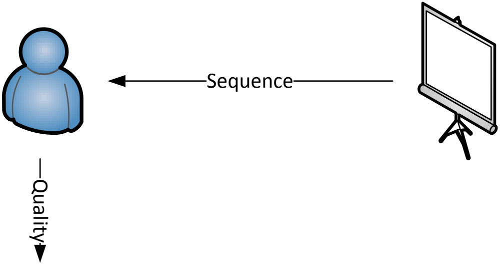
The Accuracy of Subjects In A Quality Experiment: A Theoretical Subject Model

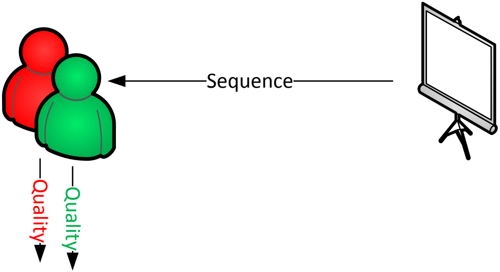
Lucjan Janowski (AGH), Margaret Pinson (NTIA)

Quality



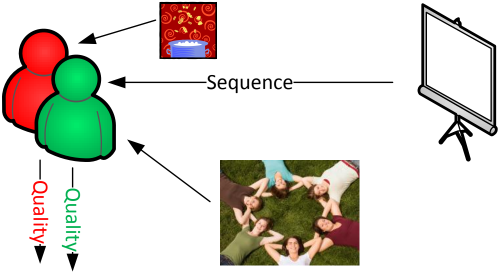
$$o_{ij} = \psi_j$$

Bias



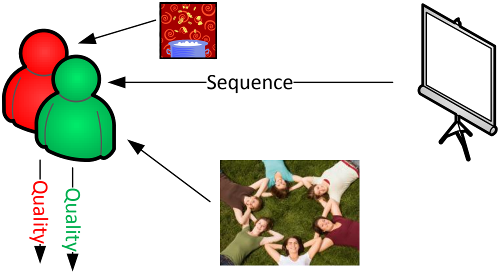
$$o_{ij} = \psi_j + \Delta_i$$

Error



$$o_{ij} = \psi_j + \Delta_i + \epsilon_{ij}$$

Can we observe users' bias?



$$o_{ij} = \psi_j + \Delta_i + \epsilon_{ij}$$

Data Source

- Name: AGH/NTIA
- Number of PVSs: 110
- Number of Subjects: 28

Data Source

- Name: AGH/NTIA
- Number of PVSs: 110
- Number of Subjects: 28

- Name: HDTV
- Number of PVSs: 834
- Number of Subjects: 144

Data Source

- Name: AGH/NTIA
- Number of PVSs: 110
- Number of Subjects: 28

- Name: vqegMM2
- Number of PVSs: 60
- Number of Subjects: 213

- Name: HDTV
- Number of PVSs: 834
- Number of Subjects: 144

Data Source

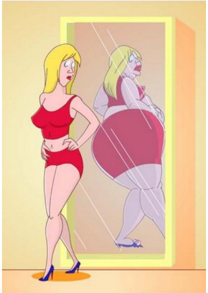
- Name: AGH/NTIA
- Number of PVSs: 110
- Number of Subjects: 28

- Name: vqegMM2
- Number of PVSs: 60
- Number of Subjects: 213

- Name: HDTV
- Number of PVSs: 834
- Number of Subjects: 144

- Name: NTIA/Verizon
- Number of PVSs: 144
- Number of Subjects: 21

What is subject bias i.e. Δ_i?



Δ increase

Unfavourable

Favourable

Image source: www.barking-moonbat.com/index.php/weblog/2006/06/P168/

Δ_i Estimation

Our subject answer model is:

$$o_{ij} = \psi_j + \Delta_i + \epsilon_{ij}$$

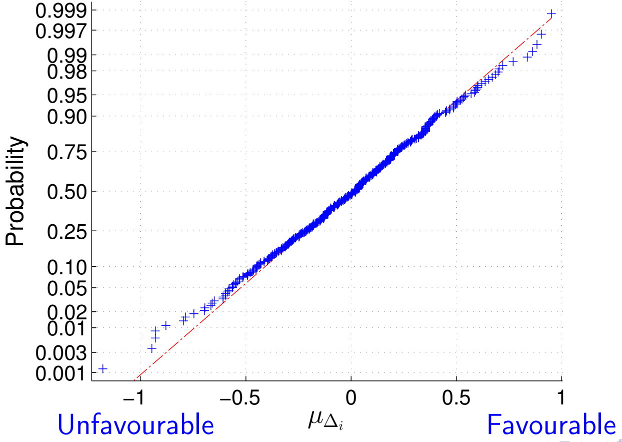
For each PVS we can estimate ψ_j by:

$$\psi_j \approx \mathrm{MOS} = \mu_{\psi_j} = \frac{1}{I}\sum_{i=1}^{I} o_{ij}$$

Knowing ψ_j we can estimate Δ_i by:

$$\Delta_i \approx \mu_{\Delta_i} = \frac{1}{J}\sum_{j=1}^{J}(o_{ij} - \mu_{\psi_j}) \approx \frac{1}{J}\sum_{j=1}^{J}(\Delta_i + \epsilon_{ij}) = \Delta_i + \frac{1}{J}\sum_{j=1}^{J}\epsilon_{ij}$$

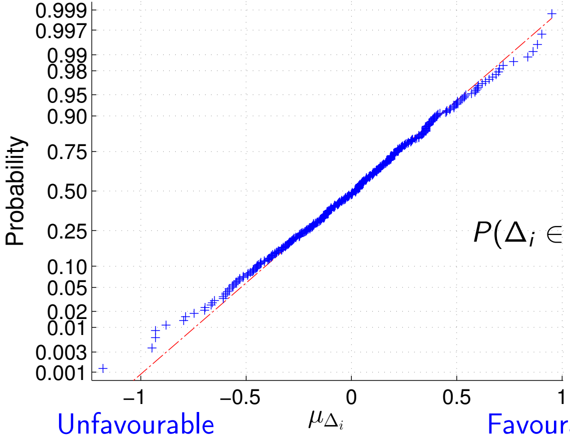
Δ_i Distribution



$E(\Delta_i) = 0$

$\sigma(\Delta_i) = 0.34$

Δ_i Distribution



Unfavourable

Favourable

$$E(\Delta_i) = 0$$

$$\sigma(\Delta_i) = 0.34$$

$$P(\Delta_i \in \underbrace{(-0.5, 0.5)}_{\text{25\% of scale}}) \approx 0.85$$

Δ_i Distribution

Dataset	Subjects	μ_{Δ_i} [min,max]
AGH/NTIA	28	[-0.59, 0.60]
vqegHD1	24	[-0.54, 0.40]
vqegHD2	24	[-0.67, 0.49]
vqegHD3	24	[-0.56, 0.88]
vqegHD4	24	[-1.18, 0.95]
vqegHD5	24	[-0.69, 0.70]
vqegHD6	24	[-0.79, 0.45]
vqegMM2	213	[-0.95, 0.90]
NTIA/Verizon	21	[-0.65, 0.60]

Are we interested in Δ_i?

- If the sequence quality is needed, the Δ influence is not needed
- If the individual opinion is investigated, Δ should be taken into account
- If the difference between users is analyzed, Δ could be the main point of investigation

Decreasing MOS Variance

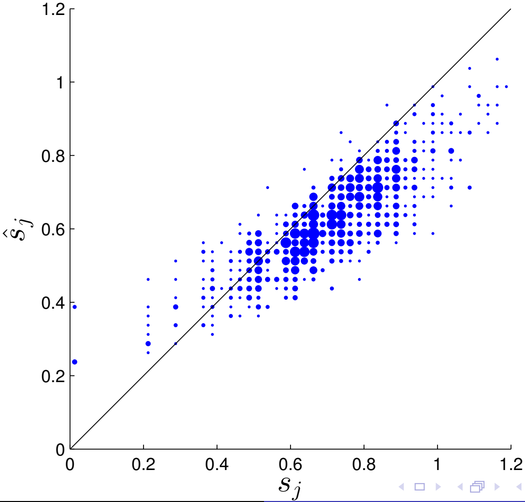
Standard deviation of the answers:

$$s_j = \sigma_i(o_{ij}) \quad (1)$$

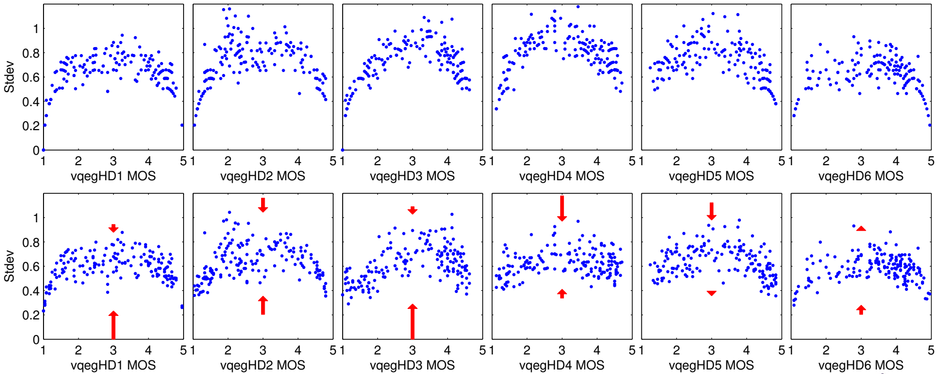
Standard deviation of non bias answers:

$$\hat{s}_j = \sigma_i(o_{ij} - \mu_{\Delta_i}) \quad (2)$$

s_j and $\hat{s}_j$ comparison, VQEG HDTV



Detail view on standard deviation change, VQEG HDTV

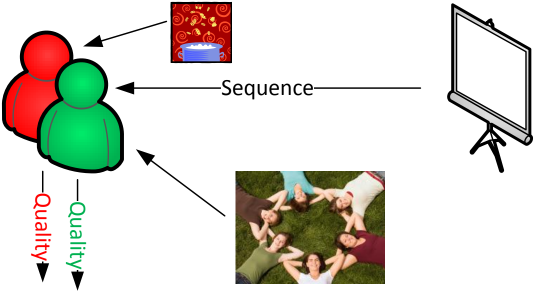


Significance test

- HDTV:
  - 2.69% increase in sensitivity (equivalent $\rightarrow$ different)
  - 0.08% decrease in sensitivity (different $\rightarrow$ equivalent)
- NTIA/Verizon:

H.264	MPEG-2	MOS stat. iden. original data	MOS stat. iden. subject bias removed
10 Mbps	18 Mbps	8 of 8	7 of 8
6 Mbps	$12\frac{1}{2}$ Mbps	7 of 8	5 of 8
$3\frac{1}{2}$ Mbps	$8\frac{1}{2}$ Mbps	4 of 8	4 of 8
2 Mbps	6 Mbps	3 of 8	3 of 8

Error Structure



$$o_{ij} = \psi_j + \Delta_i + \epsilon_{ij}$$

Error Structure

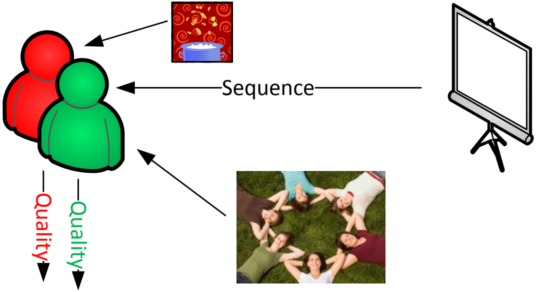
$$\epsilon_{ijr} = \alpha_i \beta_j X$$

$$\epsilon_{ijr} = \alpha_i X + \beta_j Y$$

Error Structure

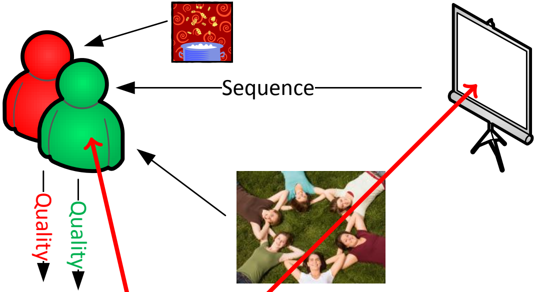
$$\epsilon_{ijr} = \alpha_i X + \beta_j Y$$

Error Structure



$$o_{ij} = \psi_j + \Delta_i + \alpha_i X + \beta_j Y \text{ where } X, Y \sim N(0,1)$$

Error Structure



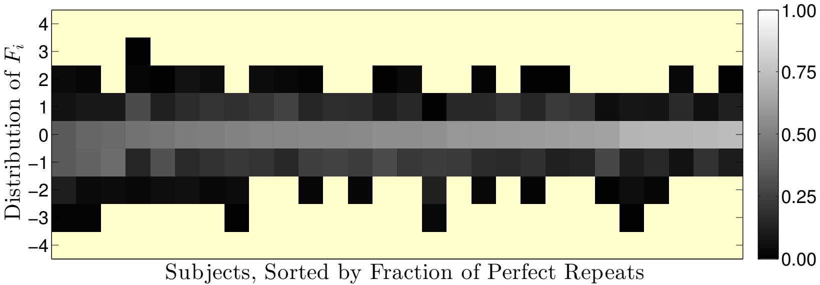
$$o_{ij} = \psi_j + \Delta_i + \alpha_i X + \beta_j Y \text{ where } X, Y \sim N(0,1)$$

Proof

?

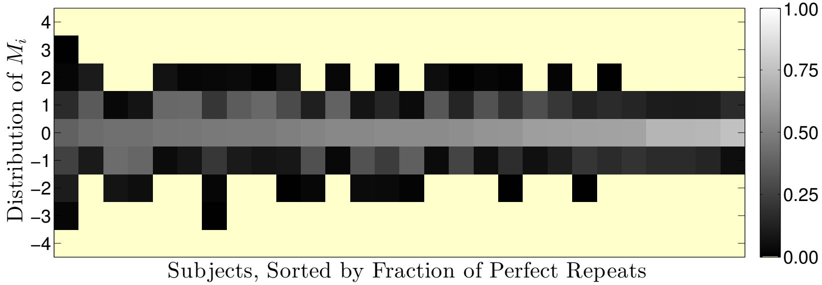
Discrete Analysis

$$F_i = \bigcup_j \bigcup_{r=2}^{6} \{o_{ijr} - o_{ij1}\}$$

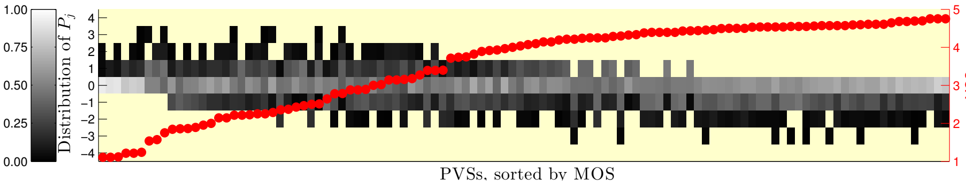


Discrete Analysis

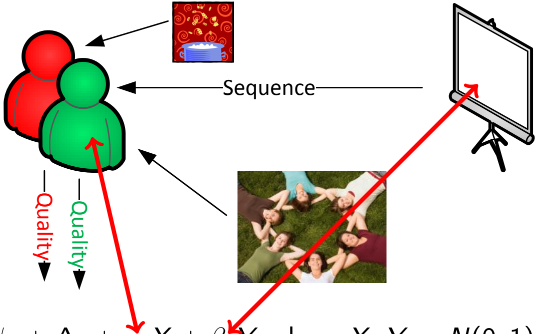
$$M_i = \bigcup_{j,r} \{o_{ijr} - m_{\psi_j} - m_{\Delta_i}\}$$



Discrete Analysis

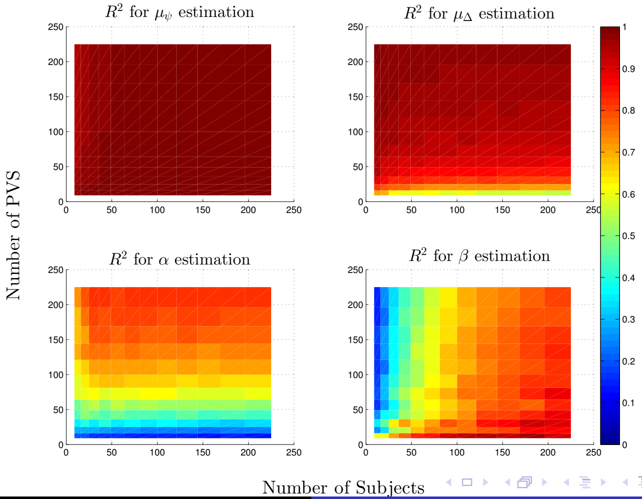


Continuous Model Analysis

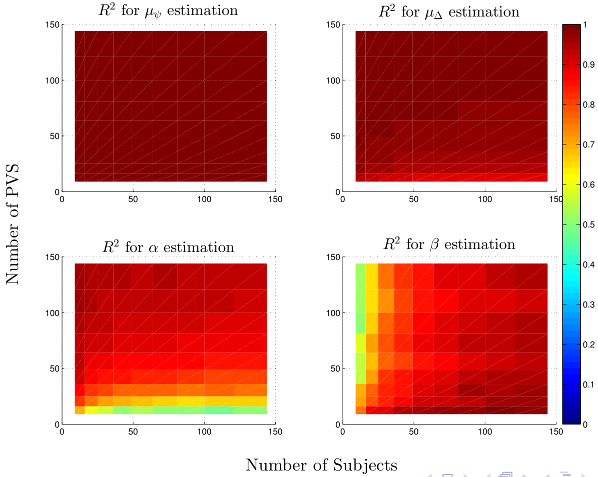


$$o_{ij} = \psi_j + \Delta_i + \alpha_i X + \beta_j Y \text{ where } X, Y \sim N(0, 1)$$

α and β Estimation Accuracy

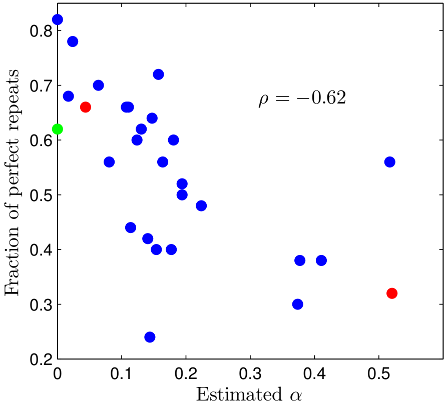


α and β Estimation Accuracy



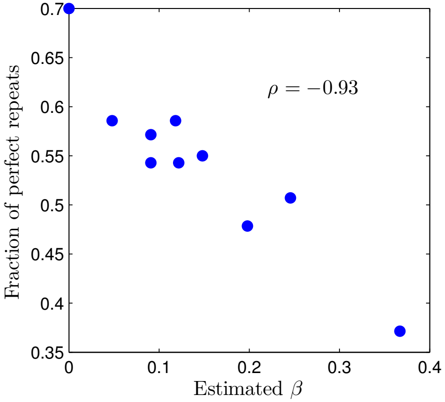
α Accuracy

#subjects = 28, #PVS = 10, #repetitions = 6.



β Accuracy

#subjects = 28, #PVS = 10, #repetitions = 6.



Conclusions

- When the analysis focuses on MOS comparisons, then Δ_i should be removed. The NTIA/Verizon dataset provides an example. Most subjective tests use this type of MOS analysis, and thus would benefit from removing Δ_i.
- When the analysis compares objective and/or subjective data with user descriptions (e.g., from blogs, forums, or questionnaires), then MOS and subject bias should be taken into consideration.
- When the analysis focuses on subject behavior, then the analaysis could focus only on Δ_i. The vqegMM2 dataset provides an example.

Conclusions

- Subject scoring is a random process:
  - accept it
  - do not try to fix it
- Expected error covers $\pm$ 25% of the scale
  - It could cover difference between 5 point ACR and continuous
  - It could explain difficult in detecting some factors like demographics
- We can use less subjects but with repeating ratings for each clip
- Subject model can be used to predict how many subjects is need to obtained statistically significant conclusion, e.g. analyzing demographics

Questions

More details can be found in:

L. Janowski and M. Pinson, "The Accuracy of Subjects in a Quality Experiment: A Theoretical Subject Model," in *IEEE Transactions on Multimedia, vol. 17, no. 12, pp. 2210-2224, Dec. 2015.*