Where do people look when using combined rating scales?

Robert Schleicher[1], Marie-Neige Garcia[1], Robert Walter[1], Sebastian Schlüter [2], Sebastian Möller[1], Alexander Raake[1]

[1]Deutsche Telekom Laboratories, TU Berlin [2] Technische Universität Berlin (TU Berlin)

Overview

This poster adresses the question to what extent people pay attention to the elements of rating scales, specifically the verbal labels and numbers on a scale used in video quality tests.

Background

Video quality tests

To compare compression algorithms and to analyze transmission errors regarding their impact on the perceived quality of video streams, standardized test setups have been developed by organizations like the International Telecommunication Union (ITU), which also specify the assessment procedures including the assessment scales to be used.

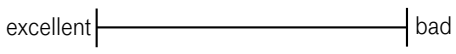
Rating scales

Common types of rating scales are (Svensson, 2000):

- **Verbal Descriptor Scale** (top): verbal categories

[ ] excellent
[ ] good
[ ] fair
[ ] poor
[ ] bad

- **Visual Analogue Scale** (middle): continuous line



- **Graphic Rating Scale** (bottom): combination of both

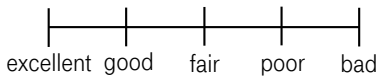


Figure 1: Different rating scales.

While categories prevent the test subject expressing intermediate ratings and may be perceptual non-linear (Zielinski, 2007), visual analogue scales might lead to confusion regarding what point on the line corresponds to what quality level.

The Graphic Rating Scale aims at overcoming these drawbacks by combining both, a continuous line and verbal descriptors (Svensson, 2000). Still, it is not clear whether this line is perceived as continuous or whether the verbal labels in combination with the vertical dividers introduce a quasi-discretization of the line perception, which might lead to quantization effects in ratings (Zielinski et al, 2007).

The scale offered by ITU Recommendation ITU-T P.910 shows additionally numbers next to the dividers to emphasize the continuous and linear character of the scale.

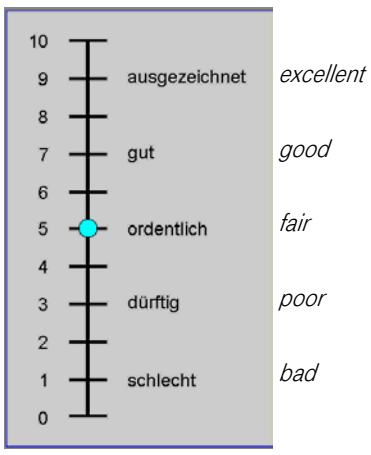


Figure 2: The German version of the ITU-P.910 scale used in the test *(English labels)*

Methods

The data reported here were collected as part of an experiment on video quality perception. The general setup is depicted in Figure 3. More information on the experimental details as well as the results regarding audiovisual quality can be found in Garcia et al. (2011).

Figure 3: Experimental setting of the video quality test.

Hardware: native HD (1920x1080) 42' LCD screen , 108 cm away from subject. Eyelink II head-mounted eyetracker (SR Research), binocular recording at 250 Hz. Drift correction after each trial.

Stimulus material: 12 standard color video excerpts depicting various TV scenes (duration 8.8-17 s), all scaled to 720x576 pixels (= 34x27 cm on screen), remaining screen greyed out (see picture), in 3 degradation levels. Rating scale was scaled to 350x576 pixels (= 16.5x27 cm on screen)

Task: subjects watched the 3*12 the clips +3 trainings clips and rated the quality subsequently to each presentation on the ITU-scale presented in the center of the screen using a PC mouse. Order of presentation was randomized.

Subjects: n=27 students (13 female, mean age 26.2+/-5.4 years) participated.

Results

Ratings

- complete scale used by subjects, but clear anchor effects of line divider: ratings that correspond to a divider on the ruler were given more frequently.
- no explicit anchor or quantization effect for rating steps that are labelled with a verbal descriptor

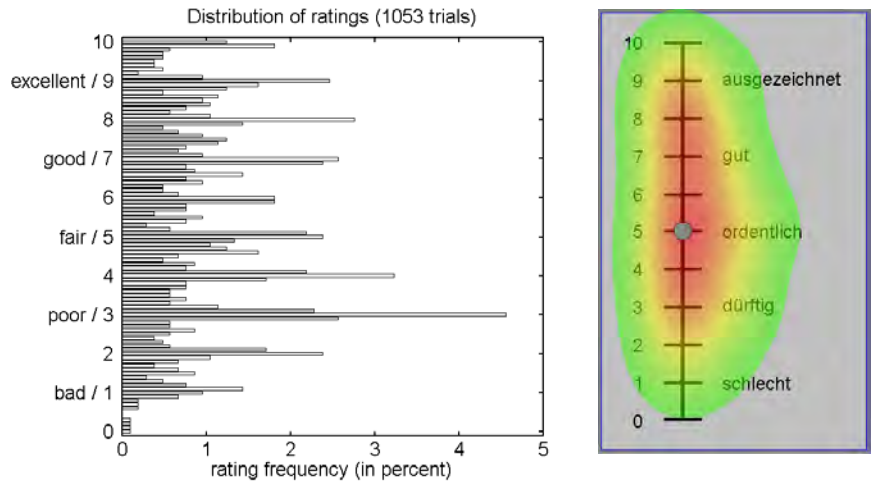


Figure 4: results of rating trials. **Left:** distribution of given ratings across the scale. **Right:** fixation heat map across all subjects displaying relative duration of fixations >150ms.

Fixations

- overall **Fixation heat map** (figure 4 right): clear preference for the ruler line where the actual rating had to be made with the mouse, but no obvious difference between fixations on the numbers and the verbal labels.

Detailed analysis of fixation behavior:

- **Total dwell time**: subjects spent 72.7% on the line where the ratings had to be made, 23.4% on the verbal labels, but only 2.4% on the numbers (see figure 5 left).
- **First fixation in a rating trial**: target of first fixation in a rating trial was in 59% on the verbal labels, in 37% on the line, and only in 4% on the numbers.

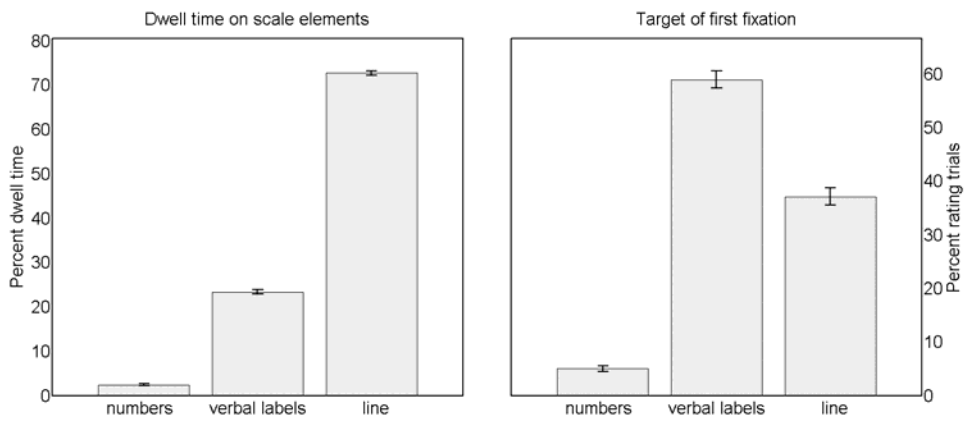


Figure 5: fixations during rating. **Left:** Percentage of total dwell time during a trial on the three scale elements. **Right:** Percentage of target of first fixation for each scale element.

An repeated-measure analysis of variance (ANOVA) confirmed the statistical significance of the depicted differences for both measures ($F_{1,29}$=267, p<0.01 for %dwell time; $F_{1,22}$=190, p<0.01 for 1st fixation; dfs Greenhouse-Geisser corrected). Post-hoc comparisons were significant for all conditions (p<0.01, Sidak-adjusted).

Summary and Conclusion

- While subjects used the complete scale for rating, there was a quantisation or anchor effect of the line dividers.
- The anchor effect is not more pronounced for steps with a verbal label compared to steps with divider/number only.
- Subjects attended the numbers much less than the verbal labels or the line itself in terms of total dwell time as well as percentage of first fixation
- Thus the meaning of the scale appears to be derived mainly from the verbal labels.
- Based on the little proportion the numbers were fixated, it is unclear whether they convey additional meaning/information.

Acknowledgements: We would like to thank Dipl-Inf. Martin Haller, TU Berlin for his help in setting up the experiment.

References

Garcia, M. N., Schleicher, R., & Raake, A. (2011). Impairment-Factor-Based Audiovisual Quality Model for IPTV: Influence of Video Resolution, Degradation Type, and Content Type. *EURASIP Journal on Image and Video Processing, 2011*(Article ID 629284), 1-14.

ITU-T Rec.P.910 (2008). Subjective video quality assessment methods for multimedia applications. Geneva, CH: International Telecommunication Union

Svensson, E. (2000). Comparison of the Quality of Assessments Using Continuous and Discrete Ordinal Rating Scales. *Biometrical Journal, 42*(4), 417-434.

Zielinski, S., Brooks, P., & Rumsey, F. (2007). *On the Use of Graphic Scales in Modern Listening Tests.* Paper presented at the 123rd Convention Audio Engineering Society, New York City, NY.

Contact

robert.schleicher@tu-berlin.de