Quality and User Experience (2021) 6:2
https://doi.org/10.1007/s41233-020-00042-1

RESEARCH ARTICLE

Towards speech quality assessment using a crowdsourcing approach: evaluation of standardized methods

Babak Naderi[1] · Rafael Zequeira Jiménez[1] · Matthias Hirth[2] · Sebastian Möller[1,3] · Florian Metzger[4] · Tobias Hoßfeld[4]

Received: 25 May 2020

Abstract
Subjective speech quality assessment has traditionally been carried out in laboratory environments under controlled conditions. With the advent of crowdsourcing platforms tasks, which need human intelligence, can be resolved by crowd workers over the Internet. Crowdsourcing also offers a new paradigm for speech quality assessment, promising higher ecological validity of the quality judgments at the expense of potentially lower reliability. This paper compares laboratory-based and crowdsourcing-based speech quality assessments in terms of comparability of results and efficiency. For this purpose, three pairs of listening-only tests have been carried out using three different crowdsourcing platforms and following the ITU-T Recommendation P.808. In each test, listeners judge the overall quality of the speech sample following the Absolute Category Rating procedure. We compare the results of the crowdsourcing approach with the results of standard laboratory tests performed according to the ITU-T Recommendation P.800. Results show that in most cases, both paradigms lead to comparable results. Notable differences are discussed with respect to their sources, and conclusions are drawn that establish practical guidelines for crowdsourcing-based speech quality assessment.

Keywords Speech quality assessment · Crowdsourcing · Validity · Reliability · P.808

Introduction

Quality of Experience (QoE) research concentrates on understanding user requirements towards systems or services, as well as their perceptions and judgments. Traditionally, QoE studies have addressed systems or services for multimedia content creation, transmission, and rendering. This includes systems for audio presentation, for video transmission, or for speech-based communication. In order to obtain quantitative metrics of QoE, subjective experiments are commonly conducted, in which representative groups of users judge multimedia content presented under controlled test conditions. Standardized guidelines exist for such experiments, e.g. in the Recommendations of the P-series of the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T), or in the Recommendations of the BS- and BT-series of the Radiocommunication Sector (ITU-R). These guidelines describe the requirements towards test participants, test design, set-up, procedure and analysis, as well as the laboratory environment in which tests should be carried out.

The new crowdsourcing (CS) paradigm offers access to the workforce of anonymous users of the Internet, e.g., for

✉ Babak Naderi
babak.naderi@tu-berlin.de

Rafael Zequeira Jiménez
rafael.zequeira@tu-berlin.de

Matthias Hirth
matthias.hirth@tu-ilmenau.de

Sebastian Möller
sebastian.moeller@tu-berlin.de

Florian Metzger
florian.metzger@uni-wuerzburg.de

Tobias Hoßfeld
tobias.hossfeld@uni-wuerzburg.de

[1] Quality and Usability Lab, Technische Universität Berlin, Berlin, Germany

[2] User-centric Analysis of Multimedia Data Group, Technische Universität Ilmenau, Ilmenau, Germany

[3] Speech and Language Technology, German Research Center for Artificial Intelligence (DFKI), Berlin, Germany

[4] Chair of Communication Networks, University of Würzburg, Würzburg, Germany

Published online: 22 November 2020

carrying out tasks which require human intelligence. QoE assessment is such a task for which CS has been discovered as a new means for collecting quality judgments. In a CS paradigm, Internet workers accomplish the tasks from their computer or mobile device. This way, it is possible to reach out to a large and widespread pool of diverse users, at much lower costs than it would be possible in a laboratory setting. In turn, the conditions under which experiments are carried out in the crowd are far less controlled, both with respect to the test participants (crowdworkers), the test procedure, as well as the test set-up and environment. The work described in [1] provides a good summary on the use of CS for QoE assessment.

A classical category of systems where quality assessment has played a major role for more than a century are telephony systems. In such systems, the speech signal to be transmitted can be affected, and thus degraded, by the network and terminal devices. The degradation may impact the listening as well as the interaction (conversation) situation, and can be quantified in listening-only or conversation tests. Traditionally, speech quality assessments are conducted by following a listening-only paradigm in a laboratory (lab) room that is shielded against noise and reverberations and is equipped with professional audio equipment. Methods and guidelines for the subjective assessment of speech quality with Absolute Category Rating (ACR) or comparison rating paradigms in a lab are defined in ITU-T Recommendation P.800 [2].

As they require the availability of test facilities and human participants, lab studies are rather expensive and time-consuming. Due to the artificial lab environment, small differences between speech stimuli could be quantified that would otherwise remain imperceptible under standard service usage conditions. The necessity for participants to access the test facilities limits the demographic user characteristics which can be covered in a single test. Thus, despite showing a high sensitivity and reliability of the results, lab tests might show a rather poor ecological validity, in that their results are not representative of everyday service usage.

Parts of these limitations might be overcome by making use of the CS paradigm. It is possible to access a larger and demographically balanced group of users at lower costs, however restricted to internet-connected and -affine users. Quality evaluations are commonly carried out under normal service usage conditions, with standard user equipment. This way, the ecological validity may be largely increased, however at the expense of a rather poor control over the test set-up, procedure, environment, and participants. To limit the impact of poor experimental control, the new ITU-T Rec. P.808 [3] for speech quality assessment in CS has been established, presenting guidelines to follow in order to reach comparable results.

Considering the advantages and inconveniences of both paradigms, the question arises which of the methods leads to higher validity and/or reliability of results. Whereas established methods for analyzing reliability of speech quality measurements are largely available, their validity is more difficult to quantify. Obtaining fully valid quality judgments would require having test participants which are fully representative of the later user group in all relevant aspects (perceptual characteristics, prior experience, expectations, usage purposes, etc.). Whereas CS-based assessment might come closer to this goal by making use of a larger pool of users, there is scarce open data available on the diversity of users, devices and listening situations which actually make use of telephony services. Thus, it cannot be proven that CS-based assessment has an advantage in this criterion. On the other hand, lab tests are still the most frequently used method for speech quality assessment, and thus could be seen as some type of "gold standard" for judging the validity of speech quality assessments. We will take this latter perspective in the remainder of this paper while still acknowledging the lack of data, which would allow to substantiate the validity of both approaches.

It is the aim of this paper to provide a detailed analysis of the differences and commonalities between laboratory-based and crowdsourcing-based quality assessments. For this purpose, we use the listening-only quality of transmitted speech rated in an ACR paradigm as an example, as this is the most popular rating method used in practice. Three standard speech quality assessment databases, which have been collected in three different lab environments, have been compared to three different CS experiments using the same stimuli. Each study followed the guidelines given in ITU-T Rec. P.800 for the laboratory setting, as well as ITU-T Rec. P.808 for the CS setting, thus minimizing the test differences as best as possible. The CS quality judgments are compared to the lab judgments, and differences between lab and crowd are analyzed on a per-test and per-condition basis. In addition, an estimation of the efficiency of both methods is given in terms of time and money.

The paper is organized as follows: Section "Related works" provides an overview of related work. Section "Recommended method" gives an overview of the currently recommended methods assessing speech quality in a lab or CS environment. Section "Evaluation" describes the three CS experiments. Section "Methodology" describes our approach to a quantitative comparison, and Section "Results" presents the results obtained with this methodology. Finally, Section "Discussion and future works" draws some conclusions regarding the applicability of CS for speech quality assessment, as well as regarding the test procedures to be adopted for reaching a high comparability of results.

Related works

Previous work has shown that CS can provide reliable measurements of the Quality of Experience (QoE) for a variety of multimedia applications [4]. In [5], the authors proposed a novel approach to identify image degradation through test questions and reliability checks in CS. [6] used CS as well to collect ground truth ratings of the aesthetic appeal of images. Work in [7] employed CS to evaluate users' perceived quality in video streaming applications, and [8] carried out perceptual audio evaluation tasks. Research in [9] addressed the QoE of a teleconference system via a CS and lab test. The authors found significant differences between the QoE ratings collected through both methods, and found also differences in terms of reliability and efficiency. Moreover, several bodies of research successfully employed CS to collect speech quality scores of a standard speech database (e.g. [10–12]). Due to the adoption of CS as a test paradigm, the analysis of the differences between the lab-based and CS-based tests has been of interest in multiple studies [13].

Moreover, the widespread acceptance of CS for QoE research has led to the creation of different CS frameworks to facilitate the execution of multimedia quality assessment experiments. Often, these frameworks permit closing the gap between the basic functionalities offered by the CS platforms and more complex study setups. The following paragraphs describe a few of these frameworks.

The *"Quadrant of Euphoria"* was a Web-based platform that aimed at easing the process of QoE assessment of images, audio, and video stimuli. It was first introduced and evaluated by [14]. However, authors did not use any standardized methodology (e.g., P.800 [2], BT.500-11 [15] or P.910 [16]) when implementing this platform, and no further adoption of this framework has been seen in the literature.

"crowdMOS" is another framework for audio, video, and image quality assessment studies. It was proposed at [17] as an open-source project that could be deployed easily on a web server. It implements three subjective test methodologies, ACR following P.800 [2] for speech, MUSHRA (multi-stimulus test with hidden reference and anchor) [18] for audio and ACR following P.910 [16] but for image quality assessment. "crowdMOS" was employed successfully in [17]. The authors demonstrated the validity and reliability of the framework by conducting multiple studies and contrasting the results to the Blizzard text-to-speech competition. And in [19] for image quality assessment where the results were compared to the LIVE subjective quality image dataset [20].

QualityCrowd is a Web-based CS framework, initially designed for video quality assessment [21], and later extended to support different testing methodologies as well as image and audio stimuli. This framework has been used in multiple studies varying from tasks related to image [22, 23], video [24, 25], and adaptive video streaming over HTTP [26].

Lastly, *BeaqleJS* is a JavaScript-based framework for subjective audio quality evaluations [27]. It implements two testing methodologies, i.e., MUSHRA following [18] and the so-called "*ABX*" test where three items (named A, B, and X) are presented to the listener, and X is randomly selected to be either the same as A or B.

This work aims at evaluating the Recommendation P.808 [3] in terms of reliability and validity. With this goal in mind, three frameworks were implemented by three different experimenters following the guidelines defined in P.808. These frameworks were used for conducting speech quality assessment tests with different datasets and on different CS platforms.

Recommended method

The ITU-T Rec. P.808 [3] provides advice on conducting subjective speech quality assessment tests with a CS approach. It focuses on the listening-only tests and gives details on an ACR implementation. Given the differences between CS- and lab-based experiments, the recommendation addresses the database structure, the design of the experiment, the listening-test procedure, as well as the analysis of the results. In the following a summary of each section of the recommendation is given.

Dataset structure

It should be considered that the listening devices of crowdworkers cannot be assumed as known and identical across all workers. Thus, the source material should be prepared with variability of the listening devices in mind. Besides that, there is no difference between the preparation of source materials for lab tests and that for CS tests.

Design of experiment

The recommendation focuses on two aspects, namely the requirements of the test platform and the test duration. The CS test can either be implemented using functionalities provided by the employed CS platform, or in a separate infrastructure and using the CS platform solely to recruit test participants. The CS platform should provide enough potential participants who meet the conditions specified in the next sections. Normally, a CS micro-task takes a couple of minutes to complete. Thus it is recommended to split an experiment session into a chain of tasks in the rating job,

each task a couple of minutes long (i.e., it should contain 5 to 15 stimuli). However, a crowdworker may perform just some of the tasks. Consequently, some participants may not rate the entire set of stimuli available in the database. That increases the error variance caused by individual differences. Depending on the database structure, it is recommended to either use a balanced blocks experimental design [28] or provide monetary bonuses for workers who perform a sequence of tasks, ideally evaluating 50% or more of the entire dataset. Meanwhile, collecting more votes per stimulus (compared to the corresponding lab experiment) is recommended.

Listening test procedure

Listening session

It is recommended to divide the session into three sections: Qualification Job, Training section and Rating Job. Participants should perform them in a consecutive order.

Qualification job This job should inform crowdworkers about the purpose of the study and check if they are eligible to participate in the study considering the conditions explained below. It is recommended to use the platform's built-in functionalities to make this job only accessible to workers with a solid track record and a good performance in many jobs. A randomly selected group of crowdworkers, who satisfied the prerequisites by giving their answers to this task, should be invited to participate in the experiment. Note that this job should be performed by a large group of crowdworkers to be able to screen for a target group of participants. The participants should represent the population of target group. It is recommended that they have normal hearing ability and are native speakers (or represent a native-level fluency) of the language under study.

Training job In the training job test instructions should be given to the participants and they should be asked to rate a preliminary list of stimuli on the provided scale. Stimuli presented in the training section should cover the entire range of quality to be expected in the test. No suggestion on the quality of stimuli should be made, and the order of stimuli should be randomized. After completing the training section access to the rating job should be temporarily granted to the worker. The job provider may consider to integrate the training job into the rating job as an additional section which will only be shown when needed. Ideally, the training should be repeated after 60 minutes.

Rating job In the rating job, first the setup of the crowdworker should be evaluated including listening environment, system and audio volume. Then crowdworkers should be asked to listen to a set of stimuli and rate them. The number of stimuli in each rating session should be limited (e.g. 10 to 15 stimuli) to avoid long sessions. The stimuli set should contain gold standard questions which are designed following the recommendation. The order or presentation of stimuli in one set should be randomized on the presentation time and it is recommended to load the entire stimuli set in advance.

Listening environment, system, and level

Crowdworkers should be asked to perform their task in a quiet and non-distractive environment. Three approaches are suggested to test if participants follow the instructions: 1. use microphone signal to estimate the sound pressure level in the environment, 2. ask the workers to rate the background noise in their environment, 3. ask crowdworkers to select speech samples with better quality from pairs of samples (pairs represent a marginal difference in quality which can only be recognize in a suitable environment).

In addition, participants should wear a binaural headphones unless the experimenters decide otherwise. The usage of binaural headphones shall be validated in the beginning of each rating job. Meanwhile, workers should be asked to set the volume of their listening device to a comfortable level when listening to a sample speech file. Afterwards they cannot change it.

Gold standard question

A gold standard question is a question which has an answer known to the experimenter. Such a question should not be visually different from other questions. The effort for a crowdworker to conceal cheating should be as high as the effort of providing a reliable answer. For the speech quality assessment task it is recommended to create a set of trapping stimuli and randomly use one of them in each test set. In order to create a trapping stimulus, a message should be recorded with a speaker who otherwise did not partake in the provided speech material. It is recommended to use a message like "*This is an interruption. Please select the answer X to confirm your attention now.*" where X can be an item from the opinion scale (e.g., X = bad, or fair in the ACR test). Five variations of this message (one for each opinion scale item) should be created. A trapping stimulus will be created by appending a variation of the message to the first seconds of a randomly selected stimulus from the dataset under test (see [29] for more details about trapping questions and their effects).

Data analysis and reporting of results

In the data screening step unreliable submissions should be removed. They include all submissions in which the listening environment was not suitable, or the listening system is not used as expected, or one or more gold standard questions are answered incorrectly. Furthermore, the remaining

Table 1 Selected datasets from the ITU-T Rec. P.863 competition pool, used for evaluation

	Dataset 401	Dataset 501	Dataset 701
Title	Psytechnics P.OLQA SWB 1	SwissQual P.OLQA SWB 1	DOLBY
Creation year	2008	2008	2013
Test method	P.800, ACR		
Number of conditions	48	50	72
Files per condition	24	4	16
Votes per file	8	24	8
Votes per condition	192	96	128
Listeners	32	24	32
Design	6 talkers (3m, 3f)	4 talkers (2m, 2f)	4 talkers (2m, 2f)
Language	British English	German, Swiss pronunciation	American English
# of samples	1152	200	1152
Listened through	Sennheiser HD25-1	Grado SR60	Sennheiser HD 600
Presentation level	73 dB(SPL) at ERP for −26 dBov signals		

submissions should be examined for unexpected patterns (e.g. no variance in ratings, or potential outliers) and unexpected user behaviours (e.g. very short working time). Finally, the MOS values for each stimulus and condition should be reported followed by the number of votes and the standard deviation.

Evaluation

In the following, three studies are described which were conducted to evaluate the ITU-T P.808 Recommendation [3]. First, the datasets used in the studies are explained, and each crowdsourcing test is described in detail. Afterwards, the results of the comparisons between crowdsourcing approach and the laboratory experiments are presented.

Datasets

Access to the pool of the ITU-T Rec. P.863 [30] competition datasets was kindly provided for this study. From the datasets available in the pool three were selected, namely 401, 501 and 701, each with a different language and study design, each includes variable types of degradations and degradation combinations. Each set was prepared on the basis of the ITU-T Rec. P.800 [2] specification. Table 1 summarizes the source materials and laboratory-based ratings provided by the corresponding contributors.

Crowdsourcing test

Three experimenters conducted the crowdsourcing tests, each using one of the above-mentioned datasets and following the ITU-T Rec. P.808. In the following, each test is described in detail.

CS 401

This experiment was conducted using the Amazon Mechanical Turk[1] (MTurk) platform. MTurk is a well-known crowdsourcing platform and provides a globally distributed crowdsourcing workforce.[2] However, as reported in 2016 [31], workers based in India and the USA make up the largest share. In this study the platform's own infrastructure is used. Jobs were designed with HTML and JavaScript. As proposed by the recommendation, a multi-step job design was followed.

Qualification job As suggested in the recommendation, platform-provided conditions were used to make sure that only workers with a history of good performance could participate in this job (i.e. overall approval rate > 98%, and accepted jobs > 500). In addition, only workers from the US were able to perform this job and no language screening tests were employed. For listening impairment test, the adapted version of a digit-triplet test [32] was used. There, five stimuli with a signal-to-noise ratio (SNR) of −11.2 dB were used. Workers were instructed to listen to each stimulus and type in the three numbers they heard. They could listen to each stimulus as many times as they wanted (the number was recorded for further analysis). The SNR of −11.2 dB was chosen to reach a high true positive rate, while previous studies suggested to use a threshold of −9.3 dB SNR for German, −11.2 dB SNR for Dutch, and −10.5 dB SNR for French digit-triplet tests in order to find normally hearing participants [33]. From 327 participants, 133 workers were not eligible to continue because of the result of the hearing

[1] https://mturk.com/.

[2] Our attempt to reach German speaking crowdworkers in Feb. 2018 was unsuccessful.

test, and 7 other workers were removed for other reasons like self-reported hearing impairment and inadequate listening devices. From the remaining 187, 100 were randomly selected (following the proposed distribution of gender and age in the draft recommendation) and given access to the next job.

Training job As the temporal qualification condition was not supported by this crowdsourcing platform, it was decided to merge the training job and rating job into one single job and implement a temporal qualification using browser cookies. As a result, the training job was a section inside the rating job, which became visible or hidden depending on the value of the browser's cookie. Each time the training section was shown, the value was updated accordingly.

Rating job 10 stimuli (and one trapping stimulus) were assessed in one rating task. Each task was compensated with US$0.50 and additional bonuses conditional to the quality of the worker's performance and the quantity of their work. In the training section five pre-selected stimuli were presented to the workers. The quality of these stimuli covers the entire range of the scale. The training certificate (cookie) expired after one day. For the environmental test we used the modified Just-Noticeable-Difference in Quality (JNDQ) method detailed in [34]. We selected four pairs of stimuli in the way that each pair was 0.6 MOS apart (as assessed in the laboratory test). In each question, the worker should select the stimulus with better quality from the presented pair (or state that there is no detectable difference). The threshold of 0.6 MOS was selected on the basis of the results of laboratory experiments using an adaptive psychophysical (staircase) method, where normally hearing subjects participated in a study that was conducted in a typical, calm room condition and using a common listening device [34]. In addition, a math question was used to check if workers were able to correctly listen to stereo sound. Workers were also forced to listen to each stimulus until the end before casting their vote. Workers were able to listen to each stimulus as many times as they wished, the number of repeats was recorded. The dataset contains 1152 (48×24) files, divided into 116 sets of stimuli (each containing 10 randomly selected files). It was planned to collect 10 votes per set.

Data screening In total, 1160 response packages (each with 10 votes) from 71 unique workers were collected. From them, 12 response packages with a wrong answer either to the math question or to the trapping questions were removed. In addition, 106 other response packages were removed as workers failed in the environmental screening test. Overall, 10,420 votes from 68 workers were accepted and used for further analyses.

CS 501

For this crowdsourcing test, we used *clickworker*[3], a German-based crowdsourcing platform with most workers from Germany, Austria, and Belgium. Clickworker did not support audio playback (as of April 2018). Thus, an HTML JavaScript-based framework[4] was implemented to administer the test to the workers, combined with a Node.js server for data collection.

The crowdsourcing study was designed with ITU-T Rec. P.808 in mind and contained three phases, i.e., *Qualification*, *Training*, and *Rating*. In the following, we outline the differences in each of the phases to those defined in the Recommendation.

Qualification job The *Qualification* phase in our crowdsourcing study included three German spoken passages to evaluate the workers' knowledge of the German language. Once they heard each audio file, they were asked to select the correct statement (related to the passage heard) out of three that were provided.

Training job Similarly to the previous test, the training job started with a stereo test. A short math exercise where the spoken digits panned between the left and right audio channels was used. In addition, workers were presented with a short hearing impairment test. 10 different audio files were created containing white noise at different frequencies ranges. Listeners were presented with four-octave filters around the frequencies 500 Hz, 1 KHz, 4 KHz, and 8 KHz, at a level of −46 dBov, −66 dBov and −76 dBov (just for 1 KHz and below). Workers were then asked whether they heard anything in each audio file. Furthermore, they listened to five speech stimuli (taken from the dataset) that covered the entire MOS range, so they could get to know what to expect on the rating part and become familiar with the scale. After correctly answering the math trapping question, workers were assigned an hour-long time frame in which they could perform the rating job.

Rating job The rating phase included 15 stimuli. Work in [35] points out that it is desirable to offer tasks with fewer speech stimuli in order to increase listener retention and decrease the study turnaround time. Additionally, one trapping question (created as explained in the P.808 Rec.) was inserted randomly within the first five stimuli and one between the 10th and the 15th speech sample. After listening to all stimuli, workers were asked to state through a slider, how fatigued they felt. When listeners failed any of the trapping questions, the access to the rating job was revoked. Also, we recorded environmental background noise when workers played the first and the ninth sample

[3] https://www.clickworker.com.

[4] https://gitlab.com/zequeira/SQAT-Cr.git.

(each 7.5 seconds long). Workers were not able to provide their opinion on the scale unless they first listened to the speech sample. They could not go forward until the audio was played completely and an option selected on the scale. And they could listen to each speech sample as many times as they wished.

Data screening We collected a total of 5245 ratings from 64 unique listeners. All of them answered the trapping question correctly. No crowdworker was removed because of the hearing test as:1. there was no guarantee that workers reported "hearing" a noise actually heard it, and 2. hearing or not hearing the noise could be due to the volume level of the device. In addition, 136 ratings were identified as extreme outliers (beyond an outer fence of a boxplot) and removed.

CS 701

The third study was conducted on the Microworkers platform.[5] Microworkers is an international crowdsourcing platform with about 1.5 million workers. Similarly to Clickworkers, Microworkers did not directly support tasks with audio playback at the time of the study in August 2018. Therefore, an external Web page was created and used for the tests. To point the worker to the Web page, we included a link to the external test page in the task description. This link is personalized by the platform, so that we were able to track the workers. Personalized payment codes were generated on the external page to avoid unauthorized task submissions on the platform. The external web page was designed according to the ITU-T P.808 Recommendation as follows.

Qualification job Instead of a qualification job with audio stimuli, this study used a self-assessment of the workers. The workers provided feedback on their hearing capabilities (The choices were: "I have a normal hearing ability.", "I have difficulties keeping up with conversations, especially in noisy surroundings.", "I have difficulty keeping up with conversations when I am not using a hearing aid.", and "I rely on lip-reading even when I am using hearing aids.") and the surrounding noise ("Not noticeable.", "Slightly noticeable.", "Noticeable but not intrusive.", "Somewhat intrusive.", "Very intrusive."). All workers were allowed to participate in training and rating jobs, independent of their survey results, for two main reasons. First, this enables us to also collect data from users with potential hearing impairments or in noisy environments and use this data in later studies. Second, it is well known that groups of workers talk amongst each other and exchange information. Therefore, workers would be encouraged to provide incorrect answers, if they are aware that previous workers got excluded from the listening task based on their survey answers. However, for the evaluation in this paper we only used workers with normal hearing abilities and no background noise or background noise that was only slightly noticeable.

Training job After successfully completing the qualification job, workers continued with a short training job. Here, workers had to adjust their listening volume to a comfortable level and listen to two test samples to familiarize themselves with the rating interface. Additionally, workers were requested to sum up three different numbers that were spoken in the sample. The numbers alternated between the left and right stereo channel to ensure the workers' playback device supported stereo audio. After completing the qualification and training job, workers were allowed to complete up to 10 ratings jobs. The workers were only requested to repeat the qualification and training job after more than one hour had passed.

Rating job During the rating job, the workers had to rate 10 stimuli and complete one attention check. The attention check appears to be a regular stimuli but the audio recording states that this is an attention test and despite of the current quality a predetermined rating should be submitted. We ensured that the rating to be submitted does not match the actually perceived quality of the audio file, e.g., the quality of the recording was of low quality but the spoken instruction told the worker to rate it as perfect. Workers hard to listen to the complete stimulus before submitting a rating but were allowed to listen to a stimulus multiple times. Each worker could complete 10 rating jobs at maximum.

Data screening In total, 197 workers successfully completed the tasks on Microworkers and completed 1032 sessions, each with 10 stimuli. We excluded 5 workers due to self reported hearing impairments. Besides those workers, no further workers had to be excluded due to noisy environments. However, 141 workers provided incorrect answers to the attention check and the stereo audio test. A closer look at the answers revealed that a large number of workers did not sum up the digits in the stereo test but instead individually entered the digits. As this points towards a general issue with the phrasing of this reliably check, we decided to include all workers that either provided the correct sum of the digits or the correct sequence of digits. With this constraint we removed 48 workers who failed the attention and the audio test.

Table 2 summarizes the crowdsourcing tests discussed above and gives statistics about the ratings per stimulus.

Methodology

The variables and notions on the collected user ratings in the lab, via crowdsourcing and via simulations are introduced in Table 3. In particular, the notion on MOS

[5] https://microworkers.com.

Table 2 Summary of conducted crowdsourcing tests

	CS 401	CS 501	CS 701
Original dataset	401	501	701
Crowdsourcing test			
Experimenter group	QUL1	QUL2	WUE
Year	2018	2018	2018
Crowdsourcing platform	MTurk	Clickworker	Microworkers
External framework	No	Yes	Yes
Duration	3 days	11 days	1 day
Number of crowd workers[6]	68	64	144
Votes per file (M/SD)	9/1.2	25.5/3.5	6.1/2.1
Votes per condition (M/SD)	217/4.8	102.2/7.3	97.1/9.0
Votes by CS worker (M/SD)	146.6/139	79.8/48.5	48.5/31.4
Files rated in one session	10	15	10
Method of CS test			
Workers were checked for...			
Being a native speaker	Filtering by location	German test	Self assessment
Listening impairments	Asking and digit triple test (threshold −11.2 dB SNR)	Web hearing test (white noise at different dB)	Self assessment
Environment	Modified JNDQ	Environment background noise recording	Self assessment
Further validity check method			
Attention questions per task	1	2	1
Sessions removed (outliers)	118	136	333

[6]Workers with one or more accepted responses to the rating job.

Table 3 Notations for the collected user ratings in the lab, crowdsourcing and simulations

Variable	Meaning
r	Number of simulation runs, it is $r = 1000$ for the numerical results
Ω	Methodology of the user rating collection; $\Omega \in \{\text{lab}, \text{cs}, i\}$ means in the lab, via crowdsourcing or via simulation, respectively, whereby i indicates the ith simulation run ($i = 1, \dots, r$)
$\mathcal{U}^{\Omega}$	Set of all users rating via Ω
$\mathcal{U}^{\Omega}_{x}$	Set of all users rating condition x via Ω
k	Number of different conditions
$\mathcal{X}^{\Omega}_{u}$	Ordered set of conditions rated by user u via Ω, it is $\mathcal{X}^{\Omega}_{u} \subseteq \{1, \dots, k\}$
$N^{\Omega}_{x,u,q}$	Number of ratings on the scale $q \in \{1, \dots, 5\}$ by user u for condition x collected via Ω
$N^{\Omega}_{x,u}$	Number of ratings by user u for condition x collected via Ω, i.e. $N^{\Omega}_{x,u} = \sum_{q=1}^{5} N^{\Omega}_{x,u,q}$
$M^{\Omega}_{x,u}$	MOS of user u for condition x via Ω; it is $M^{\Omega}_{x,u} = \sum_{q=1}^{5} q \cdot N^{\Omega}_{x,u,q} / N^{\Omega}_{x,u}$
M^{Ω}_{x}	MOS over all users u for condition x via Ω; it is $M^{\Omega}_{x} = \frac{1}{\lvert\mathcal{U}^{\Omega}_{x}\rvert} \sum_{u \in \mathcal{U}^{\Omega}_{x}} M^{\Omega}_{x,u}$
$\mathbf{M}^{\Omega}$	MOS vector for all conditions $x = 1, \dots, k$ via Ω; it is $\mathbf{M}^{\Omega} = (M^{\Omega}_{1}, \dots, M^{\Omega}_{k})$
$\mathbf{M}^{\Omega}_{u}$	MOS vector of user u for conditions $\mathcal{X}^{\Omega}_{u}$; it is $\mathbf{M}^{\Omega}_{u} = \left(M^{\Omega}_{x,u}\right)_{x \in \mathcal{X}^{\Omega}_{u}}$

values are introduced, which are the basis for several evaluation metrics. Section "Simulation: sampling of CS ratings" describes the methodology for simulating user ratings by sampling from the underlying crowdsourcing studies. Then, the evaluation metrics are introduced in Section "Evaluation metrics" to analyze the impact of the number of user ratings on validity, certainty gain, and inter-rater reliability. A summary of the notation of those metrics is provided in Table 4.

Table 4 Notations for the metrics for evaluation

Variable	Meaning
notation for any metric ψ	
n	Sample size of simulation, $n \in \{10, \dots, 200\}$
$\psi_i(n)$	Metric of simulation run i based on n samples from the full crowdsourcing dataset
$\psi(n)$	Average metric calculated over all simulation runs, $\psi(n) = \frac{1}{r}\sum_{i=1}^{r}\psi_i(n)$
$\widehat{\psi}(n)$	Power model for the metric ψ, $\hat{\psi}(n) = a \cdot n^b + c$ with parameters a, b, c
$\psi^*(n)$	Change in metric compared to minimum sample size $n_0 = 10$, it is $\psi^*(n) = \psi(n) - \psi(n_0)$
$\psi'(n)$	Derivation of the metric based on power model, $\psi'(n) = abn^{b-1}$
$\widetilde{\Psi}(n)$	Normalized derivation $\widetilde{\Psi}(n) = -bn^{b-1}/n_0^b$
validity metrics	
$\rho_i^{\text{lab}}(n)$	SRCC of the lab MOS and simulation MOS
$\delta_i^{\text{lab}}(n)$	RMSE of lab MOS and simulation MOS
$\nu_i^{\text{lab}}(n)$	RMSE of the lab MOS and the first order mapped MOS
certainty gain metrics	
$\rho_i^{\text{cs}}(n)$	SRCC of the CS MOS and simulation MOS
$\delta_i^{\text{cs}}(n)$	RMSE of the CS MOS and simulation MOS
$\Delta_i(n)$	EMD of simulation run i and CS rating distributions, averaged over all conditions
0.15	confidence interval width (CIW) averaged over all conditions
inter-rater reliability (IRR)	
$\kappa_i(n)$	SRCC between MOS of a user and the MOS of the other users, averaged over all users

Notation: user ratings and MOS values

The collected ratings in each of the above-mentioned subjective tests (i.e. 401, 501, 701) are obtained in the lab as well via crowdsourcing. Then, $N_{x,u,q}^{\Omega}$ reflects the number of ratings on the scale $q \in \{1, \dots, 5\}$ given by user u for the degradation condition x. Thereby, Ω indicates the used methodology for the data collection which are subjective experiments in the lab, via crowdsourcing or via simulations, $\Omega \in \{\text{lab}, \text{cs}, i\}$, with i denoting the i-th simulation run. The simulations are repeated $r = 1000$ times and $i \in \{1, \dots, r\}$. The set of all users rating is $\mathcal{U}^{\Omega}$ and k conditions are rated. Thus, $N_{x,u,q}^{\Omega}$ is provided for $q \in \{1, \dots, 5\}$, user $u \in \mathcal{U}^{\Omega}$ for the degradation condition $x \in \{1, \dots, k\}$.

MOS per user

The **MOS of a user** u for condition x via Ω is

$$M_{x,u}^{\Omega} = \sum_{q=1}^{5} q \cdot \frac{N_{x,u,q}^{\Omega}}{N_{x,u}^{\Omega}} \tag{1}$$

which is based on the number of ratings $N_{x,u}^{\Omega}$ by user u for condition x collected via Ω, i.e. $N_{x,u}^{\Omega} = \sum_{q=1}^{5} N_{x,u,q}^{\Omega}$.

For each user u, the ordered set of conditions rated by u via Ω is $\mathcal{X}_u^{\Omega} \subseteq \{1, \dots, k\}$. Users typically only rate a subset of all conditions, but may rate the same condition several times. To be more precise: in the lab and crowdsourcing tests on one hand a user may rate the same condition several times (i.e. ratings for different stimuli but which refer to the same degradation condition) and on the other hand, a user may not rate a certain condition at all.

The **vector of MOS values of a user** u for conditions $\mathcal{X}_u^{\Omega}$ is as follows.

$$\mathbf{M}_u^{\Omega} = \left(M_{x,u}^{\Omega}\right)_{x \in \mathcal{X}_u^{\Omega}}. \tag{2}$$

MOS over all users

The **MOS over all users** u for condition x via Ω is obtained by averaging over all users that rated that condition.

$$M_x^{\Omega} = \frac{1}{\left|\mathcal{U}_x^{\Omega}\right|} \sum_{u \in \mathcal{U}_x^{\Omega}} M_{x,u}^{\Omega}. \tag{3}$$

The **MOS vector** is then composed of all conditions $x = 1, \dots, k$.

$$\mathbf{M}^{\Omega} = (M_x^{\Omega})_{x \in \{1,\dots,k\}} = (M_1^{\Omega}, \dots, M_k^{\Omega}). \tag{4}$$

Simulation: sampling of CS ratings

Based on the above-mentioned subjective data collected via crowdsourcing, simulations were conducted by randomly

sampling with replacement from the collected votes. For each degradation condition x, a fixed number of votes, $n \in \{10, \ldots, 200\}$, were sampled in two steps. First, n users were drawn following the empirical probability distribution $P(U = u|x)$, which, given the condition x, estimates a probability that the user u provides a rating for condition $x \in \{1, \ldots, k\}$:

$$P(U = u|x) = \frac{N^{\text{cs}}_{x,u}}{\sum_{u \in \mathcal{U}^{\text{cs}}_x} N^{\text{cs}}_{x,u}}, \text{ for } u \in \mathcal{U}^{\text{cs}}_x. \tag{5}$$

Then, for each selected user u and condition x, the individual vote is sampled from the user rating distribution of user u for condition x:

$$P(Q = q|x, u) = \frac{N_{x,u,q}}{\sum_{q=1}^{5} N_{x,u,q}}, \text{ for } q \in \{1, \ldots, 5\}. \tag{6}$$

The simulation m was repeated $r = 1000$ times, and the mean and 95% confidence interval (CI) of each metric were calculated for further evaluation. The confidence intervals were too small to be visible in the resulting figures due to the large number of simulation runs and are therefore omitted. For example, the average width of the CI is only 0.0007 when comparing the validity of the lab and the crowdsourcing MOS results in Figure 3a.

Evaluation metrics

For the evaluation of the impact of the number of samples, different metrics are proposed which are categorized into three categories. *Validity metrics* compare the sampled crowdsourcing results for a fixed number of samples per condition with the results from the lab study. Spearman rank correlation coefficient (SRCC) and root mean squared error (RMSE) are used for the quantification. *Certainty gain metrics* evaluate the impact of the number of samples when comparing it to the full crowdsourcing dataset. Again, SRCC and RMSE are used. Furthermore, the distributions are compared in terms of Earth Mover's Distance (EMD). Another certainty metric is the length of the confidence intervals of the MOS values of the simulations, which are solely based on the sampled ratings. *Inter-rater reliablity metrics* (IRR) are assessing the SRCC of the rating of a user with the rating of the other users. Before those metrics are formally defined the general notion for an arbitrary metric is introduced.

Notion for metrics

In the following, a metric ψ is considered. Then, $\psi_i(n)$ considers the simulation run i based on n samples from the full crowdsourcing dataset. Averaging over all r simulation runs for a fixed number n of samples is denoted as

$$\psi(n) = \frac{1}{r} \sum_{i=1}^{r} \psi_i(n). \tag{7}$$

The function $\psi(n)$ is fitted with a *power model* of the form

$$\hat{\psi}(n) = a \cdot n^b + c \tag{8}$$

with parameters a, b, c. Table 7 shows the power model parameters for the metrics investigated in this paper. The goodness of fit (GoF) in terms of RSME between the simulation data $\psi(n)$ and the power model $\hat{\psi}(n)$ as well as the coefficient of determination R^2 show an excellent fit. All RMSE values are close to zero, while R^2 is close to one. Hence, the power model is an excellent fit and in the result figures, almost no differences can be observed.

For the analysis of the impact of the sample size on the evaluation metrics, the absolute values are often of minor interest. In particular, the *change in the metric* compared to the minimum sample size allows a comparison between the different studies (401,501,701) with respect to the impact of the number of samples per condition. This normalized metric is defined by using the minimal sample size $n_0 = 10$ and the corresponding power model at that point as

$$\psi^*(n) = \psi(n) - \hat{\psi}(n_0). \tag{9}$$

Applying the power model leads to this change in metric, $\hat{\psi}^*(n) = a(n^b - n_0^b)$. Hence, the change in metric only depends on the parameters a and b.

For assessing how many samples are required the *derivation* of metric is of interest, as the derivation quantifies the change in the metric for a sample size of n. For the derivation, the power model is used and yields

$$\psi'(n) = abn^{b-1}. \tag{10}$$

The derivation will be mainly used for deriving the required sample size across all metrics in "Comparison across the metrics" section.

Validity metrics

The validity metrics compare the simulation results, which are based on sampling ratings from the crowdsourcing dataset, with the lab results. The validity metrics consider the lab results to be the ground truth. Please note that the term "validity" must not be misinterpreted in that sense, that the metric demonstrate the validity of the crowdsourcing data. However, the term expresses the intention to compare the sampled crowdsourcing data with the lab data that is assumed to be ground truth.

As metrics, the SRCC and the RMSE of the MOS values $\mathbf{M}^{\text{lab}}$ from the lab studies and the MOS values $\mathbf{M}^i$ of the i-th simulation run when using a fixed sample size n are used. As explained in Sect. 5.3.1, the expected SRCC $\rho^{\text{lab}}(n)$ is computed by averaging over the r simulation runs and is used in the numerical results and figures.

$$\rho_i^{\text{lab}}(n) = \text{SRCC}(\mathbf{M}^i, \mathbf{M}^{\text{lab}}). \tag{11}$$

$$\rho^{\text{lab}}(n) = \frac{1}{r}\sum_{i=1}^{r} \rho_i^{\text{lab}}(n). \tag{12}$$

Accordingly, the RMSE between the lab MOS values and MOS values of the i-th simulation run is defined as

$$\delta_i^{\text{lab}}(n) = \text{RMSE}(\mathbf{M}^i, \mathbf{M}^{\text{lab}}) \tag{13}$$

For the computation of the RMSE, first order mappings are often used [36]. The first order mapping f is obtained via linear regression between the crowdsourcing MOS values $\mathbf{M}^{\text{cs}}$ and the lab MOS values $\mathbf{M}^{\text{lab}}$. The first order mapping functions are provided in Table 6. Then, the RMSE of the first order mapped MOS values of the the i-th simulation run when using sample size n is

$$v_i^{\text{lab}}(n) = \text{RMSE}(f(\mathbf{M}^i), \mathbf{M}^{\text{lab}}). \tag{14}$$

Certainty gain metrics

The certainty gain metrics compare the simulation results with the full crowdsourcing dataset. Hence, the full crowdsourcing dataset is considered as ground truth. The evaluation considers how much gain is obtained from having various sample sizes. The SRCC and the RMSE of the simulation MOS values $\mathbf{M}^i$ and the crowdsourcing MOS values $\mathbf{M}^{\text{cs}}$ are computed.

$$\rho_i^{\text{cs}}(n) = \text{SRCC}(\mathbf{M}^i, \mathbf{M}^{\text{cs}}) \tag{15}$$

$$\delta_i^{\text{cs}}(n) = \text{RMSE}(\mathbf{M}^i, \mathbf{M}^{\text{cs}}). \tag{16}$$

Going beyond MOS values only, the distance between distributions is analyzed. To this end, the Earth Mover's Distance (EMD) is used. For two discrete probability distributions X and Y, the EMD is based on the corresponding cumulative distribution function (CDF), $X(j) = P(X \leq j)$ and $Y(j) = P(Y \leq j)$.

$$\text{EMD}(X, Y) = \sum_j |X(j) - Y(j)|. \tag{17}$$

For a condition x, the user rating distribution of simulation run i and the full crowdsourcing dataset is referred to as Q_x^i and Q_x^{lab}, respectively. The average EMD of simulation run i for a fixed sample size n is computed by averaged over all k conditions.

$$\Delta_i(n) = \frac{1}{k}\sum_{x=1}^{k} \text{EMD}(Q_x^i, Q_x^{\text{lab}}). \tag{18}$$

Finally, certainty of the results is expressed in the terms of the Confidence Interval Width (CIW) for a level of significance, $\alpha = 0.05$. Then, the CIW is averaged over all k conditions in simulation run i based on n samples,

$$w_i(n) = \frac{1}{k}\sum_{x=1}^{k} w_{i,x}(n). \tag{19}$$

with $w_{i,x}(n)$ being the CIW for condition x in simulation run i. The confidence intervals are obtained using bootstrapping as recommended in [37].

Inter-rater reliability metric

As inter-rater reliability (IRR) metric the SRCC between the MOS of a user and the MOS of the other users is derived. Hence, for each user an SRCC value is obtained. Then, the average over all users is computed. The formal definition is as follows. The SRCC between the MOS $M_{x,u}^i$ of user u and the MOS of the remaining users $M_{x,U\setminus u}^i$ over all conditions $\mathcal{X}_u^i$ is denoted as $\kappa_{i,u}(n)$ for simulation run i based on n samples. The inter-rater reliability for run i is then averaged over all users.

$$\kappa_{i,u}(n) = \text{SRCC}(\mathbf{M}^i_{\mathcal{X}_u^i,u}, \mathbf{M}^i_{\mathcal{X}_u^i,U\setminus u}) \tag{20}$$

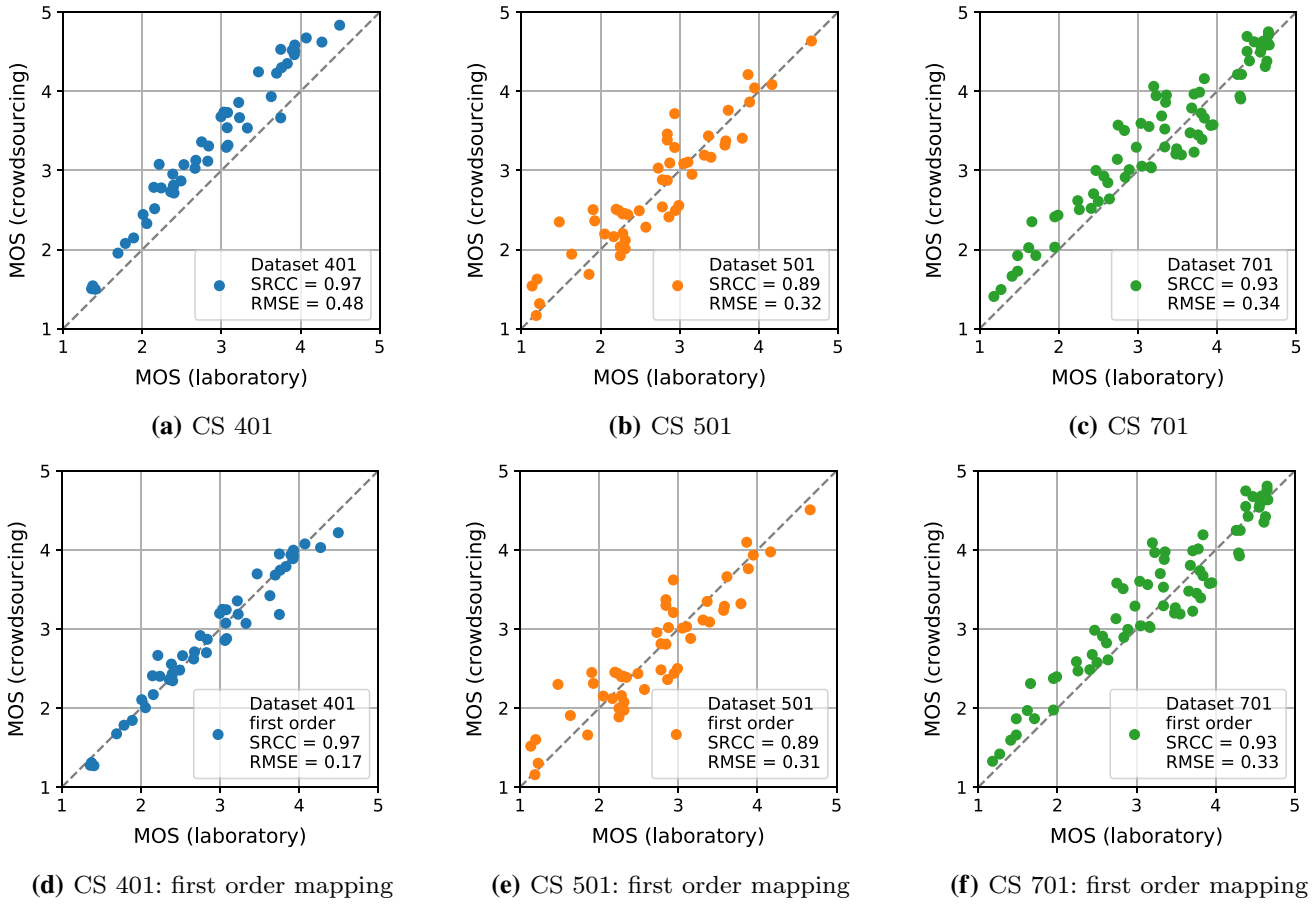
$$\kappa_i(n) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \kappa_{i,u}(n) \tag{21}$$

Results

In this chapter, we compare the lab and CS speech quality assessment tests from various perspectives. First, we compare the MOS values and SOS (Standard deviation of Opinion Scores [39]) plots from the results of both approaches. Next, we look at the validity, certainty gain, and reliability of crowdsourcing MOS as a function of the number of votes. Here we use simulations and matrices mentioned in Chapter 5. Some parts of these results have been published in [40]. In this paper, we extend them by considering larger simulation runs, more QoE metrics, and a method for aggregating result of all metrics. Later, we investigate which degradation conditions lead to a different result in the CS approach compared to the lab. And finally, we look at the efficiency of both approaches.

Table 5 Comparison between MOS values obtained in Crowdsourcing study with MOS values reported by Laboratory study using a same dataset

Dataset	#cond.	#users	avg./min. #votes per cond.	Total #votes	IRR	SRCC	SRCC TF [38]	PCC	RMSE	RMSE 1^{st} order
401	48	71	217/207	10412	0.795	0.971	0.981	0.98	0.480	0.167
501	50	64	102/83	5109	0.745	0.891	0.891	0.921	0.321	0.313
701	72	144	97/83	6990	0.777	0.930	0.919	0.949	0.337	0.332



(a) CS 401 **(b)** CS 501 **(c)** CS 701

(d) CS 401: first order mapping **(e)** CS 501: first order mapping **(f)** CS 701: first order mapping

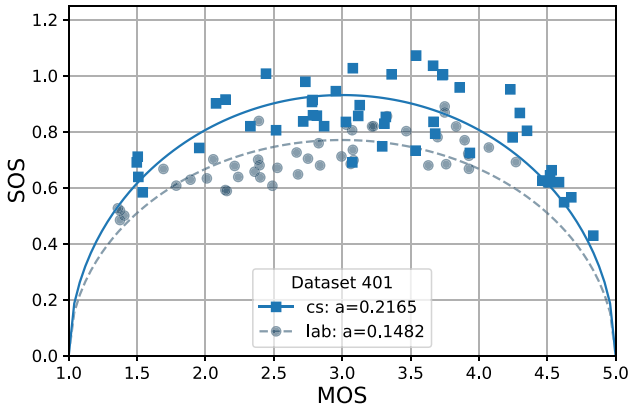
Fig. 1 Comparison between MOS values calculated per degradation conditions **a** CS401, **b** CS501, **c** CS701

Table 6 First order mapping functions adjusting the crowdsourcing MOS values to the lab MOS values based on a linear regression

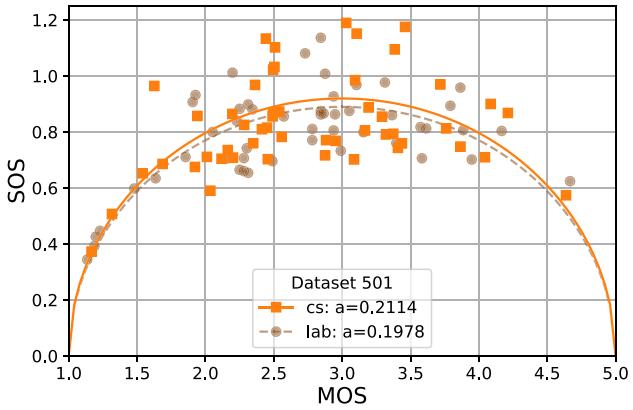
Dataset	First order mapping
401	$f(x) = 0.882347 \cdot x - 0.049968$
501	$f(x) = 0.964743 \cdot x + 0.032806$
701	$f(x) = 1.039645 \cdot x - 0.134917$

Comparison of lab and CS results

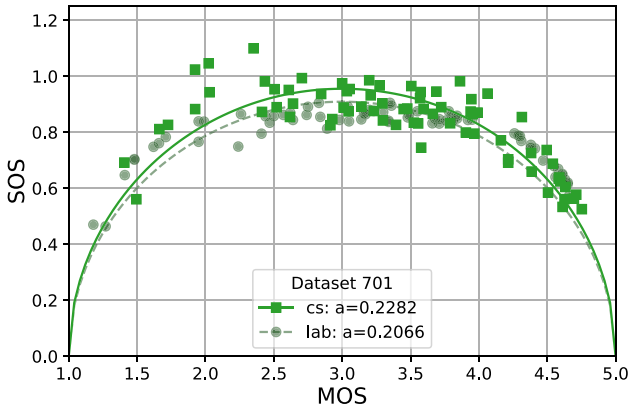
For each CS test, subjective mean opinion scores (MOS), standard deviations, and 95% confidence intervals were calculated per stimulus and condition given the accepted votes. The MOS values per condition obtained from the CS tests were compared with the values provided from the corresponding laboratory-based experiments (cf. Table 5). Results show that there is a high correlation between MOS values obtained through CS tests and those provided by the lab test, $Median = 0.930$ (cf. Figure 1). The RMSEs are also in acceptable range, $Median = 0.337$. The first-order mapping of MOS values significantly reduces the RMSE of study CS 401, ($\Delta RMSE = 0.313$). Figure 1a also shows that for CS 401, there is a bias and a different gradient between CS and lab values. For the CS 501, more deviation was observed for conditions with low to medium MOS values in contrast to the CS 701 where more variation was found in the middle to higher range of MOS values. Note that for dataset CS 401 we



(a) CS 401



(b) CS 501



(c) CS 701

Fig. 2 SOS-MOS plots and corresponding SOS parameters for the lab and the CS experiments

have more than twice the number of votes per condition. In all CS studies we found a large inter-rater reliability amongst participants, *Median* = 0.777, but not as large as lab tests, *Median* = 0.852.

SOS reflects the level of rating diversity. Following the SOS hypothesis [41] there should be a square relationship between the MOS and the SOS. Figure 2 illustrates the MOS-SOS plots (note that for CS401 no mapping is used). For all CS studies very narrow SOS $\alpha \in [0.2114, 0.2165]$ were observed. In contrast the laboratory ratings show different distributions. Ratings of the dataset 401 show the lowest standard deviation on average amongst all lab studies.

Comparison of the quality of measurement for various numbers of votes

In the lab studies, a different number of votes per condition was collected due to various study designs. However, in the CS experiments a mixed study design is applied. Therefore, we investigate the influence of the number of valid votes per condition on the quality of CS measurements by employing the simulation techniques explained in Chapter 5. We fitted power models that represent our simulation data for each metric. The coefficients of power models for all metrics are provided in Table 7. The column *GoF* indicates the Goodness of the Fit in terms of RMSE between the measurement data and the estimation of power model, while R^2 is the coefficient of determination which is the fraction of the total sum of squares that is explained by the regression.

Validity

We compared the SRCC of MOS values from lab tests with the MOS values of CS tests with various numbers of votes per condition in CS (cf. Fig. 3). We also normalized the coefficients by reporting the changes in SRCC since the starting points ($\rho^{lab}(10)$) varied for each CS test. Figure 3b shows that a coefficient increase of nearly 0.04 can be achieved by raising the number of votes per condition from 10 to 200. The SRCC of 401 and 701 datasets grew similarly, whereas the 501 gained more coefficient when the number of votes increased. The SRCC derivation plot (Fig. 3c) flattens at about 100 votes, suggesting that any change after that is not tangible. We observed identical behaviour of deviation function in all the other metrics as well.

Meanwhile, we investigated the RMSE between MOS values from lab tests and MOS values of CS tests with various numbers of votes per condition ($\delta^{\text{lab}}(n)$).

The RMSE of dataset 501, and dataset 701 change similarly when adding more votes ($\delta^{*\text{lab}}$). However when using the MOS values after the first order mapping, the changes in RMSE ($\nu^{*\text{lab}}$) of both 401 and 701 reached the same shape, whereas the RMSE of 501 did not show a tangible improvement (Fig. 4b).

Certainty gain

As explained in Chapter 5, we consider certainty gain as an indicator of how far a metric improves by using more

Table 7 Summary for all metrics: Coefficient of Power models ($y = a \cdot x^b + c$) where x is the number of votes and y is the predicted value

Model	a	b	c	GoF	R^2
ρ^{lab} 401	− 0.4178	− 1.0316	0.9749	0.0001	0.9992
ρ^{lab} 501	− 0.3562	− 0.8866	0.8912	0.0004	0.9964
ρ^{lab} 701	− 0.3917	− 1.0170	0.9304	0.0002	0.9987
δ^{lab} 401	0.6099	− 0.9638	0.4799	0.0004	0.9984
δ^{lab} 501	0.7813	− 0.9095	0.3197	0.0004	0.9993
δ^{lab} 701	0.7910	− 0.9289	0.3363	0.0004	0.9992
ν^{lab} 401	0.8185	− 0.8347	0.1645	0.0004	0.9995
ν^{lab} 501	0.7475	− 0.9097	0.3121	0.0003	0.9993
ν^{lab} 701	0.9254	− 0.9046	0.3036	0.0004	0.9994
ρ^{cs} 401	− 0.3396	− 0.897	0.9984	0.0001	0.9998
ρ^{cs} 501	− 0.4116	− 0.8063	0.9970	0.0001	0.9996
ρ^{cs} 701	− 0.3870	− 0.9221	0.9992	0.0000	0.9998
δ^{cs} 401	0.8229	− 0.5014	0.0001	0.0003	0.9999
δ^{cs} 501	0.8409	− 0.5008	0.0001	0.0003	0.9999
δ^{cs} 701	0.8401	− 0.500	0.0001	0.0003	0.9999
w 401	2.5594	− 0.4194	− 0.0562	0.0013	0.9999
w 501	2.6306	− 0.4222	− 0.0552	0.0013	0.9999
w 701	2.6290	− 0.4200	− 0.0571	0.0014	0.9999
Δ 401	0.8752	− 0.4942	− 0.0012	0.0002	0.9999
Δ 501	0.8941	− 0.4972	− 0.0009	0.0002	0.9999
Δ 701	0.9095	− 0.5026	0.0001	0.0002	0.9999
κ 401	− 1.1250	− 0.4693	0.8293	0.0037	0.9959
κ 501	− 1.4562	− 0.8412	0.7470	0.0015	0.9981
κ 701	− 3.0413	− 1.0561	0.7576	0.0027	0.9961

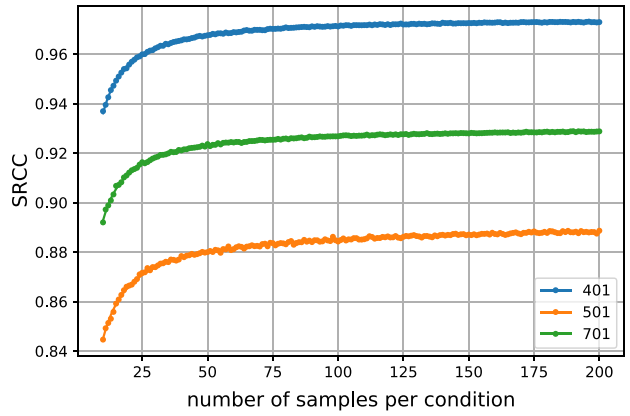
votes. Here we compare a simulated metric with the full crowdsourcing dataset. We considered SRCC, RMSE, Earth Mover's Distance (EMD), and Confidence Interval Width (CIW) metrics. Figure 5 illustrates their performance. For the SRCC, we observe that 401 and 701 change identically whereas 501 gain more coefficient by increasing the number of samples. For more than 50 samples, the curves of 401 and 701 and for about 75 samples the curve of 501 starts to flatten out. All three datasets act nearly identical for RMSE, EMD and CIW metrics. They start to flatten out after about 75 samples. A reduction in RMSE of more than 0.15 is observed when increasing the number of samples from 10 to 75. Within the lab tests, the average CIW is usually in the range of 0.3 MOS [36]. For the datasets 401, 501 and 701, the average CIW in lab is 0.20, 0.31, 0.26 based on 192, 96, 128 ratings, respectively. In CS, $n > 110$ samples are needed for $w_i(n) < 0.3$.

Inter-rater reliability (IRR)

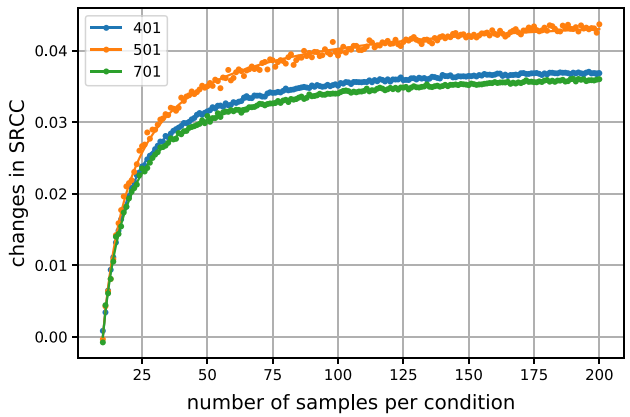
For the IRR calculation, we randomly select a portion of the test participants and compare the MOS values, resulting from their ratings, with the MOS values resulting from the entire participants. Therefore, the IRR score is related to the portion of data used. As the number of votes per condition was varied in our tests, we observed that IRR curves for dataset 501 and 701 flatten out earlier than 401, as in the 401 we had collected about two times the votes per condition than in the other datasets. To compensate this effect, we normalized the number of samples as a percentage of entire samples available in that CS dataset. Figure 6 illustrates how the IRR changes when considering the percentage of CS data used in the simulation. The plot shows that the three CS tests act similarly with a minor difference between 501 and the others. In addition, employing more than 50% of samples does not lead to a tangible change in IRR score.

Comparison across the metrics

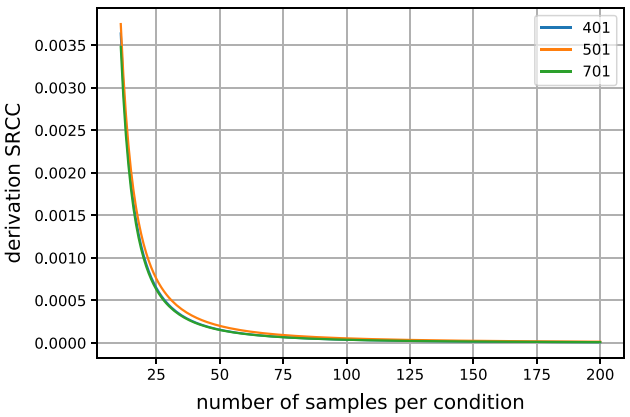
For the different metrics, the parameters of the power model $\widehat{\psi}(n) = an^b + c$ were derived and provided in Table 7. The scale parameter is a, the shape parameter is b and the offset is c. The power model allows to calculate the required number of samples, such that the preferred metric is in a certain target range. For instance, assuming that the CIW is the preferred metric and we aim to keep its maximum value to be $\tau = 0.3$.



(a) Absolute SRCC values



(b) Normalized SRCC values

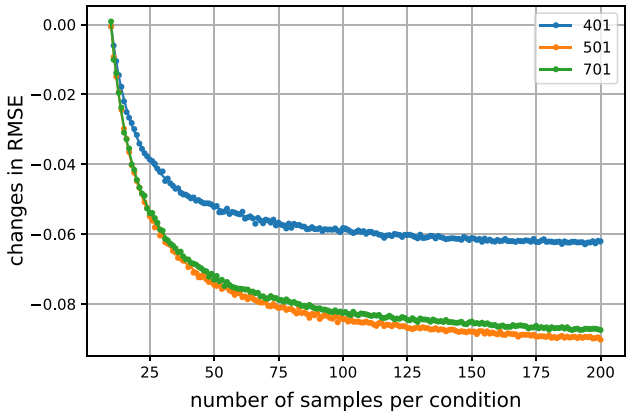


(c) Derivation SRCC values

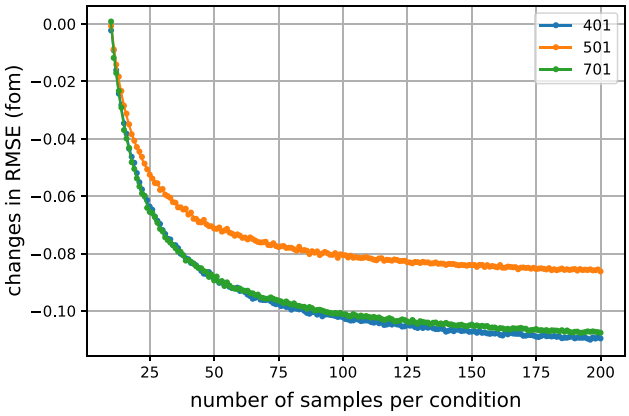
Fig. 3 Validity: SRCC of lab MOS and CS MOS, as well as power model fit

$$\hat{\psi}(n) = an^b + c < \tau \tag{22}$$

$$n \geq \left(\frac{\tau - c}{a}\right)^{\frac{1}{b}} \tag{23}$$



(a) Normalized RMSE values



(b) Normalized RMSE values after 1^{st} order mapping

Fig. 4 Validity: Changes in RMSE of lab MOS and CS MOS, as well as power model fit

The model parameters in Table 7 lead to $n \geq 111, 115, 111$ for CS 401, 501, 701, respectively. Beyond the difference of the various metrics below (or above) a certain target threshold, the relative gain of increasing the number of sample sizes is also of interest. In order to derive, when the metric $\hat{\psi}(n)$ flattens out, the derivative $\hat{\psi}'(n)$ is considered and normalized to the observed value range of the sample size, $[\hat{\psi}(n_0);\hat{\psi}(n_{\max})]$.

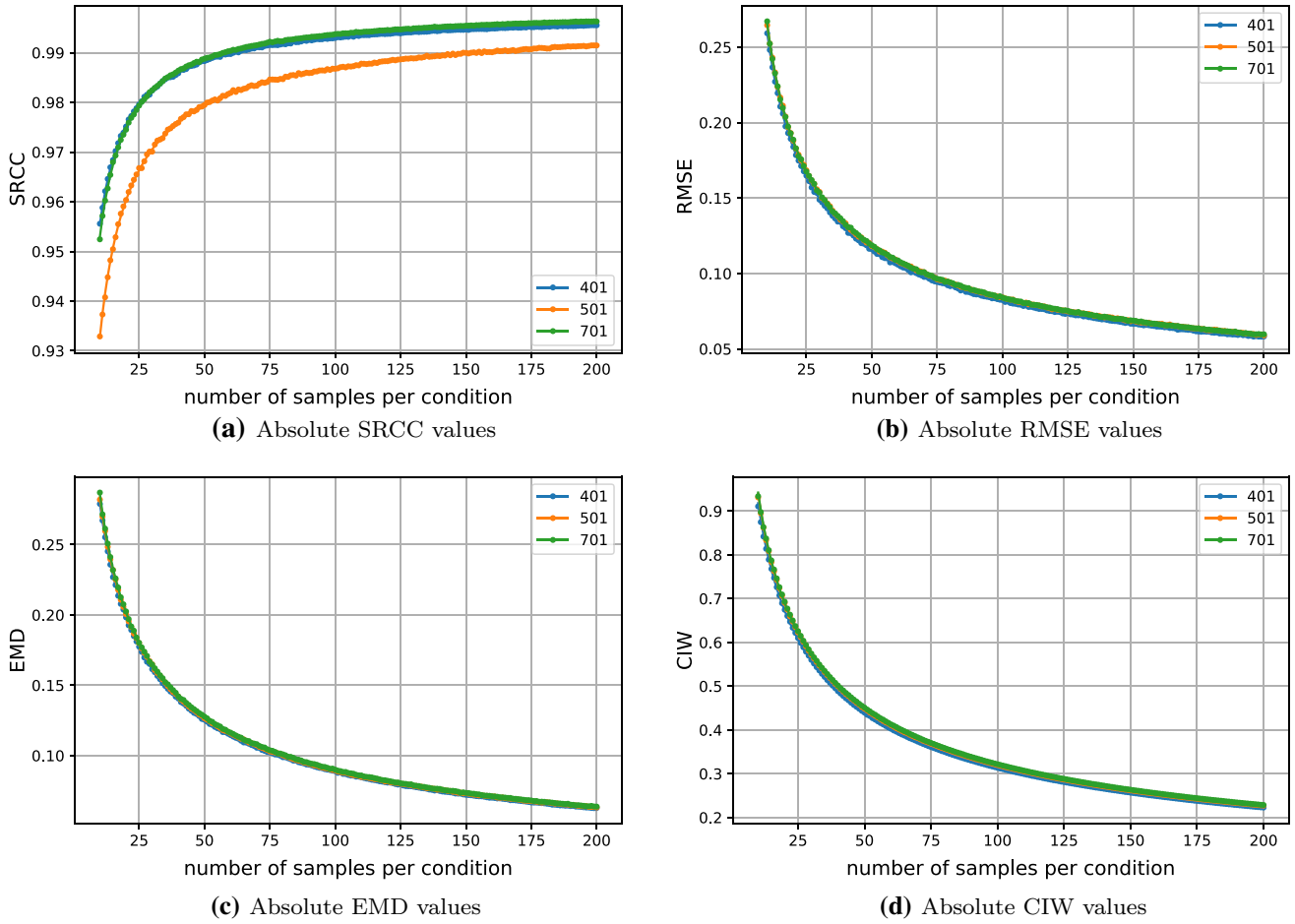
$$\widetilde{\Psi}(n) = \frac{\hat{\psi}'(n)}{\hat{\psi}(n_{\max}) - \hat{\psi}(n_0)} \tag{24}$$

Thus, $\widetilde{\Psi}(n)$ is below a certain threshold, when the sample size n exceeds a certain value.

The function $\hat{\psi}$ converges towards c, since the observed power model shape parameters are negative, i.e. $b < 0$.

$$\lim_{n\to\infty} \hat{\psi}(n) = \lim_{n\to\infty} an^b + c = \lim_{n\to\infty} \frac{a}{n^{|b|}} + c = c \tag{25}$$

The derivation of $\hat{\psi}$ is $\hat{\psi}'(n) = abn^{b-1}$. This derivation is now normalized according to Eq. (24).



(a) Absolute SRCC values

(b) Absolute RMSE values

(c) Absolute EMD values

(d) Absolute CIW values

Fig. 5 Certainty Gain: MOS comparison of sampled CS with full CS dataset

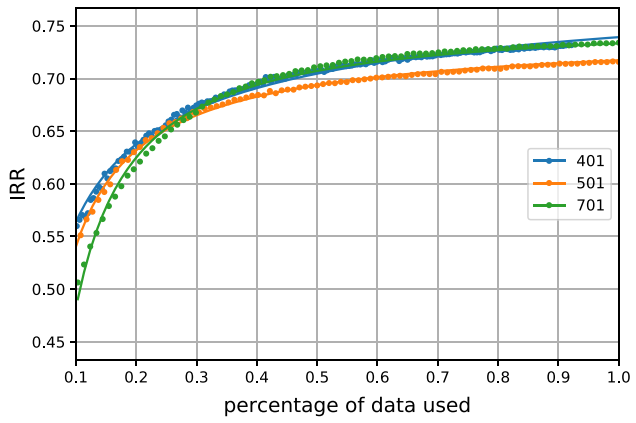


Fig. 6 Absolute Inter-Rater Reliability values, calculated using different percentage of data

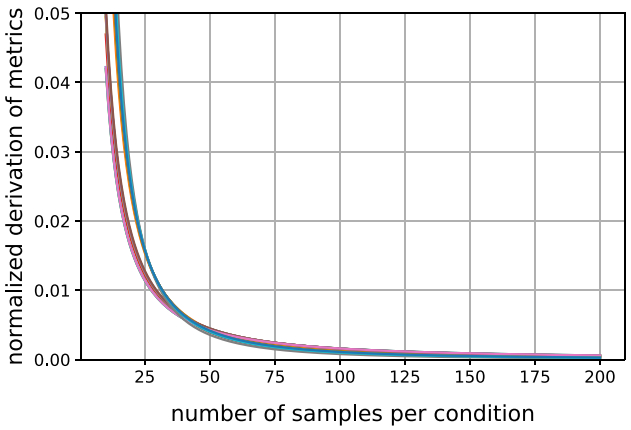


Fig. 7 Comparison of power models for the different CS datasets and all metrics (SRCC, RMS, CIW, EMD, IRR)

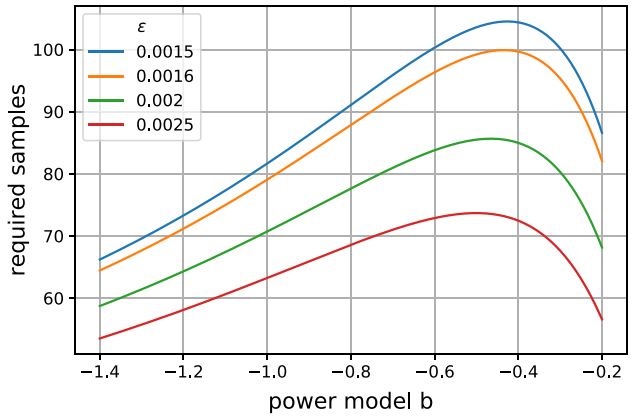


Fig. 8 Required sample size η (Eq. (26)) for a power model with given b to have the normalized derivation below a threshold ϵ

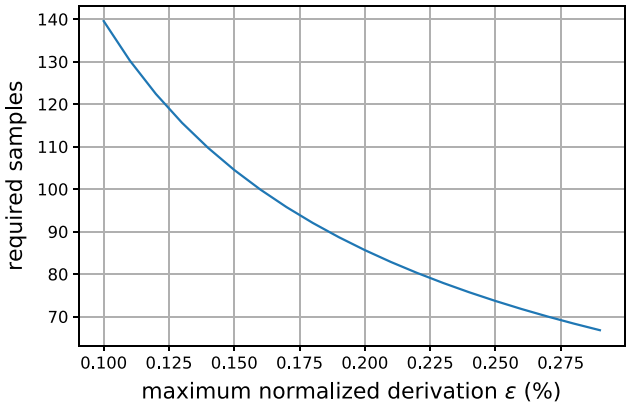


Fig. 9 Required sample size for any power model (and in particular b) to have the normalized derivation below a threshold ϵ

$$\widetilde{\Psi}(n) = \frac{\widehat{\psi}'(n)}{\widehat{\psi}(n_{\max}) - \widehat{\psi}(n_0)} = -b\frac{n^{b-1}}{n_0^b} \tag{26}$$

Figure 7 shows the normalized derivation of all metrics discussed so far for all CS datasets (401, 501 and 701). It can be seen that for all metrics and for $n > 40$ the shapes of the curves for $\widetilde{\Psi}(n)$ are similar and close to each other. For $n > 100$, the value is $\widetilde{\Psi}(n) \leq 0.0016$ for any metric and dataset.

Equation(26) can be transformed to compute the required number of samples η, such that $\widetilde{\Psi}$ is smaller than a threshold ϵ, i.e. $\widetilde{\Psi}(n) < \epsilon$.

$$\eta(\epsilon, b) = \left(-\frac{1}{b}\epsilon n_0^b\right)^{\frac{1}{b-1}} \tag{27}$$

Figure 8 shows the required sample size η that depends on b and the target threshold ϵ. In the obtained power models (cf. Table 7), it is $-1.16 < b < -0.40$. The required sample size function η has a maximum value η^* at b^* which depends on ϵ.

$$b^*(\epsilon) = -\epsilon n_0 e^{W\left(\frac{1}{e\epsilon n_0}\right)+1} \tag{28}$$

$$\eta^*(\epsilon) = \eta(\epsilon, b^*(\epsilon)) \tag{29}$$

Thus, for any power model (i.e. any negative value of b) the required sample size to ensure that ϵ is below a target threshold can be derived. Thus, if a subjective study is to be conducted, the maximum number of required samples can be derived beforehand, i.e. without knowing the subjective results and the corresponding power model. Please note that this required sample size according to Eq. (29) is an upper bound. Figure 9 shows this pessimistic approximation. For $n > 86$, $\epsilon < 0.0002$.

Comparison between degradations

The three lab tests have been carried out at three different locations, using different groups of test participants, different types of degradations (conditions), different numbers of source files per condition, and different numbers of ratings per file and per condition. Thus, it is difficult to statistically and meaningfully calculate significant differences between the means on a per test and a per-condition basis. Instead, we will use the criterion of a difference larger than 0.5 MOS for marking substantial differences between lab and CS results, using the mapped MOS for all databases. The threshold of 0.5 was chosen as it is slightly lower that the threshold chosen in the environmental test for the study CS 401, and it is also substantially smaller than the average standard deviation of all test conditions of the corresponding lab and CS test. For CS 401, only one test condition showed a substantial difference, namely a super-wideband condition with speech level attenuated by 20 dB; here, the mapped MOS in CS was substantially lower than in the lab test. For CS 501, six test conditions showed substantial differences between the mapped MOS and CS:

- Super-wideband with a low level of signal-correlated noise;
- wideband AAC coding at low bit-rate;
- narrowband EFR coding at −10 dB;
- super-wideband AAC coding with packet loss;
- narrowband AMR codec at −16 dB;
- and a VoIP wideband call with no other degradation.

In all cases, except the narrow band signals, the CS MOS was substantially higher than the lab MOS. For CS 701, 9 test conditions showed substantial differences; these included low to moderate levels of signal-correlated noise (2 conditions); super-wideband with low packet loss (3

Table 8 Comparison of efficiency between different crowdsourcing studies and the open-source P.808 toolkit[43]

Criterion	CS 401	CS 501	CS 701	P.808 Toolkit ([43])
preparation time excluding dataset preparation				
⋯ first study (building the platform)	4 weeks	≈ 8 weeks	8 weeks	1 h
⋯ any new study	1 day	1 day	1 day	< 1 h
Time for carrying out the study				
⋯ data collection duration	3 days	11 days	1 day	< 1 day
⋯ active participation of a moderator	8 h	12 h	2 h	–
Monetary expenses				
⋯ payment per vote	$0.050	EUR 0.042	$0.025	$0.050
⋯ platform overhead	40%	40%	≈ 9.5%	40%

conditions); super-wideband with moderate packet loss (2 conditions); and super-wideband with high packet loss (2 conditions). Again, in all cases, the CS MOS was substantially higher than the lab MOS.

There is a notable difference in the number of test conditions for which substantial differences occur. In CS 401, only one condition shows such differences. From a screening point of view, this test was the most elaborate one, with the strictest handling of participants. The difference is only observed for the condition with very low volume, which was rated worse in CS compared to the lab: Whereas the speech playback level is adjustable during the training session in the CS situation, it is fixed in the lab situation. This might explain why the difference occurs.

In CS 501 and CS 701, there were far more test conditions with substantial differences. This might be due to a less strict screening in those CS tests. In all cases substantial differences were observed, the CS MOS was substantially higher than the lab MOS except for narrowband signals. No particular regularity could be observed regarding the types of degradation for which this effect occurs: In CS 501 it is mostly linked to wideband or super-wideband conditions with coding; there were however numerous other such conditions in the test for which the differences were not as strong. Similarly, for CS 701 mostly signal-correlated noise and super-wideband conditions with packet loss stand out; there were also numerous other such conditions which did not show the observed differences.

In conclusion, it seems that the agreement between lab and CS results is not linked to a particular type of degradation, i.e. test condition. The only condition where such a difference could be explained was a low-volume condition, which might stem from the volume adjustment possible in the CS test. However, a stricter screening as performed in CS 401 seems to substantially decrease the number of conditions with substantial differences. Previous studies showed that presence of environmental noise at 50dB(A) SPL leads to a significant difference of the perceived quality [42].

Comparisons of efficiency

In addition to the validity and reliability of the obtained results, it is also important to compare both approaches with respect to their efficiency. This can be done only in a rather approximated way, as the lab studies reported here were carried out by commercial companies as part of their daily business, therefore precise calculations of the efforts spent are lacking. Nevertheless, it was possible to obtain reasonable estimations from three companies with respect to the efforts spent. As both approaches made use of the same speech stimuli, the effort spent preparing the stimuli is disregarded here. We can also not directly compare the efforts put in the set-up of the respective infrastructure: Whereas the time invested to build the crowdsourcing frameworks can be estimated to a range of 4 to 8 weeks (resulting in a person-day cost of around EUR 20 000–40 000), the physical set-up of the test lab will take largely more time and be more expensive (typically in the range of EUR 50 000–100 000, depending on the characteristics of the desired room and the geographical location, but the test site will also be more long-lasting.

The effort for a laboratory test can roughly be separated into [(1)] the preparation time for the test set-up (including public invitation of participants, design of test session lists, proof-listening of the stimulus material, and calibration of the audio equipment in the lab); [(2)] the time for executing and post-processing the test; [(3)] the financial compensation of test participants; and [(4)] the costs for using the test facilities. Regarding the preparation time, the estimations of the companies range between 1 and 5 business days. The execution time of the experiments depends on how many participants can be handled in parallel, as a result of the available test equipment: In case only one participant can be handled at a time, a test typically requires five working days to be completed. In the case that three or four participants can be handled in parallel, this time is reduced to approximately two days. Two more days need to be added for the post-processing of the results (statistical checks and report

generation). The costs per test participant range between EUR 35–45. With an estimated 28 participants (24 effective participants plus no-shows and outliers) this results in roughly EUR 1100–1200. The rental price for the test lab is assumed to be about EUR 4000. Assuming that a person-day is roughly charged EUR 1000, this results in an overall price of the test of roughly EUR 15 000. It has to be noted that these are purely costs, with no business-model behind. So the price which may actually be charged by a professional test lab would typically be higher.

These efforts can be compared to the ones of our crowdsourcing studies. As Table 8 shows, the time for preparing each test is estimated to be around 1 day in our studies. The time for carrying out the study ranged between 1 and 11 days in our case, but during this time period the experimenter spent only between 0.5 and 1.5 business days on the experiment. Finally, the remuneration for the participants and the test platform ranges between US$ 280 and US$ 812. No additional costs of the usage of the platform are expected. Summarizing these efforts in terms of costs, this results in approx. EUR 2550 per test.

Comparing these numbers, it is obvious that the crowdsourcing approach is significantly cheaper, especially when running many tests: Reducing the costs by a factor of 6 will seem very beneficial to test labs. Furthermore, the open-source implementation of the ITU-T Rec. P.808 [43] strongly reduces the costs of the crowdsourcing approach as well: the study preparation time is reduced to less than one hour, carrying out a study with a dataset of a typical size (i.e. about 1000 votes) normally takes a day with no need of active participation of a moderator. And post-processing and aggregation of ratings are all performed by the toolkit [43].

On the other hand, one should not forget that the investments into a crowdsourcing platform may be rather short-term, whereas the investments into a physical test lab are rather long-term. In addition, a laboratory environment can be used for many other purposes (e.g. high-quality audio testing) which would be impossible with a crowdsourcing approach.

Discussion and future works

We have conducted three speech quality tests through crowdsourcing studies to validate the ITU-T Rec. P.808. Each of the studies was conducted by a different experimenter following the exact text of the Recommendation. We used three datasets from the pool of the ITU-T Rec. P.863 competition datasets for our evaluation. Results from subjective tests in the lab, performed according to the ITU-T Rec. P.800, were kindly provided to us. We used three different crowdsourcing platforms.

Results show that there is a strong correlation between lab-based assessments, according to the ITU-T Rec. P.800, and the crowdsourcing evaluations based on the ITU-T Rec. P.808. We also investigated the effect of the number of votes per degradation condition on the validity and reliability of the P.808 approach using different metrics. Results show that about 60–100 valid votes per condition are required in the crowdsourcing tests. For each metric, we reported the model's parameters. Consequently, one can calculate the number of votes needed for their study given a specific range of their target metric.

We also investigated degradation conditions in which a substantial difference between lab and CS tests was observed. We observed a significant difference between MOS values from the lab and CS in 12% of conditions in the studies CS 501 and CS 701 but in only one condition in the CS 401 study. We interpret that the agreement between lab and CS results is not linked to a particular type of degradation. Instead a stricter screening (i.e. environment suitability test [34] and hearing ability test) as performed in CS 401 seems to significantly decrease the number of conditions with substantial differences. As shown in [34], applying the environment test leads to a reliable assessment of degradation conditions that include noise and sub-optimal presentation levels as well. However, the drawback of the environment suitability test is that it significantly increases the duration of the test session. One solution would be to use the test periodically as in [43] and ultimately extending it to a continuous monitoring solution in which a new instance of the test is required only when a substantial change in the environment is observed.

Furthermore, we compared the efficiency of the lab and our CS tests. Undoubtedly the CS approach is much cheaper and faster. The overall cost of an experiment can be reduced by a factor of 6 when using the CS approach. The newly published P.808 Toolkit [43] further reduces the cost of CS tests. Through automation only minimal work is needed by experimenter in the entire process.

During the simulation we randomly selected 10 to 200 votes by sampling with replacement from the CS tests, even though the original number of votes were 102 and 97 for CS 501 and CS 701 respectively. We believe that the number of votes in the original study are high enough to represent different groups of reliable workers. However, to be able to generalize our findings, we used resampling with replacement. One might argue that the saturation at about 100 votes, that was observed in the simulation, was due to the fact that the original study has close to 100 votes (CS 501 and CS 701). But since we have observed the same behavior from the study CS 401, which had 217 votes in the original study, we believe that the original number of votes in the CS 501 and CS 701 studies are high enough for such a simulation and resampling. In addition, we used

1000 runs of simulation and reported the average values of the matrices. The larger number of simulation runs leads to a smother scatter plot and a more accurate fit. In our previous paper [40] we used 200 runs, although similar results were observed there the fitted functions only showed minor changes. Increasing the number of runs also leads to smaller confidence interval widths.

For future work, the ITU-T Rec. P.808 can be extended to include other test methods like Degradation Category Ratings (DCR) and Comparison Category Ratings (CCR). They are suitable for validating systems with small impairments. In such a case, strict environment and setup tests should be integrated so that the participants can recognize small differences. Furthermore, the effect of language proficiency on the perceived quality of speech samples by non-native listeners should be investigated in detail. In addition, in circumstances such as a pandemic, the need for a contact-less subjective test is increased. Given the success of the ITU-T Rec. P.808, a new recommendation for evaluating the video and audiovisual quality using the crowdsourcing approach might be developed.

Acknowledgements We would like to thank the experts of the ITU-T Study Group 12 and contributors to the pool of dataset created for the ITU-T Rec. P.863 competition who kindly shared it with us for the evaluation of the ITU-T Rec. P.808. We would like to also thank Jens Berger, Ludovic Malfait, and Christian Schmidtmer for their valuable inputs. The implementation of study CS701 was written by Andre Hönnscheidt.

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

Conflicts of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Hoßfeld T, Keimel C, Hirth M, Gardlo B, Habigt J, Diepold K, Tran-Gia P (2014) Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. IEEE Trans Multimed 16(2):541–558
2. ITU-T Recommendation P.800 (1996) Methods for subjective determination of transmission quality. International Telecommunication Union, Geneva
3. ITU-T Recommendation P.808 (2018) Subjective evaluation of speech quality with a crowdsourcing approach. International Telecommunication Union, Geneva
4. Egger-Lampl S, Redi J, Hoßfeld T, Hirth M, Möller S, Naderi B, Keimel C, Saupe D (2017) Crowdsourcing quality of experience experiments. In: Archambault D, Purchase H, Hoßfeld T (eds) Evaluation in the crowd. Crowdsourcing and human-centered experiments. Springer, Cham, pp 154–190
5. Hosu V, Lin H, Saupe D (2018) Expertise screening in crowdsourcing image quality. In: 2018 Tenth international conference on quality of multimedia experience (QoMEX), pp 1–6
6. Siahaan E, Hanjalic A, Redi J (2016) A Reliable Methodology to Collect Ground Truth Data of Image Aesthetic Appeal. IEEE Trans Multim 18(7):1338–1350
7. Søgaard J, Shahid M, Pokhrel J, Brunnström K (2016) On subjective quality assessment of adaptive video streaming via crowdsourcing and laboratory based experiments. Multim Tools Appl. https://doi.org/10.1007/s11042-016-3948-3
8. Cartwright M, Pardo B, Mysore GJ, Hoffman M (2016) Fast and easy crowd sourced perceptual audio evaluation. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 619–623
9. Volk T, Keimel C, Moosmeier M, Diepold K (2015) Crowdsourcing vs. laboratory experiments - QoE evaluation of binaural playback in a teleconference scenario. Comput Netw 90:99–109
10. Naderi B, Polzehl T, Wechsung I, Köster F, Möller S (2015) Effect of trapping questions on the reliability of speech quality judgments in a crowd sourcing paradigm. In: INTERSPEECH. ISCA 2799-2803
11. Zequeira Jiménez R, Fernández Gallardo L, Möller S (2018) Influence of number of stimuli for subjective speech quality assessment in crowdsourcing. In: 2018 Tenth international conference on quality of multimedia experience (QoMEX), pp 1–6
12. Polzehl T, Naderi B, Köster F, Möller S (2015) Robustness in speech quality assessment and temporal training expiry in mobile crowdsourcing environments. In: Sixteenth annual conference of the international speech communication association
13. Gadiraju U, Möller S, Nöllenburg M, Saupe D, Egger-Lampl S, Archambault D, Fisher B (2017) Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd. In: Archambault D, Purchase H, Hoßfeld T (eds) Evaluation in the Crowd. Crowdsourcing and human-centered experiments. Springer, Cham, pp 6–26
14. Chen K-T, Chang C-J, Wu C-C, Chang Y-C, Lei C-L (2010) Quadrant of Euphoria: a crowdsourcing platform for QoE assessment. IEEE Network 24(2):28–35
15. ITU-R Recommendation BT.500-11 (2002) Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union, Geneva
16. ITU-R Recommendation P.910 (2008) Subjective video quality assessment methods for multimedia applications. International Telecommunication Union, Geneva
17. Ribeiro FP, Florêncio DAF, Zhang C, Seltzer ML (2011) 'CROWDMOS: an approach for crowdsourcing mean opinion score studies. In: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 2416–2419
18. ITU-R Recommendation BS.1534-3 (2014) Method for the subjective assessment of intermediate quality level of audio systems. International Telecommunication Union, Geneva
19. Ribeiro F, Florencio D, Nascimento V (2011) Crowdsourcing subjective image quality evaluation. In: 18th IEEE international conference on image processing, pp 3097–3100

20. Sheikh H, Wang Z, Cormack L, Bovik A (2003) Live image quality assessment database, 2003. [Online]. Available http://live.ece.utexas.edu/research/quality/
21. Keimel C, Habigt J, Horch C, Diepold K (2012) QualityCrowd - a framework for crowd-based quality evaluation. In: 2012 Picture coding symposium, pp 245–248
22. Ruchaud N, Antipov G, Korshunov P, Dugelay J-L, Ebrahimi T, Berrani S-A (2015) The impact of privacy protection filters on gender recognition. In: Tescher AG (Ed) Applications of digital image processing XXXVIII, vol 9599. International Society for Optics and Photonics. SPIE, pp 36–47
23. Korshunov P, Bernardo MV, Pinheiro AM, Ebrahimi T (2015) Impact of tone-mapping algorithms on subjective and objective face recognition in hdr images. In: Proceedings of the fourth international workshop on crowdsourcing for multimedia, ser. CrowdMM'15. Association for Computing Machinery, New York, NY, pp 39–44
24. Bonetto M, Korshunov P, Ramponi G, Ebrahimi T (2015) Privacy in mini-drone based video surveillance. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG), vol 04, pp 1–6
25. Saupe D, Hahn F, Hosu V, Zingman I, Rana M, Li S (2016) Crowd workers proven useful: a comparative study of subjective video quality assessment. In: 8th International conference on quality of multimedia experience (QoMEX)
26. Hoßfeld T, Seufert M, Sieber C, Zinner T (2014) Assessing effect sizes of influence factors towards a qoe model for http adaptive streaming. In: 2014 Sixth international workshop on quality of multimedia experience (QoMEX), pp 111–116
27. Kraft S, Zölzer U (2014) "BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality," in Linux Audio Conference. Karlsruhe, DE
28. Handbook ITU-T (2011) Practical procedures for subjective testing. International Telecommunication Union, Geneva
29. Naderi B, Polzehl T, Wechsung I, Köster F, Möller S (2015) Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm. In: Sixteenth annual conference of the international speech communication association
30. ITU-T Recommendation P.863 (2018) Perceptual Objective listening quality prediction. International Telecommunication Union, Geneva
31. Martin D, Carpendale S, Gupta N, Hoßfeld T, Naderi B, Redi J, Siahaan E, Wechsung I (2017) Understanding the crowd: ethical and practical matters in the academic use of crowdsourcing. In: Evaluation in the crowd. Crowdsourcing and human-centered experiments. Springer, New York, pp 27–69
32. Smits C, Kapteyn TS, Houtgast T (2004) Development and validation of an automatic speech-in-noise screening test by telephone. Int J Audiol 43(1):15–28
33. Buschermöhle M, Wagener K, Berg D, Meis M, Kollmeier B (2015) The german digit triplets test (part ii): validation and pass/fail criteria. Zeitschrift für Audiologie 54(1):6–13
34. Naderi B, Möller S (2020) Application of just-noticeable difference in quality as environment suitability test for crowdsourcing speech quality assessment task. In: 12th International conference on quality of multimedia experience (QoMEX). IEEE, pp 1–6
35. Zequeira Jiménez R, Mittag G, Möller S (2018) Effect of number of stimuli on users perception of different speech degradations. A crowdsourcing case study. In: IEEE international symposium on multimedia (ISM). IEEE, pp 175–179
36. ITU-T Recommendation P.1401 (2020) Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models. International Telecommunication Union, Geneva
37. HoßfeldT, Heegaard PE, Varela M, Skorin-Kapov L (2018) Confidence interval estimators for mos values. arXiv preprint arXiv:1806.01126
38. Naderi B, Möller S (2020) Transformation of mean opinion scores to avoid misleading of ranked based statistical techniques. In: 12th International conference on quality of multimedia experience (QoMEX). IEEE, pp 1–3
39. Hoßfeld T, Heegaard PE, Varela M, Möller S (2016) Qoe beyond the mos: an in-depth look at qoe via better metrics and their relation to mos. Quality User Exp 1(1):2
40. Naderi B, Hossfeld T, Hirth M, Metzger F, Möller S, Zequeira Jiménez R (2020) Impact of the number of votes on the reliability and validity of subjective speech quality assessment in the crowdsourcing approach. In: 12th international conference on quality of multimedia experience (QoMEX). IEEE, pp 1–6
41. Hoβfeld T, Schatz R, Egger S (2011) Sos: The mos is not enough! In: Third international workshop on quality of multimedia experience. IEEE, pp 131–136
42. Zequeira Jiménez R, Naderi B, Möller S (2020) Effect of environmental noise in speech quality assessment studies using crowdsourcing. In: 12th International conference on quality of multimedia experience (QoMEX). IEEE, pp 1–6
43. Naderi B, Cutler R (2020) An open source implementation of itu-t recommendation p.808 with validation. To appear in INTERSPEECH. ISCA