ITU-T SG12 Question 19
Rapporteur Meeting

Agenda

December 15, 2020

Stockholm, Sweden

- **Work items of ITU-T SG12 Question 19**
- **J.noref: Perceptual video quality measurement techniques for digital cable television in the absence of a reference**
- **J.op-tr: Methods for Optimizing Bitrates and Transmission Resolution by Considering Display Characteristics and Available Bandwidth**
- **J.src-vq: Objective Assessment Methods for Source Video Quality at the Headend**
- **J.q-uhd: Quality measurement methods for UHD services**
- **P.910rev: Subjective video quality assessment methods for multimedia applications**
- **P.913rev: Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment**

2018 November Meeting:

- **C270: Influence on multiple comparison on evaluation of objective quality metrics and the number of test subjects (Telefon AB - LM Ericsson, Deutsche Telekom AG)**
- **[Abstract] In subjective video quality tests, statistical significance comparisons are commonly used to determine whether two Mean Opinion Scores (MOS) significantly differ from each other. These analyses typically ignore the impact of multiple comparisons on significance testing, and by consequence reach conclusions that are not supported by the data. This problem can be fixed by increasing the number of test subjects to accommodate the planned analyses. More information is supplied by a recently published journal paper by Brunnström and Barkowsky[1]. This paper also describes the influence on the evaluation of objective quality metrics and shows that multiple comparisons can have large impact on the outcome.**

2018 November Meeting:

- **C270: Influence on multiple comparison on evaluation of objective quality metrics and the number of test subjects (Telefon AB - LM Ericsson, Deutsche Telekom AG)**

- **Proposal: We propose to revise the ITU Recommendations that use statistical tests for MOS comparisons. These Recs. need to describe this problem and recommend best practices. This primarily affects the Recommendations ITU-R Rec. BT.500-13, ITU-T Rec. P.910, ITU-T Rec. P.913, ITU-T Rec. P.800, and ITU-T Rec P.1401.**

- **Output: Revision of P.910 and P.913 initiated**

2019 May Meeting:

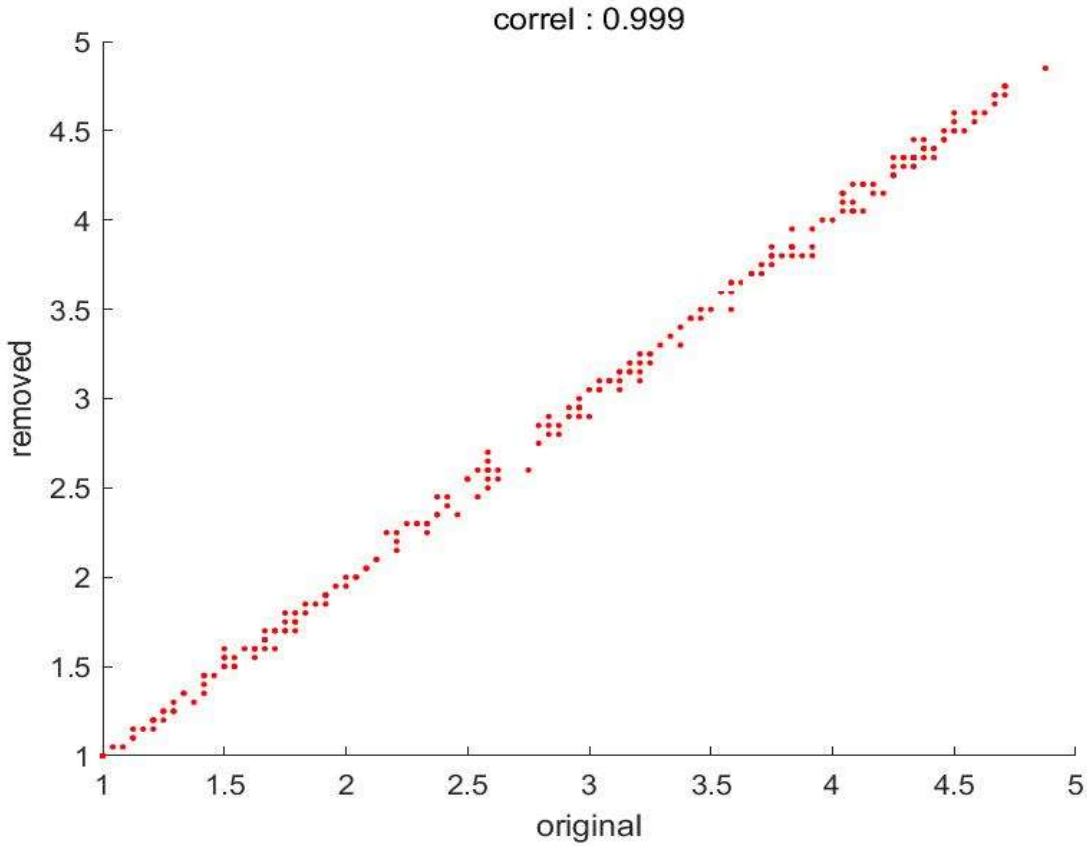
- **C348: Number of viewers and MOS values, Korea (Republic of).**

number of viewers	mean	max	min	sd
23	1.000	1.000	1.000	0.0000558
22	1.000	0.999	1.000	0.0000726
21	0.999	0.999	0.999	0.0001157
20	0.999	0.999	0.999	0.0001484
19	0.999	0.998	0.999	0.0002473
18	0.999	0.998	0.998	0.0002397
17	0.998	0.997	0.998	0.0003095
16	0.998	0.997	0.998	0.0002921
15	0.998	0.996	0.997	0.0003768
14	0.997	0.995	0.997	0.0004553
13	0.996	0.995	0.996	0.0004976
12	0.996	0.994	0.995	0.0006616
11	0.996	0.992	0.994	0.0009391
10	0.994	0.991	0.993	0.0009256
9	0.994	0.990	0.992	0.0012287
8	0.993	0.985	0.990	0.0016697
7	0.992	0.986	0.988	0.0016756
6	0.988	0.981	0.985	0.0022103
5	0.988	0.974	0.981	0.0031840
4	0.981	0.965	0.976	0.0040713
3	0.976	0.952	0.967	0.0056249
2	0.964	0.919	0.944	0.0140100
1	0.943	0.842	0.899	0.0268172

2019 May Meeting:

- **C348: Number of viewers and MOS values, Korea (Republic of).**

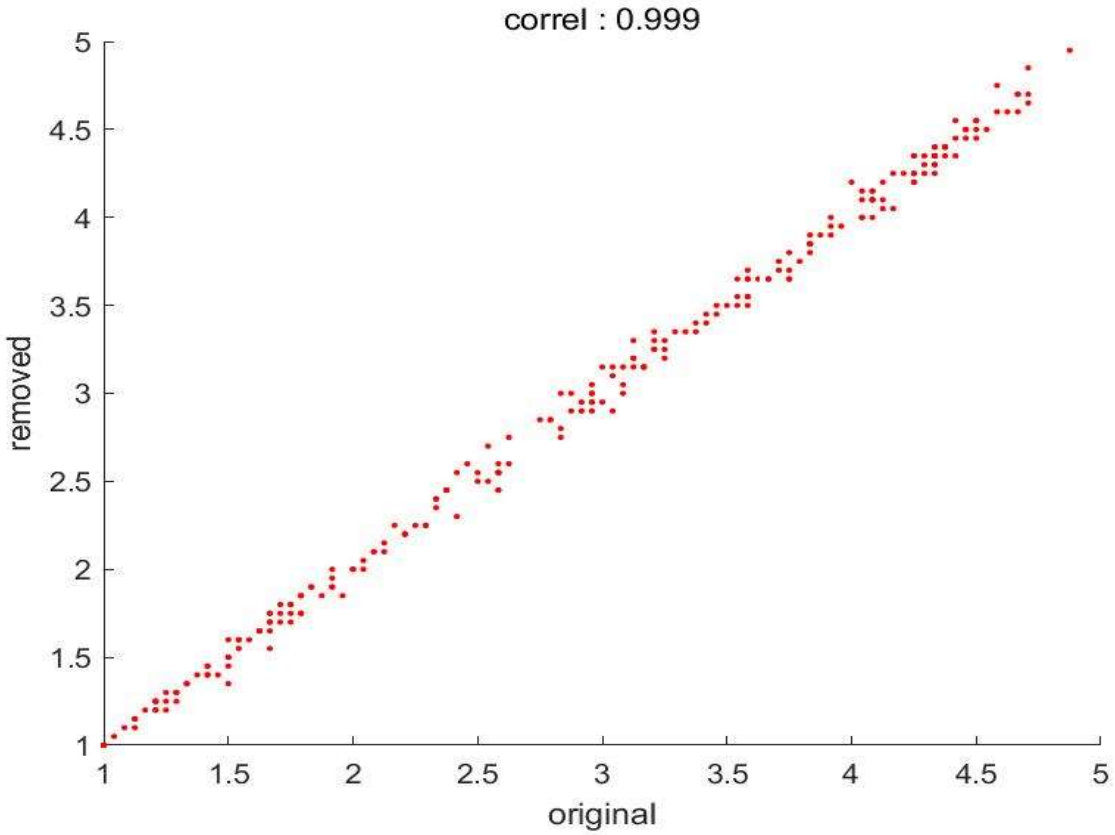
Scatter plot between the MOS scores obtained by averaging 24 viewers and the MOS scores obtained by averaging 20 viewers. (maximum correlation)



2019 May Meeting:

- **C348: Number of viewers and MOS values, Korea (Republic of).**

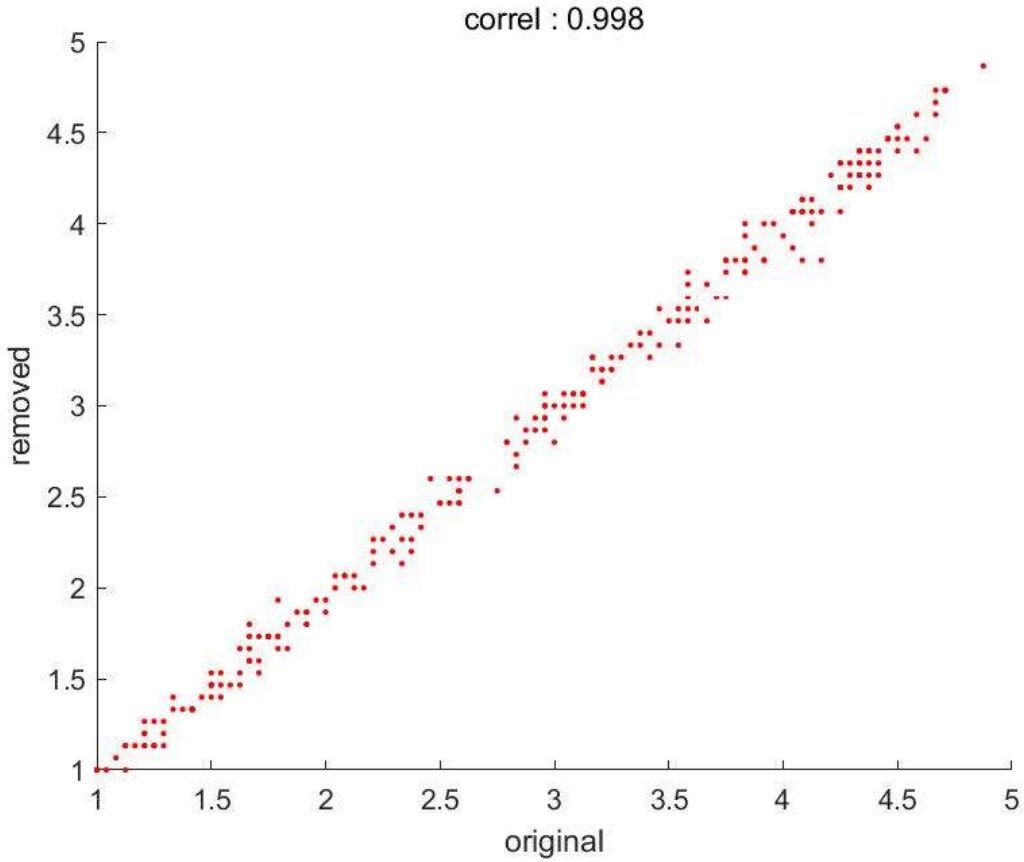
Scatter plot between the MOS scores obtained by averaging 24 viewers and the MOS scores obtained by averaging 20 viewers. (minimum correlation)



2019 May Meeting:

- **C348: Number of viewers and MOS values, Korea (Republic of).**

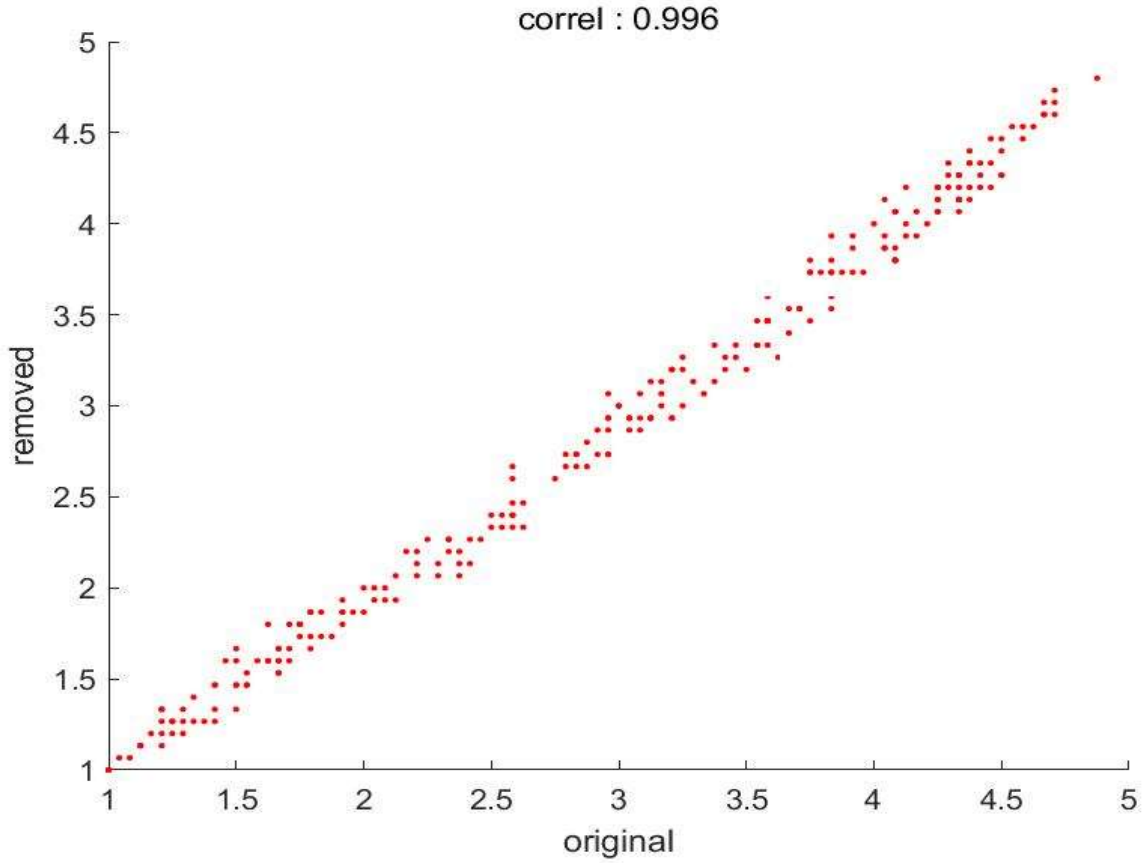
Scatter plot between the MOS scores obtained by averaging 15 viewers and the MOS scores obtained by averaging 20 viewers. (maximum correlation)



2019 May Meeting:

- **C348: Number of viewers and MOS values, Korea (Republic of).**

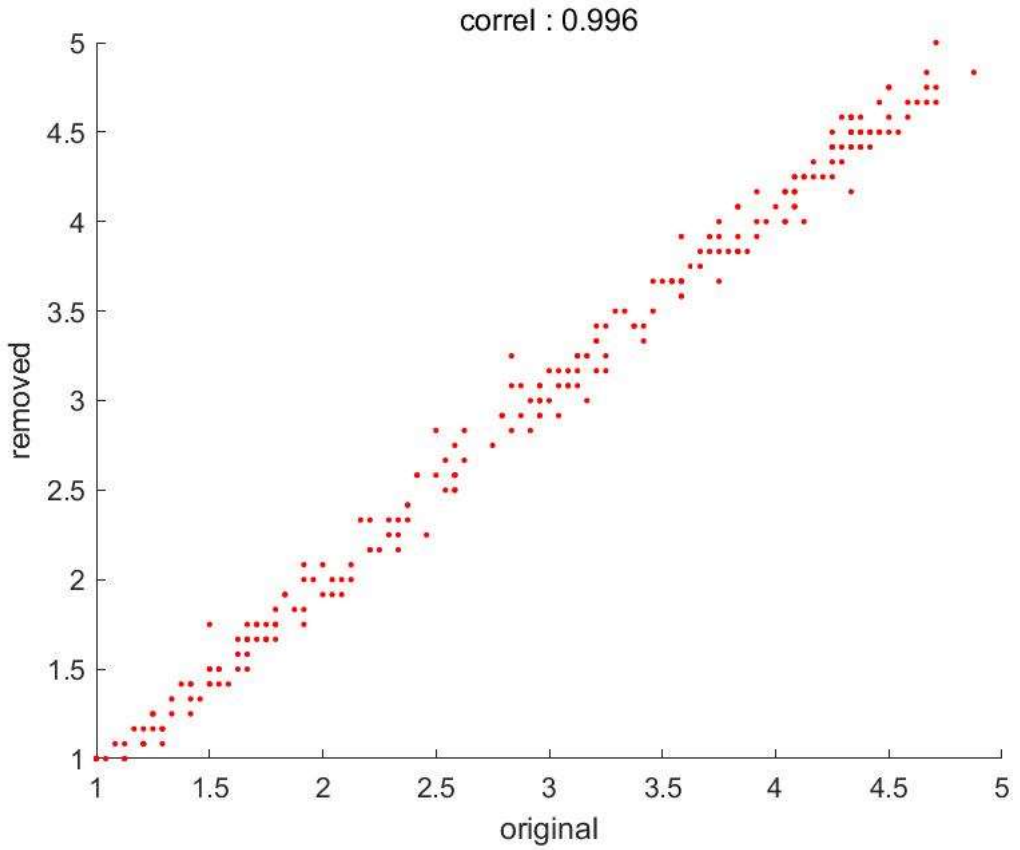
Scatter plot between the MOS scores obtained by averaging 15 viewers and the MOS scores obtained by averaging 20 viewers. (minimum correlation)



2019 May Meeting:

- **C348: Number of viewers and MOS values, Korea (Republic of).**

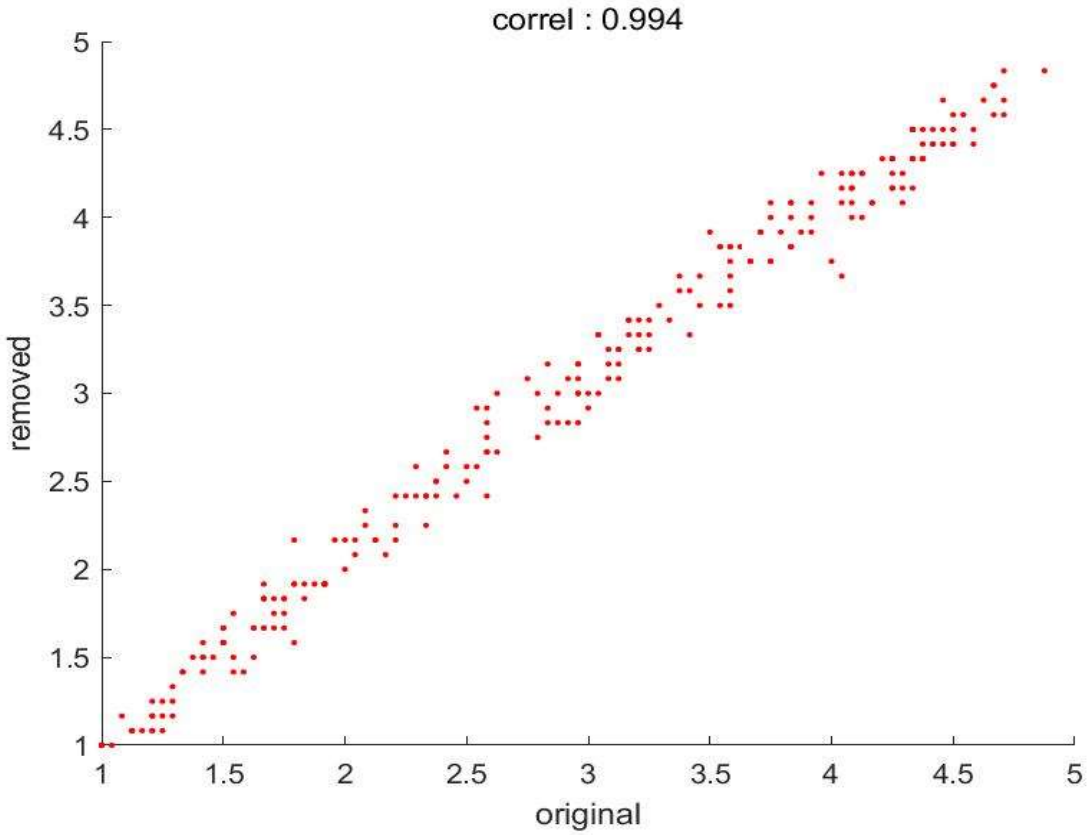
Scatter plot between the MOS scores obtained by averaging 12 viewers and the MOS scores obtained by averaging 20 viewers. (maximum correlation)



2019 May Meeting:

- **C348: Number of viewers and MOS values, Korea (Republic of).**

Scatter plot between the MOS scores obtained by averaging 12 viewers and the MOS scores obtained by averaging 20 viewers. (minimum correlation)



2019 November Meeting:

- **TD1023: LS/i on improvements on subjective experiment analysis process, VQEG**
- **[Abstract] VQEG is working on improvements on subjective experiment analysis process. We think that new estimation methods discussed within VQEG are mature enough to be standardized. We would like to propose changes to standards ITU-R BT.500 (BT.500-14 Section A1-2.3.1) and ITU-T P.913 (Section 12.4) by extending the described analysis process by methods based on the maximum likelihood estimation (MLE) approach. The advantages of the new methods are: 1) better robustness in the presence of outlier subjects in terms of parameter recovery accuracy, and 2) auxiliary information on test subjects on their bias and consistency, which provide guides on subject selection. We propose to update the two standards, motivated by the proposed new methods for screening. The rest of the liaison is for information. We plan to submit a full proposal to the next meeting, together with a white paper with technical details including a full description of the proposed MLE algorithms, extended analysis of results on a variety of public datasets, and a reference implementation.**

2019 November Meeting:

- **The number of test subjects in subjective video quality experiments, VQEG**
- **[Abstract] In subjective video quality tests, statistical significance comparisons are commonly used to determine whether two Mean Opinion Scores (MOS) significantly differ from each other. These analyses typically ignore the impact of multiple comparisons on significance testing, and by consequence reach conclusions that are not supported by the data. This problem can be fixed by increasing the number of test subjects to accommodate the planned analyses. More information is supplied by a recently published journal paper by Brunnström and Barkowsky[1].**
- **It is proposed that in the recommendations ITU-R Rec. BT.500-13, ITU-T Rec. P.910, ITU-T Rec. P.913, that suggested number of subjects be harmonized.**

2019 November Meeting:

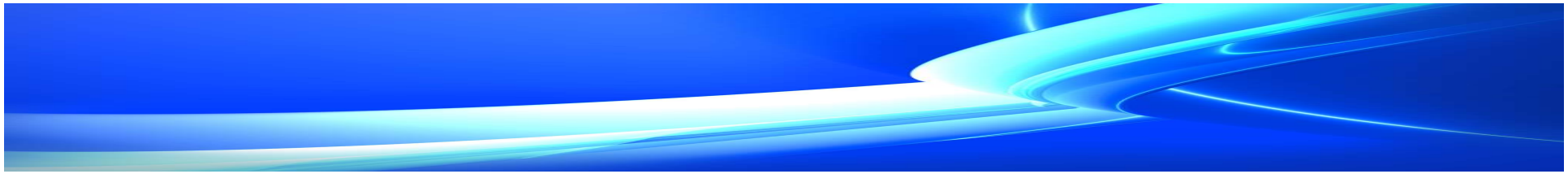
- **The number of test subjects in subjective video quality experiments, VQEG**
- **[Proposed Text] “It is critical to choose the appropriate number of subjects used in for experiments. This number depends, among other factors, on the number of comparisons planned between MOS values, and the anticipated standard deviation in the subjective scores [1]. The number to be used can estimated using power analysis and practically with the following software: https://slhck.shinyapps.io/number-of-subjects/”**

2020 April Meeting:

- **Improvements on subjective experiment data analysis process, Netflix**
- **[Abstract] VQEG is working on improvements on the subjective experiment data analysis process. We think that a new estimation method discussed within VQEG is mature enough to be standardized. We propose to update the post-screening of subjects and data analysis Sections of ITU-T Rec. P.913 with a proposed method based on a new subject inaccuracy model and maximum likelihood estimation (MLE). The new method has advantages in better model-data fit, tighter confidence intervals, better robustness against subject outliers, absence of hard coded parameters/thresholds, and auxiliary information on test subjects. This proposal is submitted together with a white paper with more extensive details, and the proposed modification to the ITU-T Rec. P.913 text. This is a joint contribution by Zhi Li (Netflix), Christos G. Bampis (Netflix), Lucjan Janowski (AGH, Poland) and Ioannis Katsavounidis (Facebook).**
- **[Report] The contribution describes a new method for screening of subjects and data analysis. The group discuss the benefits and potential risks of the proposed analysis method in various circumstances with diverse impairment types. More supporting analysis data would be desirable for the revision of the Recommendation. In particular, more comparison with the current method of P.913 using larger datasets would be helpful to understand the benefits, reliability and limits of the proposed method. Further contributions are encouraged on this topic.**

2020 April Meeting:

- **Subjective Experiment Analysis, Rohde & Schwarz GmbH & Co. KG, Ilmenau University of Technology**
- **[Abstract] This contribution presents a subject model based on observations of individual scores from subjective experiments. This model allows estimating a mean opinion score, the subject bias as well as subject variance, assessing how consistent subjects rate the quality. The goal of the contribution is to continue the discussion about subjective experiment analysis.**
- **[Report] The contribution presents a method to compute mean opinion scores. This contribution deals with a similar topic as that of C470. The group reviewed the method and further contributions are encouraged. In particular, the two methods of C470 and C487 need to be compared with more subjective datasets. Based on these further analyses and comparisons, the revisions of P.910 and P.913 will be progressed.**



THE END