(Not so) new findings about Transmission Rating scale and subjective scores

(Director's cut)

Pablo Pérez

QoMEX 2023

Financiado por la Unión Europea
NextGenerationEU

Plan de Recuperación, Transformación y Resiliencia

Do you know this equation?

$$MOS = 1 + 0.035R + R(R - 60)(100 - R) \times 7 \times 10^{-6}$$

ITU-T G.107 (12/98): The E-Model

A computational (=parametric) model for use in transmission planning (=telephony)

- E-model estimates QoE in a telephone service given some QoS values (noise, echo…)
- QoE is given in **Transmission Rating** scale $R \in [0,100]$

$$MOS = 1 + 0.035R + R(R-60)(100-R) \times 7 \times 10^{-6}$$

MOS=1 for R<0,

MOS=4.5 for R>100

Good-or-Better /

Poor-or-Worse is

$$GoB = E\left(\frac{R-60}{16}\right);\ PoW = E\left(\frac{45-R}{16}\right)$$

$$E(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

ITU-T G.107 (12/98): The E-Model

A computational (=parametric) model for use in transmission planning (=telephony)

- E-model estimates QoE in a telephone service given some QoS values (noise, echo…)
- QoE is given in **Transmission Rating** scale $R \in [0,100]$ Why?

$$MOS = 1 + 0.035R + R(R-60)(100-R) \times 7 \times 10^{-6}$$

MOS=1 for R<0,

MOS=4.5 for R>100

Why?

Good-or-Better / Poor-or-Worse is

$$GoB = E\left(\frac{R-60}{16}\right);\ PoW = E\left(\frac{45-R}{16}\right)$$

$$E(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

NOKIA

ITU-T G.107 (12/98): The E-Model

A computational (=parametric) model for use in transmission planning (=telephony)

- E-model estimates QoE in a telephone service given some QoS values (noise, echo…)
- QoE is given in **Transmission Rating** scale $R \in [0,100]$

Why?

$$MOS = 1 + \overset{0.04}{\cancel{0.035}}R + R(R-60)(100-R)\times 7\times 10^{-6}$$

MOS=1 for R<0,

MOS=~~4.5~~ 5 for R>100

Why?

Good-or-Better / Poor-or-Worse is

$$GoB = E\left(\frac{R-60}{16}\right);\; PoW = E\left(\frac{45-R}{16}\right)$$

• QoE is given in Transmission Rating scale R ∈ [0,100]

NOKIA

What did we know about Transmission Rating scale?

- Originally designed for narrowband (NB) voice
- When updating G.107 to wideband (WB) and super-wideband (SWB) voice, it was discovered that it was enough to extend the scale from 100 to 129 (WB) or 179 (SWB)

  Möller et al. (2010). Towards a universal scale for perceptual value. *QoMEX 2010*

- It works for Online Gaming models too… but not for video

  Hoßfeld et al. (2016). QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS. *Quality and User Experience*

- It was originally proposed in the Bellcore model

  Cavanaugh et al. (1976). Models for the subjective effects of loss, noise, and talker echo on telephone connections. *Bell System Technical Journal*

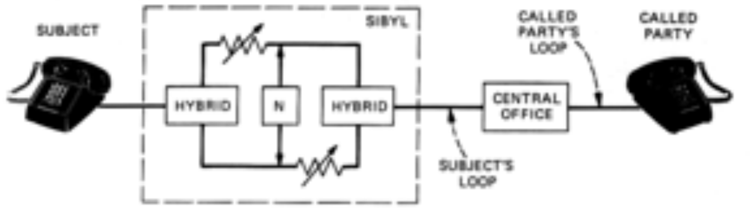
NOKIA

Subjective scores in the Bellcore Model

1. Normal distribution model

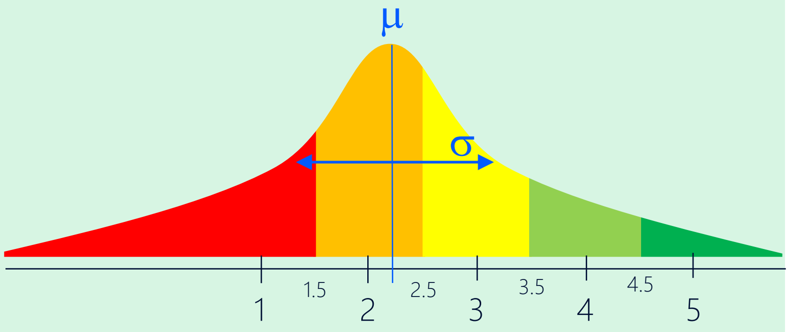
Experiment

- Phone calls are disturbed w/ attenuation & noise (HRC)
- After the call, users rate in 1-5 scale (~ACR)



Data Processing

- Scores from each HRC come from a continuous $\mathcal{N}(\mu, \sigma)$
- $MOS = MOS_Q = \sum_{i=1}^{5} i\hat{P}_i = 5 - \sum_{j=1}^{4} E\left(\frac{j+0.5-\mu}{\sigma}\right)$
- $SOS = SOS_Q$



Cavanaugh et al. (1976). Models for the subjective effects of loss, noise, and talker echo on telephone connections. *Bell System Technical Journal*

Subjective scores in the Bellcore Model

2. Constant variance within an experiment

- Bellcore model provides a single σ estimate for each experiment
- As a consequence, for each condition, SOS only depends on μ (i.e. on the MOS)

$$SOS^2 = 25 - \sum_{j=1}^{4}\left[(2j+1)E\left(\frac{j+0.5-\mu}{\sigma}\right)\right] - (MOS)^2$$

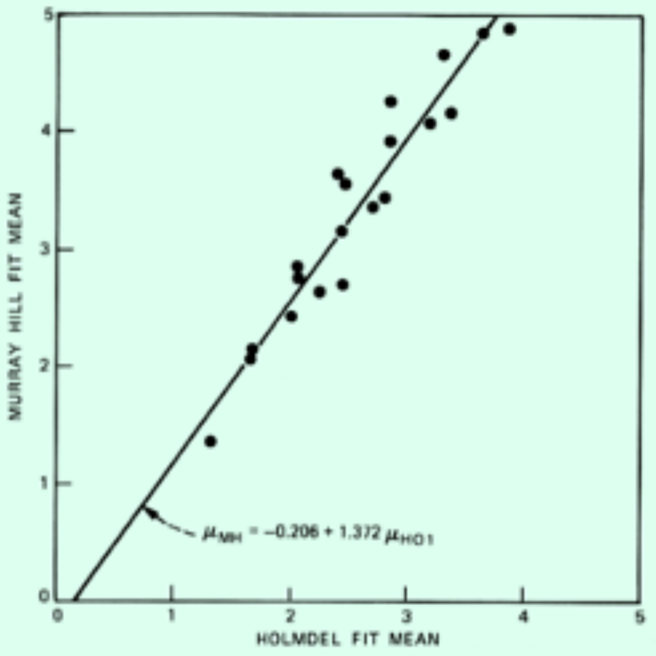
Subjective scores in the Bellcore Model

3. Comparing the results in two judgement conditions

- The same experiment done in two different years and locations yielded different results.
- However, both experiments could be linearly fitted in μ domain

- This also happened when comparing scores for narrowband (NB) and wideband (WB) telephony models.

Möller et al. (2010). Towards a universal scale for perceptual value. *QoMEX 2010*



Subjective scores in the Bellcore Model

4. The Transmission Rating scale

- To eliminate the need to have different equations for each judgement condition, a general transmission-rating scale is established
- R is a linear transformation of $\mu \rightarrow \mu = aR+b$
- R=40 and R=80 are selected for specific conditions of attenuation and noise.
- Scale is arbitrary! A **transmission-rating scale in attenuation dBs** was already used in Bellcore at that time
- When Bellcore model is proposed as input to ITU-T E-Model, a reference test condition is given

$$\mu(R) = \frac{R}{15} - 0.5; \ \sigma = \frac{16}{15}$$

ITU-T G.107 (12/98): The E-Model

A computational (=parametric) model for use in transmission planning (=telephony)

- E-model estimates QoE in a telephone service given some QoS values (noise, echo…)
- QoE is given in **Transmission Rating** scale $R \in [0,100]$ Why?

$$MOS = 1 + \sout{0.035}\ 0.04\, R + R(R-60)(100-R)\times 7 \times 10^{-6}$$

MOS=1 for R<0,

MOS=~~4.5~~ 5 for R>100

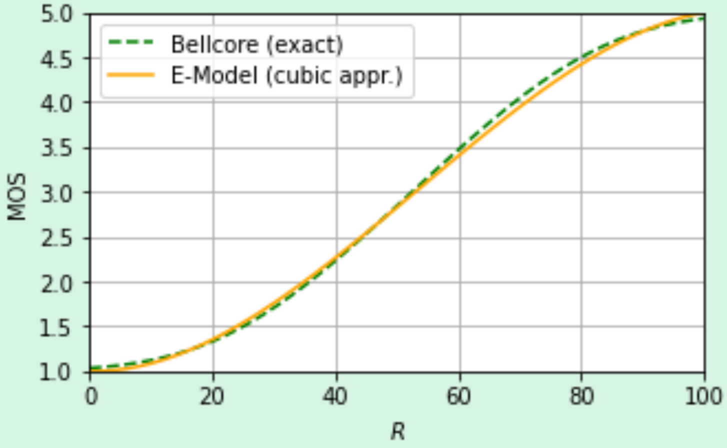
Why?

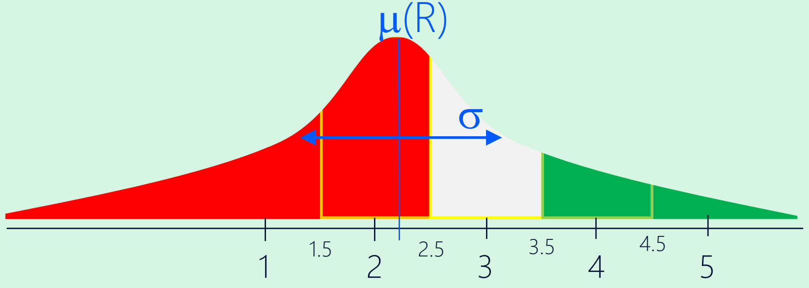
Good-or-Better / Poor-or-Worse is

$$GoB = E\left(\frac{R-60}{16}\right);\ PoW = E\left(\frac{45-R}{16}\right)$$

$$E(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

From Bellcore to the E-Model





$$GoB = E\left(\frac{\mu(R) - 3.5}{\sigma}\right);\ PoW = E\left(\frac{2.5 - \mu(R)}{\sigma}\right)$$

$$MOS = MOS_Q = \sum_{i=1}^{5} i\hat{P}_i = 5 - \sum_{j=1}^{4} E\left(\frac{j + 0.5 - \mu}{\sigma}\right)$$

$$\mu(R) = \frac{R}{15} - 0.5;\ \sigma = \frac{16}{15}$$

0.04

$$MOS = 1 + 0.035R + R(R-60)(100-R) \times 7 \times 10^{-6}$$

$$GoB = E\left(\frac{R-60}{16}\right);\ PoW = E\left(\frac{45-R}{16}\right)$$

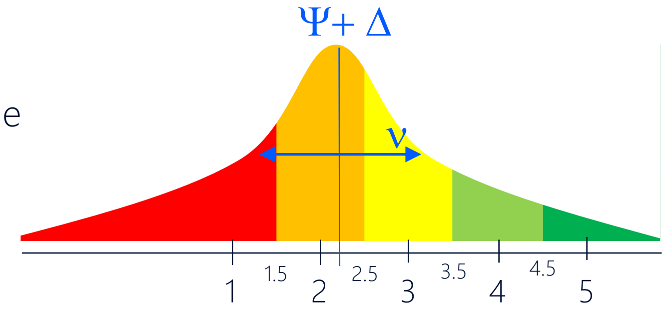
It is compatible with SoA subjective score models

1. Normal distribution model

- Subjective scores can be modeled as realizations of a random variable

$$U = \psi + \mathcal{N}(\Delta, \nu)$$

  - ψ ($=\mu$) = true quality; (Δ, ν) = subject bias and inconsistency
  - (Per-condition distribution is normal)



Li et al. (2020). A simple model for subject behavior in subjective experiments. *Electronic Imaging*

- Normal-based data model is better than the empirical distribution of scores to bootstrap subjective scores
  → to estimate QoE distribution statistics (such as GoB/PoW)

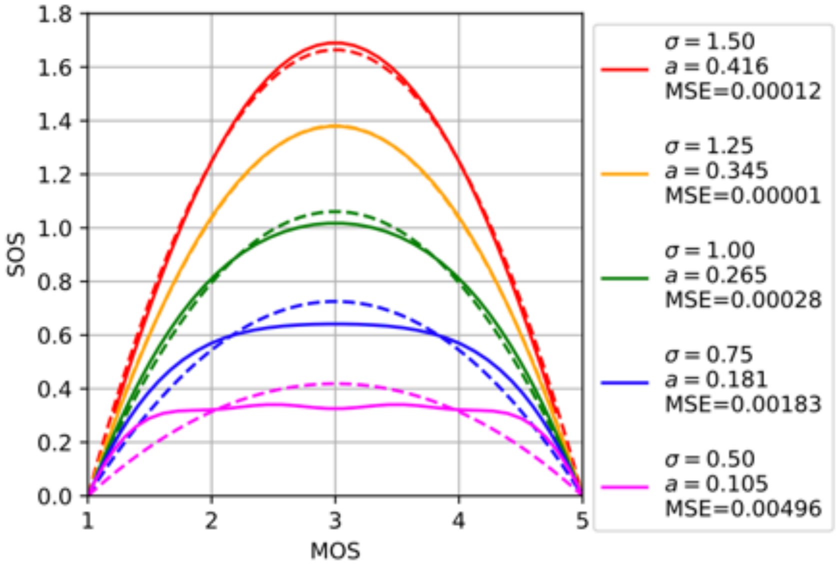
Nawala et al. (2022). Generalized score distribution: A two-parameter discrete distribution accurately describing responses from quality of experience subjective experiments. *IEEE Tr. Multimedia*

It follows the SOS hypothesis!

2. Constant variance within an experiment = SOS only depends on MOS

$$SOS^2 = a\left(-MOS^2 + 6MOS - 5\right)$$

Hoßfeld et al. (2011). SOS: The MOS is not enough! *QoMEX 2011*



- Both models (Hoßfeld *a*, Bellcore σ) provide similar results

$$\sigma = 3.01a + 0.2$$

NOKIA

It works for video under two different screen sizes!

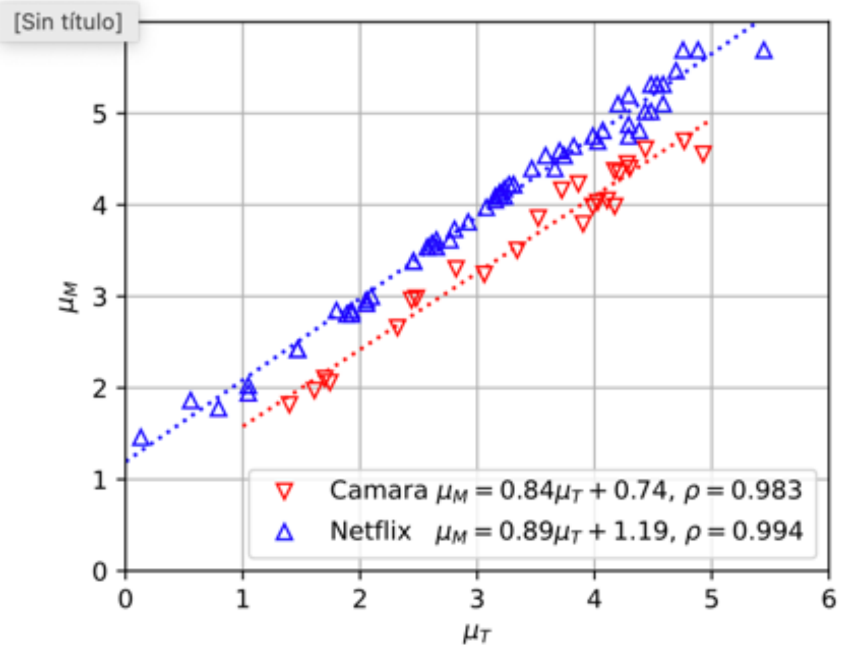
3. Comparing the results in two judgement conditions

- Two video datasets where the same content is evaluated i
  mobile vs tablet/latptop screen

  Cámara et al. (2019). Perceptually equivalent resolution in handheld devices for streaming bandwidth saving. *IEEE Signal Proc. Letters*

  VMAF (https://github.com/Netflix/vmaf), modified

- I computed μ (per condition) and σ (per experiment).
- There is a linear mapping between mobile and tablet results (ρ > 0.98)!



The transmission rating scale can be generalized

Towards a universal scale for perceptual value (again!)

CONDITIONS

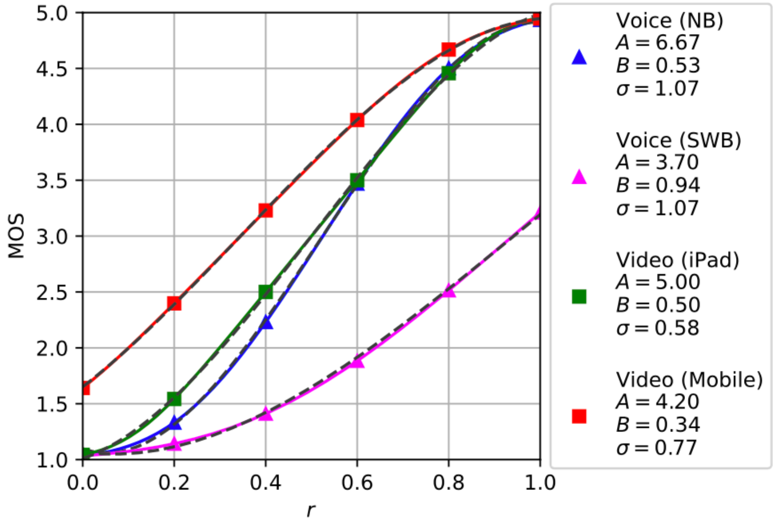
1. Scores are normally-distributed
2. Variance is constant within experiment (= SOS hypothesis)
3. Two experiments can be linearly mapped in μ-scale (works for speech and, apparently, video)

PROPERTIES

Scale is arbitrary (e.g. $r \in [0,1]$)

Cubic approximation works, but we need more parameters to cover all use cases

$$\mu(R) = A(R - B) + 3$$



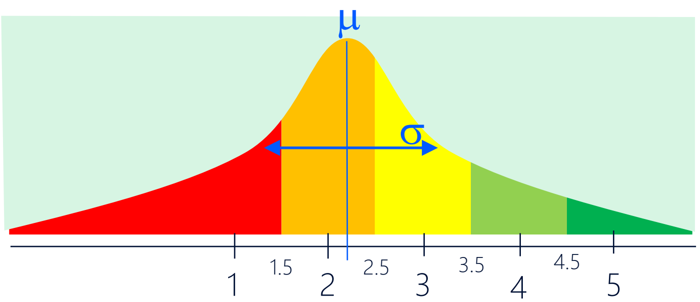
$$MOS = M_m + (M_M - M_m)\frac{R}{100}\left(1 + S\frac{(R-X)}{100}\frac{(100-R)}{100}\right)$$

More connections

We have been using Transmission Rating scales in the past (unadvertedly?)

What we have learnt today

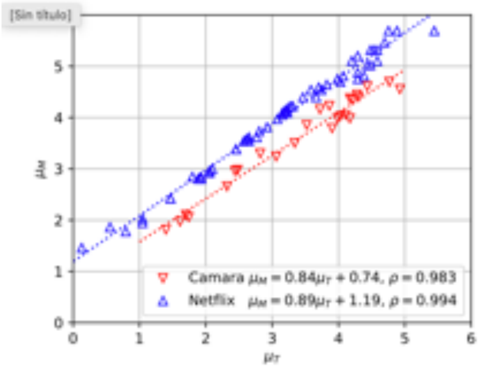
- A TR scale is just representing opinion scores by (μ, σ) instead of MOS (+SOS)
- MOS = f(μ) can be approximated by a cubic function
- TR scale seems a good way to aggregate results of two subjective experiments (more linear than MOS)



What we already knew

- Objective scores (e.g. PSNR) have better fit to MOS if mapped with a cubic function (VQEG HDTV projects)
- Objective scores have been proposed as intermediate scale to merge different subjective experiments

Pinson & Wolf (2003) An objective method for combining multiple subjective data sets



NOKIA

We should use Transmission Rating scales to define QoE (or QoMS)

…if we can confirm that video QoE satisfies the 3 conditions

BENEFITS

1. (M)OS (i.e. ACR scale) is not a property of the signal/service, but of the experiment!
    1. In particular, it depends on the range of qualities shown in the experiment
2. We could compare / aggregate / extend the results obtained in one experiment without "touching" them (e.g. extend HD quality to 4K, extend SDR to HDR)
3. The (truncated) normal representation of scores provides information about distribution of scores (not only mean) and better bootstrapping properties than the empirical distribution

CONDITIONS

1. Scores are normally-distributed
2. Variance is constant within experiment (= SOS hypothesis)
3. Two experiments can be linearly mapped in μ-scale (works for speech and, apparently, video)

IMPLICATIONS

1. For each experiment, recover (μ, σ) instead of MOS
2. μ can be <1 or >5 !!

We should use ~~Transmission Rating~~ scales to define QoE (or QoMS)

…regardless video QoE satisfies the 3 conditions or not!!

1. (M)OS (i.e. ACR scale) is not a property of the signal/service, but of the experiment!
    1. In particular, it depends on the range of qualities shown in the experiment
2. Transmission Rating was defined as an arbitrary scale that was understood by the relevant stakeholders (i.e. the rest of AT&T), as it had been used as "QoE measure" since 1930s
    1. Simply by providing anchoring points to some conditions
3. We should define QoE in **arbitrary units which are understood outside our community**
    1. E.g. **PIXELS** (HD quality vs 4K quality)
    2. Or pixels per second (to include frame rate). Or pps x bits/pixel = bits per second (to include HDR)
4. Remember that TR was defined in "SNR dBs" for narrowband speech
    1. To handle speech bandwidth, we could add Hz * dB = bits per second!
5. (Intuitively we would be defining QoE as the amount of "effective information" that the network is able to communicate)

Conclusions

What a journey! Would you join me in the next steps?

1. Transmission Rating scale can model perceptual quality, independently of test context (e.g. screen size!)
2. It worked great for speech. It can work for other use cases!
3. Many ideas were already proposed in a 1976 paper!
4. Still a lot of work to do

Do you have any question?

Do you have suggestions for the next steps? (See you in the ☕)

GOBIERNO DE ESPAÑA
MINISTERIO DE ASUNTOS ECONÓMICOS Y TRANSFORMACIÓN DIGITAL
SECRETARÍA DE ESTADO DE TELECOMUNICACIONES E INFRAESTRUCTURAS DIGITALES
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia

NOKIA

NOKIA