# Study of Rating Scales for Subjective Video Quality Assessment Using Single-Stimulus Presentation

Quan Huynh-Thu (Psytechnics Ltd)
Alexander Raake, Marie-Neige Garcia (DT Labs)
Philip Corriveau (Intel Corp)
Filippo Speranza (CRC Canada)

# Motivation of Study

- Provide data to help VQEG make informed decisions:
  - Current discussions on adequate subjective test methodology and rating scale to use in on-going and future VQEG validation projects

- Advance knowledge in the field of subjective quality assessment

# Scope and Methodology of Study

- Scope:
  - Single-stimulus presentation
  - Retrospective quality rating
- Methodology:
  - Unique set of processed videos
  - Unique test lab
  - Unique stimulus pattern presentation
  - Different rating scales:
    - 5-point discrete scale
    - 9-point discrete scale
    - 5-point continuous scale
    - 11-point continuous scale

# Experimental Design

- Video format: HD1080p
- Video length: 12 seconds
- No audio
- Test design:
  - 8 SRCs
  - 16 HRCs (incl. hidden reference condition):
    - Coding
    - Coding + transmission errors (slicing and frame freezing)
  - 8 x 16 = 128 PVSs
- Codec: H.264
- Bit rates: 2 – 16 Mbps
- PLR: 0.25 – 4 %
- 24 viewers per experiment (after post-hoc screening as per VQEG HDTV test plan)
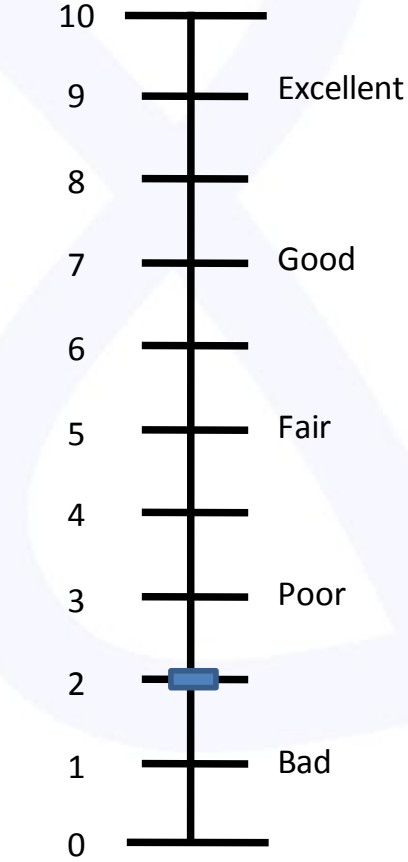
# Experimental Set-up

- Psytechnics subjective testing facilities
- Test environment conforming to ITU Rec.
- 24" LCD display, 1080p native resolution
- Viewing distance: 3H
- One viewer at a time in front of display
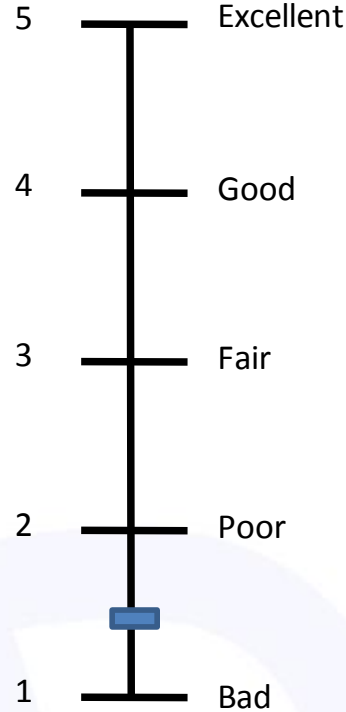- Different randomized presentation order for each subject

# Rating Scales



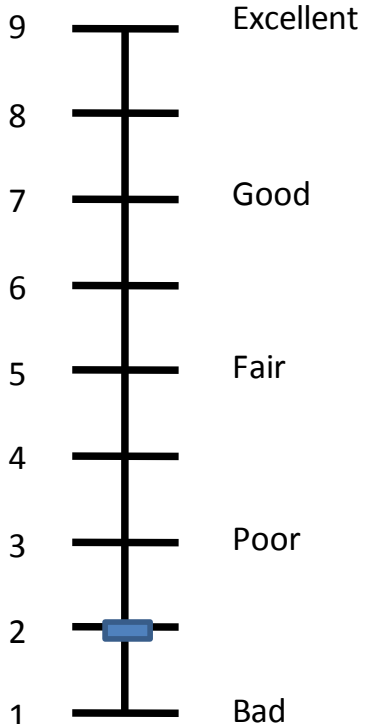Slider locks only to ticks to simulate discrete scale

| Exp1 | Exp2 | Exp3 | Exp4 |
|------|------|------|------|
| 5-point discrete | 11-point continuous | 5-point continuous | 9-point discrete |

# Test Instructions for Continuous Scales

- ## For 5-pt and 11-pt continuous scales:
  - "You can place the slider at any point on the scale"
  - "You can click on the slider and drag it to the desired position or, click on the scale at the desired position (the slider will jump directly to this position)"

- ## For 11-pt scale:
  - Did **not** instruct viewers to avoid "0" or "10"
  - Verbal descriptions of "0" and "10" presented in the written instructions but on the on-screen scale during the test
  - Viewers instructed that practice trials will present examples of "best" and "worst" qualities

The number 10 denotes the best possible quality

10

9 — Excellent

8

7 — Good

6

5 — Fair

4

3 — Poor

2

1 — Bad

0

The number 0 denotes the worst possible quality
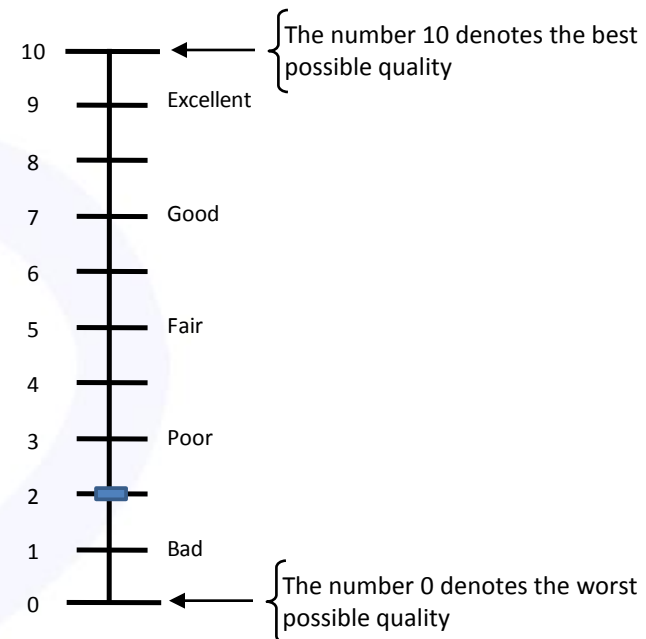
# Distributions of Ratings



5-pt discrete scale
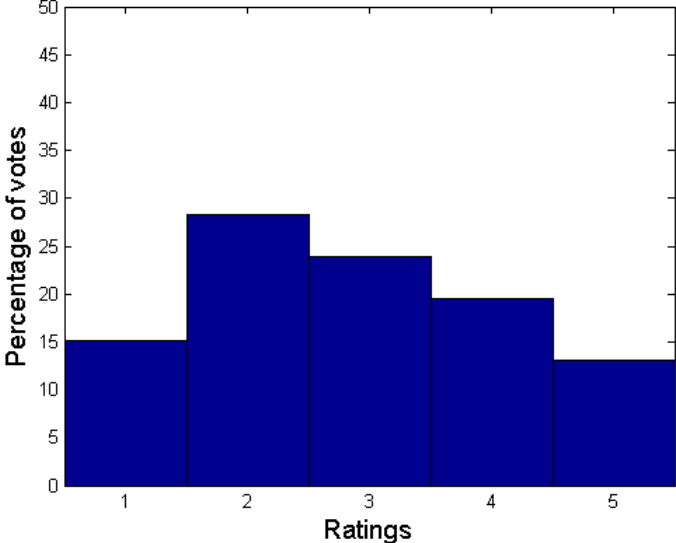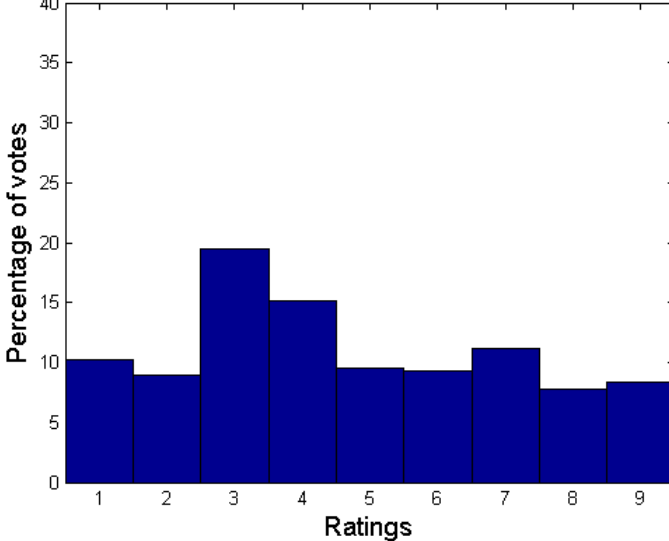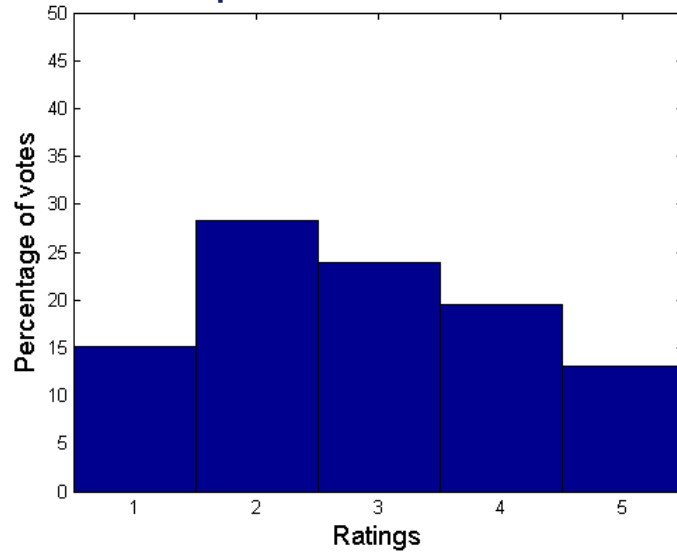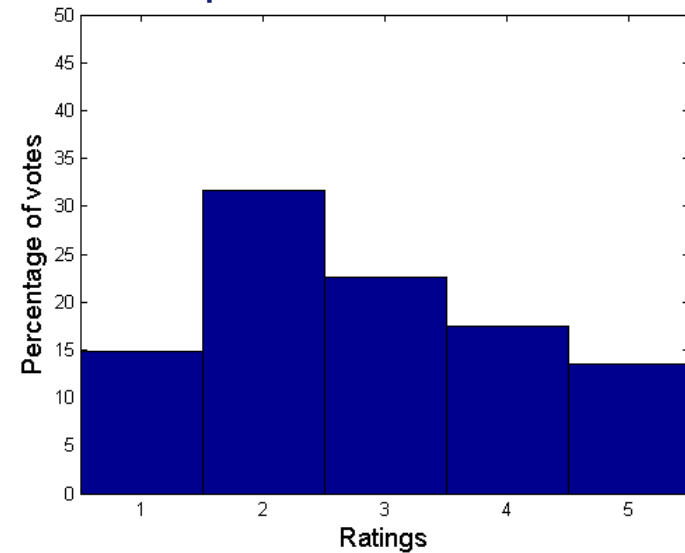
11-pt continuous scale
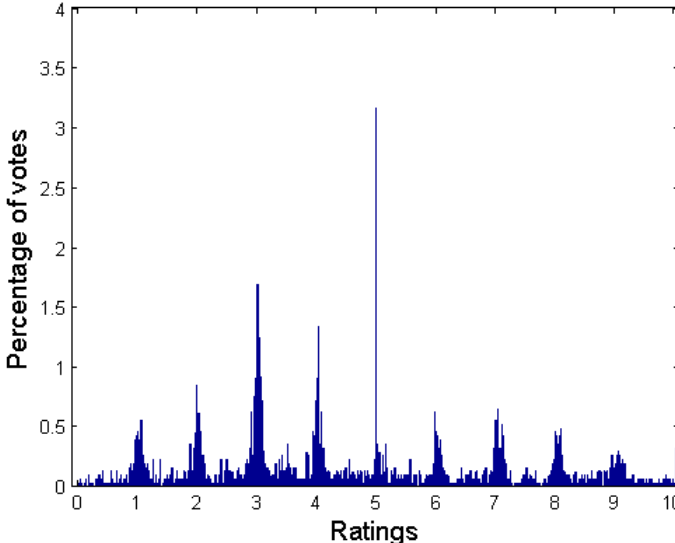
5-pt continuous scale
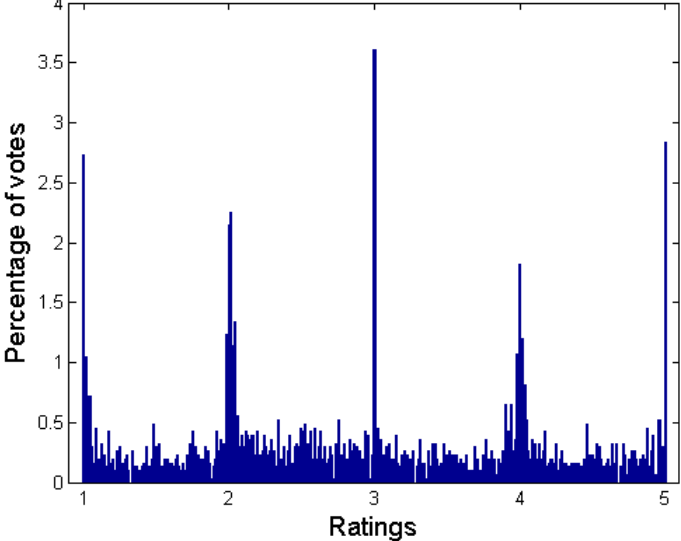
9-pt discrete scale

# Distributions of Ratings Using 5 Bins
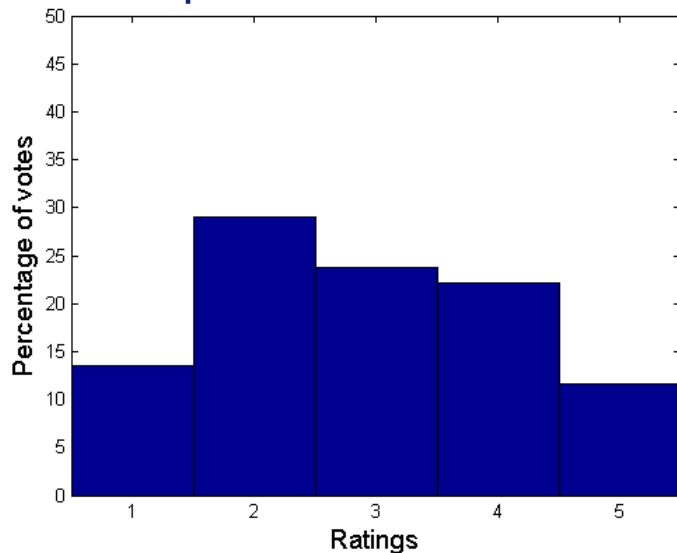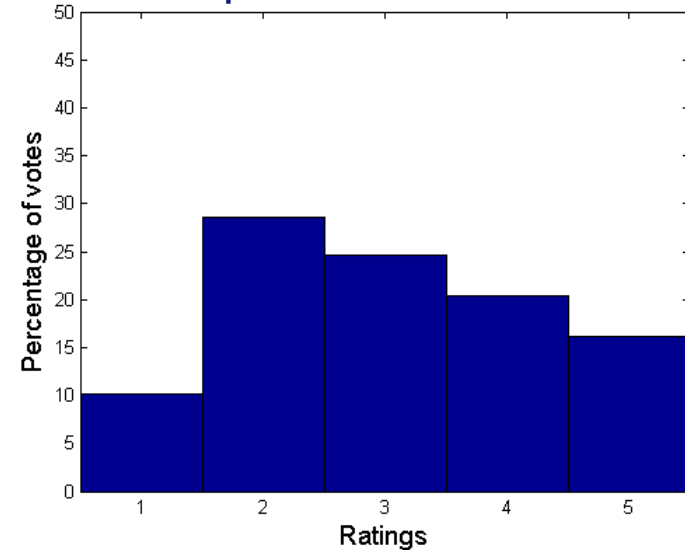
# Relationship Between Scales

- Quantization effects with continuous scales
  - Most viewers tend to align their ratings with marks and labels on continuous scales
- Re-scaling needed for comparison between scales
  - Re-scaling of all votes on same scale
  - Re-scaling using a linear transformation aligning labels between scales:
    - Exp1: stays between [1,5]
    - Exp2: $score_{map} = (score_{orig} / 2) + 0.5$
    - Exp3: stays between [1,5]
    - Exp4: $score_{map} = (score_{orig} / 2) + 0.5$
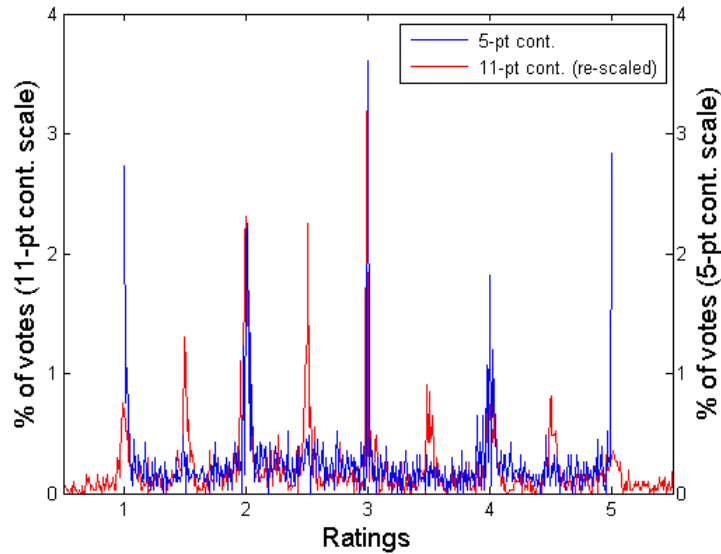
# Comparison of Distribution of Ratings

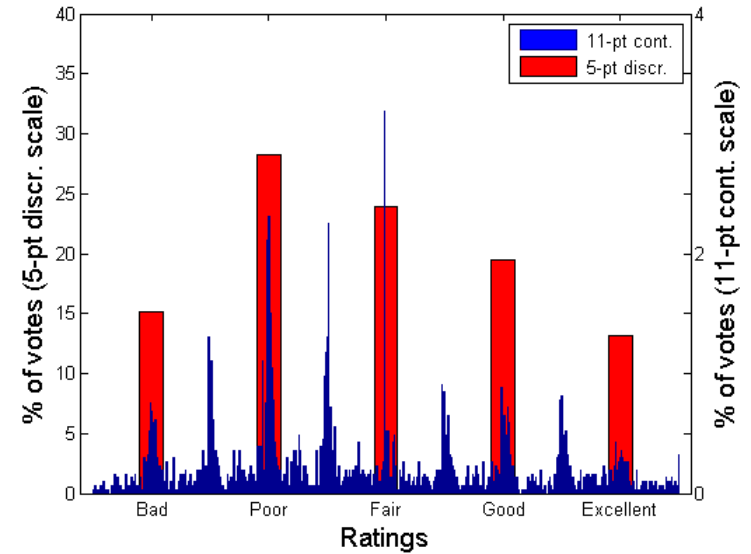## 5-pt **discrete vs. continuous** scale
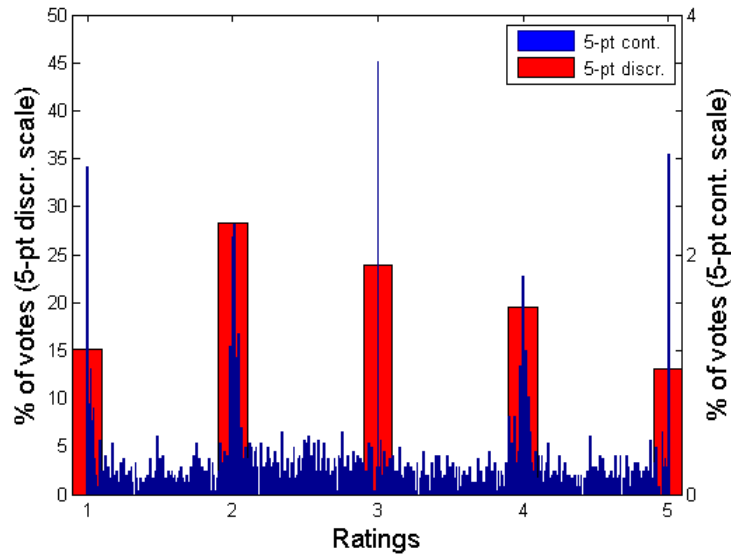


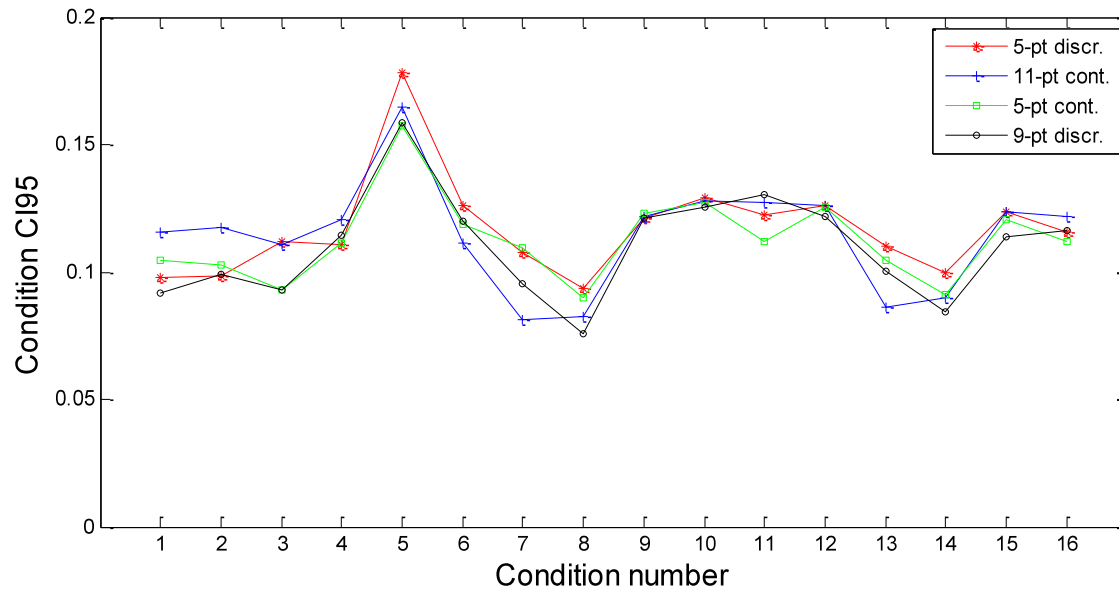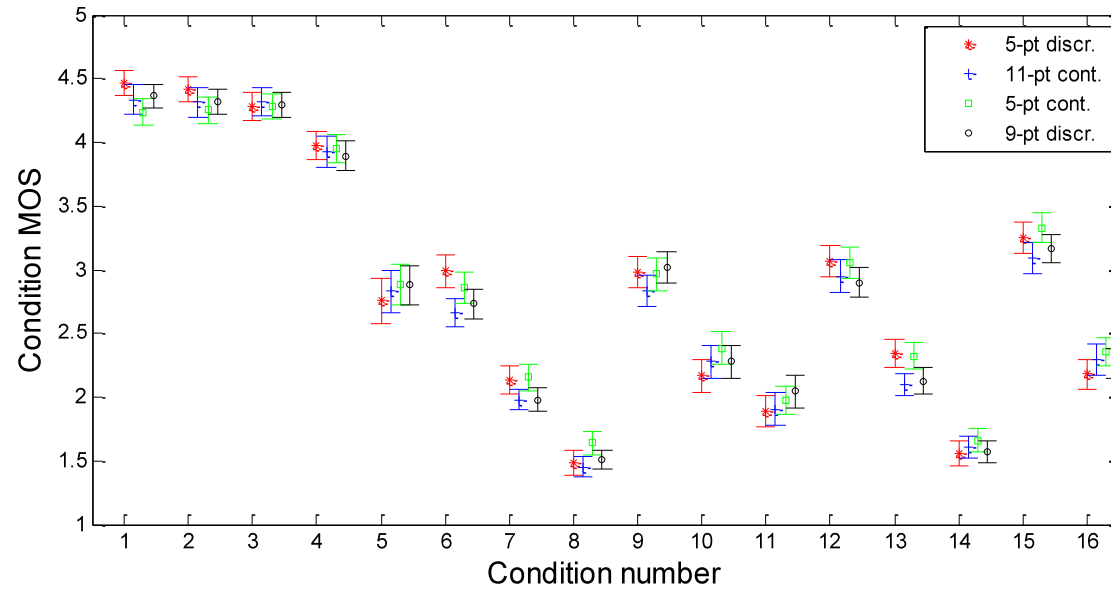## 5-pt vs. 9-pt discrete scale
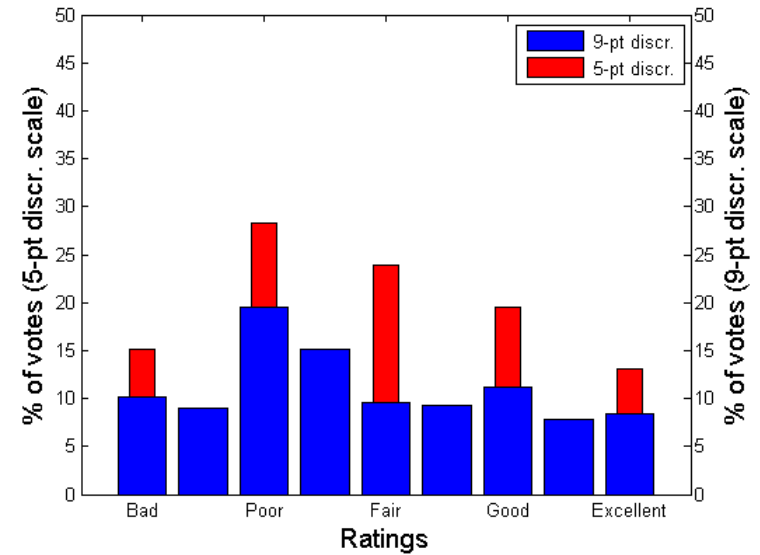


## 5-pt vs. 11-pt continuous scale



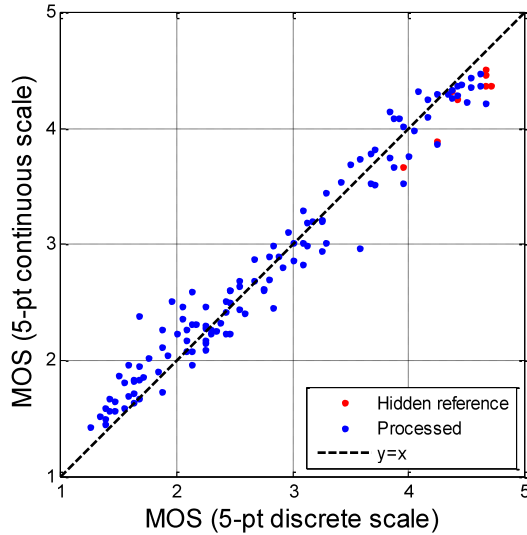## 5-pt discrete vs. 11-pt continuous scale
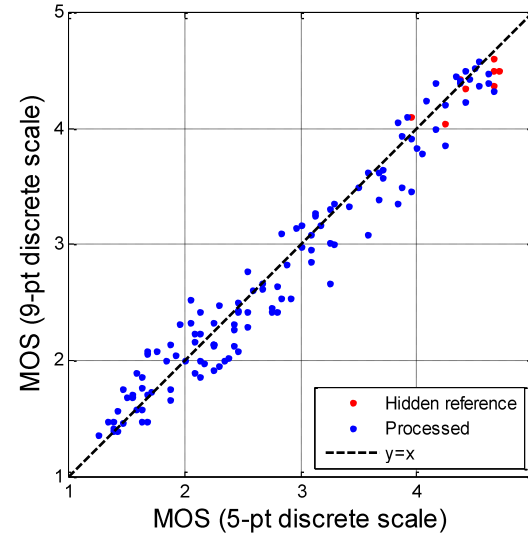
# Comparison of Condition MOS and CI

# Scatter plots of MOS
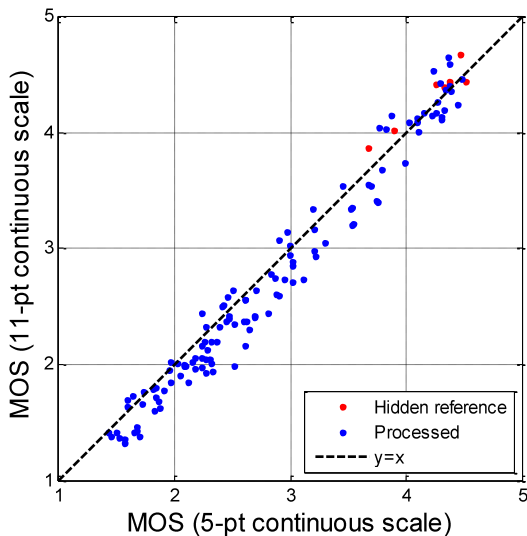
## 5-pt **discrete vs. continuous** scale
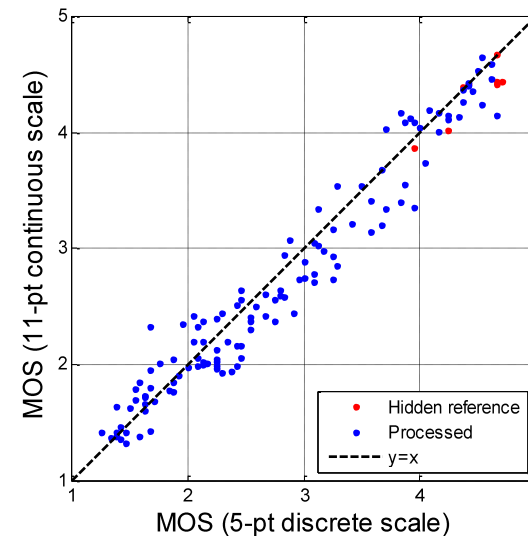


R=0.98

## 5-pt vs. 9-pt discrete scale



R=0.98

## 5-pt vs. 11-pt continuous scale
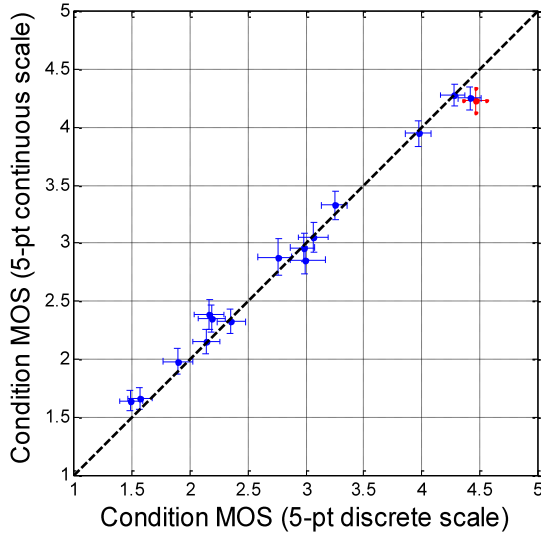


R=0.99

## 5-pt discrete vs. 11-pt continuous scale



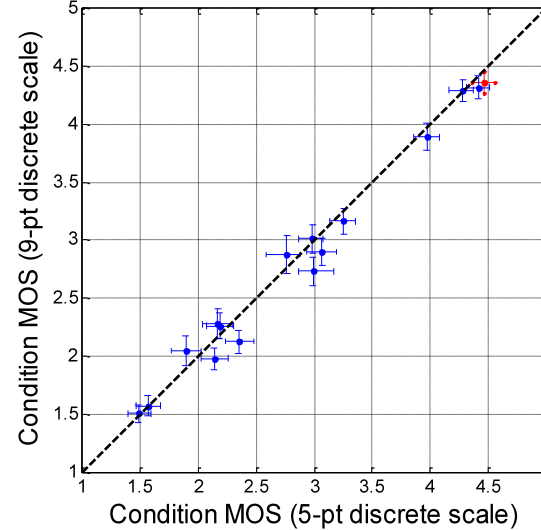R=0.98

# Scatter plots of Condition MOS
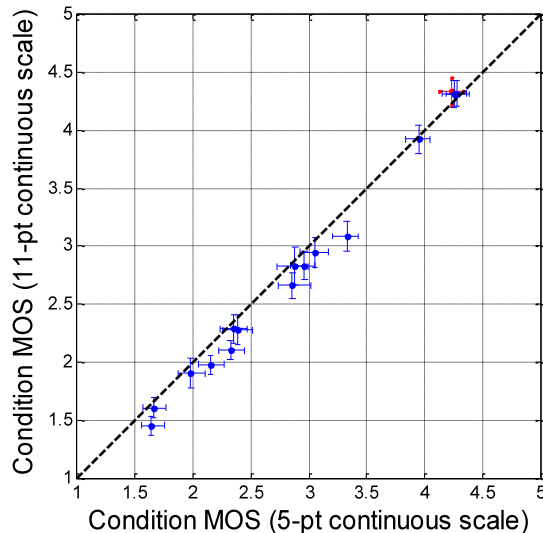
## 5-pt **discrete vs. continuous** scale



R=0.99

## **5-pt vs. 9-pt** discrete scale



R=0.99

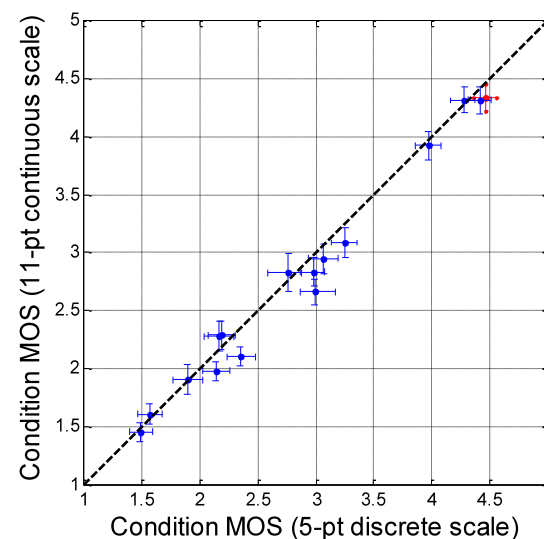## **5-pt vs. 11-pt** continuous scale



R=0.99

## **5-pt discrete vs. 11-pt continuous** scale



R=0.99

# Student T-tests and ANOVA

| | Exp1 (5-pt discr.) | Exp2 (11-pt cont.) | Exp3 (5-pt cont.) | Exp4 (9-pt discr.) |
|---|---|---|---|---|
| Mean MOS | 2.8727 | 2.8047 | 2.8936 | 2.8337 |
| Mean CI | 0.2952 | 0.2973 | 0.2932 | 0.2839 |

- No statistical differences between mean quality
- No statistical differences between mean CI

# Preliminary conclusions

- Data show that viewers tend to align their ratings with the positions of the labels on the scales
- There is no significant difference between the results obtained with the different scales

# Future Work

- Comparison using different re-scaling approaches

- More detailed analysis