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)
Problem Statement

Subjective Experiment

Data Analysis

Conclusions

Outline

1. Problem Statement
2. Subjective Experiment
3. Data Analysis
4. Conclusions
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.
Problem Statement
Subjective Experiment
Data Analysis
Conclusions

Outline

1. Problem Statement
2. Subjective Experiment
3. Data Analysis
4. Conclusions
Laboratories

<table>
<thead>
<tr>
<th>Lab</th>
<th>Country</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>USA</td>
<td>© Intel Corporation</td>
</tr>
<tr>
<td>UGhent</td>
<td>Belgium</td>
<td>Ghent University</td>
</tr>
<tr>
<td>UWS</td>
<td>Scotland (UK)</td>
<td>University of the West of Scotland</td>
</tr>
<tr>
<td>AGH</td>
<td>Poland</td>
<td>AGH University of Science and Technology</td>
</tr>
</tbody>
</table>
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 Corner
(e) 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 ($\leq 0.257$) and ($\geq 0.949$)
  - 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

1. Problem Statement
2. Subjective Experiment
3. Data Analysis
4. Conclusions
Analysis Goal

Estimate function:

\[ p = f(\delta_{MOS}) \]

where \( p \) probability of selecting a higher quality image, \( \delta_{MOS} \) the MOS difference.
Correlation

Figure: Pearson’s Correlation Coefficient between the MOS ratings of each subject and the crowd sourced ratings.
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 ($x$)

<table>
<thead>
<tr>
<th>Lab</th>
<th>PCC</th>
<th>ANOVA $p$</th>
<th>Linear fit</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>0.9632</td>
<td>0.9037</td>
<td>$y = 0.9975x - 0.0190$</td>
<td>0.9277</td>
</tr>
<tr>
<td>UGhent</td>
<td>0.8938</td>
<td>0.0917</td>
<td>$y = 1.1008x - 0.6665$</td>
<td>0.7989</td>
</tr>
<tr>
<td>UWS</td>
<td>0.9276</td>
<td>0.7165</td>
<td>$y = 1.0316x - 0.0243$</td>
<td>0.8605</td>
</tr>
<tr>
<td>All</td>
<td>0.9111</td>
<td>0.3680</td>
<td>$y = 1.0451x - 0.2429$</td>
<td>0.4379</td>
</tr>
</tbody>
</table>
Table: Correlation between the results from when the left or the right screen was selected at each lab for the same pairs of images

<table>
<thead>
<tr>
<th>Lab</th>
<th>PCC</th>
<th>Correlation</th>
<th>ANOVA p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>0.686</td>
<td>Strong</td>
<td>0.0109*</td>
</tr>
<tr>
<td>UGhent</td>
<td>0.821</td>
<td>Very strong</td>
<td>0.3349</td>
</tr>
<tr>
<td>UWS</td>
<td>0.624</td>
<td>Strong</td>
<td>0.0045*</td>
</tr>
</tbody>
</table>

* indicates a statistically significant difference.
χ² Analysis

VQEG: Interpreting VIQET scores, when can users see a difference?
\[ \chi^2 \] Analysis, Clean Data

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)

**Problem Statement**

Subjective Experiment

Data Analysis

Conclusions

**\chi^2 Analysis, Clean Data**

![Graph showing p-value per user](image-url)
$\chi^2$ Analysis, Labs for Clean Data

![Graph showing p-value per laboratory for AGH, Intel, UGent, and UWS labs.](Image)

- 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)
Probability by Image Type

![Graph showing probability of selecting higher quality image by image type: Night, Indoor, Landscape.](image-url)
The Model

\[ p = f(\delta_{MOS}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \delta_{MOS})}} \] (1)
The Model

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

<table>
<thead>
<tr>
<th>( p_h )</th>
<th>All</th>
<th>Intel</th>
<th>Ghent</th>
<th>UWS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.90</td>
<td>1.0357</td>
<td>1.0865</td>
<td>0.8324</td>
<td>1.4015</td>
</tr>
<tr>
<td>0.80</td>
<td>0.6112</td>
<td>0.6300</td>
<td>0.5384</td>
<td>0.7434</td>
</tr>
<tr>
<td>0.75</td>
<td>0.4606</td>
<td>0.4681</td>
<td>0.4342</td>
<td>0.5100</td>
</tr>
<tr>
<td>0.70</td>
<td>0.3291</td>
<td>0.3266</td>
<td>0.3431</td>
<td>0.3060</td>
</tr>
</tbody>
</table>
Regression for Different Laboratories
The Results Scattering

![Graph showing the results scattering with points plotted on a scatter plot. The x-axis represents \( \delta_{MOS} \) and the y-axis represents the probability of selecting a higher quality image. The points are color-coded by different categories such as Autumn, Beach, Bridge, Building, Corner, Evac, Flowers, Fruit, Ghent, Green, Levi, Mirror, Parking, Pipes, Tree, and others.]
Scattering Simulation

Simulation – Binomial Distribution

Pr. of choosing better vs. Delta MOS

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)

VQEG: Interpreting VIQET scores, when can users see a difference?
Scattering Per Image Type

Model with 95% confidence intervals

PhotoClass
- Autumn
- Beach
- Bridge
- Building
- Corner
- Evac
- Flowers
- Fruit
- Ghent
- Green
- Levi
- Mirror
- Parking
- Pipes
- Trees

Delta MOS

Pr. of choosing better

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)
Scattering $\delta$
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

1. Problem Statement
2. Subjective Experiment
3. Data Analysis
4. Conclusions
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 $\delta \approx 0.5$. 

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)

VQEG: Interpreting VIQET scores, when can users see a difference?
We have up to 8 noticeable differences between uncompressed and bad quality (MOS = 2?).

Our finding tells 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