University of Brasília

TU Delft

VIDEO QUALITY EXPERTS GROUP MEETING – GLASGOW, SEPT 17 2015

VIDEO QUALITY RULER: A NEW EXPERIMENTAL METHODOLOGY FOR ASSESSING VIDEO QUALITY

PEDRO GARCIA, JUDITH REDI, MYLENE FARIAS AND ALEXANDRE FIENO

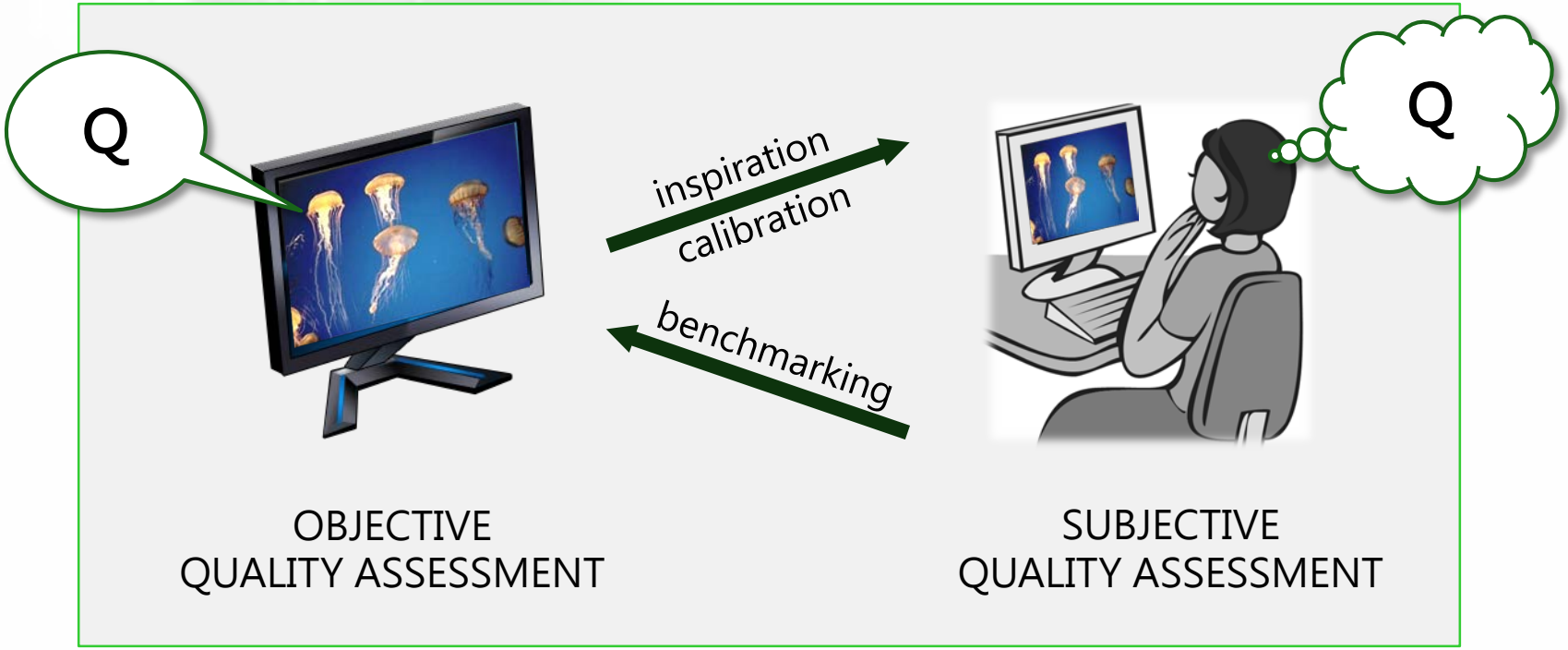
IN THIS TALK…

- Methodologies for subjective quality assessment
  - (un)reliability of subjective assessments and its effect on modelling quality perception/judgments
  - (un)reliable methodologies for video quality assessment

- A new proposal: the video quality ruler
  - Video adaptation of Keelan's image quality ruler (2000) – which comes with several challenges
  - Evaluation: (1) does it work at all, and (2) does it work better than e.g. single stimulus assessments?

- Why coming to VQEG with this
  - Interest in further exploring the potential of this method?
  - Possible application in e.g. JEG hybrid subjective evaluations?

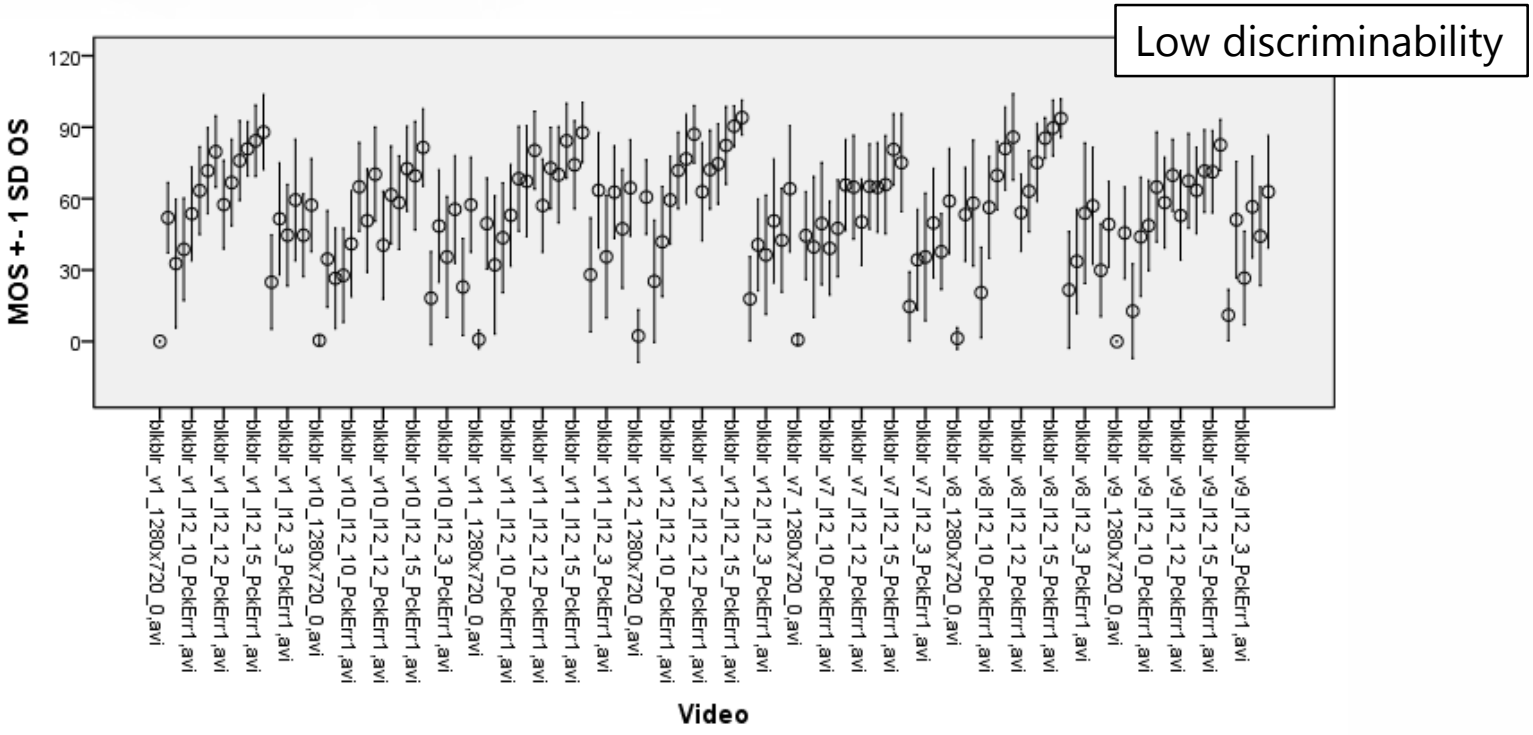
VIDEO QUALITY ASSESSMENT



If subjective assessments are unreliable,
so will be objective metrics calibrated on them.

UNRELIABLE SUBJECTIVE QUALITY ASSESSMENTS

- Multiply distorted videos (blur + compression + packet loss)
- Single stimulus evaluation
- Avg stdev around MOS: 18,79 (on a 100-point scale)

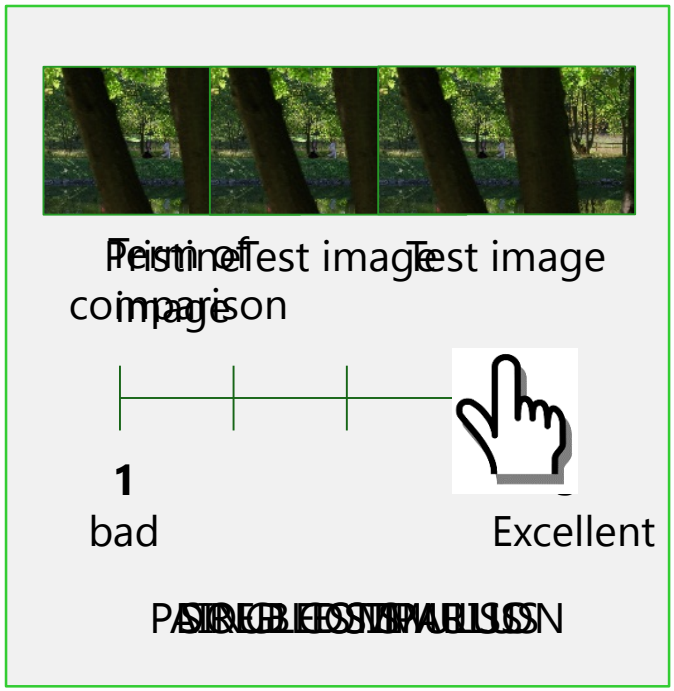


WHAT MAKES SUBJECTIVE MEASUREMENTS UNRELIABLE?

Participants sloppiness, Errors in (entering) judgment, Fuzziness (in the definition) of the attribute being assessed…

… and **Methodology**

- Single Vs Double Stimulus
  - A term of comparison helps in expressing judgments (Keelan 2000)
- Direct vs implicit scaling
  - Direct scaling methodologies (ACR, DSIS) require the association of quality with a number/semantic label
    - No actual benefit in term of std reduction
  - Implicit scaling methods (PC) require only visual comparison (and a choice)
    - Lower cognitive load (Engeldrum 2001)
    - …but time consuming!



THE QUALITY RULER METHOD

1 overall quality Just Noticeable Difference (JND)

Test image

Score:
2.5 JNDs

Quality judgment is reduced to a set of visual comparisons

For images, it significantly reduces SOS (Redi 2010)

Keelan, B., "Handbook of image quality: characterization and prediction," Marcel Dekker, Inc., New York, 2002.

A VIDEO QUALITY RULER?

Challenges

- A SQS for video quality assessment?
- Dispersion of attention between SQS and test stimuli

Images or videos?

Shared or second screen?

Simultaneous displaying?

THE VIDEO QUALITY RULER SOS

Adapted from Image Quality ruler implement

Can people score video quality having Image quality as a reference?

- 16 images
- Spans a range from 0 to 15 multivariate JNDs
  - Calibrated through a large paired comparison experiment
- Images vary in blur (as per Keelan 2000)
- Has been shown to be suitable to evaluated both blurry and differently distorted images

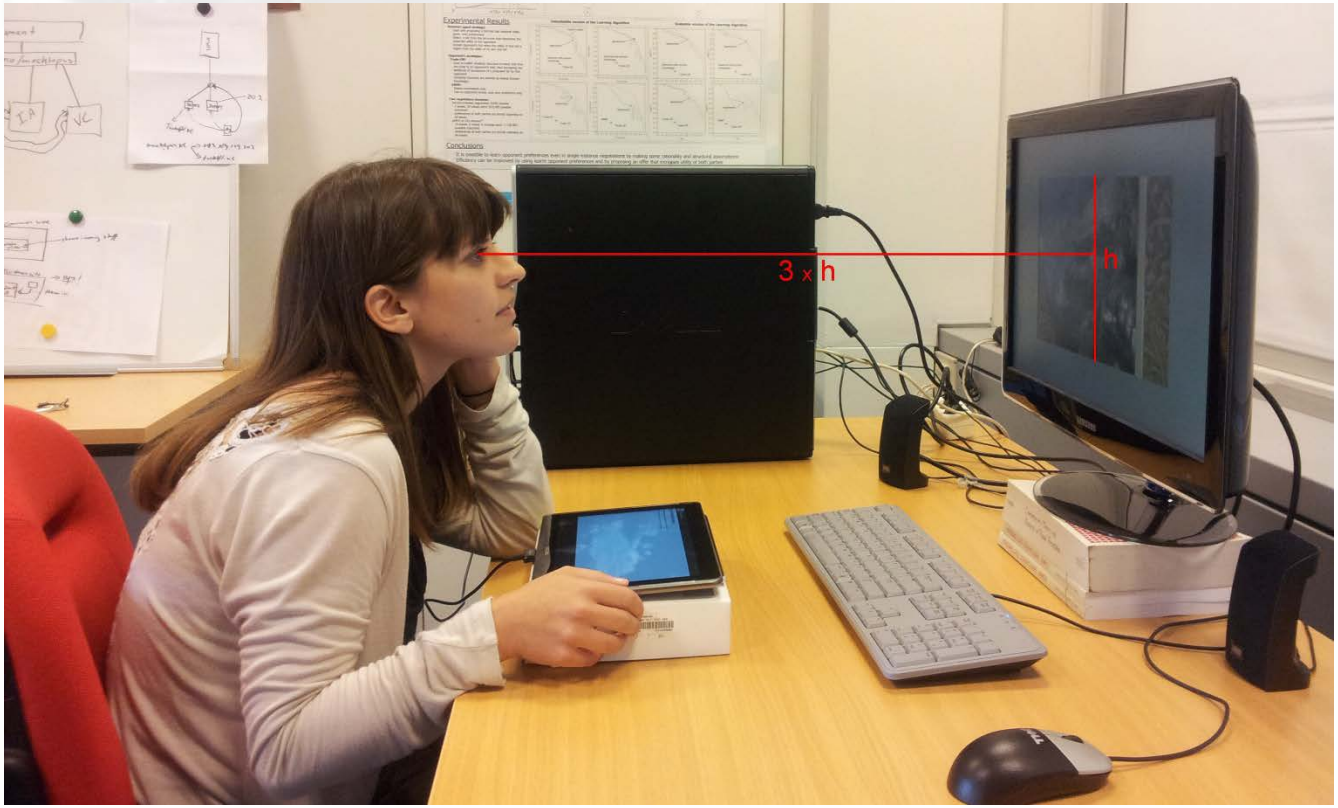
SQS PRESENTATION

- Adoption of a second screen, to avoid re-scaling video and/or ruler images
- Use of horizontally adjacent screen may have caused issues in artefact visibility related to viewing angle and distraction in the periphery

  →use of vertically adjacent screens

  →TABLET!

THE VIDEO QUALITY RULER



PROTOCOL

1. Video tutorial

2. Ruler exploration

3. Guided practicing

4. Scoring task (2 sessions)

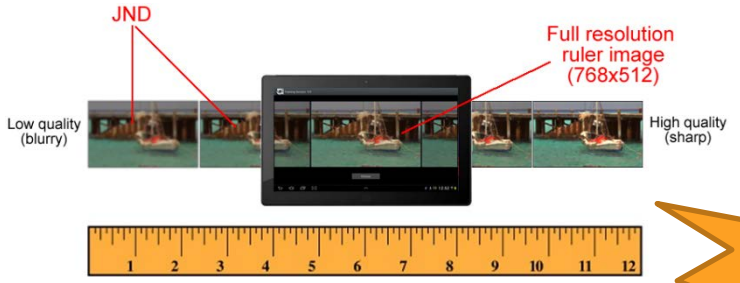
1. First video view

2. Test video looping

3. judgment

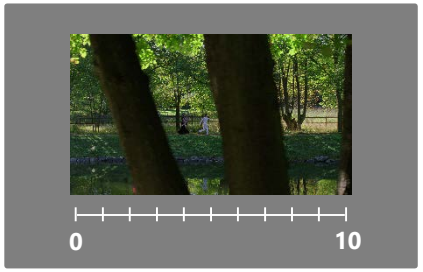
EXPERIMENT

VIDEO QUALITY RULER (VQR)



VS

SINGLE STIMULUS (SS)

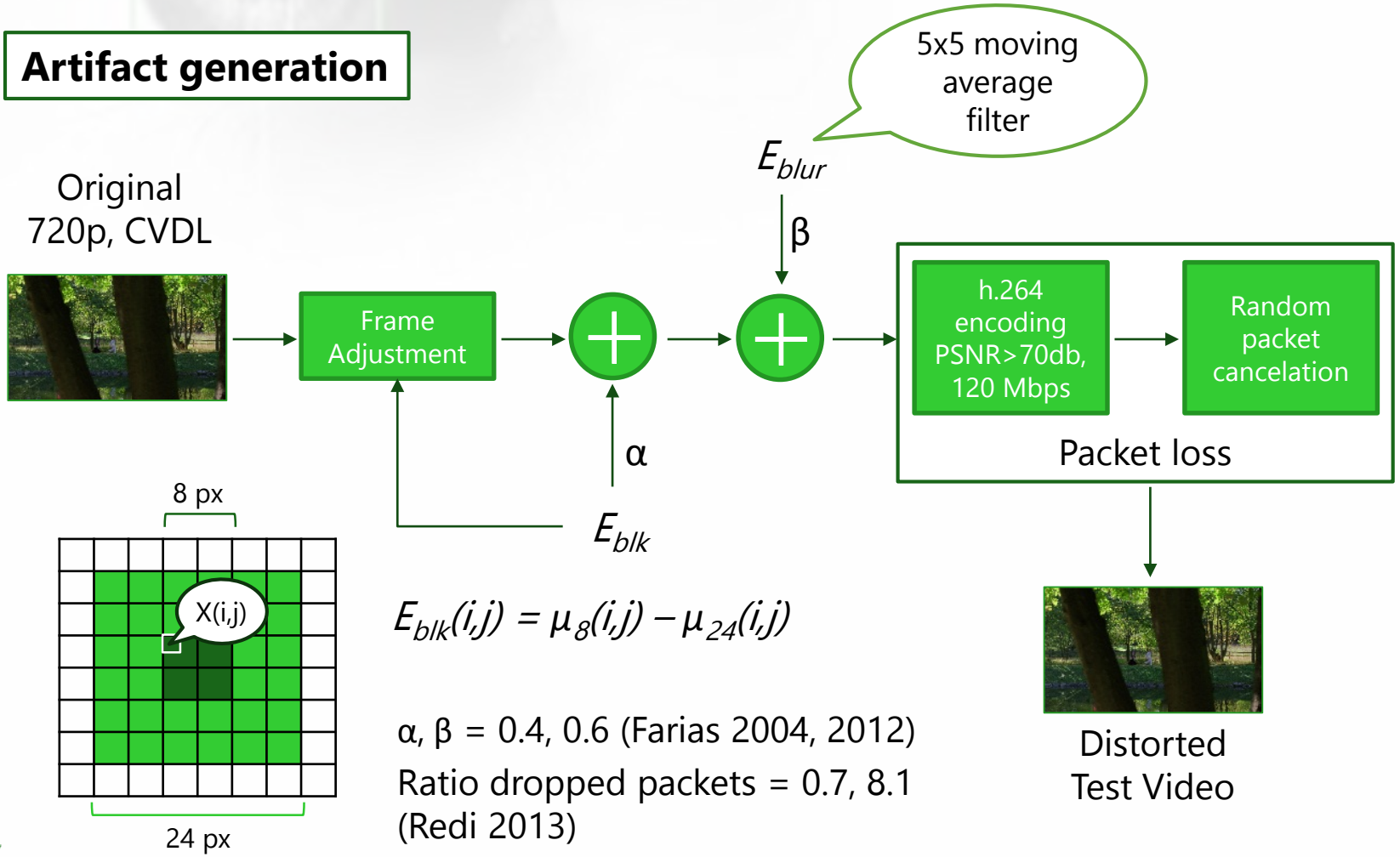


Research questions:

1. Can we measure video quality with the Video Quality Ruler at all?
2. If so, are these measures reliable at least as muthch as the ones obtained with the Single Stimulus method?

STIMULI: MULTIPLY DISTORTED VIDEOS

Artifact generation



α, β = 0.4, 0.6 (Farias 2004, 2012)

Ratio dropped packets = 0.7, 8.1 (Redi 2013)

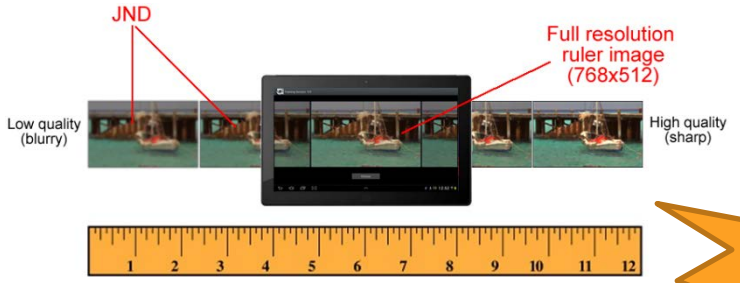
STIMULI

7 Original, 1280x720, 50fps - 49 videosin total

Combination	Packet loss	Blockiness (α)	Blurriness (β)
1	0.0	0.0	0.0
2	0.0	0.6	0.0
3	8.1	0.0	0.0
4	0.7	0.0	0.4
5	8.1	0.0	0.6
6	8.1	0.4	0.6
7	8.1	0.6	0.6

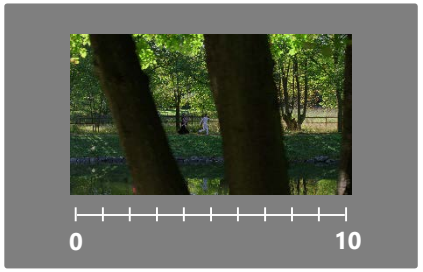
EXPERIMENT

VIDEO QUALITY RULER



VS

SINGLE STIMULUS



- Same environmental conditions
- Same display and HW
- 24 participants for SS, 17 for QR

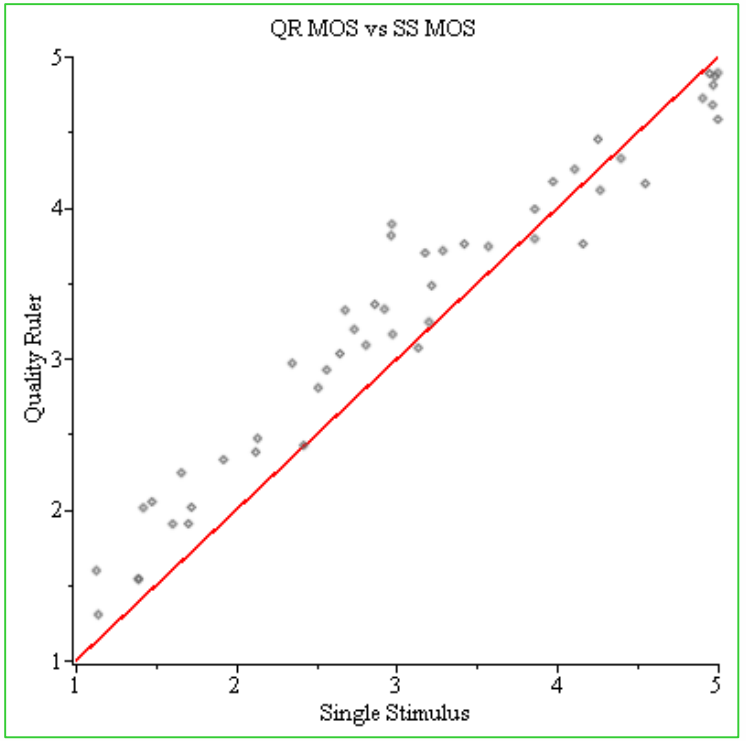
CAN WE MEASURE VIDEO QUALITY WITH THE VQR?

- Or, do SS and VQR measure the same thing?

parallel form reliability analysis

Scores linearly re-scaled in [1-5]

- Linear Correlation: 0.9663
- Spearman's Correlation: 0.9643
- Kendall's Correlation: 0.8511
- RMSE: 0.3871
- Outlier Ratio ([MOS-2σ; MOS+2σ]): 0



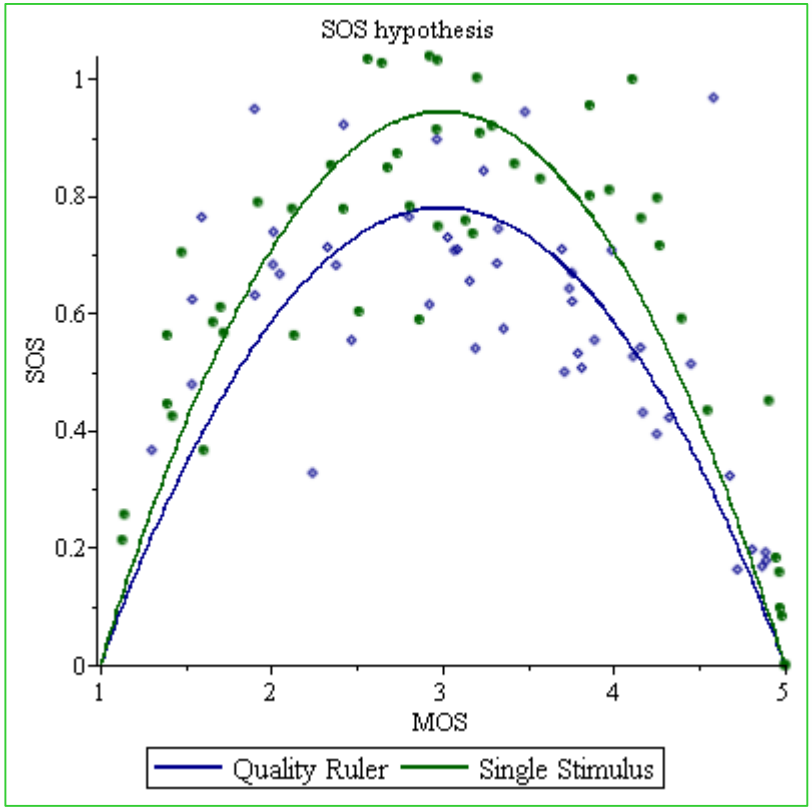
ARE VQR MEASUREMENTS RELIABLE?

Inter-subject variability analysis:

- SOS hypothesis (Hossfeld et al., 2011): measures the width of the standard deviation of opinion scores (SOS) wrt the magnitude of MOS.

$$SOS_s(j)^2 = \alpha(-MOS_s(j)^2+6MOS_s(j) - 5)$$

- The bigger alpha, the higher the inter-subject variability



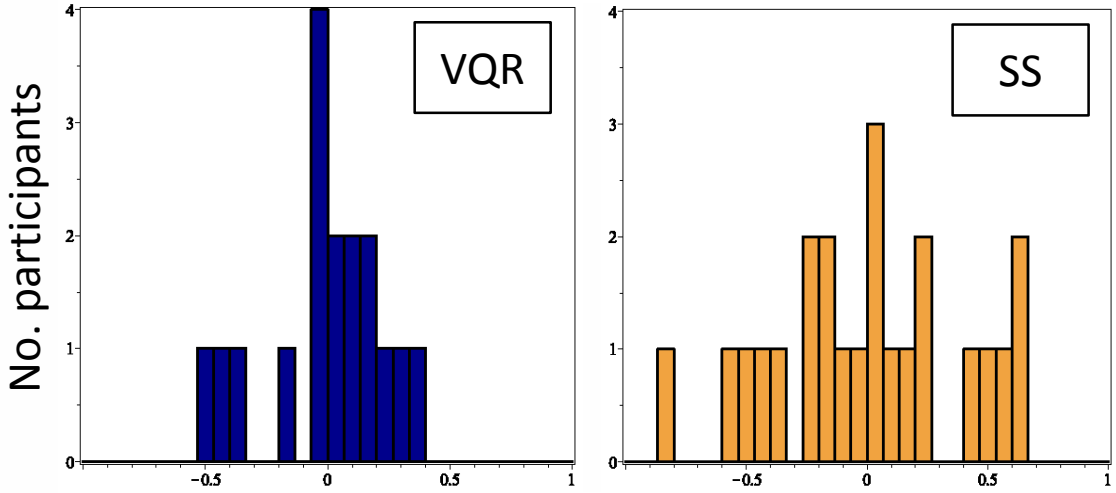
ARE VQR MEASUREMENTS RELIABLE?

Subject Bias analysis

- Models rating behaviour (Janowski and Pinson 2014). The rating expressed by user *n* for image *i* on scale *s* is expressed as:

$$OS_s(i,n) = \mathrm{MOS}_s(i) + \Delta_{n,s} + \varepsilon_{i,n,s}$$

$\Delta_{n,s}$ is the subject bias term, indicates subjectivity in the scoring scale usage. The higher, the more different is the scoring behaviour of user n from the others.



THOUGHTS

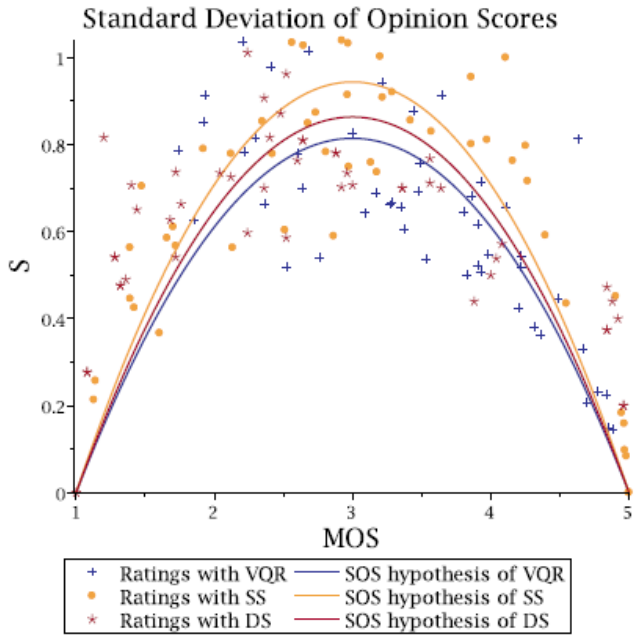
- VQR seems to be able to provide video quality assessment measures that are highly similar to those that would be obtained with a SS methodology
- And with higher reliability

- Nevertheless, VQR is more time consuming than SS, so there is a trade off between efficiency and reliability
  - in this sense, it should be compared with other Double Stimulus methodologies (e.g., DSIS)

BONUS

- We recently conducted a new experiment, to compare VQR with a double stimulus methodology (DSIS)
- Same environmental conditions as before, DSIS protocol, 5 point annoyance scale, 24 subjects
- Still work in progress, but here's a sneak peak:

VQR seems to deliver more reliable MOS than DSIS



(a) Raw subjective data

FUTURE WORK/OPEN QUESTIONS

- Comparison with other methodologies
  - SAMVIQ, Paired Comparison
- Investigation of the SQS properties
  - Are multivariate image quality JNDs equivalent to video quality JNDs? – probably not
  - Would a video-based calibration of the SQS yield more reliable results?
  - is the reliability of the tool depending on the tablet and main monitor display types – probably so, since JNDs depend on that
- Investigation of the tool properties
  - Repeatability of MOS and independence on context effects (it was proven for images, does it hold for video?)
    - Also across multiple artifacts?
- Sensitivity at high and low qualities

POSSIBLE JOINT WORK WITHIN VQEG

- If repeatability of MOS and independence of context effects is proven, then VQR would be a great asset: providing reliable MOS, in terms of multivariate JNDs, and repeatable across experiments
  - In principle, subjective quality evaluations could be run across different labs and dataset without the need of REALIGNMENT sets
  - This may be especially appealing for **JEG,** which is collecting a wide variety of videos presenting more than one artefact/distortion

- The (image) quality ruler may also be employed in **VIME**
  - "New approaches to subjective study design for the purpose of addressing emerging quality assessment needs (as market and consumer demands evolve)"
  - What about calibrating a SQS for consumer content evaluation?

- … your ideas?

j.a.redi@tudelft.nl

THANK YOU.

VIDEO PROPERTIES