the study. The latter is achieved when the worker clicks a button to navigate to the next page of the test. It is recommended to obtain consent up front in the qualification job.

Experimenters must not ask for crowdworkers' email addresses or any social media identifiers as it undermines participant confidentiality. Experimenters are also encouraged to implement a communication channel (e.g., via comments, a contact form, or forums) in order for workers to get in touch with them to report any test errors or issues they may face [1].

It is recommended to perform pilot testing to ensure that the experimental setup acts as expected and to avoid any inconvenience with the primary study. Meanwhile, insides about the appropriateness of job design including task duration and rewards can be collected.

7.4.2.1 Compensation

Similar to compensations given to participants in a laboratory setting, compensation in a crowdsourcing environment is determined by the requester. It directly influences the participants' motivation, and indirectly the quality of collected data. Compensation value must be relevant to the amount of time it takes the worker to complete a test and the complexity of the test itself.

It is recommended to consider extra rewards for workers who submit more rating tasks with high quality. In addition, an explicit instruction needs to be made so that workers are not feeling compelled to "please" the requesters, but rather answer honestly in accordance with their own perception.

7.4.3 Data screening

Within data-screening process reliability of submitted answers from workers should be evaluated. In case that a response to a rating task was collected in inadequate environment (e.g. noisy environment for SQA) or by using inappropriate equipment (wearing a single earpiece within SQA) that response should be discarded. If a worker fails any validity check question, it is recommended that the experimenter classifies that worker's votes as invalid and removes entire votes given by that worker from the study.

Furthermore, following user rating based screening mechanisms are recommended:

- Examine entire votes for each stimulus (or condition) for univariate outliers using box plots and standardized scores. As suggested by Tabachnick and Fidell [40], votes with absolute z-score

larger than 3.29 were considered to be potential outliers. All responses from a worker with more than one potential outlier (detected by either box plot or z-scores) should be removed.

- Calculate sample correlation coefficient between the MOS from worker *n* and the global MOS (condition level). If the calculated value is below 0.25, all votes from worker *n* should be removed [39].

For further discussions on screening mechanisms based on user ratings see [16].

7.5 Statistical analyses

The numerical mean (over subjects) should be calculated for each stimuli for initial inspection (e.g. effect of talker's gender in SQA) and then for each condition.

For each stimuli and condition, MOS values (or DMOS depending to the applied scale) should be accompanied by sufficient information to allow a basic statistical analysis to be performed, for example, the calculation of a confidence interval. For any given stimulus and condition, this information comprises the number of votes, the mean of the votes and the standard deviation of the votes.

Confidence intervals should be evaluated. Framework presented in ITU-T P.1401 [37] for the statistical evaluation of objective quality algorithms can be used. Furthermore, depending to the experiment design, further analysis can be performed using mixed models (individual differences as random effect) or significance tests performed by conventional analysis-of-variance techniques.

8 Specific Recommendations for individual media QoE

8.1 Speech

See Recommendation ITU-T P.808 [5].