VQEG: Interpreting VIQET scores, when can users see a difference?

Understanding user experience differences for photo quality

P. J. Corriveau (Intel), L. Janowski (AGH),
S. Katsigiannis (UWS), N. Ramzan (UWS), M. A. Saad (Intel),
J. J. Scovell (Intel), G. V. Wallendael (UGENT)

Outline

Problem Statement

A question that arises from using a tool like VIQET that produces MOS scores is whether the MOS differences produced are noticeable to a consumer.

We expect that MOS 5 is noticeably better than MOS 1 but it is not clear if MOS 3.8 is noticeably better than MOS 3.6.

Outline

Laboratories

Lab	Country	Institution
Intel	USA	©Intel Corporation
UGhent	Belgium	Ghent University
UWS	Scotland (UK)	University of the West of Scotland
AGH	Poland	AGH University of Science and Technology

Setup

- Two identical displays at each lab
- A keyboard and mouse was used to make selections.
- The distance: three times the height of the display.
- Displays: *Intel* and *UGhent*: Samsung 28" (3840 × 2160), *UWS*: Sony Bravia 55" 4K TVs (3840 × 2160), *AGH*: Samsung 40" (1920 × 1080)
- A total of 91 participants completed the study: 36 participants (19 male, 17 female) at *Intel*, 31 (27 male, 4 female) at *UGhent*, and 24 (19 male, 5 female) at *UWS*.
- Subject's age vary between 20 and 59 years old. Participants from *UGhent* and *UWS* were mostly PhD students and researchers, while participants from *Intel* were employed in various sectors and were recruited through a third party.

Subjective Test Procedure

- Phase one:
  - Two images, the same scene but were taken with different cameras **not the exact same frame**.
  - Image selection by hitting the left or right arrow followed by “Enter” on the keyboard.
  - After selection automatically changed to the next image pair
  - A random order, also randomized the images between the two displays
  - The same 220 image pairs were presented to all 91 users.
- Phase two:
  - A single full screen image on the left display
  - Five point ratings scale on the right display.
  - Total number of images 51
  - In order to maximize the number of images to be rated, not all images were the same across the three labs.

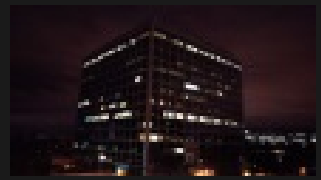
Example

Source Images

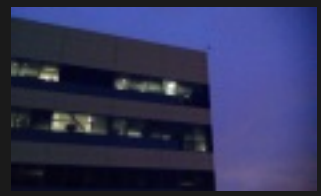
(a) AutumnMtn

(b) Beach Toys

(c) Bridge



(d) Building

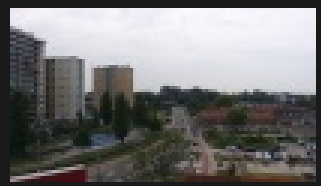


(e) Build. Corner

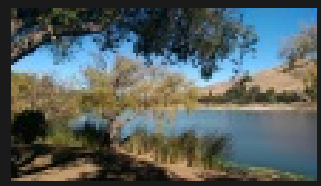
(f) Evac. Plan

(g) Flowers

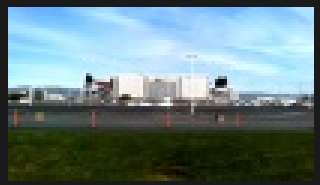
(h) Fruit



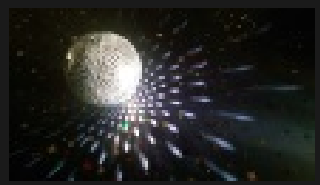
(i) Ghent



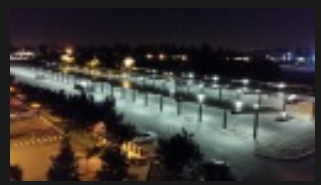
(j) Green Tree



(k) Levi



(l) Mirror Ball



(m) Parking

(n) Pipes

(o) Tree Lake

Figure: Sample images from each scene.

Images Selection

- Phase one:
  - Fourteen image pairs per scene
  - MOS from crowd sourcing study conducted by Intel
  - MOS difference (≤ 0.257) and (≥ 0.949)
  - Pairs selected to cover wide range of MOS scores and similar MOS differences for different qualities
- Phase two:
  - It was run to validate crowd sourcing study data
  - 3 images per scene, 51 in total
  - 15 of those images were the same across the lab, with 6 repetitions
  - Covering the whole range of MOS scores

Outline

Analysis Goal

Estimate function:

$$p = f(\delta_{MOS})$$

where p probability of selecting a higher quality image, δ_{MOS} the MOS difference.

Correlation

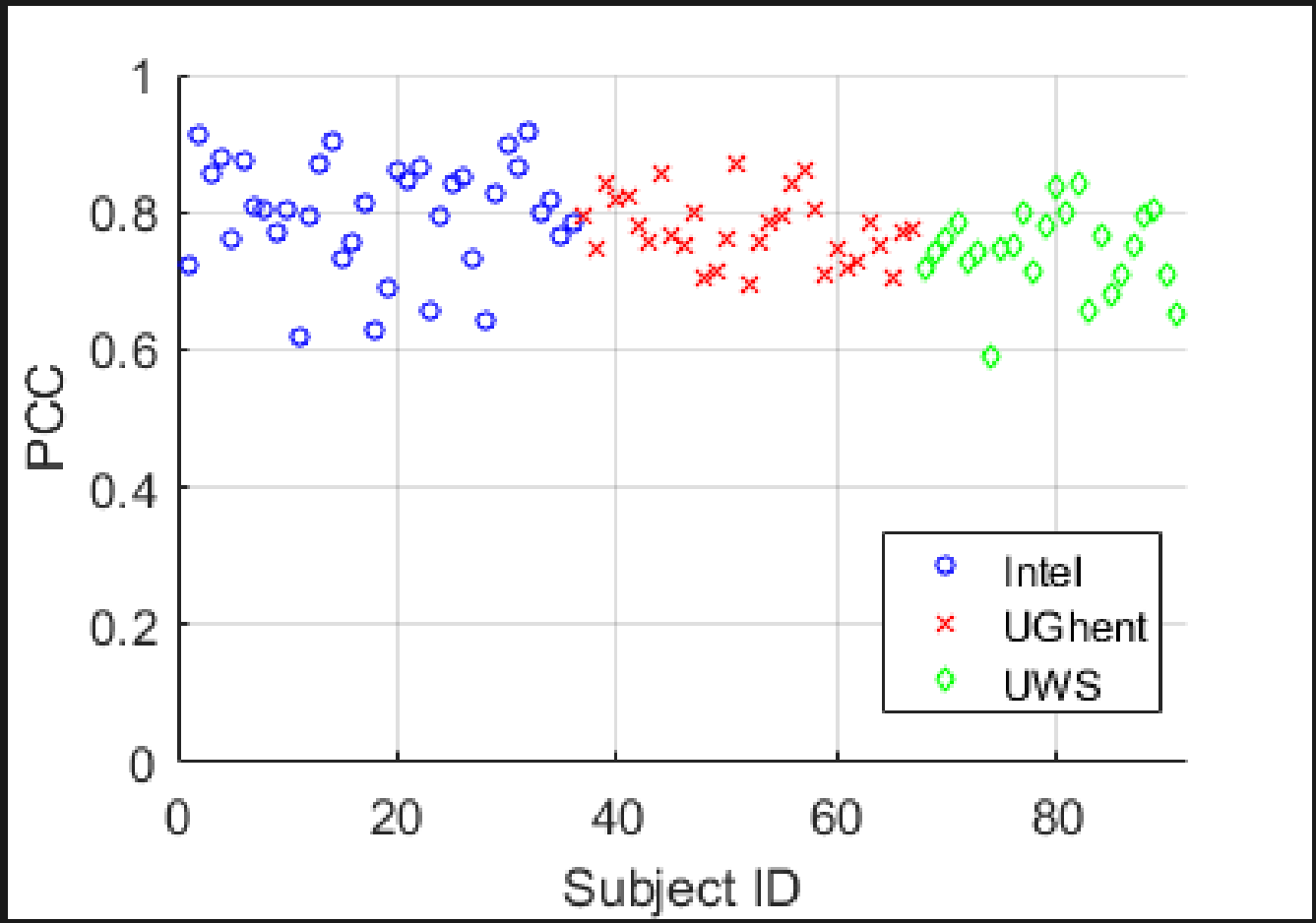


Figure: Pearson's Correlation Coefficient between the MOS ratings of each subject and the crowd sourced ratings.

Scatter Plot

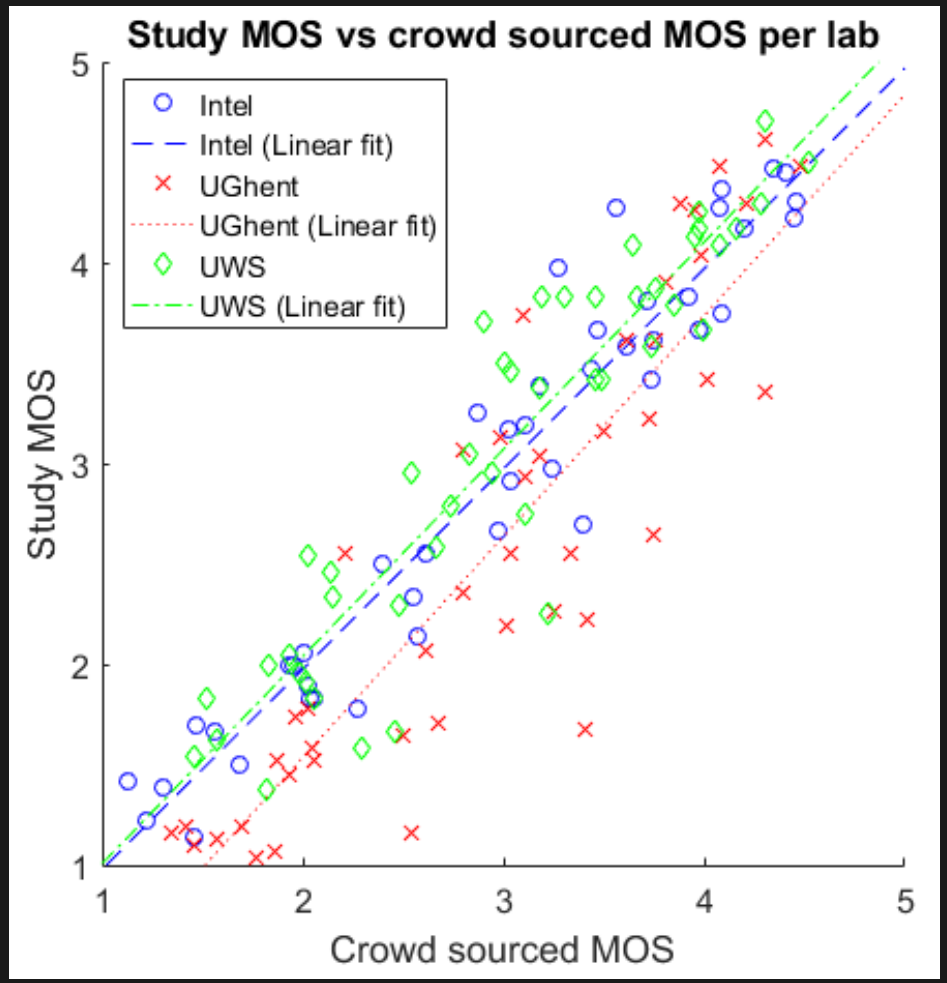


Figure: This study MOS compared to crowd sourced MOS ratings.

Linear Fitting

Table: Relationship between the MOS received for each image through this study (y) and through crowd sourcing (x)

Lab	PCC	ANOVA p	Linear fit	R^2
Intel	0.9632	0.9037	$y = 0.9975x - 0.0190$	0.9277
UGhent	0.8938	0.0917	$y = 1.1008x - 0.6665$	0.7989
UWS	0.9276	0.7165	$y = 1.0316x - 0.0243$	0.8605
All	0.9111	0.3680	$y = 1.0451x - 0.2429$	0.4379

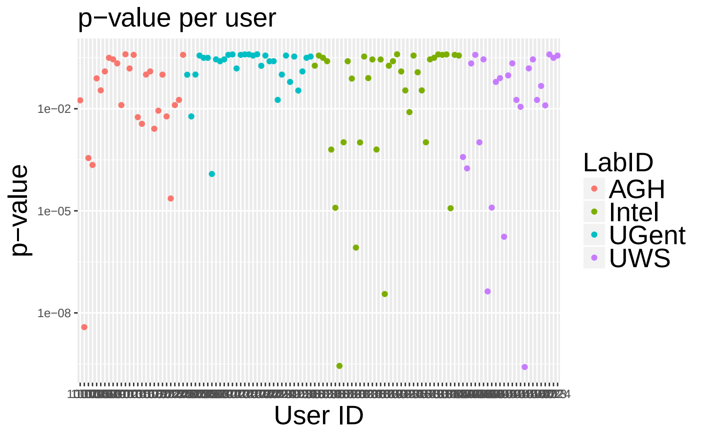
Screen Selection

Table: Correlation between the results from when the left or the right screen was selected at each lab for the same pairs of images

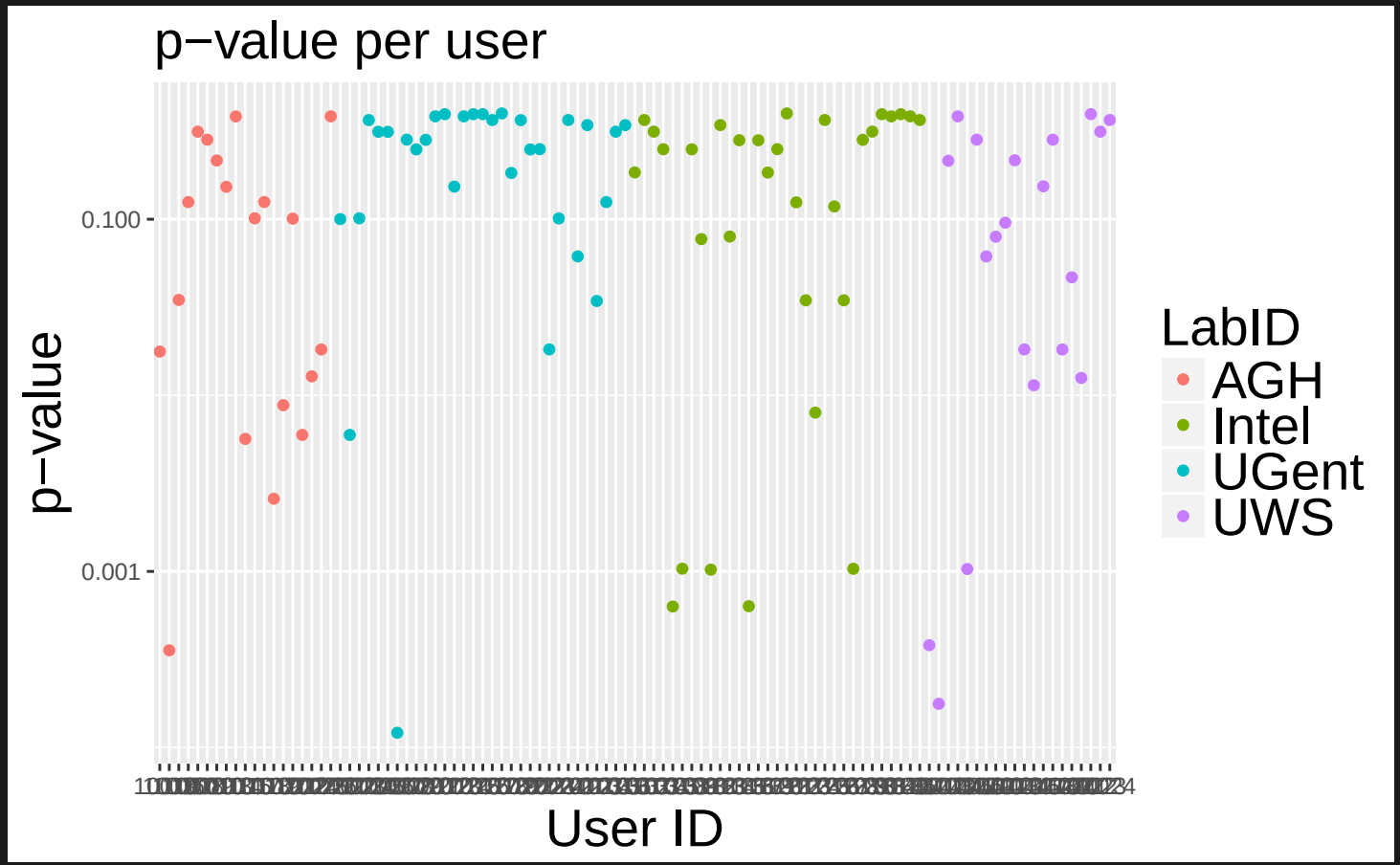
Lab	PCC	Correlation	ANOVA p
Intel	0.686	Strong	0.0109*
UGhent	0.821	Very strong	0.3349
UWS	0.624	Strong	0.0045*

* indicates a statistically significant difference

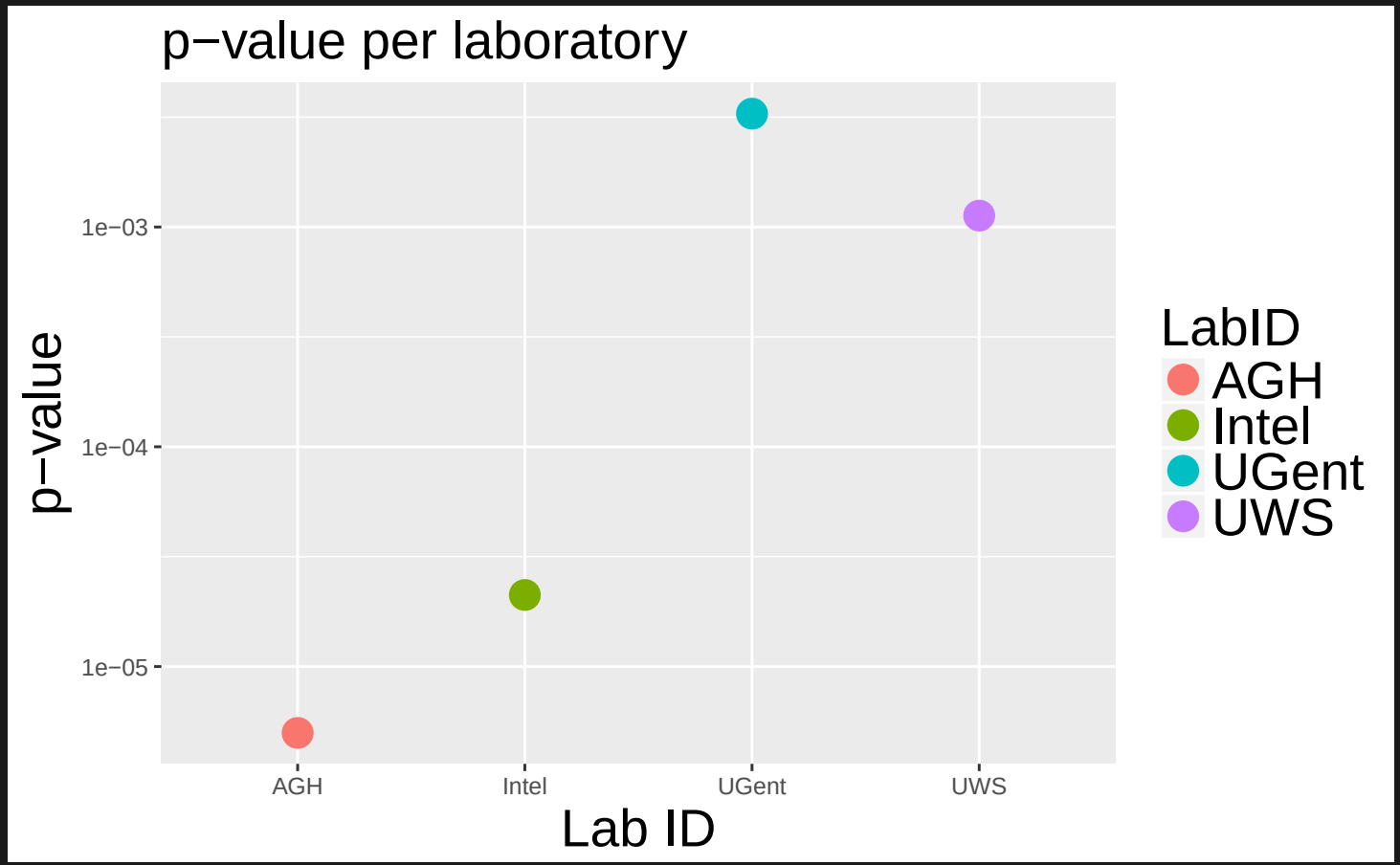
χ^2 Analysis



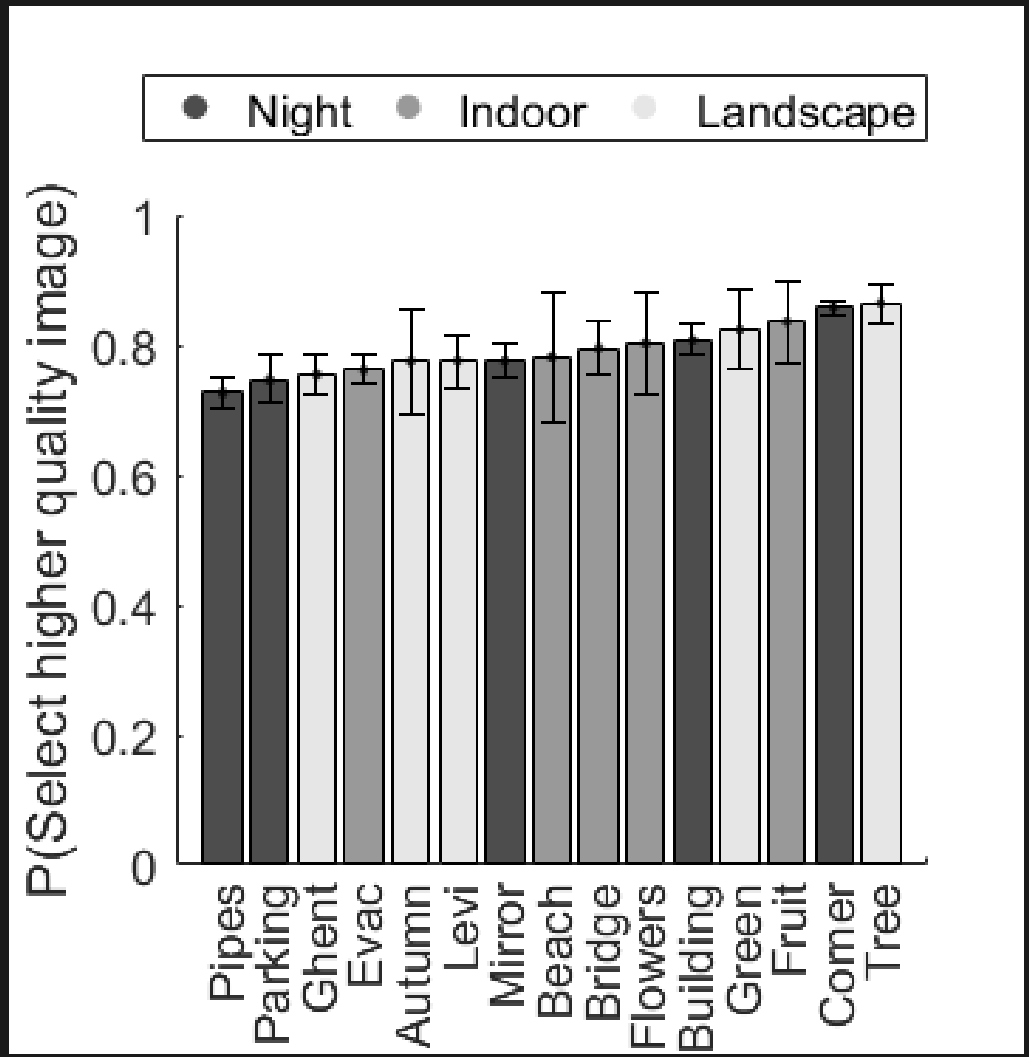
χ^2 Analysis, Clean Data



χ^2 Analysis, Labs for Clean Data



Probability by Image Type



The Model

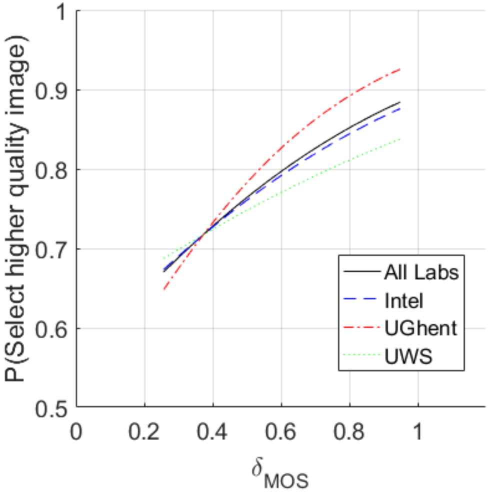
$$p = f\left(\delta_{MOS}\right) = \frac{1}{1 + e^{-\left(\beta_0 + \beta_1 \cdot \delta_{MOS}\right)}} \quad (1)$$

The Model

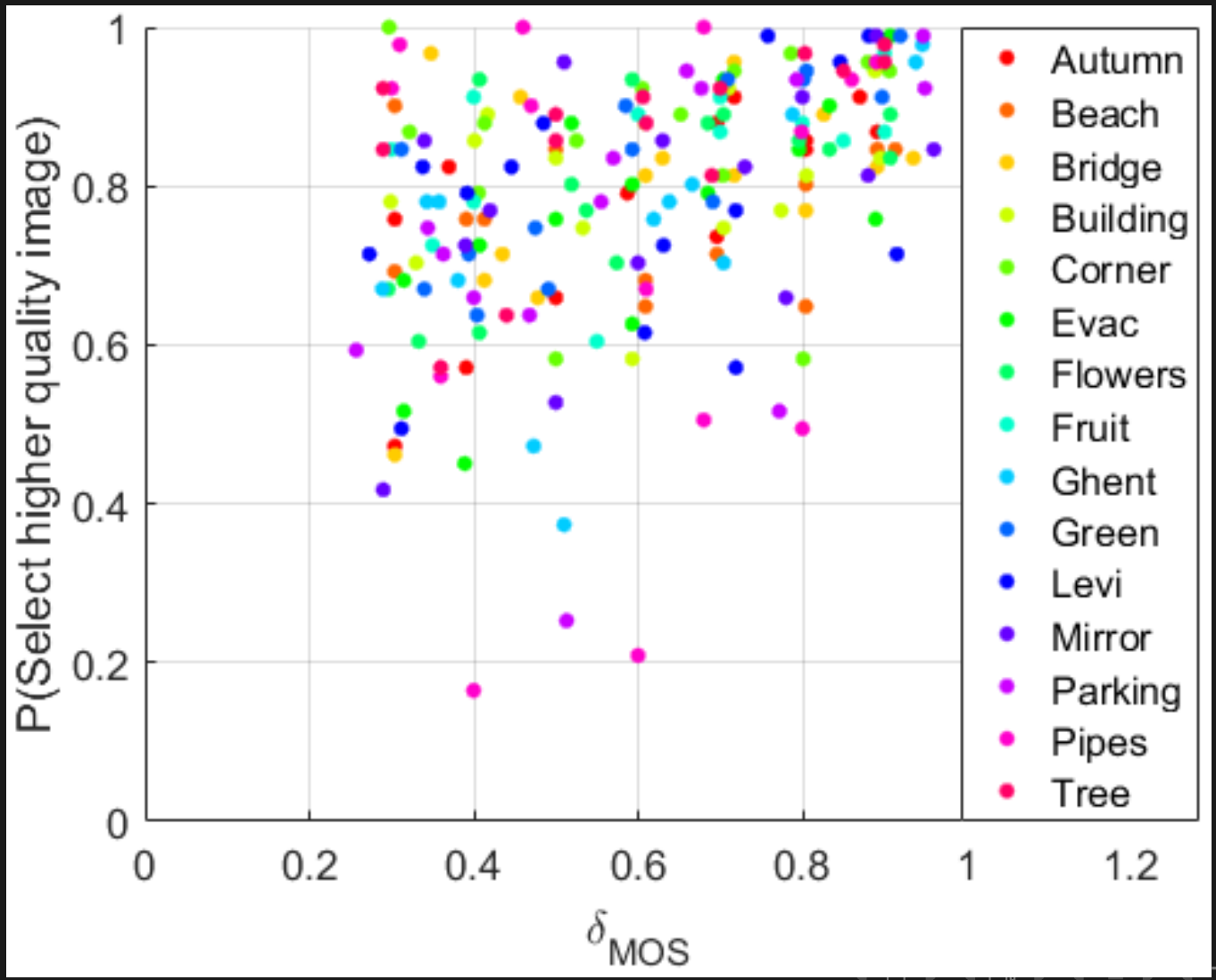
$$p = f\left(\delta_{MOS}\right) = \frac{1}{1 + e^{-\left(\beta_0 + \beta_1 \cdot \delta_{MOS}\right)}} \quad (1)$$

p_h	**All**	**Intel**	**Ghent**	**UWS**
0.90	1.0357	1.0865	0.8324	1.4015
0.80	0.6112	0.6300	0.5384	0.7434
0.75	0.4606	0.4681	0.4342	0.5100
0.70	0.3291	0.3266	0.3431	0.3060

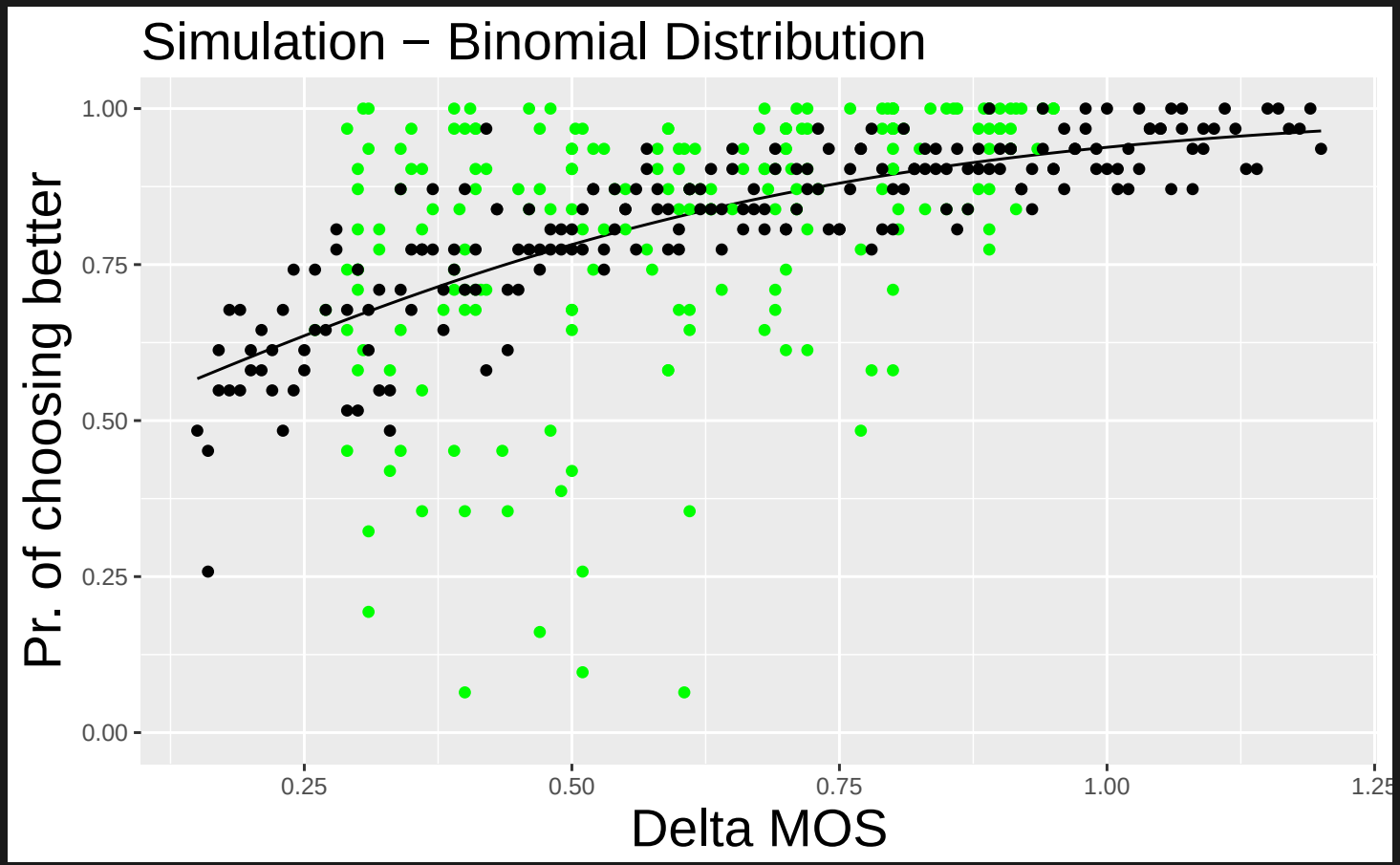
Regression for Different Laboratories



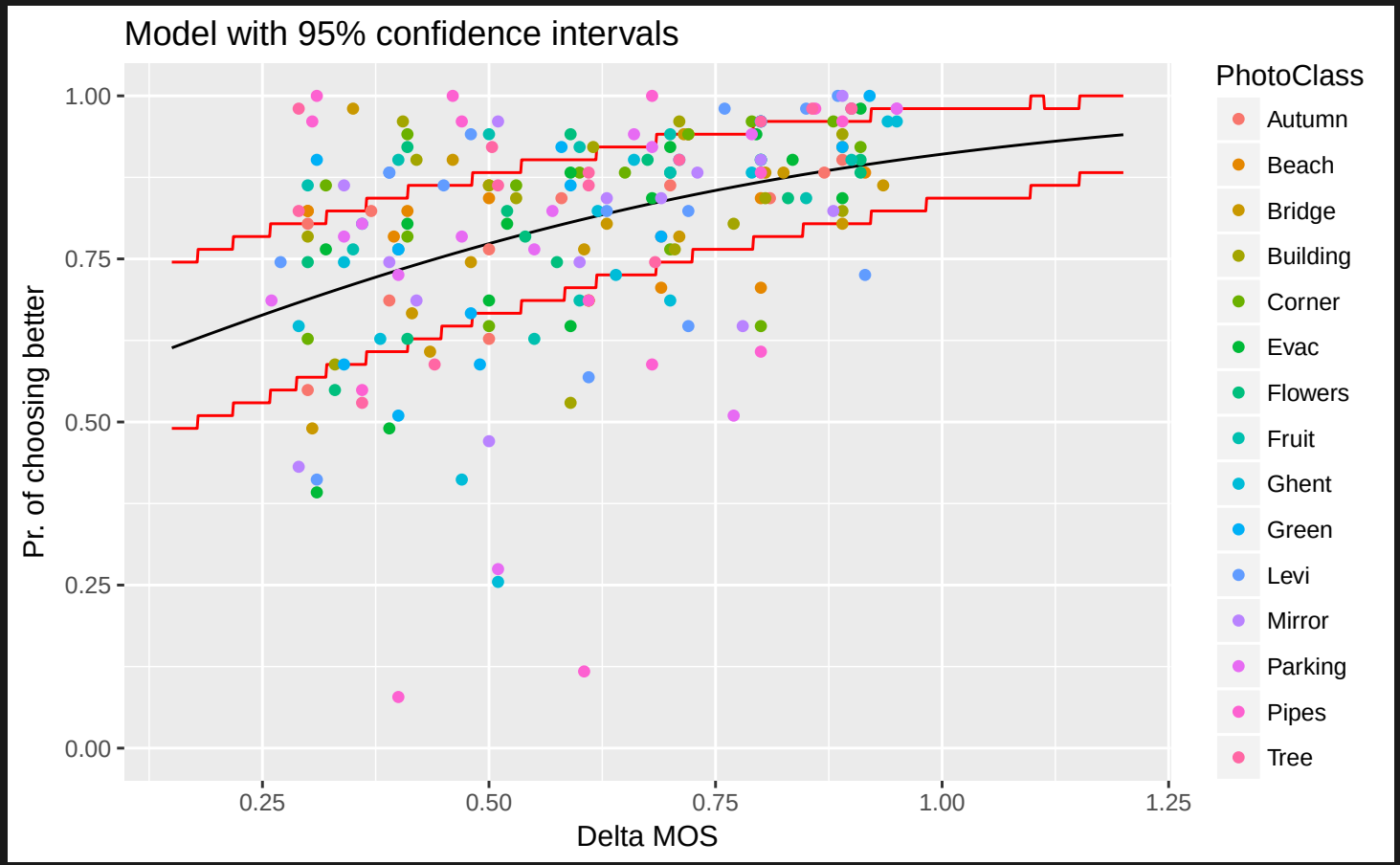
The Results Scattering



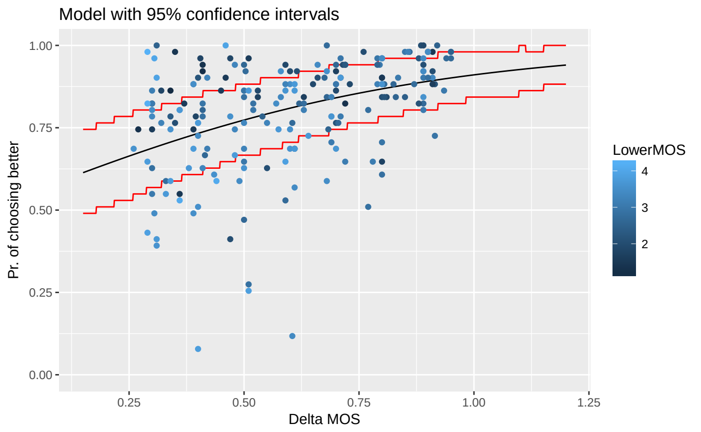
Scattering Simulation



Scattering Per Image Type



Scattering δ



Example 1

(a) Image Pipes.D, MOS: 4.200

(b) Image Pipes.W, MOS: 3.800

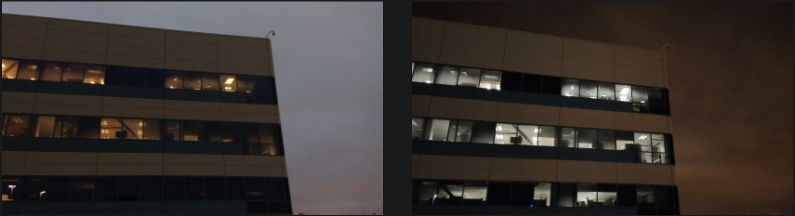
Figure: The most extreme case with $\delta_{MOS} = 0.400$ and only 16.48% of subjects choosing Pipes.D.

Example 2

(a) Image Pipes.AA, MOS: 3.990 (b) Image Pipes.DD, MOS: 3.390

Figure: The second most extreme case with $\delta_{MOS} = 0.600$ and only 20.88% of subjects choosing Pipes.AA.

Example 3



(a) Image Corner.D, MOS: 3.130 (b) Image Corner.DD, MOS: 2.833

Figure: Extremely correct case where even with small MOS difference $\delta_{MOS} = 0.297$, all subjects (100%) chose Corner.D.

Example 4

(a) Image Pipes.BB, MOS: 2.810

(b) Image Pipes.I, MOS: 2.500

Figure: Extremely correct case where even with small MOS difference $\delta_{MOS} = 0.310$, 97.80% of the subjects chose Pipes.BB.

Outline

Conclusions

- We cannot see influence of a scene type on the obtained results
- The estimated logistic regression was used to compute the percentage of people that would successfully detect the higher quality image, as a function of the MOS difference between two images
- A MOS difference of 0.46 is required in order for 75% of the people to be able to detect the higher quality image
- The experiment with different images is very difficult and for some cases the obtained results does not hold
- The detected differences between laboratories resulted in difference in the obtained MOS difference, it is interesting to understand dose differences better
- The results are solid for the MOS delta which are in the middle of the investigated scale $\delta \simeq 0.5$

Link to JND

- We have up to 8 noticeable differences between uncompressed and bad quality (MOS = 2?)
- Our finding tels that $\delta_{MOS} = 0.46$ is noticeable
- The levels would be:
  1. 5.00-4.54
  2. 4.54-4.08
  3. 4.08-3.62
  4. 3.62-3.16
  5. 3.16-2.70
  6. 2.70-2.24
  7. 2.24-1.78