# Advantages and Challenges of Online Listening Tests for Sound Quality Studies

Serkan Atamer, M. Ercan Altinsoy, Yue Zhang

*Chair of Acoustic and Haptic Engineering, TU Dresden, 01062, Dresden, E-Mail: serkan.atamer@tu-dresden.de*

## Introduction

Coronavirus pandemic negatively affected the researchers need to conduct auditory experiments. Especially the small listening cabins increases the risk of infections since the transmission through small particles in the air is highly common. This caused delays in research projects. Although the cabins usually include the air ventilation system, still the flow rates are quite low due to the background noise considerations, and there are longer waiting periods between the participants for proper air ventilation as well as the surface disinfection, which makes at the end whole testing not so feasible timewise.

Although the online auditory experiments are not a topic of Covid times, unsurprisingly they are becoming more important at the times where it is hard to conduct the auditory experiments in a way before they have been conducted. Researchers are trying in a high pace to convert the experimental studies into the online versions due to the unforeseen laboratory closures or difficulties in participant recruitments. However, experiment platform building, publishing and hosting, data acquisition and participant recruitment can be challenging at the beginning of an online experiment user.

There are advantages and disadvantages of conducting online auditory experiments. Besides being safe in the pandemic, it is possible to reach more participants in a shorter time in an online experiment. Moreover, it is possible to reach different subjects from different cultures, or market regions which makes the cross-cultural research and geographically market oriented sound quality research possible.

Of course on the other hand, there are almost inevitable problems of online auditory testing. Challenges of playback in original levels as well as the problems arising due to the not suitable frequency response of headphones of each participants are the main issues of online auditory experiments. It is also not possible to ensure proper background noise levels for participants. And also the lack of supervision might affect the results drastically, if subjects does not understand the question properly. Lastly, for some studies such as reaction time measurements, hardware latency issues might generate a huge problem which makes the reaction time measurements almost impossible for some platforms. The latency of the different online experiment platforms have been thoroughly investigated in a recent study [1] and the overview of online testing can also be found in different studies published recently. [2,3,4].

The aim of this study is to investigate, if it is possible to get reliable results in online experiments in sound quality studies for some particular kind of stimuli and some particular type of listening experiments, despite disadvantages mentioned above.

## Methodology

First assumption is that usually in the sound quality experiments, selection of stimuli and their relative comparisons effects the subject evaluation rather than the absolute evaluations of each participant for each stimulus. Hence it is assumed that, for average level, not so complicated stimulus such as vacuum cleaner sound should give reliable results also for online testing in an annoyance / disturbance evaluations. Another assumption is that, since it is not possible to control the absolute playback level for each participant, it can still be possible to use testing methods which inherently include the relative responses such as magnitude estimation.

Considering these assumptions in mind, three previous sound quality experiments from three different equipment category are selected and repeated in an online testing platform. The conducted tests are given in Table 1.

**Table 1:** Repeated Listening Experiments

| | **Test1** | **Test2** | **Test3** |
|---|---|---|---|
| Question | Annoyance | Annoyance | Loudness |
| Type | Category Scaling | Category Scaling | Magnitude Estimation |
| Stimuli | Vacuum Cleaner Sounds | Dishwasher Sounds (equally loud) | Electric Shaver Sounds |
| Number of Stimuli | 54 | 38 | 23 |

The first test was an annoyance test of vacuum cleaner noise, using the category scaling to estimate the noises from 0 (not annoying) to 100 (very annoying) by a slider with verbal anchors. Second one is the annoyance test of dishwasher noise with equally loud stimuli. In test 2, the similar category scaling was used. Dishwasher noise is normally rather challenging for sound quality testing since the levels are relatively low, and for the original level playback an isolated cabin is necessary. With this equal loudness test, it is expected to decrease the effect of non-proper level calibration of the playback signal whereas investigate the effects of other features of the sound on sound quality assessments rather than loudness itself. The last test was a loudness evaluation test, using the real shaver signals by using the method of magnitude estimation.

Three tests are conducted in a developed online testing platform. All the tests are conducted in different times with different participants. Exactly same stimuli and the same training stimuli are used in both online and offline versions. The testing methods are also kept identical. At the end, results of the online listening experiments are compared with the former offline versions, and the possible reasons of similarities and deviations are discussed.

## Online Experiment Platform

Online experiments are conducted in a web-page developed using basic HTML, CSS and JavaScript and the publishing and data collection has been done on Firebase platform which offers the free database for limited data transfer. There are advantages and disadvantages of conducting an online experiment in a self-written web-page. First of all, for a study around 40-50 subjects with almost 100 stimuli, each test is free rather than the participant allowance. Also there is no limit for experiment design since the design has solely based on your code. It is possible to make any modifications and adjustments, where it might not be the case for every other option. However, website development can be at the beginning time consuming if there are no former experience available. Also there were some browser compatibility issues. Lastly, that was not an issue in a sound quality evaluation, but proper latency could be an issue for the ones who are interested in reaction times.

## Results of the Online Listening Experiments

### Listening Test 1: Vacuum Cleaner Annoyance

Vacuum cleaner noise, is quite stationary, relatively loud – lying between 60 – 75 dBA and usually perceivable around 100 Hz to 10000 Hz. In this range they have a quite band noise characteristics, only some of them having relative strong tonal components. Some of them also have a low frequency booming tone around 100 Hz. 28 participants attended the online experiment. Subjects were asked to estimate the annoyance of 54 vacuum cleaner noises.

Figure 1 and Figure 2 show the results of the online listening test and the former listening test conducted in the laboratories. Figure 1 shows a relatively strong correlation between both tests, when the mean evaluations are considered. Figure 2 shows the mean evaluations (upper) and the standard deviations (lower) of each stimulus for each test. Here it is possible to see that the online test (blue) has higher standard deviation values than the offline version.

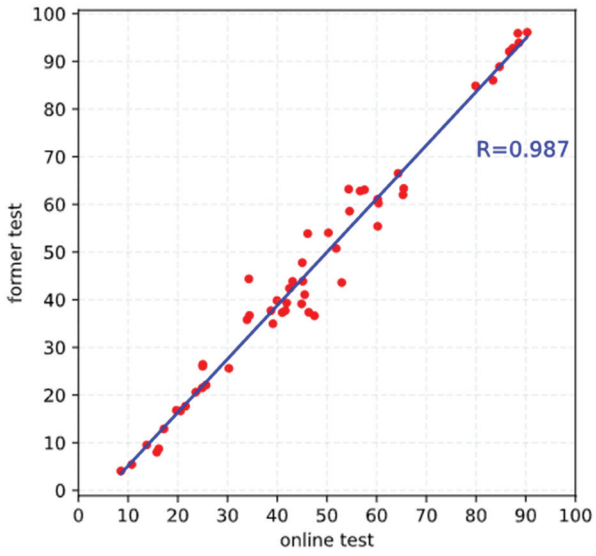


**Figure 1:** Mean scores of the each stimuli for online test and the former offline version for test 1: vacuum cleaner annoyance estimations

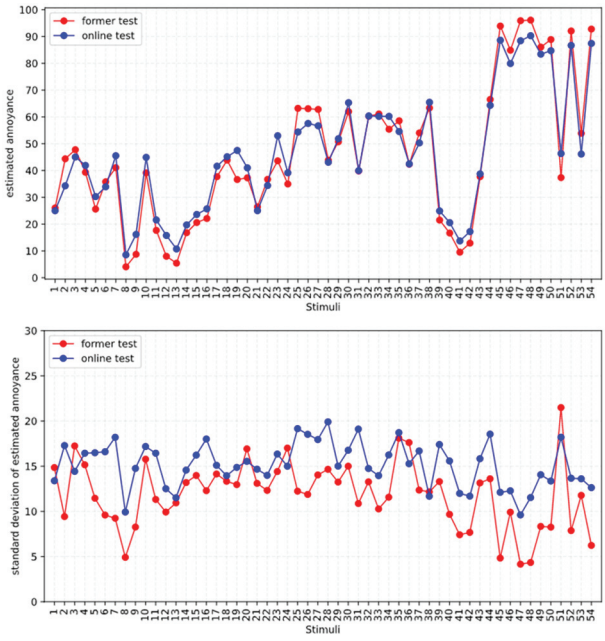


**Figure 2:** Mean annoyance evaluations (upper) and the standard deviations (lower) of each stimulus for test 1: vacuum cleaner annoyance estimations.

### Listening Test 2: Dishwasher Annoyance (Equally Loud Stimuli)

Dishwasher noise is relatively low, and the equalized A-weighted averaged levels in this test were around 30 dB(A). Washing cycle stimuli always are in the form of repetitive water splashes, usually audible around 100 – 500 Hz. Depending on the isolation material used, it is also possible to hear splashes in higher frequencies up to 4-5 kHz in some models. For some of the models, constant booming noise of 100 Hz is also present coming from the input pump. 25 participants are attended in this second online experiment. Subjects were asked to estimate the annoyance of 38 dishwasher noises.

Figure 3 and Figure 4 show the results of the both tests (online and offline). Here Figure 3 shows that the similarity between online experiment and the former experiment was quite weak. Again the standard deviations of annoyance estimations obtained from each dishwasher stimuli is relatively higher in the online experiments (Figure 4, lower). For both tests, participants' mean evaluations are around 50, and this is understandable, considering the fact that one of the strongest effect loudness kept constant. Hence the variance in the annoyance estimations are lower. However, this variation was even smaller for the online test, where the mean annoyance scores of each stimuli was always around 50, showing that they were not able to differentiate any difference between the stimuli. Here, the online test results are not reliable and not comparable with the offline version.

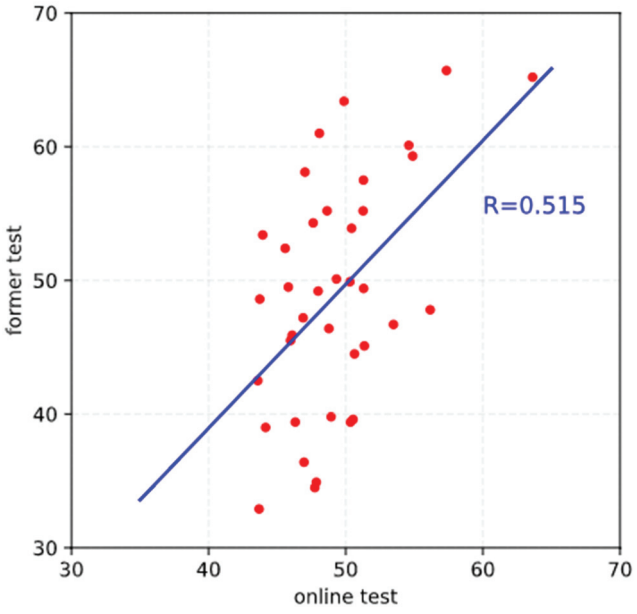


**Figure 3:** Mean scores of the each stimuli for online test and the former offline version for test 2: dishwasher annoyance estimations

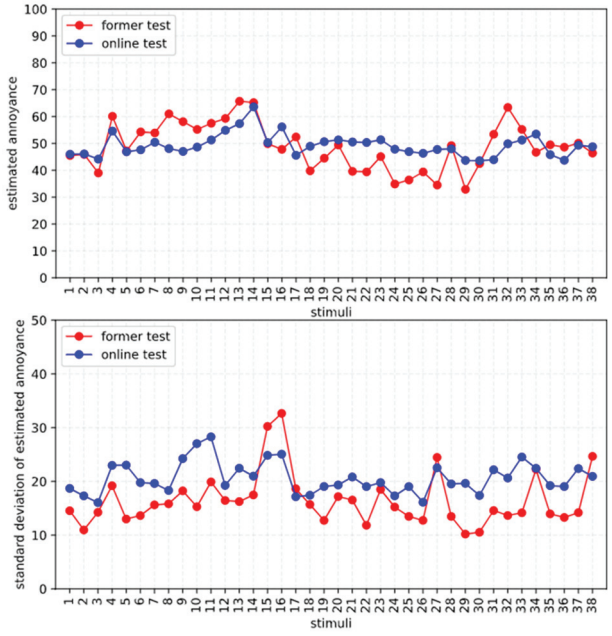


**Figure 4:** Mean annoyance evaluations (upper) and the standard deviations (lower) of each stimulus for test 2: dishwasher annoyance estimations.

### Listening Test 3: Electric Shaver Loudness

Shavers have mostly rough characteristics due to the quite close tonal components from electric motors. A-weighted sound levels are around 55 to 60 dB(A). They are also stationary, as in the case of vacuum cleaners. The quietest stimuli is selected as anchor stimuli for both online and offline tests. Participants need to evaluate the loudness of the each stimuli in comparison to the loudness of the quietest stimuli. 25 people participated in the third listening test and subjects were asked to estimate the loudness of 23 electric shaver noises.

Figure 5 and Figure 6 shows the results of both tests. The correlation between the mean evaluations of both tests are quite strong (Figure 5). Upper part of Figure 6 shows linearized mean evaluations of both tests. Here it is possible to see that the tendency is similar but the absolute evaluations are different. This particular effect can also be seen on the slope of the trend line. On the other hand magnitude estimation tests show no systematical high standard deviations in comparison to the category scaling (Figure 6, lower).

The correlation of mean evaluations between two tests show actually a very high degree, but the slope of this line shows actually a possible logarithmic bias, which can happen in the magnitude estimations [5]. This bias could happen if the same test is repeated again offline, since only the lowest range of the stimulus is used as anchor value.

In order to eliminate this effect, the same test can be conducted again with a different anchor stimuli (particularly for the loudest stimuli or the average loud stimuli) and the results could be averaged. This difference is mainly due to the inherent characteristics of magnitude estimation method, rather than the effect of online testing. Hence it can be said that the online experiment shows reliable results for test 3.

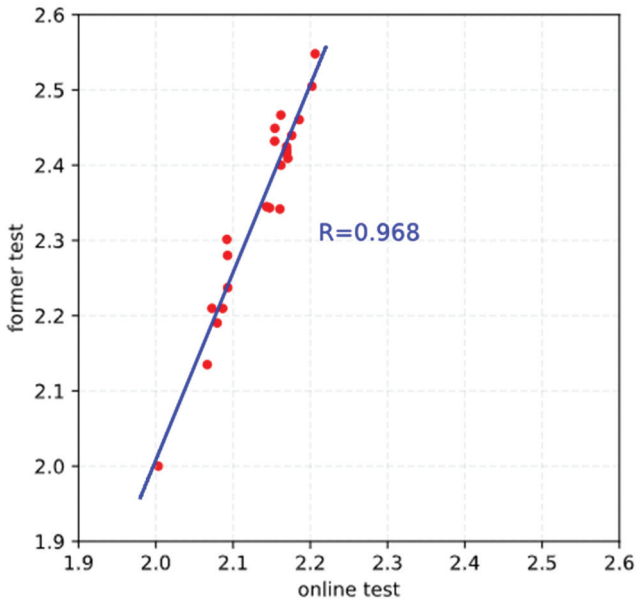


**Figure 5:** Mean scores of the each stimuli for online test and the former offline version for test 3: electric shaver loudness estimations

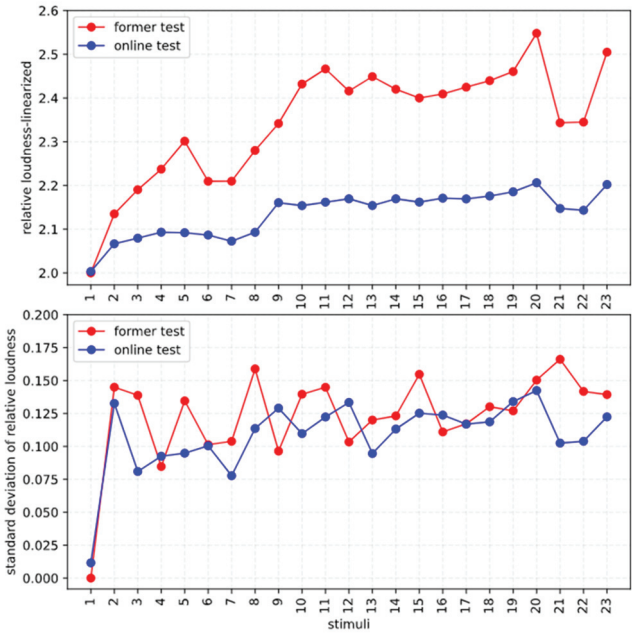


**Figure 6:** Mean annoyance evaluations (upper) and the standard deviations (lower) of each stimulus for test 3: electric shaver loudness estimations.

## Conclusions

It can be concluded that the test 1 and test 3 showed some potential for online testing. The easy to grasp characteristics of noise, their stationarity and their relative loud original levels made it possible to reach this goal. Moreover, the magnitude estimation tests can also be a good candidate for online testing since it is based on relative evaluation of the stimuli.

For the dishwasher test, it is observed that since the participants were allowed to set their own levels for playback, usually a high noise portion, normally in real listening levels not audible is, become audible. Figure 7 shows one example frequency content of a dishwasher stimuli. After 2000 Hz is actually almost inaudible, but if the participants increase the volume too much, then this pretty high noise share makes it almost impossible to concentrate on the other features of the sound. This higher noise content blurred the main effect and the participants were unable to differentiate. As a solution, a low pass filter could be applied to the signal before testing, but the cut off frequency of the filter needs to be fixed properly, so that not any important information is lost for some stimuli changing the timbre.

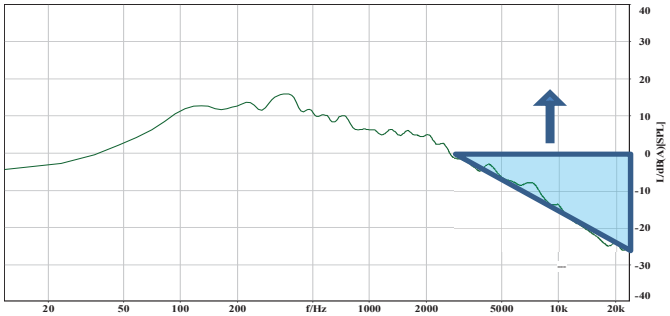


**Figure 7:** One example frequency content of a dishwasher stimuli. After 2000 Hz is actually almost inaudible,

To conclude, online listening tests show great potential, but for some specific cases and specific equipment for sound quality studies. It is better to conduct a small pre-study, online vs. offline, to check the reliability before investing more time and money. And always it needs to be kept in mind that the standard deviations might be higher in online version.

The original playback volume is an inevitable issue, though there are already some ideas to overcome this. Such as asking for a 9 band loudness equalization of a participant. Participant needs to assign the loudness of different signals in comparison to a 1 kHz tone, to make it equally loud [6]. Or there are some ideas about playing back a usual conversation, and the participant is asked to bring the level of this conversation to a normal conversation level. Another study suggests a brief psychophysical test for determining whether online experiment participants are wearing headphones [7]. All of these approaches are relatively strong but of course cannot solve the original playback level or the proper background level problem totally.

## References

[1] Bridges, David, et al. "The timing mega-study: Comparing a range of experiment generators, both lab-based and online." PeerJ 8 (2020): e9414.

[2] Grootswagers, Tijl. "A primer on running human behavioral experiments online." Behavior research methods 52.6 (2020): 2283-2286.

[3] Sauter, Marian, Dejan Draschkow, and Wolfgang Mack. "Building, hosting and recruiting: A brief introduction to running behavioral experiments online." Brain sciences 10.4 (2020): 251.

[4] Woods, Andy T., et al. "Conducting perception research over the internet: a tutorial review." PeerJ 3 (2015): e1058.

[5] Poulton, E C.: Models for biases in judging sensory magnitude. In: Psychological bulletin 86 (1979),

[6] Nicholas Jillings, Brecht De Man, David Moffat and Joshua D. Reiss, "Web Audio Evaluation Tool: A Browser-Based Listening Test Environment," 12th Sound and Music Computing Conference, July 2015.

[7] Woods, Kevin JP, et al. "Headphone screening to facilitate web-based auditory experiments." Attention, Perception, & Psychophysics 79.7 (2017): 2064-2072.