The following text of proposed edits to ITU-T Rec. P.913 includes edits from the May 2022 VQEG meeting. Additional comments appear in the GoogleDoc version, at <https://docs.google.com/document/d/1of8ZsB2nZpi8fn5AOD5r9BfgjuGI4X7L/edit>

Proposed Changes to ITU-T Rec. P.913

This document contains proposed edits to ITU-T Rec. P.913, to be discussed in the SAM meetings. Rec. P.913 encourages updated methods and modified protocols in response to new technologies and specialized testing needs. Thus, it seems a more appropriate venue than P.910 or BT.500. Put simply, P.913 encourages researchers to modify protocols as necessary for their work, as long as they clearly document and explain any deviations from P.913.

**Clause 6.5 “Duration of stimuli**

First paragraph currently reads:

“The methods in this Recommendation are intended for stimuli that range from 5 to 20 s in duration. Sequences of 8–10 s are highly recommended. For longer durations, it becomes difficult for viewers to take into account all of the quality variations and score properly in a global evaluation. The temporal forgiveness effects become important when the time duration of a stimulus is high, see [b-Hands].”

Recommend either change 5 s minimum to 4 s minimum, based a series of four tests conducted with 4 s media. The 5 s minimum seems to have been chosen based on previously conducted tests rather than any implementation of shorter videos. The following subjective tests use 4 s video presentation:

* Margaret H. Pinson, "[ITS4S: A Video Quality Dataset with Four-Second Unrepeated Scenes](https://www.its.bldrdoc.gov/publications/3194.aspx)," NTIA Technical Memo TM-18-532, February 2018
* Margaret H. Pinson, "ITS4S2: An Image Quality Dataset With Unrepeated Images From Consumer Cameras," NTIA Technical Memo TM-19-537, April 2019
* Margaret H. Pinson, "ITS4S3: A Video Quality Dataset With Unrepeated Videos, Camera Impairments, and Public Safety Scenarios," NTIA Technical Memo TM-19-538, April 2019
* Margaret H. Pinson; Samuel Elting, "ITS4S4: A Video Quality Study of Camera Pans," NTIA Technical Memo TM-20-545, December 2019

Similarly, recent tests for adaptive streaming have successfully used ACR and the techniques described in P.913 on long sequences (e.g., 1 or 5 minutes).

We recommend changing the above paragraph above to read as follows. New and modified text is marked in **red**.

“The methods in this Recommendation are intended for stimuli that range from **4** to 20 s in duration. Sequences of 8–10 s are highly recommended.

“Caution must be taken when using stimuli with durations longer than 20 s. For longer durations, it becomes difficult for viewers to take into account all of the quality variations and score properly in a global evaluation. The temporal forgiveness effects become important when the time duration of a stimulus is high, see [b-Hands]. [b-Robitza, 2015], [b-Robitza, 2018], [b-Raake, 2020], [b-Baram ???], and ITU-T Rec. P.809 demonstrate the successful application of the 5-level ACR method for stimuli of 30 second, 1 minute, and 5 minutes duration within the context of quality model development. Tests with 5 minutes duration media can be very enjoyable if the media have audio [b-Brunnström ???].”

[b-Robitza, 2015] Robitza, W., Garcia, M.-N., and Raake, A. (2015). At Home in the Lab: Assessing Audiovisual Quality of HTTP-based Adaptive Streaming with an Immersive Test Paradigm. In Seventh International Workshop on Quality of Multimedia Experience (QoMEX). Costa Navarino.

[b-Robitza, 2018]    Robitza, W., Göring, S., Raake, A., Lindegren, D., Heikkilä, G., Gustafsson, J., List, P., Feiten, B., Wüstenhagen, U., Garcia, M.-N., Yamagishi, K., Broom, S. (2018). HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P.1203 – Open Databases and Software. In 9th ACM Multimedia Systems Conference. Amsterdam.

[b-Raake, 2020]    Raake A. et al. (2020). Multi-Model Standard for Bitstream-, Pixel-Based and Hybrid Video Quality Assessment of UHD/4K: ITU-T P.1204, IEEE Access, vol. 8, pp. 193020-193049.

[b-Brunnström ???]    ????

[b-Barman, 2018]       Barman et al.. (2018). An Evaluation of Video Quality Assessment Metrics for Passive Gaming Video Streaming. In Proceedings of the 23rd Packet Video Workshop (PV '18). Association for Computing Machinery, New York, NY, USA, 7–12. https://doi.org/10.1145/3210424.3210434

**Within Clause 7.2 “Acceptable changes to the methods”, Clause 7.1.1 “Absolute category rating method”**

**Proposed new text appended to 7.1.1.1 “Comments”**

Proposed new text is marked in red. The remaining text already appears in P.913.

“The ACR method produces a high number of ratings in a brief period of time.”

“ACR ratings confound the impact of the impairment with the influence of the content upon the subject (e.g., whether the subject likes or dislikes the production quality of the stimulus).”

“The following statistics characterize the expected precision of a 5-level ACR test. Let us define the subjective test’s confidence interval (ΔSCI) as the minimum difference in MOS values where 95% of stimuli pairs are statistically different (according to the Student’s *t*-test using a 95% confidence level). A subjective test’s ΔSCI is measured using subsets of stimuli pairs that have similar MOS differences (e.g., 0.1 +/- 0.05, 0.2 +/- 0.05, or 0.3 +/-0.05).

Measurements in [b-Pinson, 2020] indicate that the ACR method rarely yields ΔSCI below the following values: 0.5 for 24 subjects, 0.7 for 15 subjects, 1.1 for 9 subjects, and 1.5 for 6 subjects. Unexplained factors in the experiment design and implementation may produce ΔSCI up to the next category of subjects or higher (e.g., 24 subject tests typically have ΔSCI  between 0.5 and 0.7). The size of the confidence interval depends on the post subject screening method. More advanced data cleansing methods (like Clause 12.6) may reduce the confidence interval.”

**Within clause 7.2 “Acceptable changes to the methods”,**

**Propose new clause 7.2.3 “Skip option”**

The proposed new text is:

“Any rating scale may be supplemented with a “skip” option. When selected, the stimuli will not be rated by that subject. Subjects are encouraged to use the “skip” option if they are briefly inattentive and did not observe the stimuli. The stimuli will either not be rated (i.e., a missing value) or put back into the playlist, to be randomly presented later in the session. The recorded data must indicate that the subject used the “skip” option, regardless of the method.

The “skip” option is highly recommended for short video sequences (e.g., 4 s or 5 s duration) and the FOWR protocol (see Clause 9.5). Table 5 of [b-Pinson ITS4S], Table 6 of [b-Pinson ITS4S3], and Table 3 of [b-Pinson ITS4S4] provide statistics on the frequency of subjects using the “skip” option for 4 s stimuli.

Alternatively, subjects can provide the confidence of their rating. More information on this solution can be found in the field of psychometrics, such as [b-Fleming] and [[b-Maniscalco]](https://doi.org/10.1037/rev0000045), and Signal Detection Theory (SDT). [b-Robitza 2014] provides statistics for a “skip” option expressed as a 5-level scale where subjects state their confidence in having provided a reliable rating.

Any rating scale may be supplemented with a “watch again” option. When selected, the stimuli will be presented again to the subject, before they rate the stimuli. The recorded data must indicate that the subject used the “watch again” option. The “watch again” option ensures that each subject will view and rate each stimulus. The “watch again” option may impact ratings, because some subjects will watch again to check details. By contrast, the “skip” option makes sure that each subject rates each stimulus after viewing it only once.

The dataset report must include the frequency at which the “skip” and “watch again” were used. Some concerns have been raised that the “skip” and “watch again” options could complicate difficult tasks, confounding results and possibly impacting the ratings.”

[b-Pinson ITS4S]    Margaret H. Pinson, "[ITS4S: A Video Quality Dataset with Four-Second Unrepeated Scenes](https://www.its.bldrdoc.gov/publications/3194.aspx)," NTIA Technical Memo TM-18-532, February 2018.

[b-Pinson ITS4S3]    Margaret H. Pinson, "[ITS4S3: A Video Quality Dataset With Unrepeated Videos, Camera Impairments, and Public Safety Scenarios](https://www.its.bldrdoc.gov/publications/3220.aspx)," NTIA Technical Memo TM-19-538, April 2019.

[b-Pinson ITS4S4]    Margaret H. Pinson; Samuel Elting,  "[ITS4S4: A Video Quality Study of Camera Pans](https://www.its.bldrdoc.gov/publications/3233.aspx) ," NTIA Technical Memo TM-20-545, December         2019.

[b-Robitza 2014]    W. Robitza and H. Hlavacs, "Assessing the validity of subjective QoE data through rating times and self-reported confidence," 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), 2014, pp. 297-302, doi: 10.1109/QoMEX.2014.6982335.

[b-Fleming]    Stephen Fleming and Nathaniel D. Daw (2017) “Self-Evaluation of Decision-Making: A General Bayesian Framework for Metacognitive Computation,” Psy-chological Review, 124 (Jan 2017), 91–114.<https://doi.org/10.1037/rev0000045>

[[b-Maniscalco]](https://doi.org/10.1037/rev0000045) Brian Maniscalco and Hakwan Lau (2012) “A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings,” *Consciousness and Cognition* 21 (2012), 422–430.

Statistics for use of the “skip” option are as follows. [b-Pinson ITS4S] was a long test of 600 stimuli in a controlled environment. 33% of subjects never used the "skip" option and the subject who used “skip” the most often used this option for 2.3% of the stimuli. [b-Pinson ITS4S3] was short tests of 100 stimuli in a public environment, where 68% of subjects never used the “skip” option and the subject who used “skip” the most often used this option for 10% of the stimuli. [b-Pinson ITS4S4] was another short test of 100 stimuli in a public environment, where 76% of subjects never used the “skip” option and the subject who used “skip” the most often used this option for 6% of the stimuli. In [b-Robitza 2014], the “skip” option was expressed as a level of confidence. The incidence rate of “inconfident” was 3.8% and “very inconfident” was 0.3%. All four studies indicate that the “skip” option will not be abused.

**Clause 9.1 “Number of subjects”**

Sixth paragraph currently reads:

“The number of subjects in an experiment can be reduced, if each subject scores each PVS multiple times. One rating from each of 24 subjects should yield approximately the same accuracy as three ratings from each of nine subjects. This technique would not be appropriate when the goal of the experiment is to characterize differences among subjects, see [b-Janowski].”

The proposed new text is:

“The number of subjects in an experiment can be reduced, if each subject scores each PVS multiple times. For such purpose it is recommended to use the FOWR subject protocol as described in Clause 9.5.”

**Propose new Clause 9.5 “Few observers with repetitions (FOWR) subject protocol”**

This clause describes the few observers with repetitions (FOWR) protocol for subject selection. With FOWR, a small number of subjects rate the same set of stimuli repeatedly, on different days. For most applications, the FOWR protocol has 4 subjects rating all stimuli 4 times on subsequent days. This is referred to as the 4 × 4 FOWR protocol. If accurate agreement is required, the FOWR protocol can be increased to 5 subjects scoring 5 times (5 × 5 FOWR protocol) or 6 subjects scoring 5 times (6 × 5 FOWR protocol). See [b-Perez].

The 4 × 4 FOWR protocol is not unacceptably worse than a 15-subject test. The 5 × 5 FOWR protocol and 6 × 5 FOWR protocol are not unacceptably worse than a 24-subject test.

The FOWR method allows a small team to make a quick and reasonably accurate quality assessments, when the time and expense of subject recruitment is non-viable. The subjects could be a team members or colleagues who work for the same company but on different projects or for different teams. It is important monitor these subjects. For example, ask whether the participant knows the technical side of the system and make sure they do not focus too much on their technical insights. Subjects should be asked to use the system as a naïve subject. The FOWR protocol is recommended for pilot studies (to indicate trending), for pre-tests, and when an objective metric is not available (e.g., new technologies, camera capture).

There are intrinsic limitations on the FOWR protocol, particularly with respect to its capacity for agreement, as subject bias cannot be accurately characterized or compensated. This technique would not be appropriate when the goal of the experiment is to characterize differences among subjects, see [b-Janowski].

[b-Perez]    Perez P., Janowski L., Garcia N., Pinson M., Subjective Assessment Experiments That Recruit Few Observers With Repetitions (FOWR), *IEEE Transactions on Multimedia*, doi: 10.1109/TMM.2021.3098450.

**Propose new Clause 10.6 “Conventional vs unrepeated scene designs”**

The conventional experiment design contains a full matrix of source stimuli and conditions of interest. That is, all source stimuli are processed through all video processing chains (e.g., codec, encoder, bit-rate, coder settings, network errors, and decoder). This allows statistically significant comparisons between the codecs, encoding options, and network conditions. Subjects are exposed to the same source stimuli many times during the experiment, which may increase boredom. The conventional experiment design may be impractical or undesirable for some areas of research (e.g., when evaluating camera capture impairments).

In the unrepeated scene experiment design, each subject views each source stimuli only once. An unrepeated scene experiment can be structured similarly to a full matrix design. When examining the video quality of cameras, the same set of scenes could be photographed with different cameras. When using long stimuli to evaluate technologies like HTTP adaptive streaming, full-length content (like music videos or sports games) could be divided into 5-minute stimuli and one segment from each content associated with each processing chain. Caution should be taken when using full-length content that conveys a story or has temporal changes to the spatial-temporal complexity, because different segments could influence quality ratings. Unrepeated scene experiment designs may increase realism, reduce subject boredom, and prevent subjects from memorizing stimuli. Data analyses may be more difficult because the stimuli and condition variables are confounded, leading to decreased statistical power when evaluating the impact of conditions on quality ratings. See [b-Janowski, 2019].

Unrepeated scene designs have been used successfully within the context of creating the recommendation series ITU-T Rec. P.1203 and ITU-T Rec. P.1204. Results from tests for the development of the ITU-T P.1203 recommendation have been presented in [b-Robitza, 2015] and [b-Robitza, 2018].

Unrepeated scene experiment designs are important for immersive tests, which focus the subject on the system’s intended usage scenario. Immersive tests prioritize realism and try to match the sensory experience of the target application. Choosing subject matter, impairments, and stimuli playback mechanisms that mimic the target application. The subject may be instructed to keep the intended application in mind, while rating stimuli. Examples of immersive tests can be found in [b-Pinson ITS4S3] and [b-Robitza, 2015]. Immersive tests may increase the ecological validity of the results at the expense of introducing additional confounding variables.

Immersive tests typically include audio, because consumers rarely watch videos with no sound. The use of audiovisual stimuli to evaluate video-only impairments has consequences, because the overall audiovisual quality (*AVQ*) can be predicted from the video-only quality (*VQ*) and the audio-only quality (*AQ*). [b-Pinson, 2011] indicates that the product

*AVQ* ≅ *VQ* × *AQ*,

provides a simple and reasonably accurate estimation. Some studies support a more general form with additive terms:

*AVQ* ≅ *w0* + *w1* × *VQ* + *w2* × *AQ* + *w3* × *VQ* × *AQ*,

but there is little agreement on the relative magnitude of the weights (*w0*, *w1*, *w2*, and *w3*), which also depend on the exact application, e.g., video streaming versus video conferencing. If video impairments are to be studied, it is recommended to select stimuli with high quality audio.

[b-Janowski, 2019]    Janowski L., Malfait L., Pinson M. (2019), Evaluating experiment design with unrepeated scenes for video quality subjective assessment, Quality and User Experience, 4, 2.

[b-Robitza, 2018]    Robitza, W., Göring, S., Raake, A., Lindegren, D., Heikkilä, G., Gustafsson, J., List, P., Feiten, B., Wüstenhagen, U., Garcia, M.-N., Yamagishi, K., Broom, S. (2018). HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P.1203 – Open Databases and Software. In 9th ACM Multimedia Systems Conference. Amsterdam.

[b-Robitza, 2015]     Robitza, W., Garcia, M.-N., and Raake, A. (2015). At Home in the Lab: Assessing Audiovisual Quality of HTTP-based Adaptive Streaming with an Immersive Test Paradigm. In Seventh International Workshop on Quality of Multimedia Experience (QoMEX). Costa Navarino.

[b-Pinson, 2011    Margaret H. Pinson; William J. Ingram; Arthur A. Webster,         "[Audiovisual Quality Components: An Analysis](https://www.its.bldrdoc.gov/publications/2565.aspx)," IEEE Signal *Processing Magazine*, vol.28, no.6, pp.60-67, Nov. 2011 [doi: 10.1109/MSP.2011.942470](http://dx.doi.org/10.1109/MSP.2011.942470)

**Propose new Clause 12.7 “Disagreement rate for lab-to-lab and method-to-method comparisons”**

This clause contains guidelines for lab-to-lab comparisons (i.e., when the subject pool for a single experiment is split among two or more laboratories) and method-to-method comparisons (i.e., when the same stimuli are rated with two different test methods).

Given a subjective test, we will choose all pairs of stimuli, **A** and **B**, where both stimuli were rated by the same subjects and the stimuli are drawn from the same dataset. An occasional missing rating is acceptable. We will use the MOS values and the paired stimuli Student’s *t*-test to compare the rating distributions for **A** and **B** at the 95% confidence level. For each lab’s subjects, we will decide whether **A** is better than, equivalent to, or worse than **B**.

We will then tally the frequency of the four possible classification types, defined below. The confusion matrix is presented in Table 1.  We will not apply a correction factor for the number of comparisons, because this would change the distribution of conclusions. That is, we want to estimate all four incidence rates, not the overall likelihood of type 1 or type 2 errors.

* *Agree Ranking*    Both labs conclude that quality of **A** is better than the quality of **B**, or both labs conclude that the quality of **A** is worse than the quality of **B**
* *Agree Tie*        Both labs conclude that **A** and **B** have statistically equivalent quality
* *Unconfirmed*    One lab can rank order the quality of **A** and **B** but the other lab concludes that **A** and **B** have statistically equivalent quality
* *Disagree*        The labs reach opposing conclusion on the quality ranking of **A** and **B**

Table 1 — Confusion Matrix for Different Subjective Test Labs or Different Test Methods

|  |  |  |
| --- | --- | --- |
|  |  | **Subjective Test 1** |
|  |  | **Better** | **Equivalent** | **Worse** |
| **SubjectiveTest 2** | **Better** | Agree Ranking | Unconfirmed | Disagree |
| **Equivalent** | Unconfirmed | Agree Tie | Unconfirmed |
| **Worse** | Disagree | Unconfirmed | Agree Ranking |

Calculate the *disagree* incidence rate as a percentage of all pairs of stimuli. Based on statistics provided by [b-Pinson, 2020], *disagree* incidence rates above 0.31% are unusual enough to warrant investigation and *disagree* incidence rates above 1.0% indicate a method-to-method difference or lab-to-lab difference.

The *agree ranking*, *agree tie*, and *unconfirmed* incidence rates are impacted by the range of quality, test method, and number of subjects in the test.

[b-Pinson, 2020]    Pinson M. (2020). Confidence Intervals for Subjective Tests and Objective Metrics That Assess Image, Video, Speech, or Audiovisual Quality, *NTIA Technical Report 21-550*.

**Propose new Clause 10.7 “Framework for Evaluating Specific Tasks”**

This clause contains guidelines for designing tests that investigate the quality requirements of specific tasks.

ITU-T Rec. P.912 describes methods that are suitable for video tasks where there is a right and wrong answer (e.g., reading a license plate). P.912 was inspired by speech intelligibility tests, where there is a right and wrong answer (phonemes or words spoken). With a two dimensional audio signal, we can compute the signal to noise ratio (SNR) between the desired signal (speech) and the undesired signal (e.g., background noise).

The methods in P.912 cannot be used for many visual and audiovisual tasks. We cannot compute SNR for video. For many tasks, there is no threshold below which video becomes “useless.” For example, even a very low quality surveillance video can show whether someone is present in an area. Instead of a usability threshold, we must establish the relationship between video quality and the value of a video for the specific task. This relationship may be influenced by many factors (e.g., the subject’s expertise and market expectations).

The following framework allows the methods described in this Recommendation to be used, when analyzing the quality needs of a specific task. First, select stimuli and impairments that are typical for the task. Accurate examples (real or simulated) will help subjects notice problems that would hinder the task. Second, the instructions must encourage subjects to consider that specific task when rating the stimuli. Third, the subjects must have some knowledge of the task. Experts may yield increased accuracy.

Subjective tests conducted with this framework can be found in the following papers. [b-Kumcu, 2017] analyzes denoising and laparoscopic surgery, using various rating methods. [b-Martini, 2014] investigates the needs of physicians when assessing ultrasounds. [b-Kara, 2016] analyzes willingness to pay. [b-Pinson ITS4S3] analyzes camera quality needs of first responders using 5-level ACR.

[b-Martini, 2014]     Razaak, M., Martini, M.G. and Savino, K., 2014. A study on quality assessment for medical ultrasound video compressed via HEVC. IEEE Journal of biomedical and health informatics, 18(5), pp.1552-1559.

[b-Kara, 2016]    Kara, P.A., Kovács, P.T., Martini, M.G., Barsi, A., Lackner, K. and Balogh, T., 2016. From a different point of view: How the field of view of light field displays affects the willingness to pay and to use. QoMEX 2016.

[b-???]        (Netflix study)

[b-Pinson ITS4S3]    Margaret H. Pinson, "[ITS4S3: A Video Quality Dataset With Unrepeated Videos, Camera Impairments, and Public Safety Scenarios](https://www.its.bldrdoc.gov/publications/3220.aspx)," NTIA Technical Memo TM-19-538, April 2019.

[b-Kumcu, 2017]    A. Kumcu, K. Bombeke, L. Platiša, L. Jovanov, J. Van Looy and W. Philips, "Performance of Four Subjective Video Quality Assessment Protocols and Impact of Different Rating Preprocessing and Analysis Methods," *in IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 48-63, Feb. 2017, doi: 10.1109/JSTSP.2016.2638681.