AccAnn:
A new QoE measurement methodology and its application on video streaming

Jing Li, Lukas Krasula, Zhi Li,
Yoann Baveye, Patrick Le Callet

UNIVERSITÉ DE NANTES

NETFLIX

QoE in video streaming

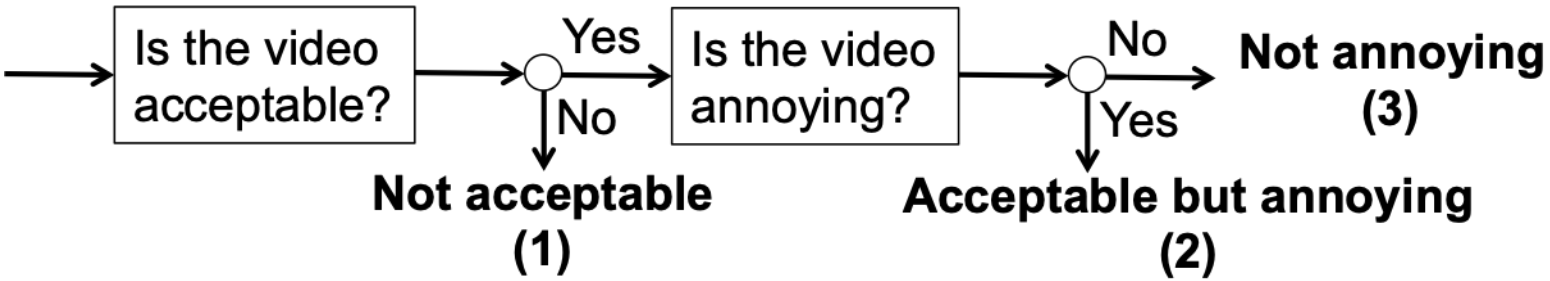
For video streaming service provider

- Exact quality rating is not that important

- It is more interesting to know the lower bound of user's QoE, i.e.,
  - Below which the video quality is not acceptable?
  - Above which the video quality is satisfying?

Measuring Acceptability/Annoyance

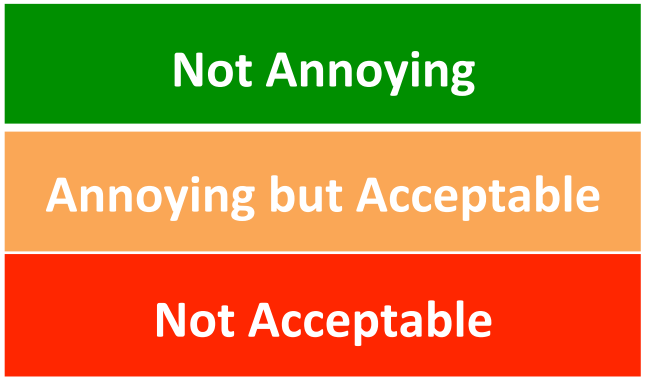
Traditional Multi-step method:



Satu Jumisko-Pyykk¨o and Miska M. Hannuksela, “Does context matter in quality evaluation of mobile television?,” in Int. Conf. on Human Computer Interaction with Mob. Devices and Serv., 2008.
D. Khaustova, Objective Assessment of Stereoscopic Video Quality of 3DTV, Ph.D. thesis, University of Rennes, 2015.

AccAnn: A new subjective QoE assessment methodology

- **Objective**: detecting the Acceptability and Annoyance threshold
- Viewers are asked to provide their opinion on the QoE in terms of *



* D. Khaustova, Objective Assessment of Stereoscopic Video Quality of 3DTV, Ph.D. thesis, University of Rennes, 2015

User profile assumption

- User profile → user's expectation
- In reality, it's hard to get diverse profiles to make analysis
- Question:

Can we simulate user's profile by assigning a "role" to the observers and evaluate his/her expectation accordingly?

Experiment instruction

- You are a basic/premium user, watch on TV/ Tablet

  - Basic user costs 6 euros/month
  - Premium user costs 12 euros/month

- The video is **not acceptable** when its quality is **not sufficient for the price you are paying**. Such quality would make **you think about changing** the service or provider.
- The video is **annoying** when its quality is acceptable (**would not** make you think about changing the service) but **not sufficient to satisfy** your expectations.
- **Not annoying** video there fore **satisfies** your expectations about the services.

Experiment

- Test video sequences:
  - 10 Full HD source videos, 10 seconds
  - 4 quality levels (Netflix's per-tile encode optimization)+ 1 reference (no distortion)
  - In total 49 videos in the test (1 was missing during playlist generation)
- Two devices:
  - TV (Philips 46PFL9705H Full HDTV 46')
  - Tablet (Samsung Galaxy Tab A6 10.1', Full HD)

Experiment

- Subjects:
  - 33 naïve observers
  - Each observer is assigned a “profile”:
    - Basic user, costs 6 Euros/month
    - Premium user, costs 12 Euros/month
  - 17 Basic users and 16 Premium users
  - Make **Acceptability**/**Annoyance** judgment based on their profile assumption

Experiment

- Test environment and procedure
  - Each observer evaluated the videos on two devices (at different time).
  - Watching Tablet in a “home-like” environment
    - Free viewing distance, sit on a leather sofa with any position they wanted
  - Watching TV in 3h viewing distance
  - Room illumination： ITU-R BT.500
  - Each test duration： ~13 minutes/observer

Experimental results

- **3 - Not Annoying**
- **2 - Annoying but Acceptable**
- **1 - Not Acceptable**

	Pvs 1	Pvs 2	Pvs 3	Pvs 4	…	Pvs N
Obs 1	1	3	2	1		3
Obs 2	2	3	2	2		2
Obs 3	1	3	3	1		3
…						
Obs M	2	2	3	2		2

Results analysis

- Besides Mean Opinion Score…

How to analyze the AccAnn data

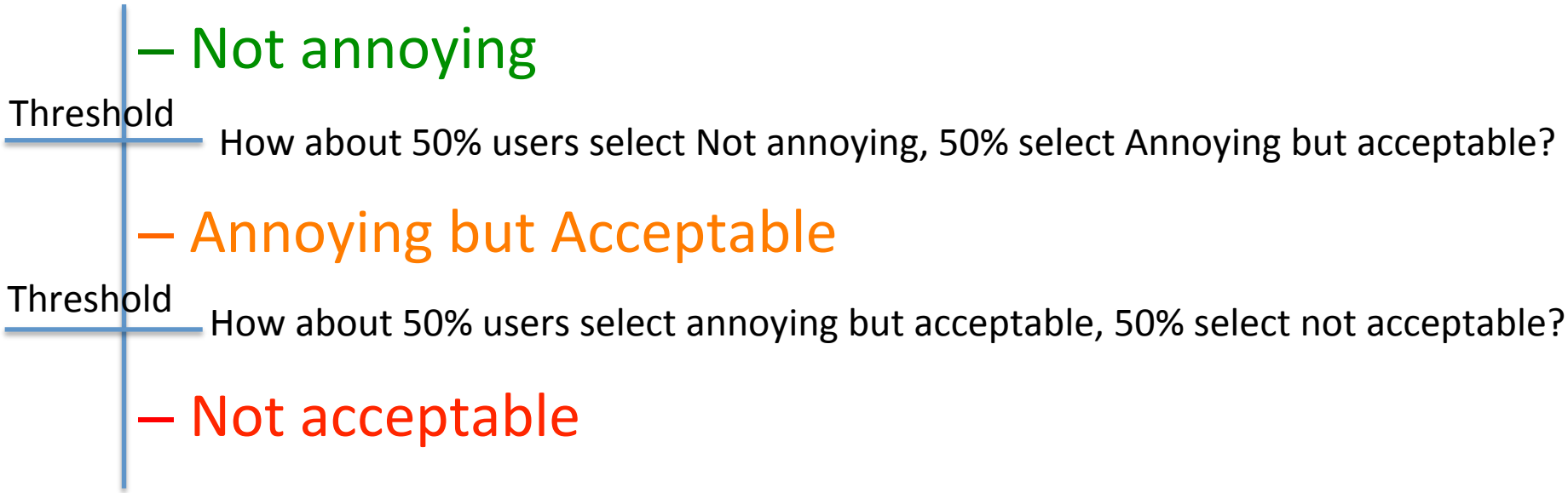
- Acceptability/Annoyance is not a score, but a category:
  - Not annoying

    How about 50% users select Not annoying, 50% select Annoying but acceptable?
  - Annoying but Acceptable

    How about 50% users select annoying but acceptable, 50% select not acceptable?
  - Not acceptable

How to analyze the AccAnn data

- Acceptability/Annoyance is not a score, but a category:



How to analyze the AccAnn data

- Acceptability/Annoyance is not a score, but a category:

Using exact text:
Barnard's exact test, or
Fisher's exact test
e.g., For video A,
15 obs select not annoying
18 obs select annoying but accept.

Input:

15	18
18	15

output: p-value = 0.6
→ No significant difference
→ Video A:
unsure about its annoyance
for sure about its acceptability

How to analyze the AccAnn data

- Acceptability/Annoyance is not a score, but a category:

3	Not annoying
2.5	Unsure about annoyance (threshold)
2	Annoying but Acceptable
1.5	Unsure about acceptability (threshold)
1	Not acceptable

Application I:

Benchmarking of the state of the art video quality metrics

Evaluation

Evaluated metrics:

- PSNR
- PSNRHVS[Ponomarenko2007]
- SSIM[Wang2004]
- VIFp[Sheikh2006]
- VQM[Pinson2004]
- VQM_VFD[Wolf2011]
- VMAF[Li2016]

Evaluation methods:
PLCC
ROCC
between Objective score and
AccAnn categories.

3	Not annoying
2.5	Unsure about annoyance (threshold)
2	Annoying but Acceptable
1.5	Unsure about acceptability (threshold)
1	Not acceptable

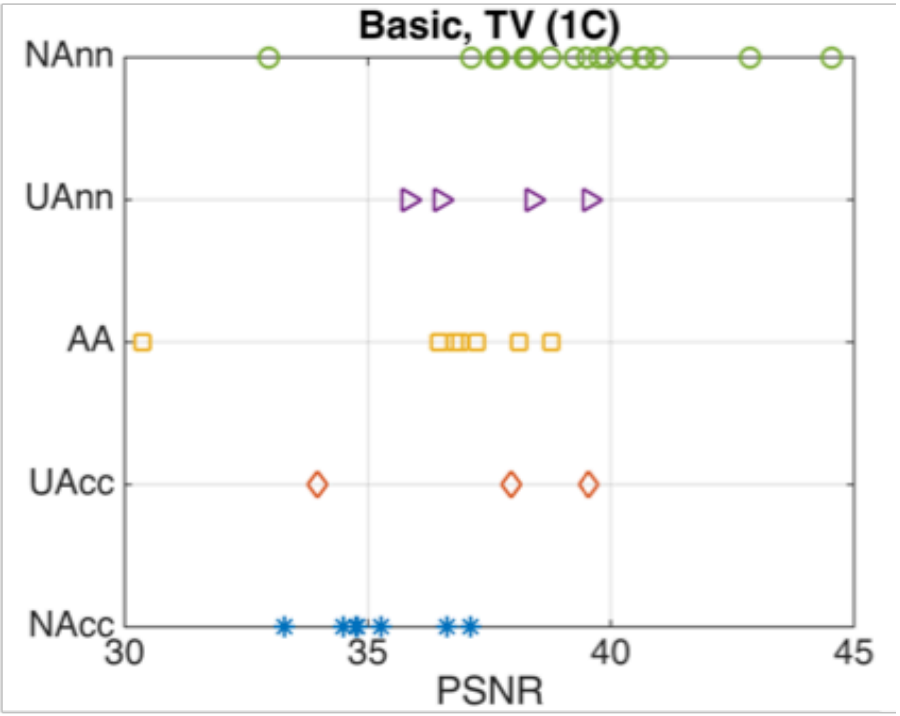
Performance

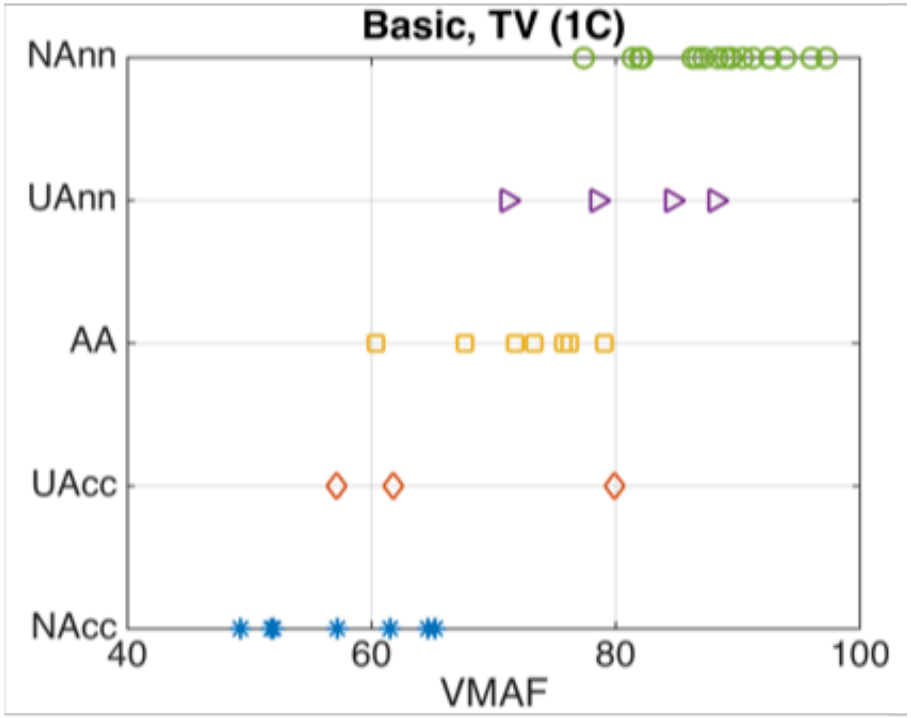
PLCC				
Scenario	1C	1D	2C	2D
PSNR	0.6647	0.7238	0.7211	0.7597
PSNRHVS[45]	0.7184	0.7691	0.7480	0.8117
SSIM[38]	0.4679	0.4975	0.6401	0.5910
VIFp[46]	0.6685	0.6489	0.6657	0.6444
VQM[41]	0.8482	0.8637	0.8307	0.9069
VQM-VFD[42]	0.9013	0.8989	0.8362	0.9227
VMAF[44]	**0.9028**	**0.9075**	**0.8796**	**0.9289**

SROCC				
Scenario	1C	1D	2C	2D
PSNR	0.5988	0.6417	0.6785	0.7275
PSNRHVS[45]	0.6395	0.6763	0.7026	0.7704
SSIM[38]	0.5549	0.5638	0.6231	0.5923
VIFp[46]	0.5866	0.5974	0.6352	0.6085
VQM[41]	0.6864	0.6810	0.7233	0.7976
VQM-VFD[42]	0.7194	0.7172	0.7389	0.8059
VMAF[44]	**0.8128**	**0.8206**	**0.8379**	**0.8800**

1C: TV, Basic
1D: Tablet, Basic
2C: TV, Premium
2D: Tablet, Prem

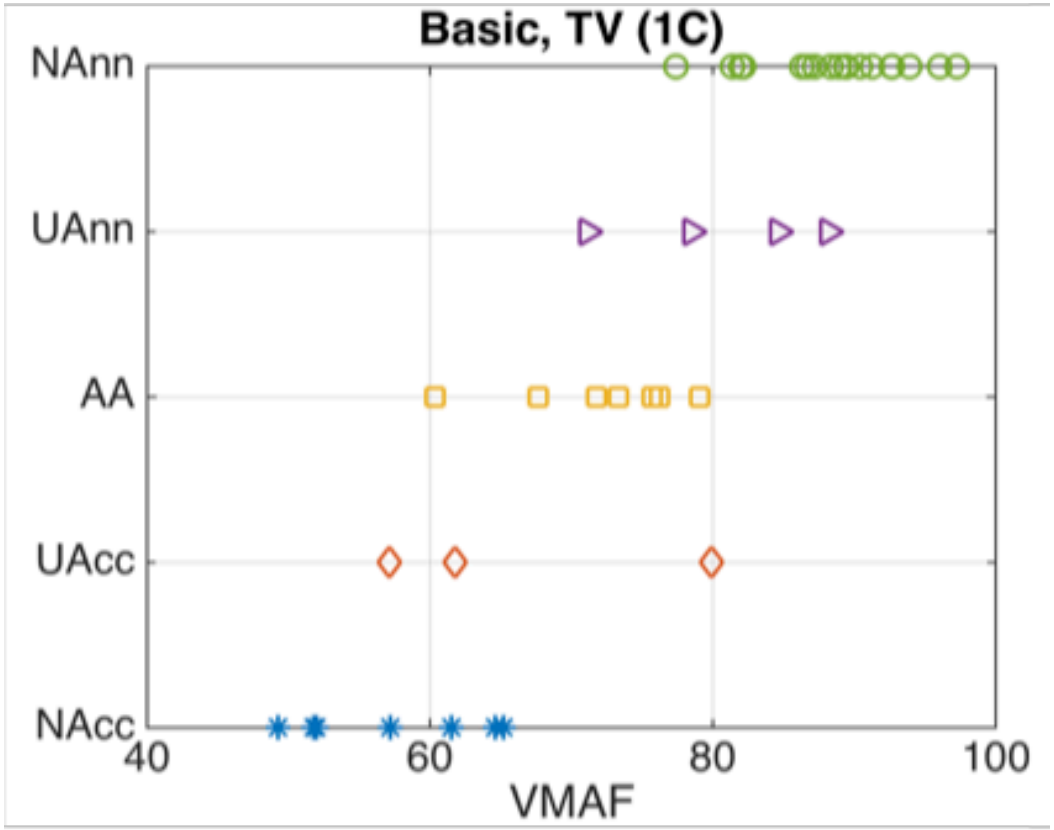
Scatter plot



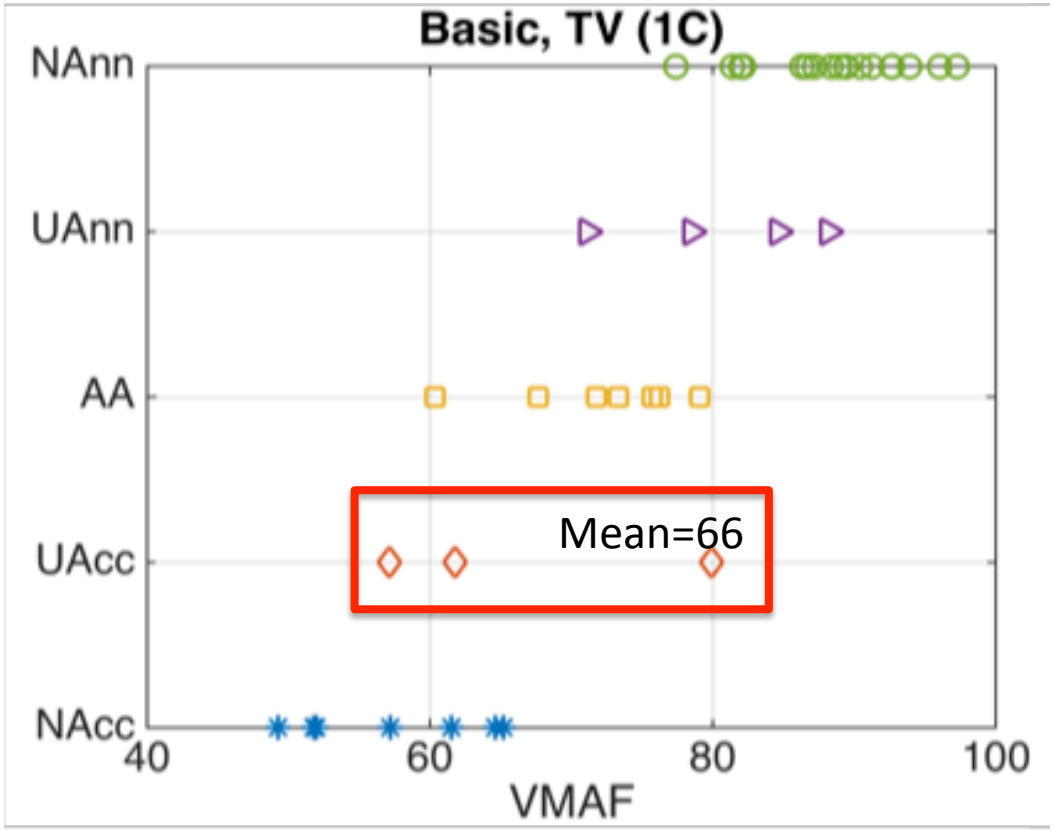


NAnn: Not annoying
UAnn: Unsure about annoyance
AA: Annoying but Acceptable
UAcc: Unsure about acceptability
Nacc: Not acceptable

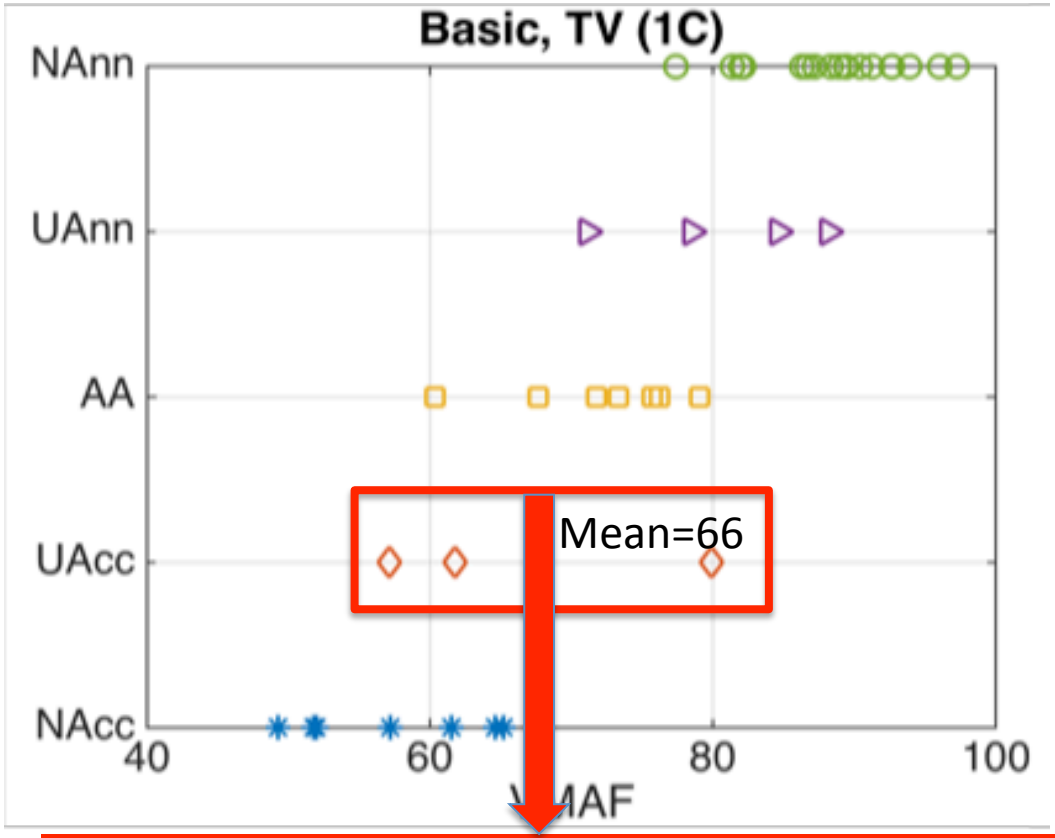
Thresholds of objective metric for Acceptability/Annoyance



Thresholds of objective metric for Acceptability/Annoyance

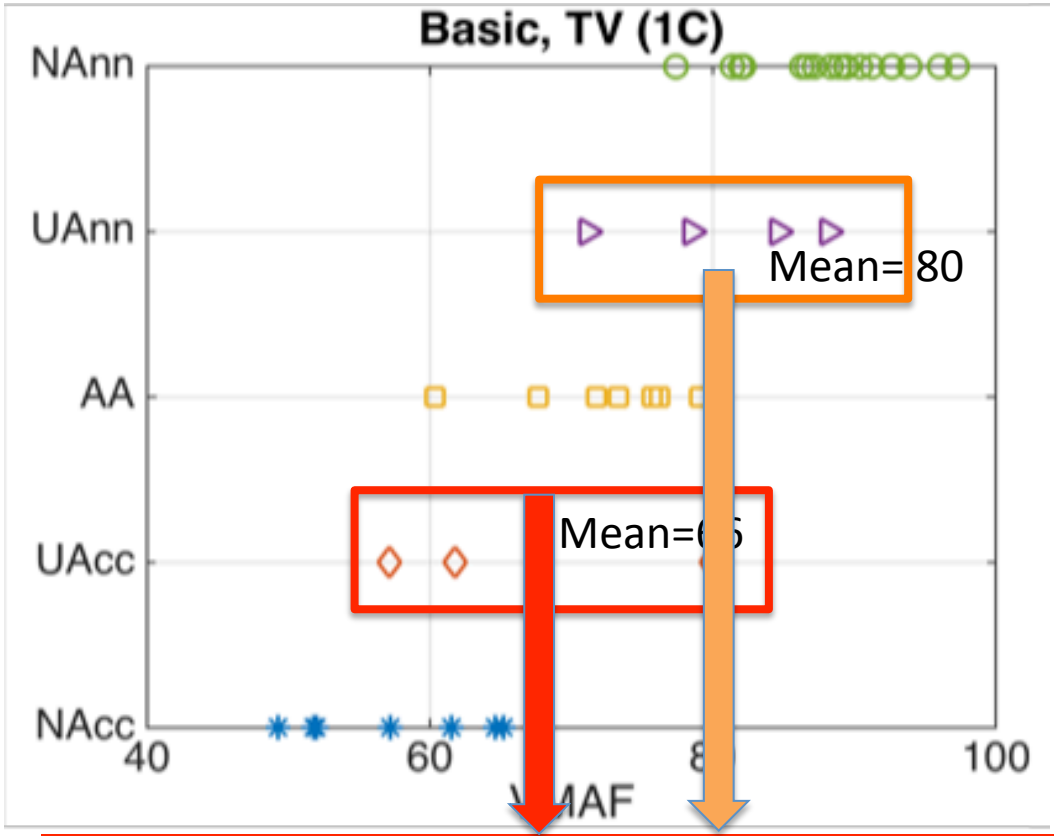


Thresholds of objective metric for Acceptability/Annoyance



VMAF score 66 is considered as the Acceptability threshold below which the video streaming service is not acceptable

Thresholds of objective metric for Acceptability/Annoyance



VMAF score 80 is considered as the Annoyance threshold above which the users may satisfy the service

Thresholds for VMAF

	TV, Basic users	Tablet, Basic users	TV, Premium users	Tablet, Premium users
Annoyance threshold	80	80	85	87
Acceptability threshold	66	58	74	71

- Users have higher tolerance on Tablet than on TV
- Premium users are more picky than Basic users

- Note: Basic and Premium users are assigned roles, not the real case.
  → user profile assumption in this case really works (please refer to our paper to see more details).

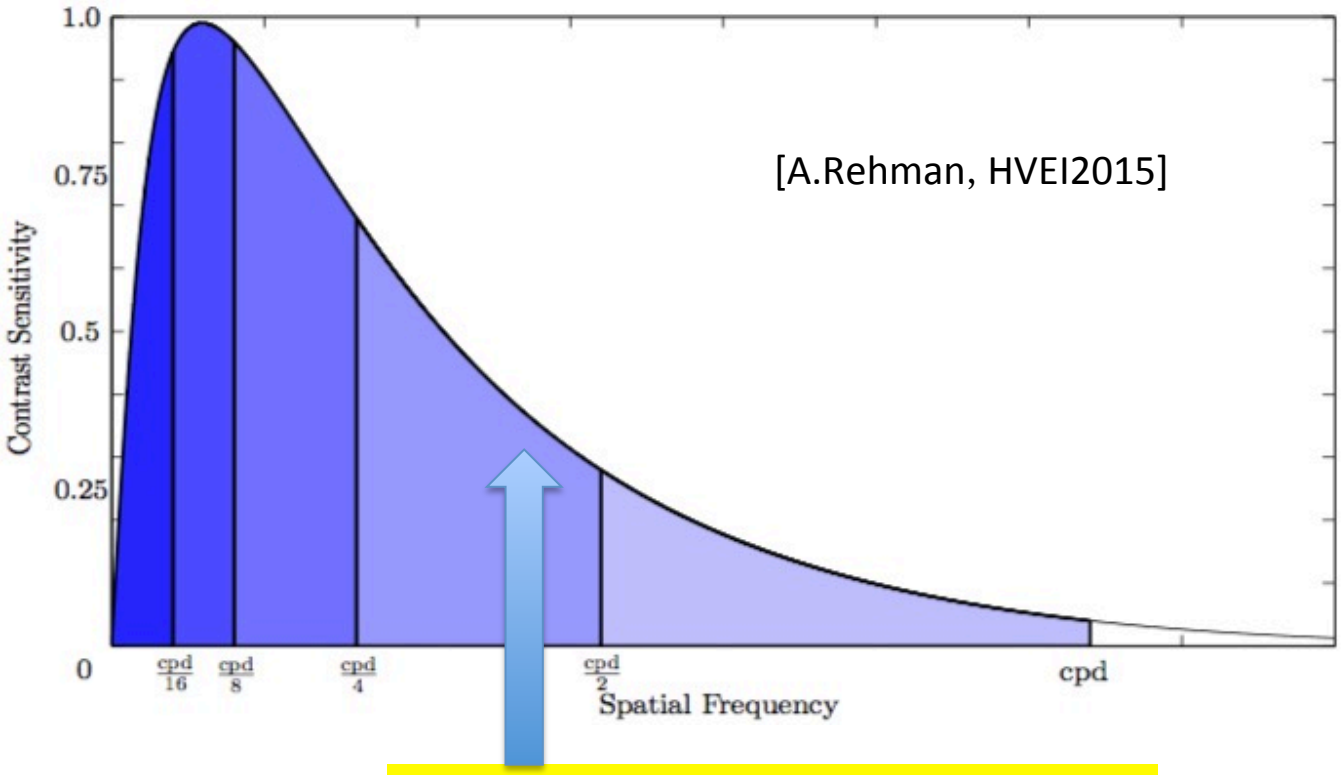
Application II:

Quantifying the Influence of Devices on Quality of Experience for Video Streaming

Influence of device on QoE

Perceptual quality

Other factors



[A.Rehman, HVEI2015]

Contrast sensitivity is determined by:
Average or range of viewing distance
Size of screen
Screen resolution
Screen contrast
Illumination of viewing environment
Viewing angle
…

Experiment

- 49 PVS
- Subjects:
  - 33 naïve observers
  - Each observer is assigned a “profile”:
    - Basic user, costs 6 Euros/month
    - Premium user, costs 12 Euros/month
  - 17 Basic users and 16 Premium users
  - Make **Acceptability**/**Annoyance** judgment based on their profile assumption
- Two devices:
  - TV (Philips 46PFL9705H Full HDTV 46’)
  - Tablet (Samsung Galaxy Tab A6 10.1’, Full HD)
- **Each observer evaluated the videos on two devices** (at different time).

Quantifying the influence of device

- Eliminated By Aspects (EBA) model [Tversky1972]

When we make choice between items

(a) the common characteristics of the considered choice set are eliminated, as any discriminating choice cannot be based on them ;
(b) a characteristic is randomly selected and all the options not having this characteristic are eliminated. The higher the utility of a characteristic is, the larger the probability of selecting this characteristic is ;
(c) if remaining options still have specific characteristics, one turns over at the first stage. In the contrary, if the residual choices have the same characteristics, the procedure ends. If only one option remains, it is selected. In the contrary, all remaining options have the same probability to be selected.

Quantifying the influence of device

- Eliminated By Aspects (EBA) model[Tversky1972]
  - Each video has its own quality attribute: $u(q_i)$

    $i = 1, 2, \ldots, 49$
  - Each video is shown on TV or Tablet: $u(d_i)$

    $d_i = d_{TV}$ or d_{tab}
  - The probability that observer prefers video *i* over video *j* is:

$$P_{ij} = \frac{u(q_i) + u(d_i)}{u(q_i) + u(d_i) + u(q_j) + u(d_j)}$$

EBA model

- Converting AccAnn score to Pair Comparison

  For an observer *s*

  – If score i > score j, $pcm_s(i,j) = 1$

  $pcm_s(j,i) = 0$

  For all observers → $M_{ij} = \sum_s pcm_s$

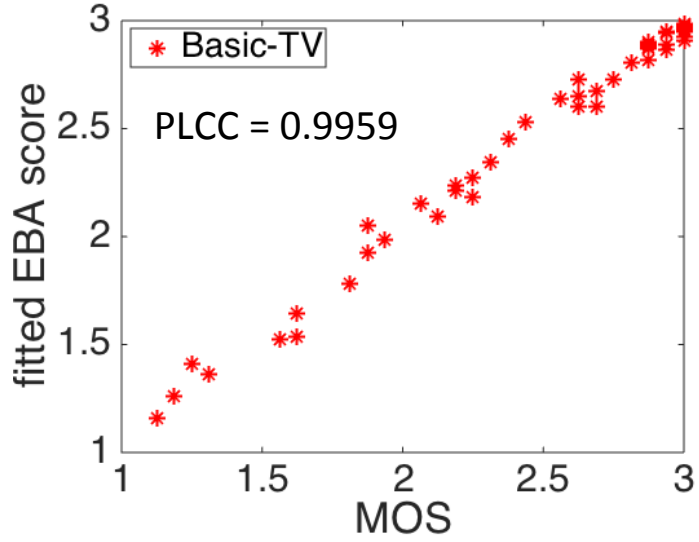
  Likelihood Function:

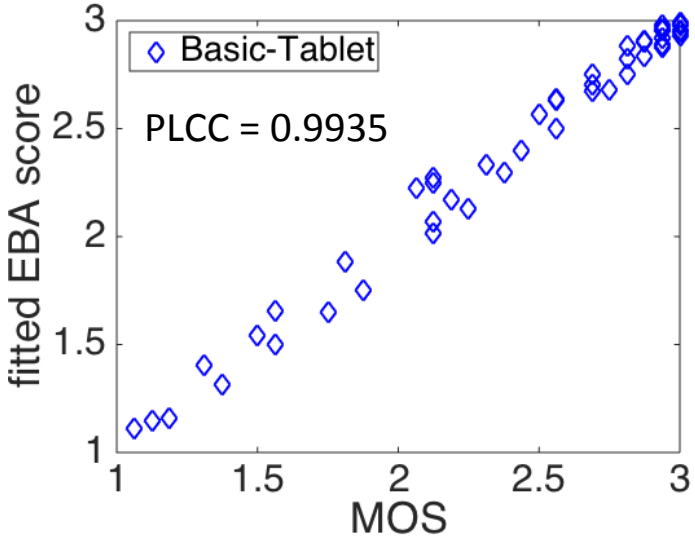
$$L = \prod_{i<j} p_{ij}^{M_{ij}} (1 - p_{ij})^{M_{ji}}$$

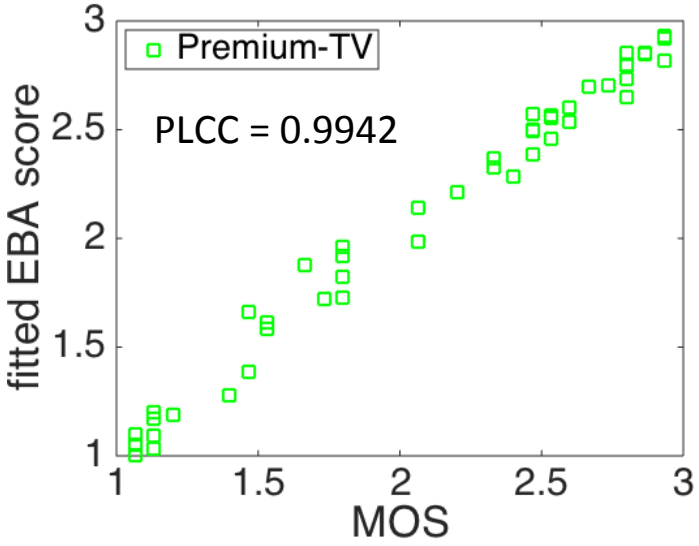
$$P_{ij} = \frac{u(q_i) + u(d_i)}{u(q_i) + u(d_i) + u(q_j) + u(d_j)}$$

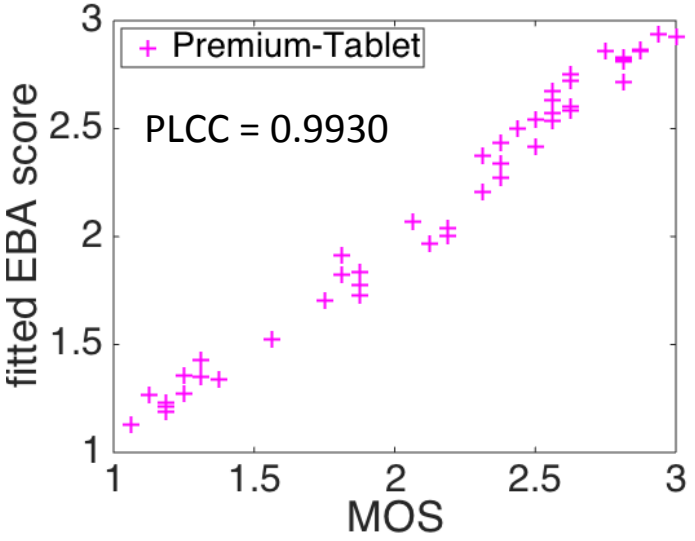
- Estimating $u(q_i)$ and $u(d_i)$ by MLE
  – Recover the true quality of video sequence and the **influence from device** (TV and Tablet)

Results: recovered AccAnn score using EBA model





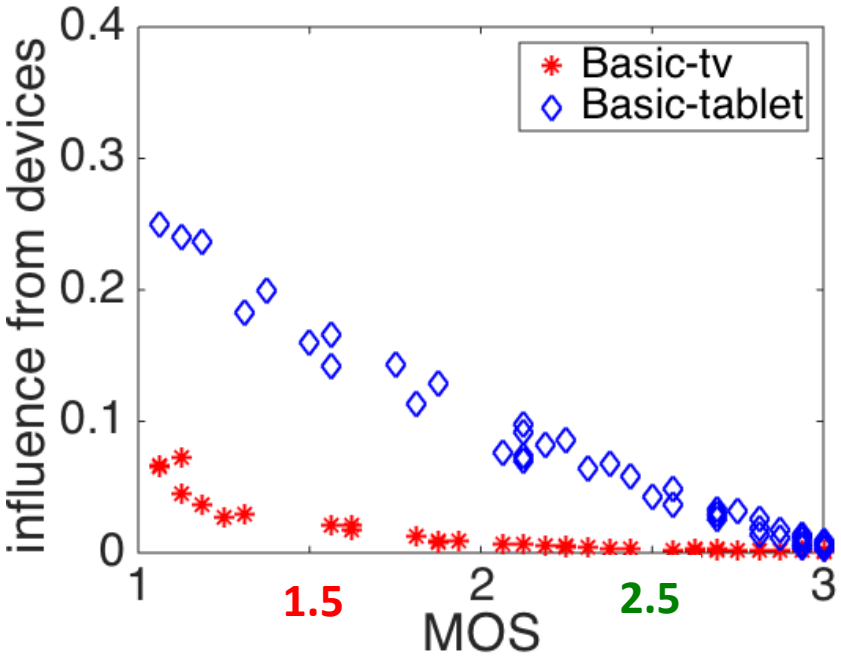


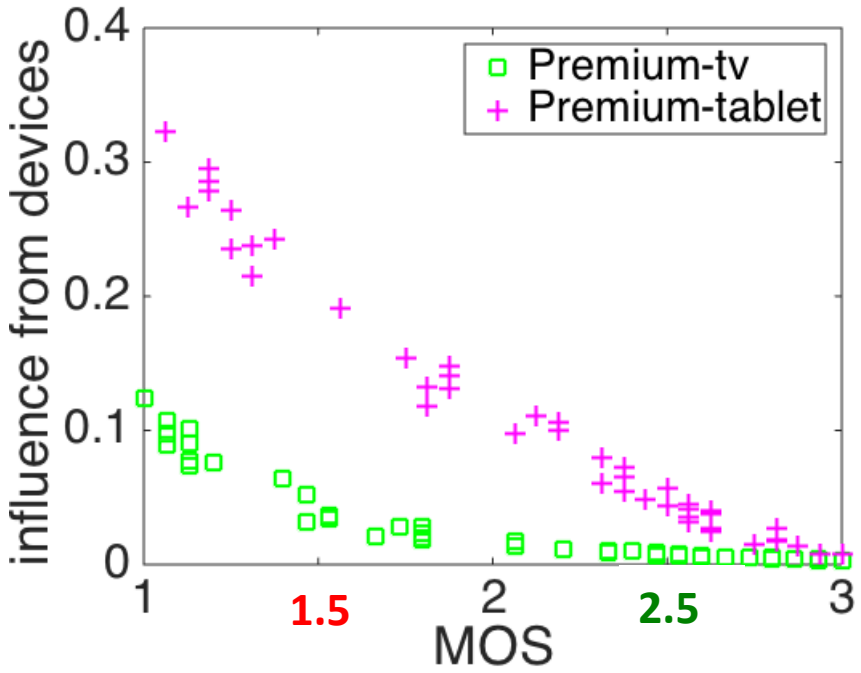


Note: MOS here is mean opinion score rather than AccAnn category

Results: influence from device $u(d_i)$

For y-axis, higher value means higher QoE

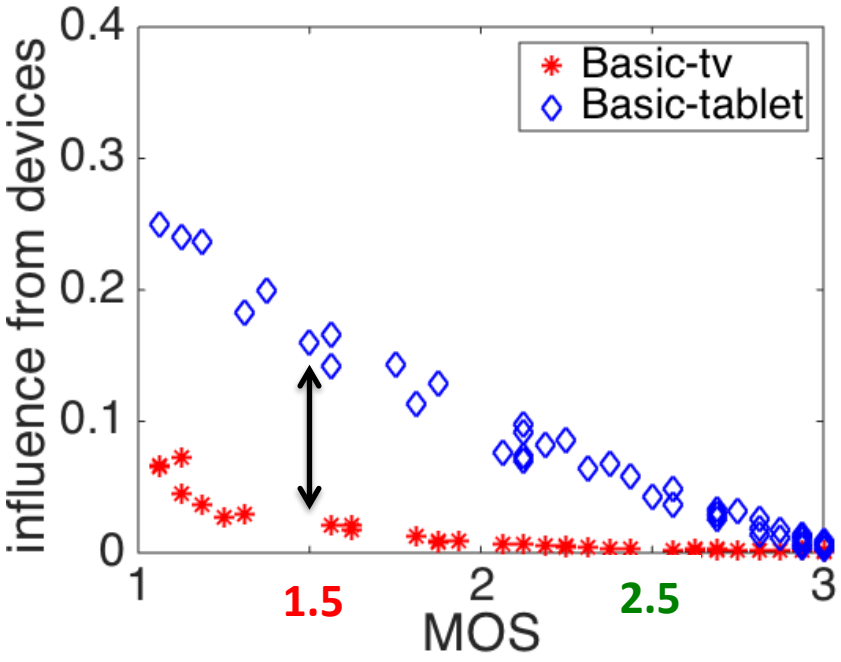


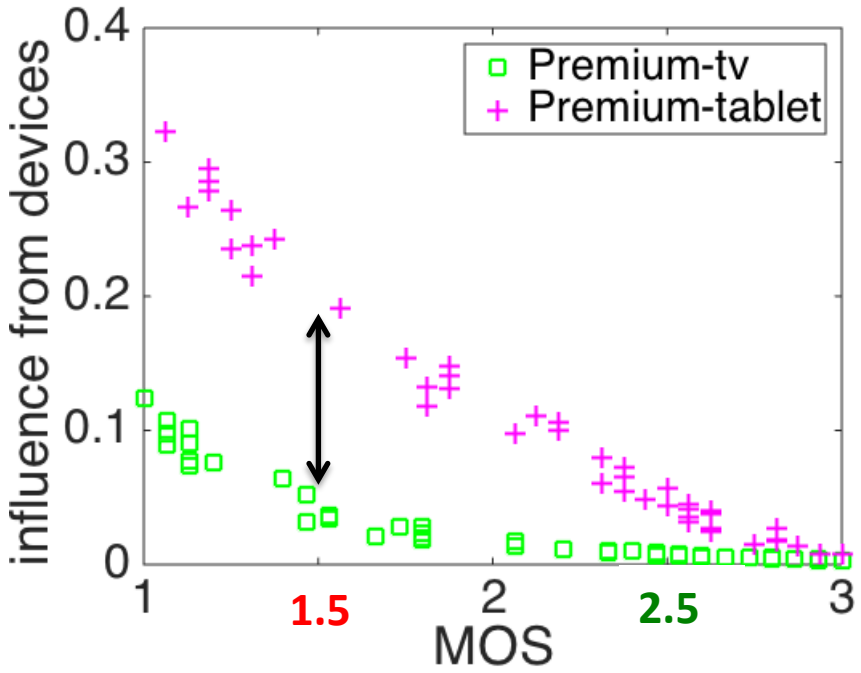


The influence of device on Acceptance/Annoyance is QoE dependent:

Results: influence from device $u(d_i)$

For y-axis, higher value means higher QoE



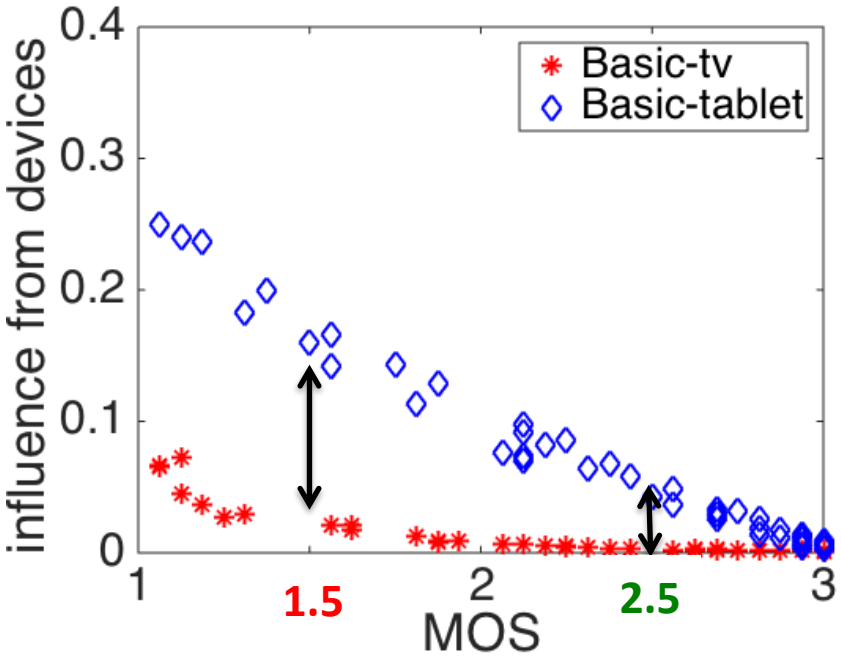


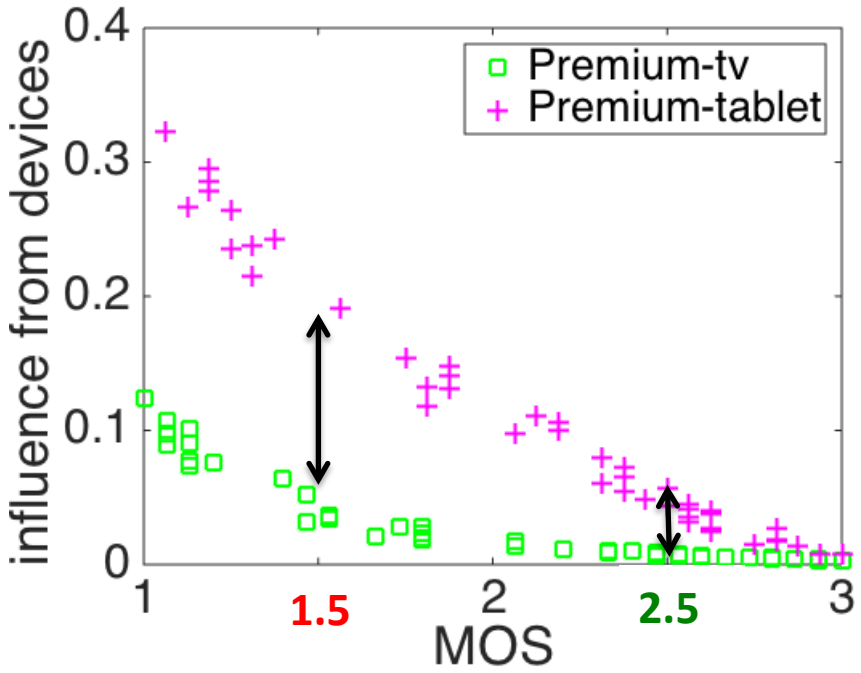
The influence of device on Acceptance/Annoyance is QoE dependent:

- The observers using Tablet had higher tolerance on **Unacceptability threshold** of the video sequence (MOS = 1.5) than watching on TV.

Results: influence from device $u(d_i)$

For y-axis, higher value means higher QoE





The influence of device on Acceptance/Annoyance is QoE dependent:

- The observers using Tablet had higher tolerance on **Unacceptability threshold** of the video sequence (MOS = 1.5) than watching on TV.

- The influence of devices on the **thresholds of Annoyance** (MOS = 2.5) was less than that of Unacceptability.

Conclusion

- A new QoE assessment methodology: AccAnn
  - Evaluate Acceptance/Annoyance of video
  - More efficient than the traditional multi-step approach
  - User profile can be assigned
- Threshold of objective quality metrics on Acceptability/Annoyance
  - VMAF 66 (acceptability) and VMAF 80 (annoyance) for TV, Basic users
- Influence of device on AccAnn
  - Influence is not constant, depending on quality
    - High quality video: small/little influence from device
    - Low quality video: Tablet shows better experience than TV
- Important information for service providers:
  - how much difference could be made on video encodes for different devices?
- Train objective quality metric
  - based on recovered device-neutral AccAnn score + adapt to difference devices

Corresponding work to this presentation

- Jing Li, Lukas Krasula, Yoann Baveye, Zhi Li, Patrick Le Callet, "**AccAnn: A New Subjective Assessment Methodology for Measuring Acceptability and Annoyance of Quality of Experience"***, IEEE Trans. Multimedia*, *accepted, 2019*

- Jing Li, Lukas Krasula, Zhi Li, Yoann Baveye, Patrick Le Callet, **"Quantifying the influence of devices on Quality of Experience for Video streaming**", *PCS 2018*.

Thank you very much!