INTERNATIONAL TELECOMMUNICATION UNION

**TELECOMMUNICATION STANDARDIZATION SECTOR**

STUDY PERIOD 2017-2020

**COM 12-C 039**

**STUDY GROUP 12**

**Original: English**

**Question(s):** 7/12 Geneva, 10-19 January 2017

**CONTRIBUTION**

**Source:** Deutsche Telekom AG

**Title:** Speech quality assessment in crowdsourcing: Comparison with laboratory and recommendations on task design

**Purpose:** Discussion

| | | |
|---|---|---|
| **Contact:** | Babak Naderi<br>Telekom Innovation Labs, TU Berlin<br>Germany | Tel: +49 30 8353 54221<br>Fax: +49 30 8353 58409<br>Email: babak.naderi@tu-berlin.de |
| **Contact:** | Sebastian Möller<br>Telekom Innovation Labs, TU Berlin<br>Germany | Tel: +49 30 8353 58465<br>Fax: +49 30 8353 58409<br>E-mail: sebastian.moeller@tu-berlin.de |

**Keywords:** ITU-T SG12; crowdtesting; reliability check; trapping question; crowd vs lab

**Abstract:** This contribution reports result of a speech quality assessment study conducted on a crowdsourcing platform and compares them with laboratory data to evaluate different reliability checking methods that can be used in crowdsourcing task design. Recommendations on trapping questions and how to design a crowdsourcing speech quality assessment task to increase the reliability of responses are given which may be considered in the future Rec. P.CROWD. The aim is to encourage experts to apply the proposed methods in future studies to further validate their effects, and ultimately add them to the corresponding part of the Recommendation text. This contribution is based on [1].

## 1. Introduction

Crowdtesting cannot be considered as a direct implementation of laboratory testing methodologies in an Internet-based environment [2], due to factors it inherits from the nature of crowdsourcing. A range of differences and aspects which influence crowdtesting results, considering speech quality assessment tasks, have been presented in COM-12-386 to Q.7/12 experts in the last Study Period. One crucial aspect, well-known to the crowdsourcing community, is that the results from paid micro-tasks are often noisy due to corrupted responses submitted by participants who are not concentrated enough while working, or who do not work as instructed due to a variety of reasons [3]. For example, instructions may be misunderstood; the crowd workers may be interrupted or may take a break while carrying out the study; they may split their attention between the study and parallel activities [4]; or they try to maximize their monetary benefit with minimum effort. Previous research proposed different approaches, such as the use of gold standards, majority voting, and behavioural logging, to evaluate results of crowdsourcing tasks in a post-processing step [5][6]. In the context of quality assessment tasks, researchers have used additional methods like content

questions and consistency tests [7], [8]. 'Trapping questions', i.e. questions with a known answer, allow the experimenter to identify inattentive or wilfully cheating users.

Based on the aforementioned research, the current contribution investigates 1) whether trapping questions increase the quality of responses collected in a crowdsourcing environment using a listening only test (LOT) quality assessment task, and 2) which type of trapping questions work best.

## 2. Study design

Three different types of trapping questions are designed based on previous studies and tested in a LOT quality assessment task using a standard database. Correlations between MOS collected in the crowdsourcing studies and laboratory ratings, and the Root Mean Square Deviation (RMSD), are used to evaluate the influence of different trapping questions. In the following, each part is explained in detail.

### 2.1. Database (SwissQual)

For the experiment, we used stimuli from the SwissQual 501 database from the ITU-T Rec. P.863 competition, which has been kindly provided by SwissQual AG, Solothurn. This database includes variable types of degradations and degradation combinations. The stimuli were produced according to the ITU-T Rec. P.863 specifications. Overall, 200 stimuli are arranged to carry 50 conditions. Each condition describes one degradation or a combination of degradations and each is composed of four stimuli (with the same degradation) recorded by four speakers with four different German sentences. The conditions represent degradations like different audio bandwidths (narrowband 300-3400 Hz, wideband 50-7000 Hz, super-wideband 50-14000 Hz), signal-correlated as well as uncorrelated noise, ambient background noise of different types, temporal clipping, speech coding at different bitrates, packet loss of different temporal loss profiles, different frequency distortions, as well as combinations of these degradations. The database contains 24 quality ratings from German natives per stimulus, which were obtained in a lab environment in accordance with ITU-T Rec. P.800. The resulting MOS per stimulus and test condition serve as a reference.

### 2.2. Crowdsourcing labelling procedure

Labelling was processed using the *Crowdee* mobile crowdsourcing platform of the Quality and Usability Lab at Technische Universität Berlin, Germany[1]. This platform provides the opportunity for app-based crowdsourcing assessments including crowdsourcing speech recordings in the field. Workers are from various places in Germany. The app to be used by the workers is based on the Android operation system and is freely available in the Google Play Store.

The study implemented several connected series of micro tasks. The workers (participants) were presented three individual jobs: (1) Qualification, (2) Training, (3) Speech quality assessment.

Workers were welcomed by a qualification test of about 2-3 minutes. This test included inquiries on any hearing impairment and prior experiences with quality assessment, as well as a technical headset and audio playback test. We asked the workers to seek a quiet place, and controlled the playback volume after asking to adjust it to a comfortable level. We imposed the usage of two-eared headsets such that jobs could only be started when connecting it, and validated its usage by short math exercises with digits panning between left and right in stereo. After approval, workers were assigned to one of the different treatments (i.e. 3 trapping question mechanisms and one control group) of the study and automatically given access to a second job, namely training. In the training,

[1] https://www.crowdee.de

we presented 12 types of degradations, which are the first 12 anchor stimuli in the SwissQual database, and made the user explicitly aware of the presented degradation types, e.g. added noise, band limitations, packet loss, and different distortion types. Audio-files were pre-loaded on the worker's device and could be played multiple times for training. In this test, training expired after one day, and workers were forced to re-train before continuing when the one-day period had expired. Immediately after training, they were given access to the speech quality assessment jobs.

The jobs were structured as follows: First, the workers were asked to verify the background noise level and to report on their current level of tiredness. After that, the workers were either presented with five stimuli in case no trapping questions were used, or with six stimuli (five stimuli plus one trapping stimulus) in case trapping questions were employed. They were asked to rate the quality of each of the non-trapping stimuli. For the trapping stimuli, the answer scales may differ (cf. Sec. 2.3). Accordingly, each job contained either 5 or 6 stimuli, and workers were free to do up to 40 jobs in a row (i.e. rate all 200 stimuli of the database), or to pause in between tasks at their discretion. Workers were forced to listen to the entire stimulus before they could rate it and proceed to the next one. Ratings were collected on a 5-point ACR scale congruent to the scale used in the lab study.

### 2.3 Trapping questions

Three different groups representing different configurations of speech-related trapping questions were compared to a control group with no trapping question (*Trapping T0 - No Trapping*). For each studied group, the speech quality assessment jobs were slightly altered by adding one additional stimulus, which was modified. From the original dataset, 40 different stimuli were randomly selected (different speakers, and different degradation conditions) to build a reference trapping stimuli dataset. The dataset was manipulated to create different types of audio trapping questions:

**Trapping T1 - Motivation Message:** For the first group of trapping stimuli, a message was recorded with a speaker not being part of the speech material to be judged in German. It was appended to the first four seconds of each of the 40 trapping stimuli. The message was as follows: "*This is an interruption. We - the team of Crowdee - like to ensure that people work conscientiously and attentively on our tasks. Please select the answer 'X' to confirm your attention now*."

In this group, the trapping question was visually identical to the other speech labeling questions, but the trapping stimulus request workers to choose a specific answer option (e.g. *X = Poor*, or *Fair*).

**Trapping T2 - Low Effort:** In the second group of trapping stimuli, one or more animal sounds were inserted in the middle and at the end of each stimulus from the reference trapping stimuli dataset. In this case, the trapping question was also visually different, as a different rating scale was used: Workers were asked to indicate whether they recognized the animal sound in the stimulus in a multiple-choice answer format. Here, workers can provide a true answer for the trapping with low effort; full concentration in entire assessment job is not required.

**Trapping T3 - High Effort:** In the last group of trapping questions, the same stimuli created for the second group (*Trapping 2*) were used, but the stimuli were presented together with the ACR rating scale, which was employed for all other stimuli. In addition, a multiple choice question was added at the end of the job (i.e. after being presented with five non-trapping stimuli plus one trapping stimulus). Workers were asked an additional question at the end, namely they were told to specify all the animal sound(s) that they recognized in any of the previous stimuli. As for *Trapping T2*, the trapping question was visually recognizable, but in case the worker was inattentive while rating the other stimuli, he/she would need to review all previous stimuli again, to find out the correct answer. In this case, the effort to conceal cheating is high.

*Trapping T1* emphasized the importance of highly reliable responses to the workers; the objective to employ this kind of trapping questions was to evaluate the hypothesis that participants put more effort in the job if they are aware of the value of their work. With *Trapping T2* and *Trapping T3* we examined the assumption that the likelihood of cheating decreases if the effort to conceal cheating is as high as the effort to accurately complete a task [9].

For all groups, the position of stimuli in the task was shuffled by the crowdsourcing platform, using the randomization function provided by *Crowdee*.

## 3. Data collection and results

The study was conducted using *Crowdee* for 18 days. After submitting an answer to the qualification job, a worker was automatically assigned to one of the four study groups (*T0, T1, T2, T3*) by the platform. As a result, each worker could only perform the speech quality assessment jobs designed for their study group. In each study group, 24 ratings for each 200 stimuli in the database were collected (instead of T0 group which workers did not finish all tasks during study time). Overall, 179 workers (87 f, 92 m, $M_{age}$= 27.9 y., $SD_{age}$=8.1 y) participated in the study. Based on the trapping questions, 49 responses were rejected (*T1* =2, *T2* =1, *T3* = 46). Note that, the responses from the task containing wrongly answered trapping question were removed and other judgments from the same worker in tasks with correctly answered trapping questions were used.

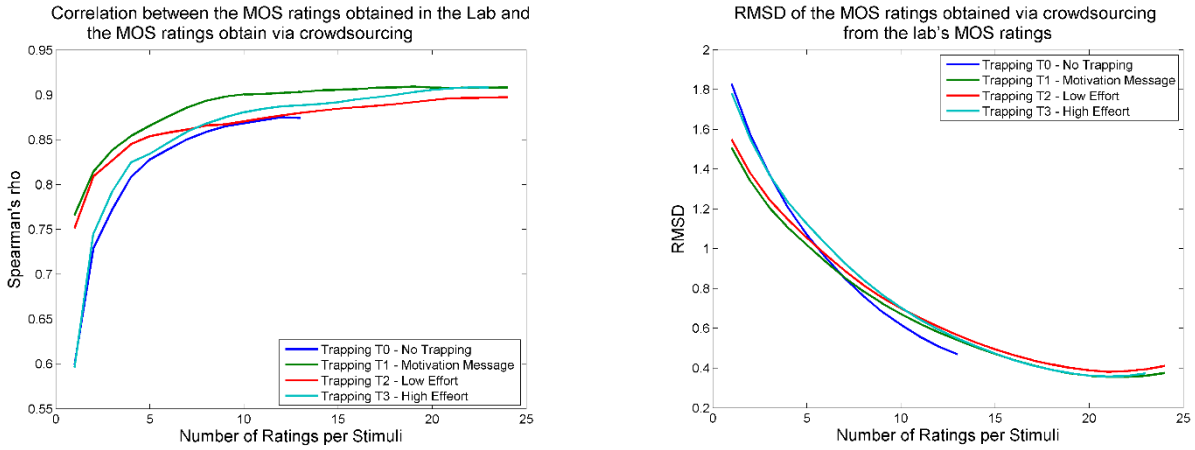
For all crowdsourcing groups (*T0, T1, T2, T3*), MOS values were calculated for each of the 200 stimuli of the database. Four different Spearman's rank-order correlations, one for each group, were computed to determine the relationship between the MOS ratings obtained from the crowd and the MOS ratings obtained in the lab. In addition, the Root Mean Square Deviations (RMSD) from the laboratory MOS ratings were calculated for the MOS ratings of the four crowdsourcing groups.

| **Group** | **$r_s$** | **p-value** | **RMSD** |
|---|---|---|---|
| *Trapping T0 - No Trapping* | 0.886 | < 0.001 | 0.426 |
| *Trapping T1 - Motivation Message* | 0.909 | < 0.001 | 0.375 |
| *Trapping T2 - Low Effort* | 0.897 | < 0.001 | 0.411 |
| *Trapping T3 - High Effort* | 0.909 | < 0.001 | 0.390 |

Table 1. Correlation between the MOS ratings obtained in the lab and the MOS ratings obtained via crowdsourcing (N=200).

Strong positive correlations with the MOS ratings obtained in the lab were observed for all groups, regardless of the kind of employed trapping question. However, for *Trapping T1* the highest correlation as well as the lowest RMSD was observed; thus, the results indicate the best performance for this group.

Figure 1: Changes in correlation and RMSD depending on number of ratings**.**




In the next step, for each of the 50 conditions, we calculated the 95% confidence intervals[2] (CIs) of the mean ratings obtained in the crowd. Again, the calculations were conducted separately for each group (*T0*, *T1*, *T2*, *T3*).

Based on this data, we checked for *which* and for *how many* conditions the CIs of the crowdsourcing ratings did not overlap with the CIs of the ratings obtained in the lab. Again the results (cf. Table 2) showed best results for *Trapping 1*: for 35 of the conditions an overlap of the CIs was observed. For both, *Trapping T2* and *Trapping T3*, for 30 conditions the CIs of the means were overlapping with CIs of the lab means. Poorest performance was observed for *Trapping T0.* Next, we examined if the number of overlapping and non-overlapping conditions for *T1*, *T2*, and *T3* differed from the control condition *T0*. A $\chi^2$ - test indicated a statistically significant difference between *T0* and *T1*, $\chi^2$= (1, N = 50)=5.15, *p*=.023.

Accordingly, the results obtained with *Trapping T1 - Motivation Message* are more consistent to lab results than the results obtained with *Trapping T0 - No Trapping.*

| **Group** | **N of CIs lower** | **N of CIs higher** | **N of CIs overlapping** |
|---|---|---|---|
| *Trapping T0 - No Trapping* | 17 | 6 | 27 |
| *Trapping T1 - Motivation Message* | 13 | 2 | 35 |
| *Trapping T2 - Low Effort* | 17 | 3 | 30 |
| *Trapping T3 - High Effort* | 16 | 4 | 30 |

Table 2. Number of condition (out of 50) with the CIs of the crowd means being lower, higher and overlapping with the CIs of the lab means.

In addition, the results show that for 9 of the 50 conditions, the MOS in crowdsourcing studies are significantly different from the lab study (≈ 0.5 on the MOS scale) in all groups (Table 3). An explorative analysis of the data showed that especially narrow-band (NB) speech files tend to be rated with a lower quality in the crowdsourcing study (38% of the NB conditions provoke a

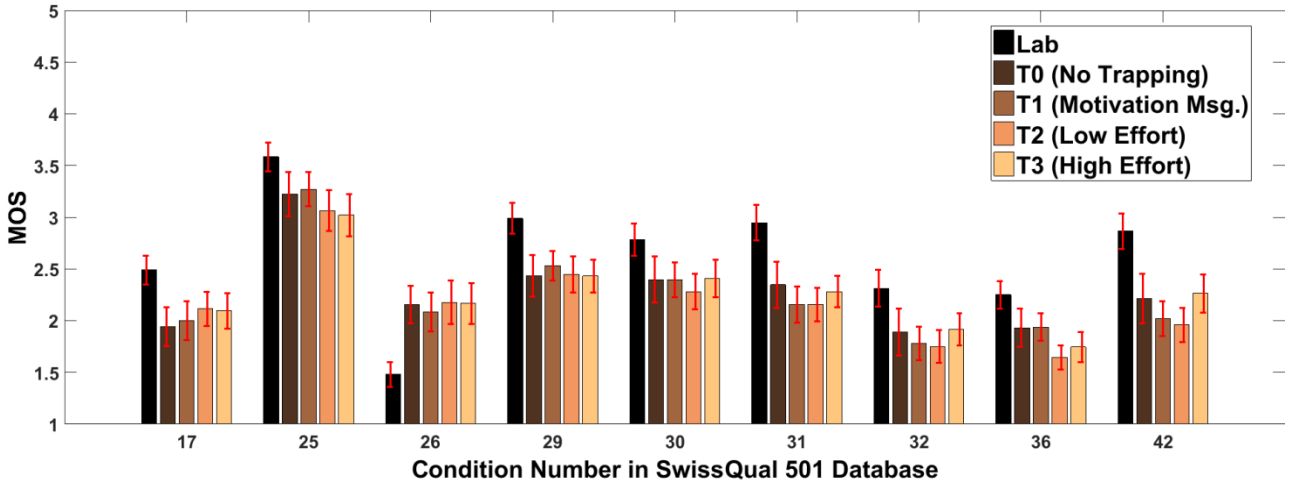
[2] Note that, CIs offer several advantages compared to p-values [10]: Like p-values, they can be used to estimate the statistical significance of an effect (e.g. non-overlapping 95% CIs indicate a difference on the p<.01 level). Furthermore they can be used to compare different studies and they also provide information regarding the precision of the measure (the wider the CIs the lower the precision of the estimate).

significantly lower rating in the crowdsourcing than in the laboratory). These results are in line with [11] where NB conditions showed a significantly lower rating. Hence, with respect to the robustness of the quality ratings in crowdsourcing experiments compared to lab studies, the results indicate that some specific condition characteristics provoke a different rating.

| # | Description (yellow = training samples) | NB | WB | SWB | Codecs | PLC | Pkt loss | BGN | live | sim. |
|---|---|---|---|---|---|---|---|---|---|---|
| 17 | AMR-WB Mode 2 (12.65 kbps) + Noise 16dB SNR + -16dB | | 1 | | 1 | | | 1 | | 1 |
| 25 | AMR-NB Live L2M + Bad channel + No DTX DL + Nokia chipset | 1 | | | 1 | | | | 1 | |
| 26 | AAC LC low bitrate (WB) | | 1 | | 1 | | | | | 1 |
| 29 | EFR Live M2L + ac. Recording | 1 | | | 1 | | | | 1 | |
| 30 | EFR Live M2L + +5dB + ampl. clipping + ac. recording | 1 | | | 1 | | | | 1 | |
| 31 | EFR Live M2L + -10dB + ac. Recording | 1 | | | 1 | | | | 1 | |
| 32 | EFR Live M2L + -20dB + ac. Recording | 1 | | | 1 | | | | 1 | |
| 36 | AMR-NB Live M2M + Noise 16dB SNR - Phone NS + DTX UL + QC chipset | 1 | | | 1 | | | 1 | 1 | |
| 42 | Video Call live AMR + QualComm chipset + -16dB | 1 | | | 1 | 1 | 1 | | 1 | |

Table 3. Descriptions of the selected conditions with significant differences between lab and crowd. NB: Narrowband; WB: Wideband; SWB: Super Wideband; PLC: Packet-loss-concealment; BGN: Background noise; sim: simulated

Figure 2: Degradation conditions with significant differences between lab and crowdsourcing test.




## 4. Discussion and next steps

In this contribution, the influence of different types of trapping questions on the reliability of speech quality crowdtesting is examined. In all groups with trapping questions (T1, T2 and T3) all obtained data (correlations, RMSD, number of conditions for which the CIs of the means were overlapping with the CIs of the lab means) tended to be more consistent to the lab data compared to the data obtained in the group without any trapping question.

**Rec. 1: Use trapping stimuli to increase the reliability of speech quality crowdtesting.**

Best results were observed for the type of trapping question, for which a recorded voice was presented in the middle of a random stimulus. The voice explained to the workers that high quality responses are important, and asked them to select a specific item to show their concentration. A possible explanation for the effect of this kind of trapping questions is that they communicate the importance and the value of their work to the crowd workers.

**Rec. 2: Use trapping stimuli with motivational messages to encourage workers and acknowledge their effort.**

For 9 degradations there were significant differences between MOS collected in the lab and all crowdsourcing groups. Results show that around 38% of the NB conditions provoke a significantly lower rating in crowdsourcing studies. Therefore, care should be taken when analyzing and interpreting NB conditions in crowdsourcing experiments. Further investigations are required to find out the source of this effect (environmental noise, using mobile phone for playback, unexperienced crowd workers).

**In order to proceed with the draft P.CROWD, we would like to kindly ask ITU-T SG12 experts to discuss the mentioned recommendations during the upcoming SG12 meeting, and to apply them in their own upcoming crowdtesting studies, on different platforms.**

## 5. References

1. Naderi, B., Polzehl, T., Wechsung, I., Köster, F., Möller, S.: Effect of Trapping Questions on the Reliability of Speech Quality Judgments in a Crowdsourcing Paradigm. In: 16th Ann. Conf. of the Int. Speech Comm. Assoc. (Interspeech 2015). ISCA. pp. 2799–2803 (2015).
2. Hoßfeld, T., Hirth, M., Redi, J., Mazza, F., Korshunov, P., Naderi, B., Seufert, M., Gardlo, B., Egger, S., Keimel, C.: Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force Crowdsourcing. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003 Qualinet) (2014).
3. Naderi, B., Wechsung, I., Möller, S.: Effect of Being Observed on the Reliability of Responses in Crowdsourcing Micro-task Platforms. In: (in preparation) (2015).
4. Reips, U.-D.: Standards for Internet-based experimenting. Exp. Psychol. 49, 243 (2002).
5. Dai, P., Rzeszotarski, J., Paritosh, P., Chi, E.H.: And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions. In: Proc. of 18th ACM CSCW (2015).
6. Gadiraju, U., Kawase, R., Dietze, S., Demartini, G.: Understanding malicious behavior in crowdsourcing platforms: The case of online surveys. In: Proceedings of CHI (2015).
7. Hoßfeld, T., Seufert, M., Hirth, M., Zinner, T., Tran-Gia, P., Schatz, R.: Quantification of YouTube QoE via crowdsourcing. In: Multimedia (ISM), 2011 IEEE International Symposium on. pp. 494–499. IEEE (2011).
8. Hoßfeld, T., Keimel, C.: Crowdsourcing in QoE Evaluation. In: Quality of Experience. pp. 315–327. Springer (2014).
9. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing User Studies with Mechanical Turk. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 453–456. ACM, New York, NY, USA (2008).
10. Cumming, G., Finch, S.: Inference by eye: confidence intervals and how to read pictures of data. Am. Psychol. 60, 170 (2005).
11. Polzehl, T., Naderi, B., Köster, F., Möller, S.: Robustness in Speech Quality Assessment and Temporal Training Expiry in Mobile Crowdsourcing Environments. In: 16th Ann. Conf. of the Int. Speech Comm. Assoc. (Interspeech 2015). ISCA. pp. 2794–2798.

_____