ON THE PROPERTIES OF SUBJECTIVE RATINGS IN VIDEO QUALITY EXPERIMENTS

Stefan Winkler

Symmetricom
San Jose, CA 95131, USA

ABSTRACT

Subjective video quality experiments are classical statistical measurements, and as such the mathematical tools for their analysis are well understood. However, there remain certain practical aspects that are rarely discussed, yet are essential to the design of efficient experiments. Using numerical simulations as well as real-life data from a multitude of subjective experiments, this paper attempts to shed more light on the distribution and variability of subjective ratings, the effects of discrete rating scales, and the number of subjects needed.

Index Terms— Subjective experiments, video quality measurement, Mean Opinion Score (MOS)

1. INTRODUCTION

In the field of signal processing and telecommunication, speech quality measurement has quite a long history. More recently, quality assessment has been extended to audio and video as well [1]. Procedures and standards for subjective assessment of speech, audio and video have been around for many years. Due to the proliferation of digital audio and video content, conducting subjective experiments to measure its quality has become relatively commonplace.

There are questions that often arise in the design of a subjective experiment or in the analysis of subjective data, for which there may be accepted guidelines and procedures, or choices that are made based on the intuition of the experimenters; at the same time, it is surprisingly difficult to find scientific evidence to substantiate the implications and trade-offs of these choices. One issue that several studies have explored is continuous- vs. single-rating methods [2, 3]. Other examples of such questions include:

- Which rating scale should be used?
- How many subjects are needed?
- Do the subjective ratings indicate any problems in terms of accuracy or reliability?

We address these issues by investigating aspects related to the distribution and variability of the ratings. We use numerical simulations as well as real-life data from a multitude of subjective experiments to better understand the statistics behind the data.

The paper is organized as follows. In Section 2, we simulate subjective ratings to evaluate the theoretical behavior of such data. In Section 3, we analyze data from real-life subjective experiments and compare them to the simulations. Section 4 concludes the paper.

2. SIMULATED DATA

The premise is that media quality is a statistical quantity. Every subject's rating thus represents a sample of a distribution. This distribution can be characterized by various parameters, most importantly the Mean Opinion Score (MOS), i.e. the rating of a certain clip averaged over all subjects.

2.1. Standard Deviation and Confidence Intervals

The confidence interval of the mean of a distribution of subjective ratings is given by:

$$\pm\, t_P \frac{\sigma}{\sqrt{N}}, \tag{1}$$

where N is the number of subjects, σ is the estimated standard deviation, and t_P is based on the t-distribution.[1] P specifies the probability or confidence that the mean lies within the given interval; common values for P are 95% or 99%. As an example, $t_{95\%} = 2.093$ for $N = 20$.

Naturally, the confidence interval becomes smaller when we have more subjects, simply due to the higher N and lower t (the value of t also decreases with more degrees of freedom $N-1$). Let us instead look at the standard deviation σ, which describes the variability inherent in the data, irrespective of the number of samples.

We first simulate subjective ratings using a random number generator. Assuming a Gaussian distribution of the subjective ratings, we choose a standard deviation of 15 on a rating scale of 0-100 (in other words, we express the standard

[1] The t-distribution should be used whenever the variance is unknown and has to be estimated from the sample data. For large sample sizes, the t-distribution approaches the normal distribution; for small sample sizes, the t-distribution is more tail-heavy. As a result, one has to extend farther from the mean to cover a given percentage of the area.

deviation in percent of the scale). This value is not untypical of real subjective experiments, as we will see below in Section 3.

Depending on the number of subjects and the actual ratings produced, the estimated standard deviation will vary. The results are shown in Figure 1. It can be seen that the average standard deviation rapidly approaches the actual value of 15; in fact, for 10-15 subjects the relative difference is below 2% already. Of course, the value of σ depends on the actual samples – this variation also decreases with N, as shown by the gray area around the average value in the plot. As expected from Eq. (1), the confidence interval size continues to decrease even when σ is nearly constant.

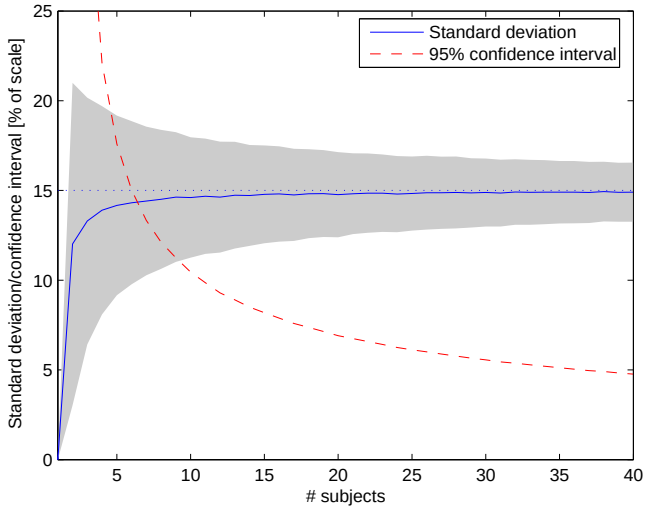


Fig. 1. Average standard deviation σ (solid blue line) and 95% confidence interval size (dashed red line) as a function of the number of subjects N. The gray area indicates the standard deviation of the distribution of σ.

2.2. MOS Variability

It is also instructive to look at the influence of the subjective rating scale on the variability of the MOS. Here we simulate the following scales:

- Continuous 0-100 scale.
- Discrete 0-100 scale.
- Discrete 11-point scale (0-10).
- Discrete 9-point scale (1-9);
- Discrete 5-point scale (1-5).

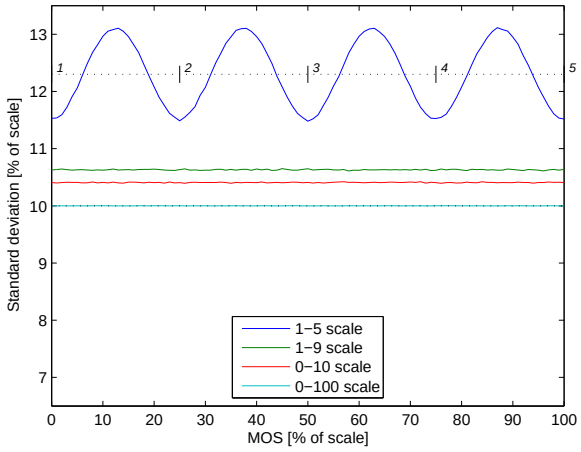
The 0-100 scale is defined in ITU-R Recommendation BT.500 [4] and is normally used for Double Stimulus Continuous Quality Scale (DSCQS) testing and Single Stimulus Continuous Quality Evaluation (SSCQE). The continuous version is representative of any continuous quality scale for our purposes. The 11-point and 9-point scales are defined in ITU-T Recommendation P.910 [5] and are used for Absolute Category Rating (ACR) testing. The 5-point scale is perhaps the most well-known. It is defined in several ITU standards [4, 5, 6] and is commonly used for Double Stimulus Impairment Scale (DSIS), Degradation Category Rating (DCR), and Absolute Category Rating (ACR) testing.

We map the samples from the continuous Gaussian distribution to discrete scales by scaling them to the appropriate range (e.g. divide by 4 and add 1 to get to the 1-5 scale), followed by a rounding operation. This has an interesting effect on the standard deviation of MOS, which is shown in Figure 2(a): The standard deviation not only increases as the number of discrete ratings on the scale goes down, it also varies with MOS due to the discretization of the ratings. Using the example shown, a $\sigma = 10$ on the 0-100 scale equates to a σ oscillating around 12–13% on the discrete 1-5 scale. Both effects are most pronounced for the 1-5 scale; the variation virtually disappears for the finer scales, while some offset remains. The discrete and continuous 0-100 scales are almost indistinguishable.

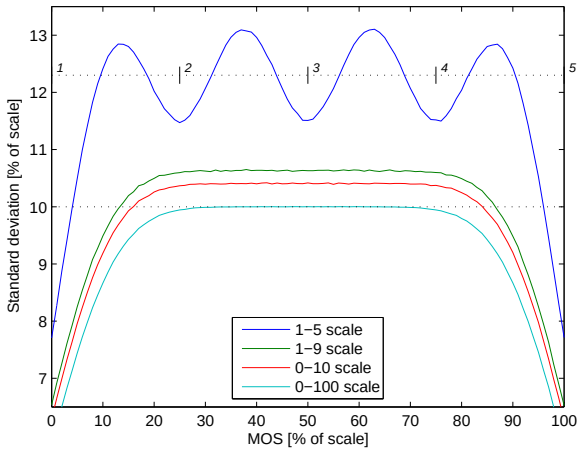
Figure 2(a) shows results without clipping of the ratings, i.e. the ratings are allowed to go beyond the ends of the scales to maintain the shape of the distribution across the MOS scale. Clipping ratings to the ends of the scale has the effect illustrated in Figure 2(b,c) – the standard deviation is significantly reduced towards the ends of the scale. As we will see in Section 3, this effect is found in data from real subjective experiments as well, where it can be even more pronounced.

The offset (i.e. the increase of σ for discrete scales compared to a continuous scale) and the amount of variation depend on the value of the standard deviation itself. This can already be seen comparing Figure 2(b) and (c), where the significant variability evident for the 1-5 scale with $\sigma = 10$ becomes hardly noticeable for $\sigma = 15$, and the offset reduces from 2–3% to less than 2% over the baseline σ.

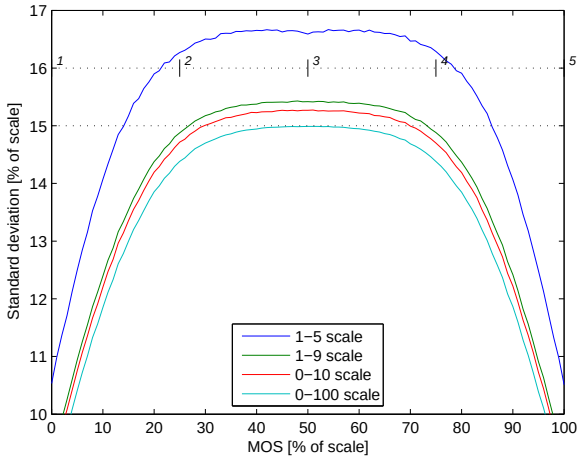
Figure 3 attempts to illustrate the general behavior of this phenomenon. Essentially, the variation that is present for small standard deviations (as indicated by the gray patches in the plot) disappears almost completely once σ passes a certain threshold. For the 1-5 scale, this threshold is $\sigma \approx 15$; for the 0-10 and 1-9 scales, it is $\sigma \approx 5$ (the 1-9 scale is not shown in the figure, as it is very close to the 0-10 scale). As we will see in Section 3 below, the standard deviation for real subjective experiments is generally above these thresholds. The offsets (e.g. 1-2% of the scale for the 1-5 scale) remain even for larger values of σ, i.e. discrete scales intrinsically have somewhat larger standard deviations than a continuous scale. In practice however, the effect is only significant for the 1-5 scale.



(a) No clipping, $\sigma = 10$



(b) Clipped to range, $\sigma = 10$



(c) Clipped to range, $\sigma = 15$

Fig. 2. Standard deviation as a function of rating scale and MOS. For easier comparison, all values are shown in % of the respective scales (the 1-5 scale is mapped into the plots for reference). σ denotes the standard deviation of the original continuous samples without clipping.

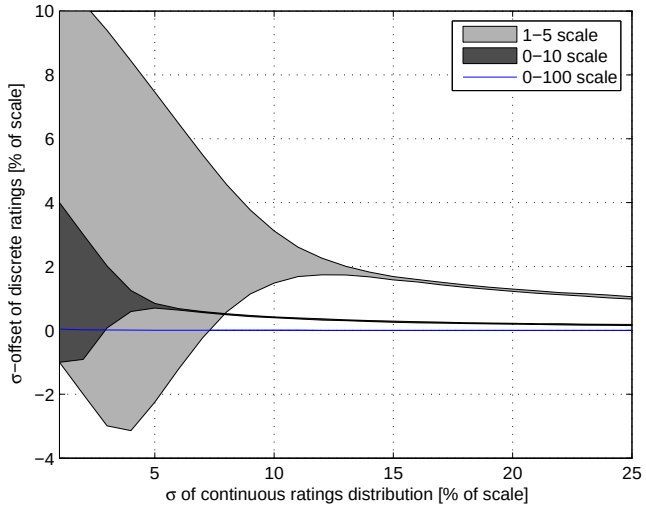


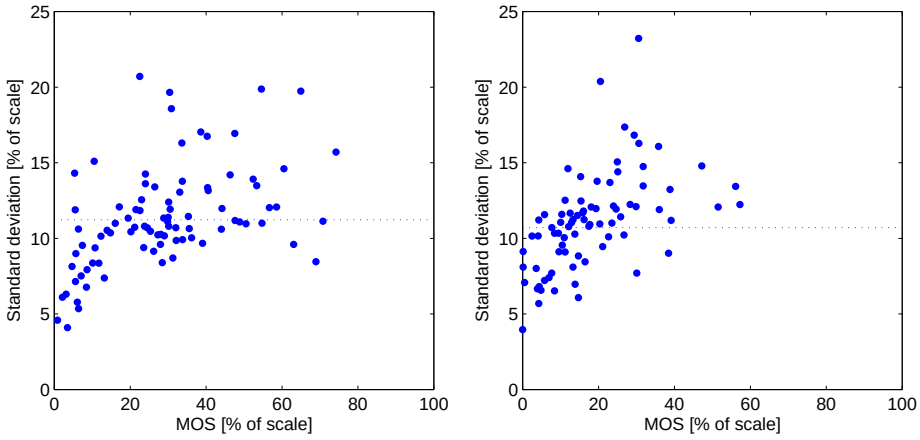
Fig. 3. Range of standard deviation offsets of various rating scales over σ of the continuous scale. No clipping was assumed for all scales for clarity, cf. Figure 2(a).

3. EXPERIMENTAL DATA

In this section, we look at rating data from a variety of subjective experiments. We focus the analyses on the standard deviation and MOS.

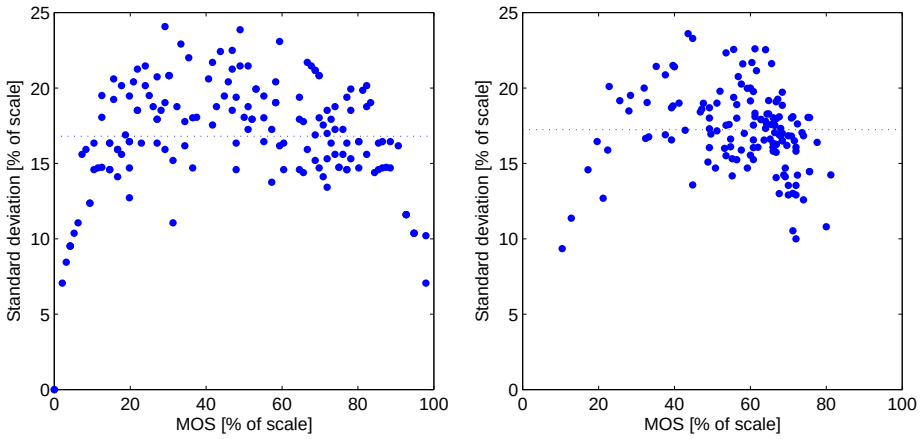
3.1. Data Sets

- Experiment #1 (streaming): Medium-size (CIF) clips, compression and network losses. 11 source sequences, 7 test conditions. DSIS method, discrete 5-point scale, 20 subjects [3].
- Experiment #2 (VQEG Multimedia test, cif01 set): Medium-size (CIF) clips, compression only. 8 source sequences, 16 test conditions. ACR method, discrete 5-point scale, 24 subjects [7].
- Experiment #3 (IPTV): SD format, compression only. 11 source sequences, 14 test conditions. ACR method, discrete 11-point scale, 27 subjects (unpublished).
- Experiment #4 (audiovisual): QCIF format video with stereo audio track. 6 audiovisual source sequences, 10/7/8 video/audio/audiovisual test conditions in separate sessions. ACR method, discrete 11-point scale, 24 subjects [8].
- VQEG Full Reference Television (FR-TV) Phase I: SD format, compression and transmission impairments. 20 source sequences, 16 test conditions. 8 labs, separate low- and high-quality sessions. DSCQS method, discrete and continuous 100-point scales, 17–27 subjects (depending on lab) [9].



(a) VQEG FR-TV I, lab 1, low quality

(b) VQEG FR-TV I, lab 1, high quality



(c) Experiment #2 (5-point ACR)

(d) Experiment #3 (11-point ACR)

Fig. 4. Standard deviation vs. MOS for different experiments (VQEG plots show normalized standard deviation vs. DMOS). The dotted lines indicate the average standard deviation over all clips.

3.2. Standard Deviation and MOS

The standard deviation is plotted as a function of MOS in Figure 4. The values for the VQEG FR-TV Phase I clips have been corrected for the fact that DSCQS data (differential MOS or DMOS) is a subtraction of two distributions, which increases the standard deviation by a factor of $\sqrt{2}$.

As we pointed out previously [10], the standard deviation is typically highest around the middle of the MOS range and decreases towards the ends of the scale. This behavior can be observed for most experiments, independently of the specific rating scale used. Data set #2 shown in Figure 4(c) is a particularly nice example, as the ratings span the entire quality range. This matches well with the simulations discussed in Section 2.2, where this effect was shown to result from clipping the ratings at the ends of the scales. The VQEG FR-TV Phase I experiment comprised mostly high-quality sequences, which is evident from the plots as well, yet the underlying shape remains similar.

3.3. Standard Deviation across Experiments

Comparing the average standard deviation between experiments is somewhat problematic, because they use different clips and subjects, and the VQEG data in particular suffers from a limited quality range. In order to reduce at least the influence of the boundary effects observed in the data towards the ends of the scale, we only consider clips with a MOS in the middle 25% of the range (e.g. from 2.5 to 3.5 on the 1-5 scale) in the following discussions.

Figure 5 compares the raw standard deviation across experiments. For each experiment, the average, minimum, and maximum standard deviation across clips is shown. The values for the VQEG FR-TV Phase I experiments have again been normalized by $1/\sqrt{2}$. Note that the averages are higher than the ones shown in Figure 4, because only clips with mid-range MOS are considered here.

The DSCQS experiments do have a relatively low (normalized) standard deviation of around 10–15, whereas the

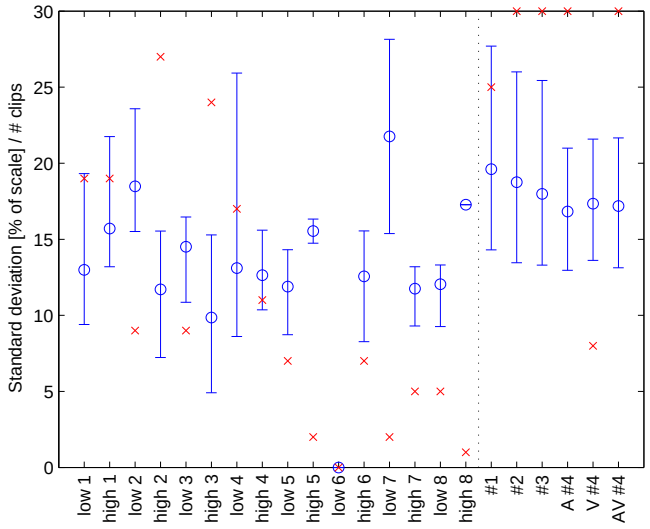


Fig. 5. Standard deviation for all experiments for clips with mid-range MOS. The circles indicate the average across clips per experiment, the error bars indicate the lowest and highest standard deviation in each experiment. The red crosses show the number of test clips with mid-range MOS. The data to the left of the dotted line are from VQEG FR-TV Phase I.

standard deviation is generally higher for the 11-point and 5-point data (around 15–20% of the respective scales). This difference is too large to be explained purely by the difference in rating scales established in the simulations (cf. Section 2.2). However, as mentioned before, a direct comparison is made difficult by the fact that these data are from different experiments with very different test material, different observers, conducted in different labs. Furthermore, as was already mentioned above, some of the VQEG FR-TV Phase I data is so biased towards low DSCQS scores that there are very few data points with mid-range MOS (the number of clips in that range is shown in the plot as well, as an indicator of the reliability).

3.4. Number of Subjects

Now let us look at the dependence of MOS and standard deviation on the number of subjects (again we only consider clips for which the MOS is in the middle 25% of the range).

First we randomly pick $M = 1 \ldots N$ out of the total number N of subjects in each experiment. We compute the MOS and the standard deviation and average them over all possible combinations of M out of N subjects. We further normalize the standard deviation by the largest value obtained for each clip (typically that is the standard deviation for $M = N$). These curves are plotted in Figure 6. Despite the large number of different experiments (22) and clips (364) used for this plot, the curves match quite well, all asymptotically approaching 1. This illustrates that the distribution of subjective ratings is very similar across a large variety of rating scales, test material, labs, etc. Note that there are different numbers of subjects in each experiment (see Figure 7 below), which likely contributes to the variability of these curves.

For comparison, Figure 6 also shows the simulated standard deviation from Section 2.1 again. It approaches the final value a bit faster than the data from the experiments, which may indicate that the distributions of subjective ratings are not perfectly Gaussian.

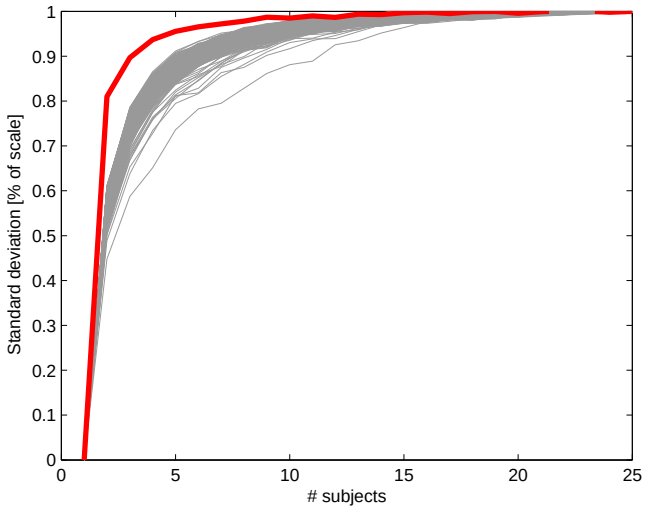


Fig. 6. Standard deviation (normalized to highest value) as a function of the number of subjects, across all experiments for clips with mid-range MOS (gray lines). The thick red line indicates the simulated standard deviation from Figure 1.

How many subjects are needed to capture the variability of the data with sufficiently high accuracy? In an attempt to quantify this number, we look at how fast the standard deviation and the MOS approach the "real" (final) values. For every clip with mid-range MOS in each experiment, we find the minimum number of subjects M where the standard deviation is reasonably close to its final value (the standard deviation for all N subjects). The threshold for "reasonably close" depends on requirements; here we choose 5% of the final value. Furthermore, we determine the minimum number of subjects M where the MOS variation (in terms of all possible combinations of M out of N subjects) falls below a certain threshold ($\sigma < 5\%$ of the scale). The results are shown in Figure 7,where average, smallest and largest M over the clips considered in each experiment are indicated by error bars. The minimum number of subjects according to this criterion is surprisingly constant across all experiments – it hovers around 10 subjects. For comparison, the total number of subjects that participated in each experiment are also shown in the plot.

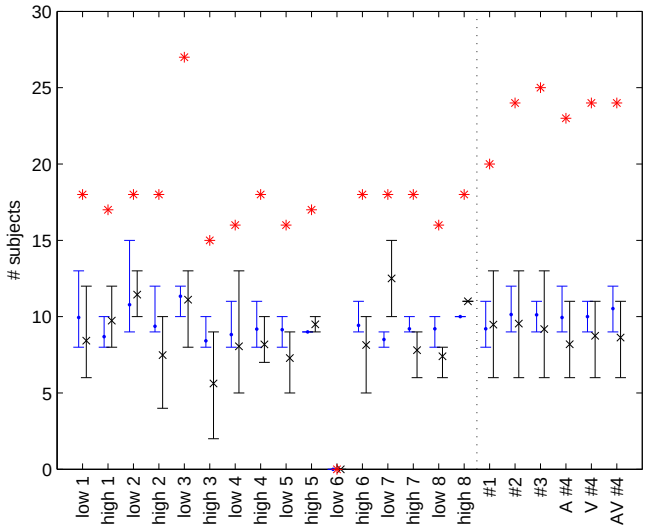


Fig. 7. Minimum number of subjects M where the standard deviation is within 5% of the final value (blue dots) and the MOS variation (σ) falls below 5% of the scale (black crosses). The error bars indicate the average/smallest/largest M across clips with mid-range MOS for each experiment. The stars show the total number of subjects. The data to the left of the dotted line are from VQEG FR-TV Phase I; the 'low 6' experiment has no clips with mid-range MOS.

4. CONCLUSIONS

Despite their relative simplicity, the simulations and analyses of the data described in this paper have revealed interesting aspects of subjective ratings and their distributions.

We quantified the influence of the rating scale on the standard deviation of the ratings. The impact of clipping of the ratings was also observed in the experimental data and matches well with our simulations. While discretization of the scale leads to an increase of the standard deviation in the simulations, practical proof of this effect remains inconclusive, as the differences between experiments are too large for reliable direct comparison of standard deviations. Data sets with subjective ratings using the different scales for the same test material would be needed for further analysis.

We also found evidence that number of subjects may not need to be as high as generally assumed; in fact, the minimum of 15 recommended by ITU appears to be a very reasonable suggestion.

5. REFERENCES

[1] Stefan Winkler, *Digital Video Quality – Vision Models and Metrics*, John Wiley & Sons, 2005.

[2] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Proc. SPIE Visual Communications and Image Processing*, Lugano, Switzerland, July 8–11, 2003, vol. 5150, pp. 573–582.

[3] Stefan Winkler and Ruth Campos, "Video quality evaluation for Internet streaming applications," in *Proc. SPIE Human Vision and Electronic Imaging*, Santa Clara, CA, January 21–24, 2003, vol. 5007, pp. 104–115.

[4] ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Geneva, Switzerland, 2002.

[5] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, Geneva, Switzerland, 1999.

[6] ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia applications," International Telecommunication Union, Geneva, Switzerland, 1998.

[7] VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of multimedia quality assessment," September 2008, Available at http://www.vqeg.org/.

[8] Stefan Winkler and Christof Faller, "Perceived audiovisual quality of low-bitrate multimedia content," *IEEE Transactions on Multimedia*, vol. 8, no. 5, pp. 973–980, 2006.

[9] VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," April 2000, Available at http://www.vqeg.org/.

[10] Stefan Winkler and Frédéric Dufaux, "Video quality evaluation for mobile applications," in *Proc. SPIE Visual Communications and Image Processing*, Lugano, Switzerland, July 8–11, 2003, vol. 5150, pp. 593–603.