Experiment Designs to Avoid Scene Reuse

Lucjan Janowski (AGH), Margaret Pinson (NTIA),
Ludovic Malfait (Dolby)

Outline

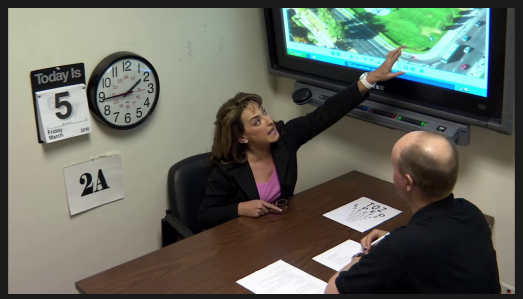
Problem Statement

- I cannot show subjects the same scene twice.
  How will this impact my experiment?
- I like the full matrix design (SRC x HRC)
  Why should I change?

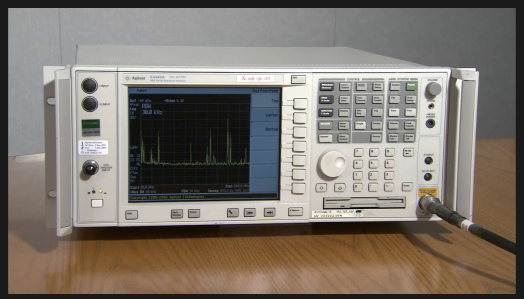
Traditional Experiment Design (SRC)

	SRC1	SRC2	SRC3
HRC1			
HRC2			
HRC3			

Related Sequence Design (RSRC)

	RSRC1	RSRC2	RSRC3
HRC1			
HRC2			
HRC3			

Coding Difficulty Sequence Design (CD-SRC)

	CD-SRC1	CD-SRC2	CD-SRC3
HRC1			
HRC2			
HRC3			

Outline

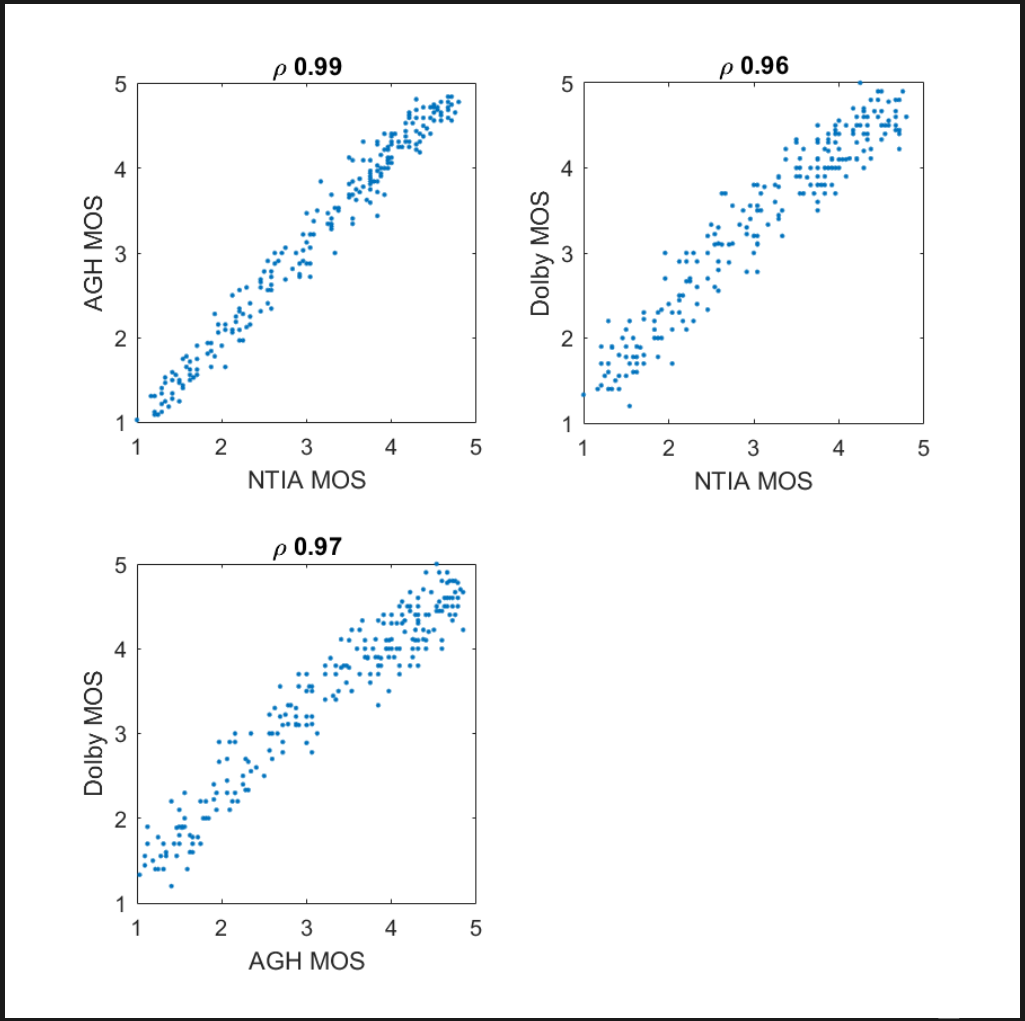
AGH/NTIA Dataset

- Preliminary
- Subjects: NTIA
- Designed to answer a different question
- Each HRC has constant quality

Novel Dataset

- Subjects: AGH (32), NTIA (24), and Dolby (11 $\rightarrow$ 24)
- MPEG-2, H.264, H.265
- Low, medium, high bitrate
- Sessions
  1. SRC
  2. RSRC
  3. CD-SRC
  4. RSRC
  5. CD-SRC
  6. Random

Lab to Lab Novel



Subject Screening

Typical screening rules ← We did this

- 0.75 Pearson correlation threshold
- 1 subject

Screen each session separately ← Not this

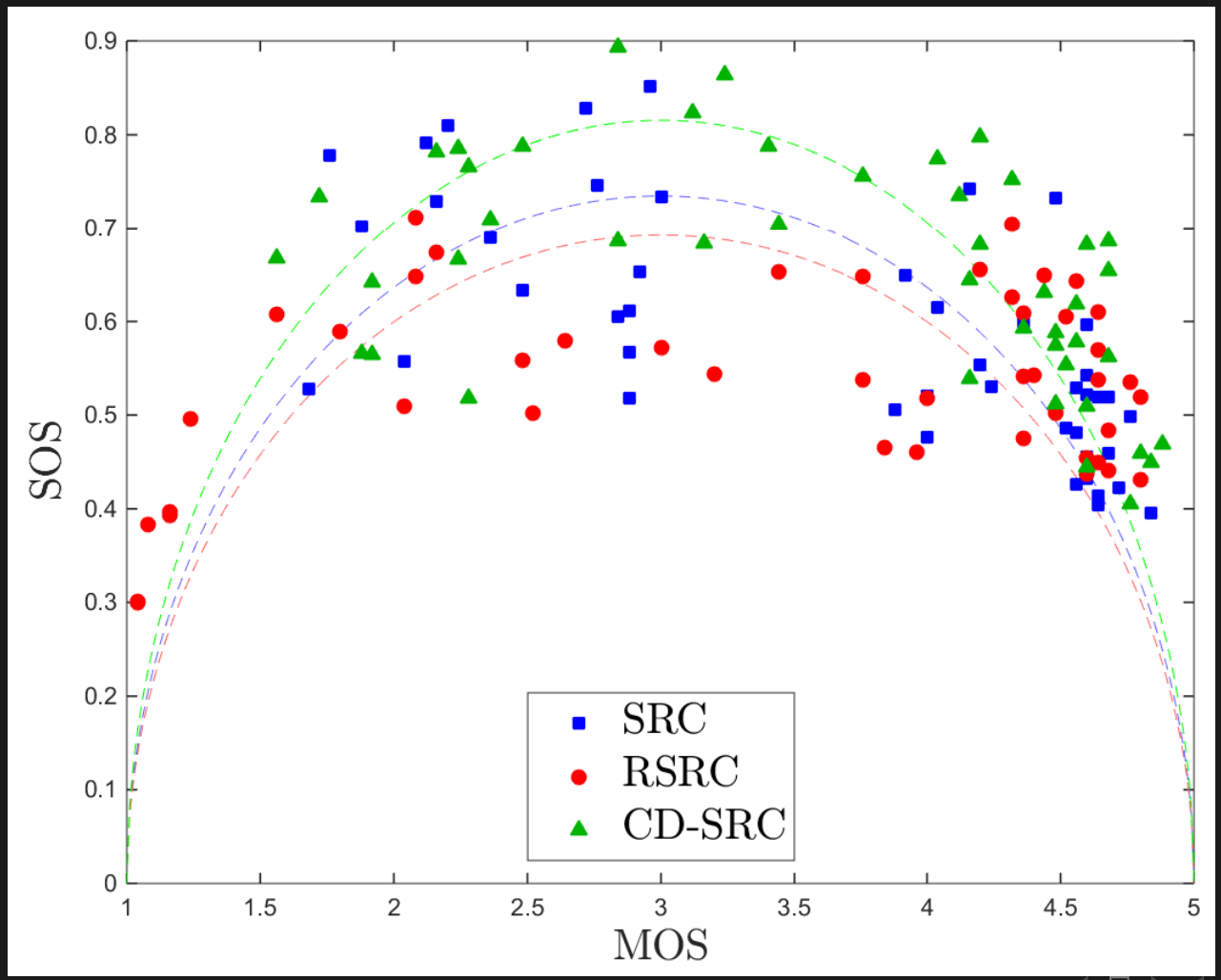
- 1 subjects for SRC
- 2 subjects for RSRC1
- 5 subjects for CD-SRC1
- 2 subjects for RSRC2
- 7 subjects for CD-SRC2
- 5 subjects for Random

Mean and Stdev of Sessions
Novel

	SRC	RSRC1	MOS CD1	RSRC2	CD2	Rand
NTIA	3.00	3.35	3.07	3.18	3.10	3.14
AGH	3.06	3.40	3.13	3.30	3.25	3.33
Dolby	3.27	3.53	3.37	3.48	3.35	3.50

	SRC	RSRC1	SOS CD1	RSRC2	CD2	Rand
NTIA	0.72	0.73	0.77	0.70	0.79	0.77
AGH	0.69	0.66	0.68	0.70	0.70	0.72
Dolby	0.77	0.67	0.85	0.76	0.80	0.81

Hoßfeld, Schatz and Egger (HSE) Coefficient (HSE) AGH/NTIA



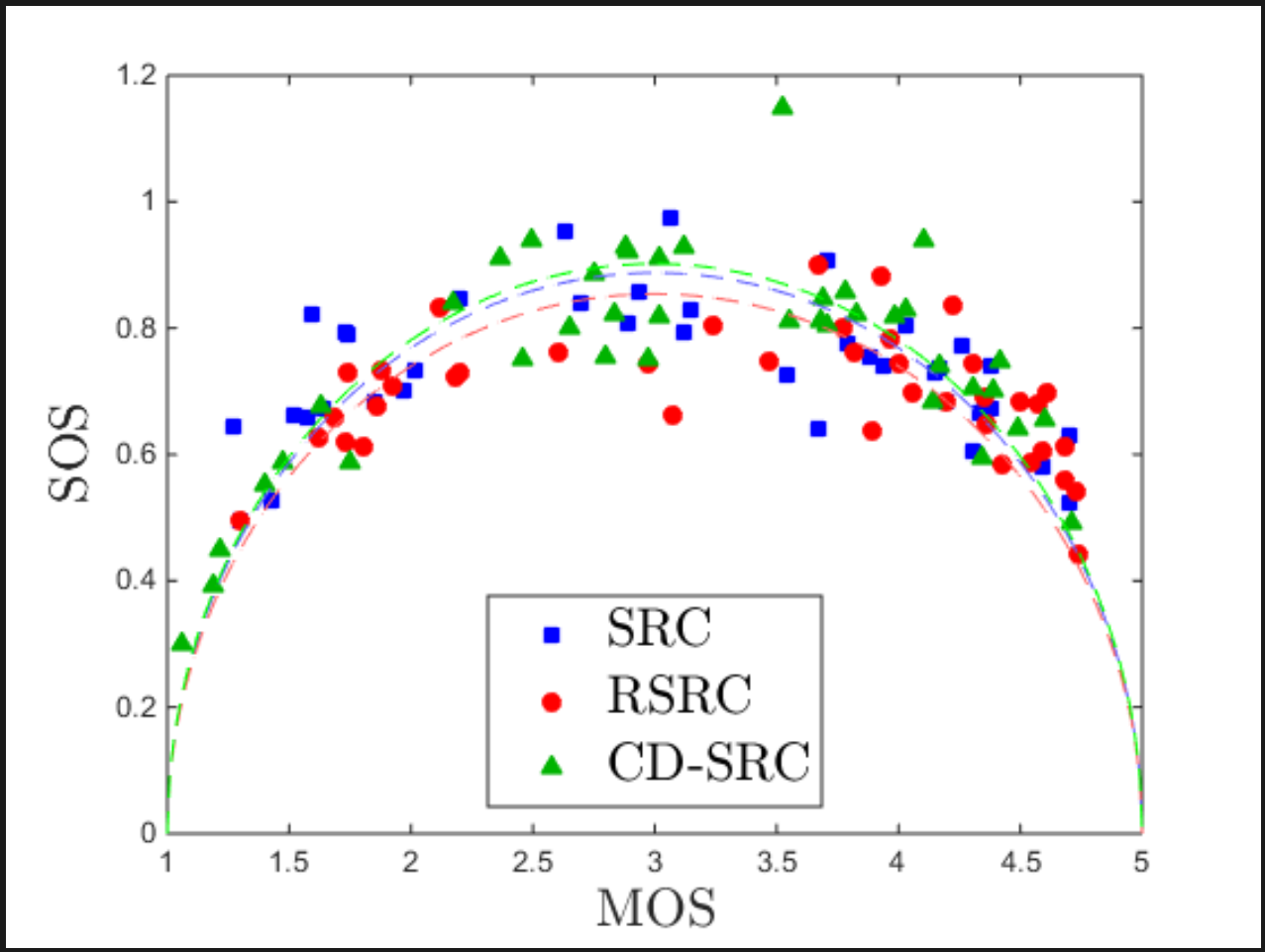
HSE for:

SRC = 0.143

RSRC = 0.138

CD-SRC = 0.161

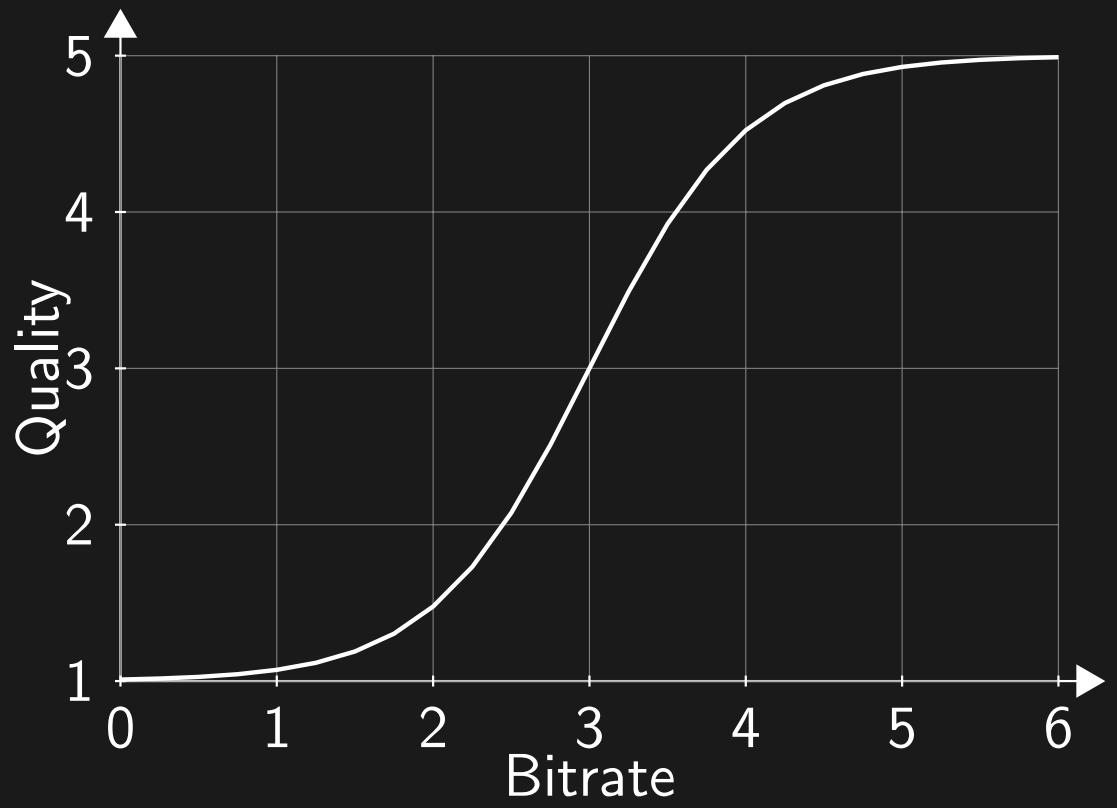
Hoßfeld, Schatz and Egger (HSE) Coefficient (HSE) Novel



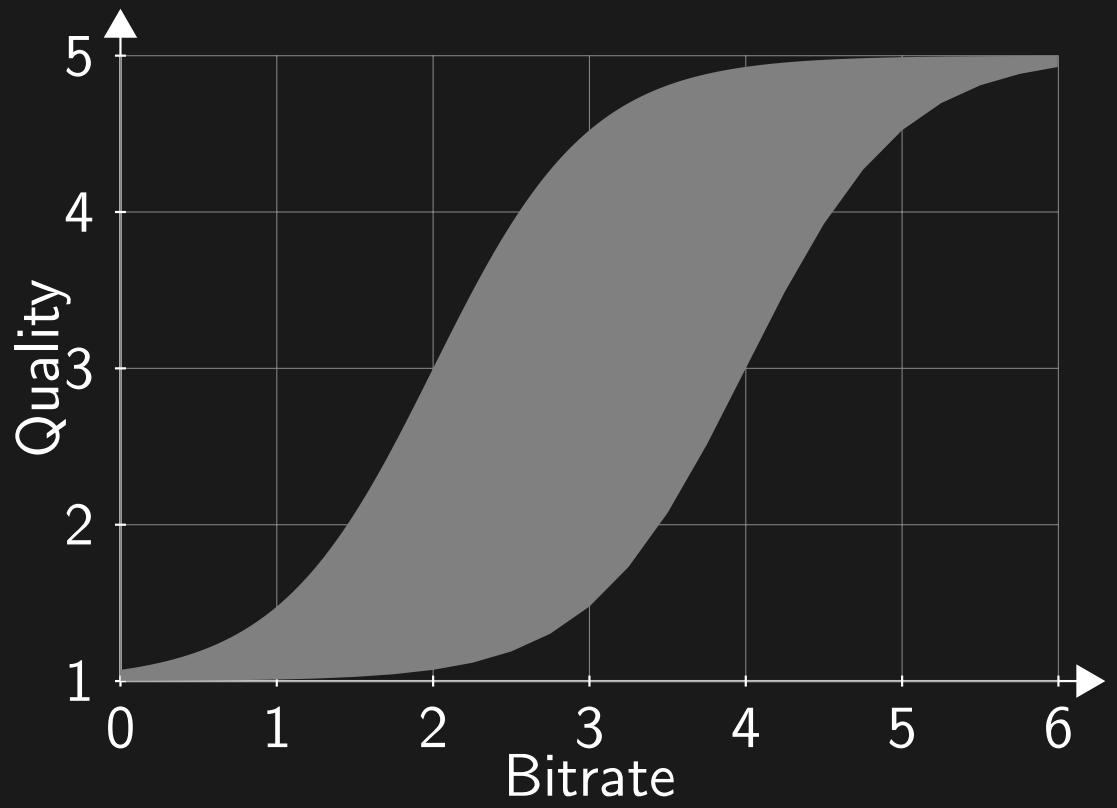
HSE for:

SRC	= 0.197
RSRC	= 0.182
CD-SRC	= 0.203

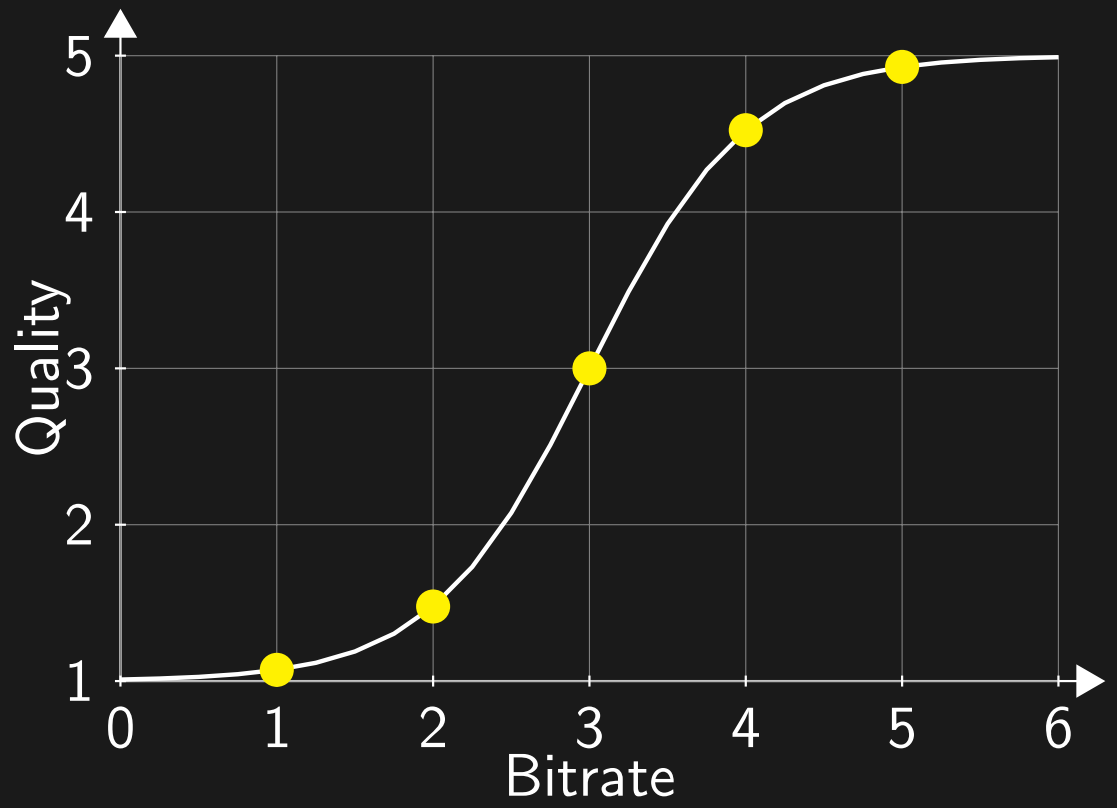
Theoretical Relationship MOS & Bitrate



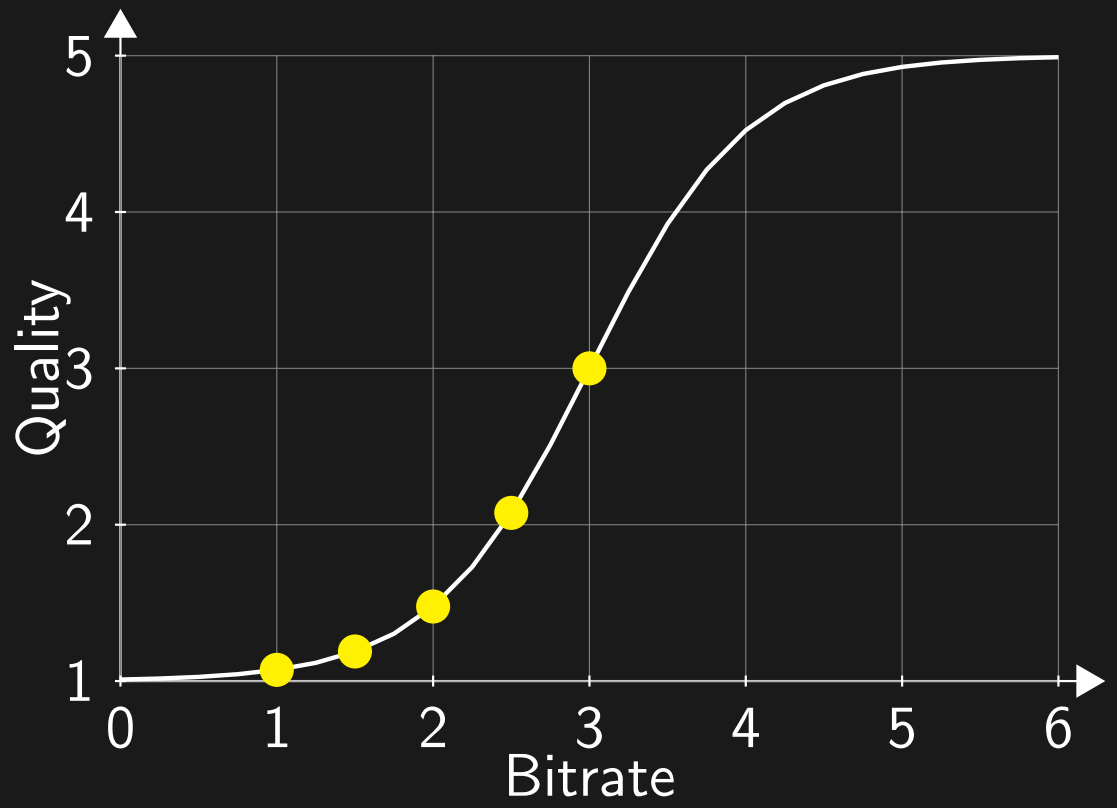
Theoretical Relationship MOS & Bitrate



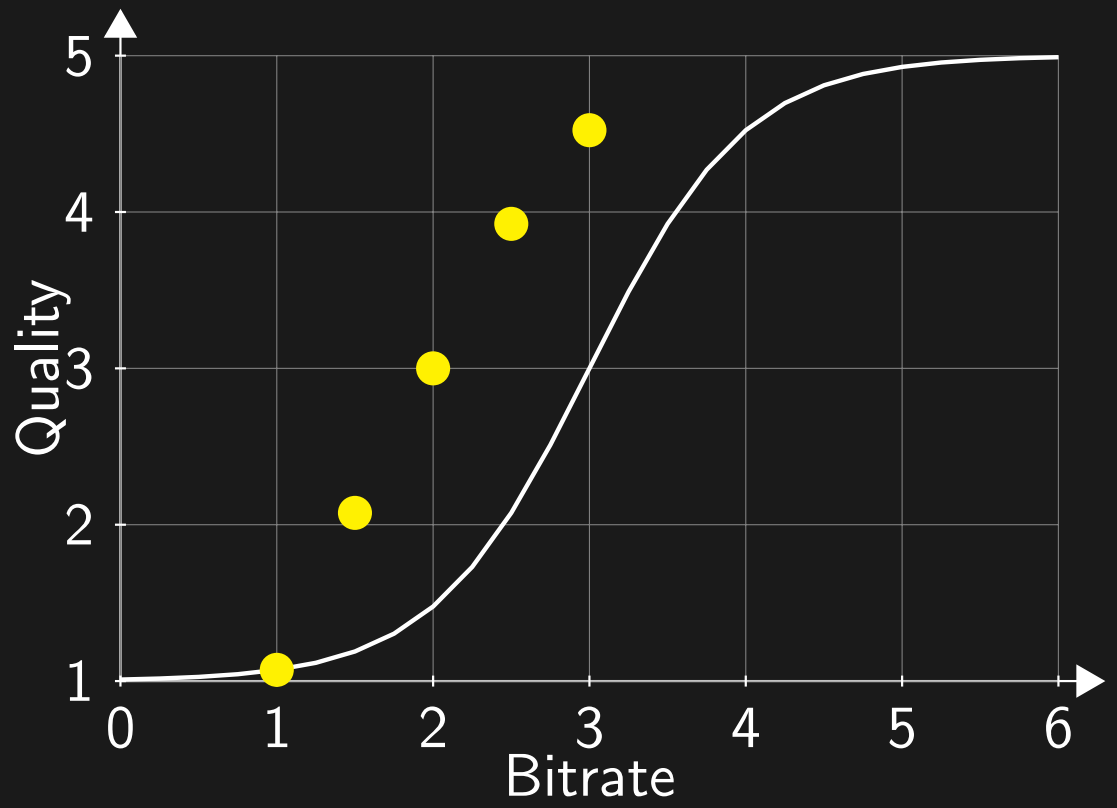
Theoretical Relationship MOS & Bitrate



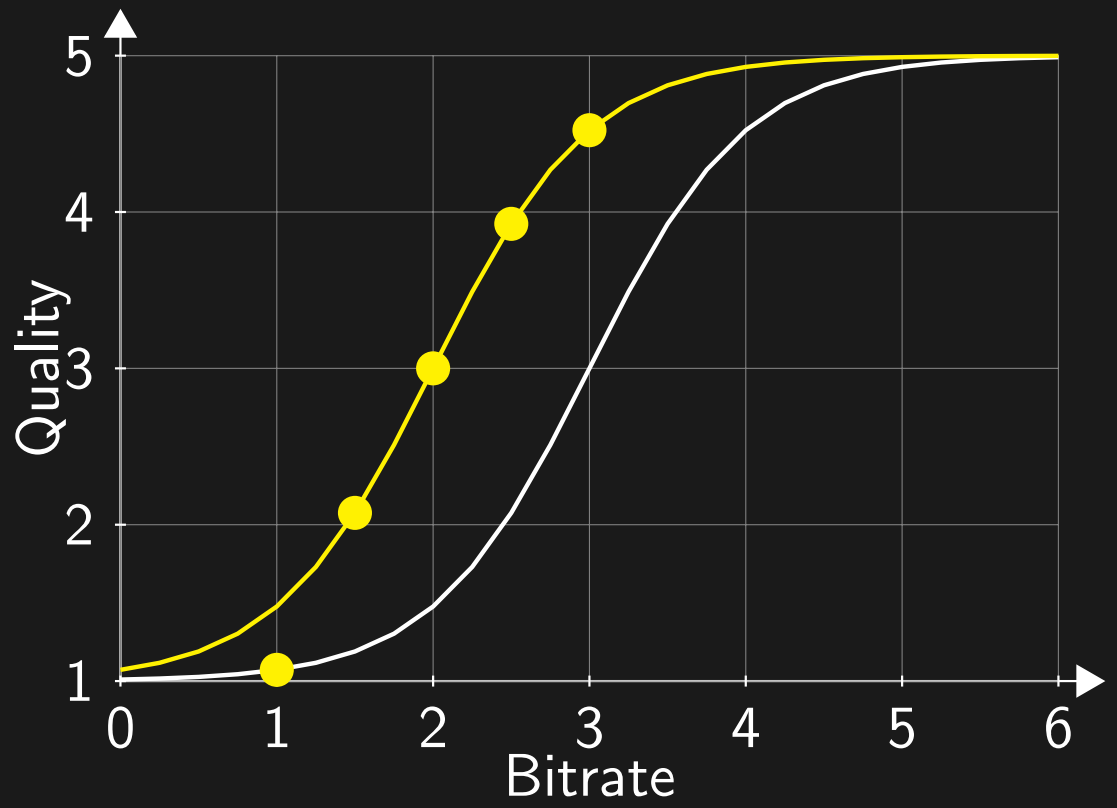
Theoretical Relationship MOS & Bitrate



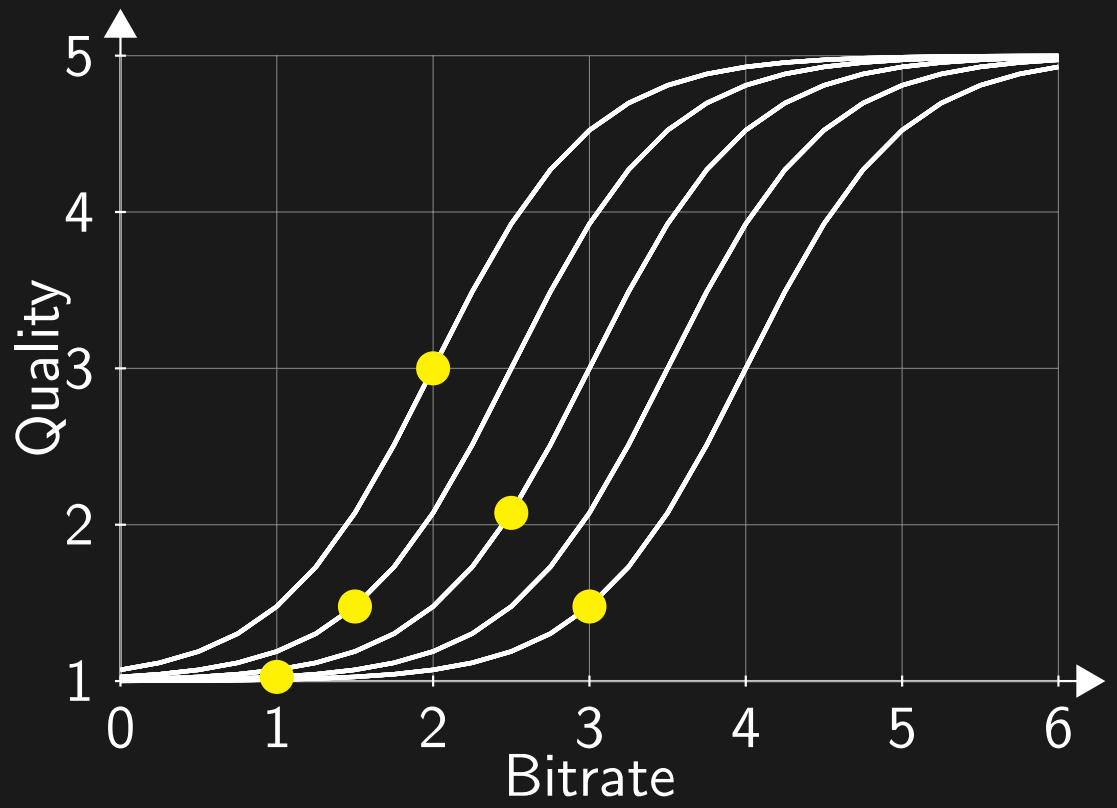
Theoretical Relationship MOS & Bitrate



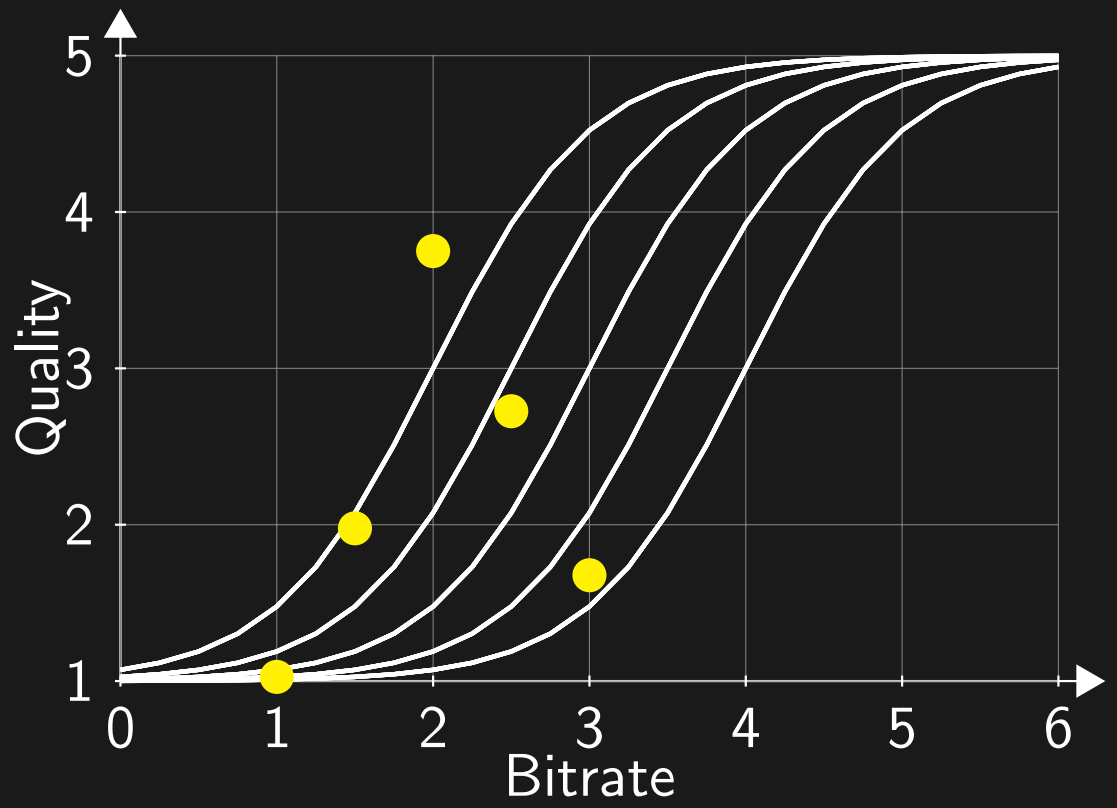
Theoretical Relationship MOS & Bitrate



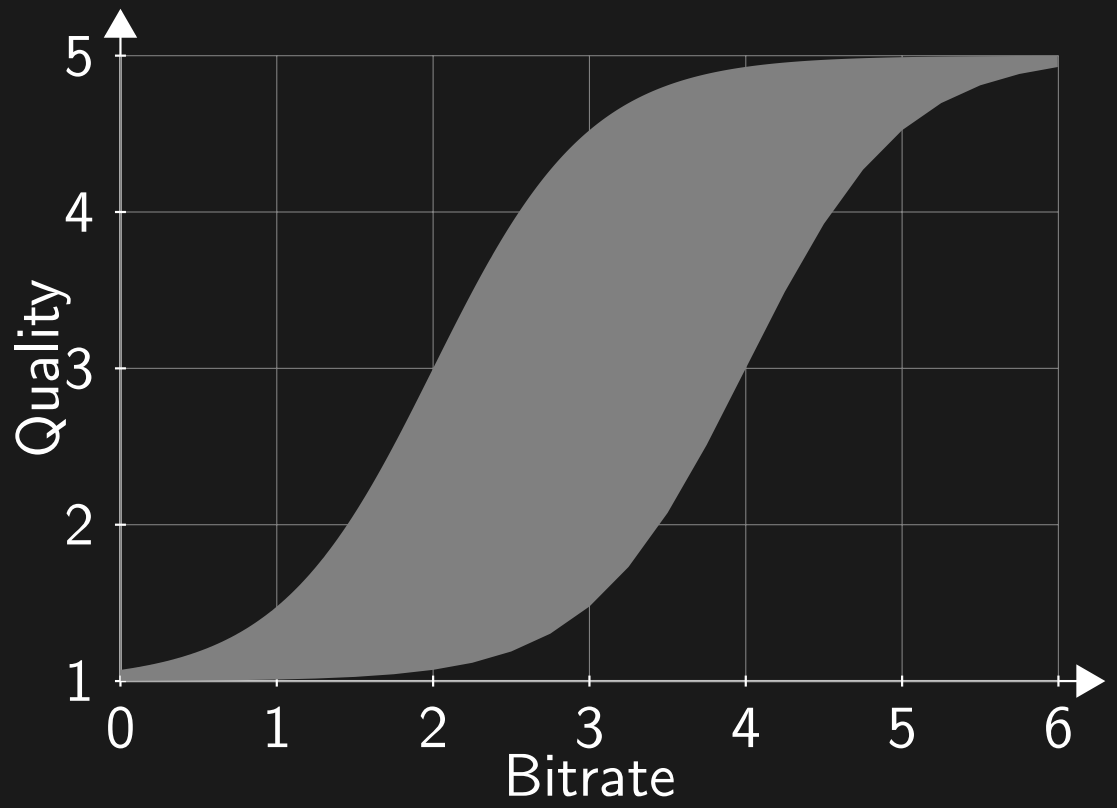
Theoretical Relationship MOS & Bitrate



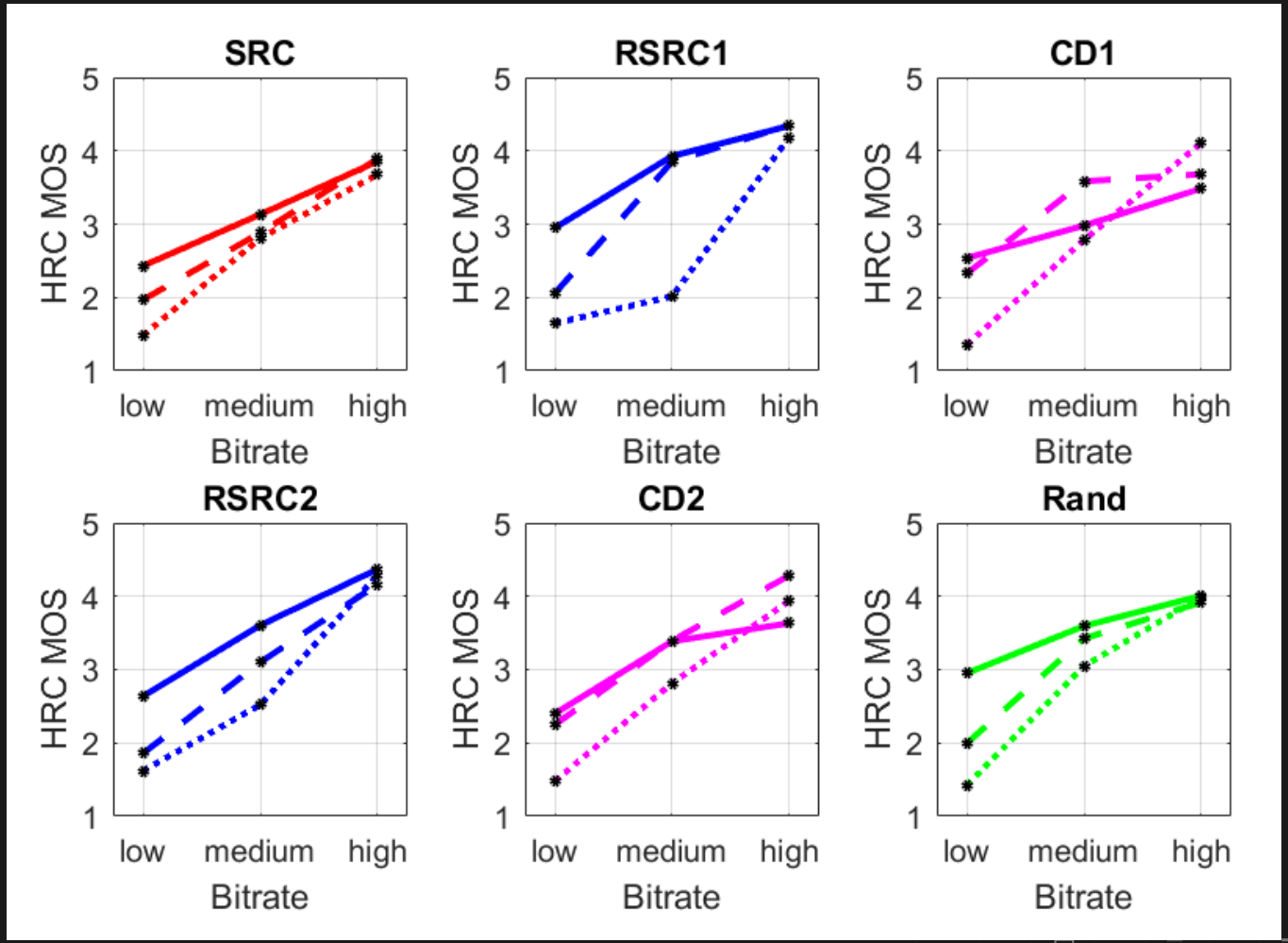
Theoretical Relationship MOS & Bitrate



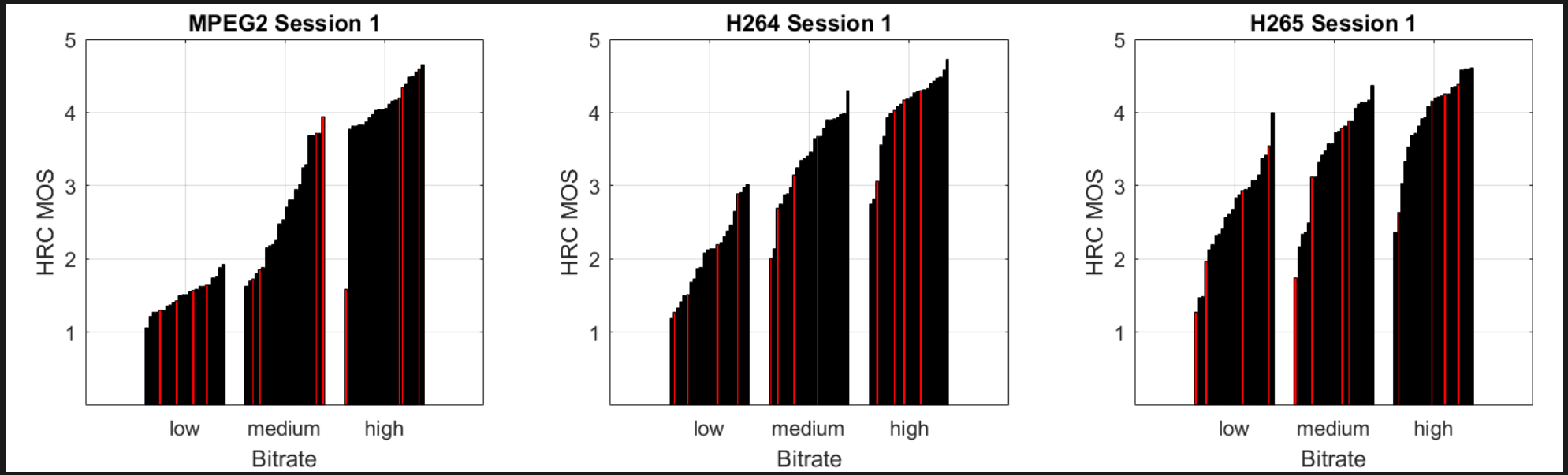
Theoretical Relationship MOS & Bitrate



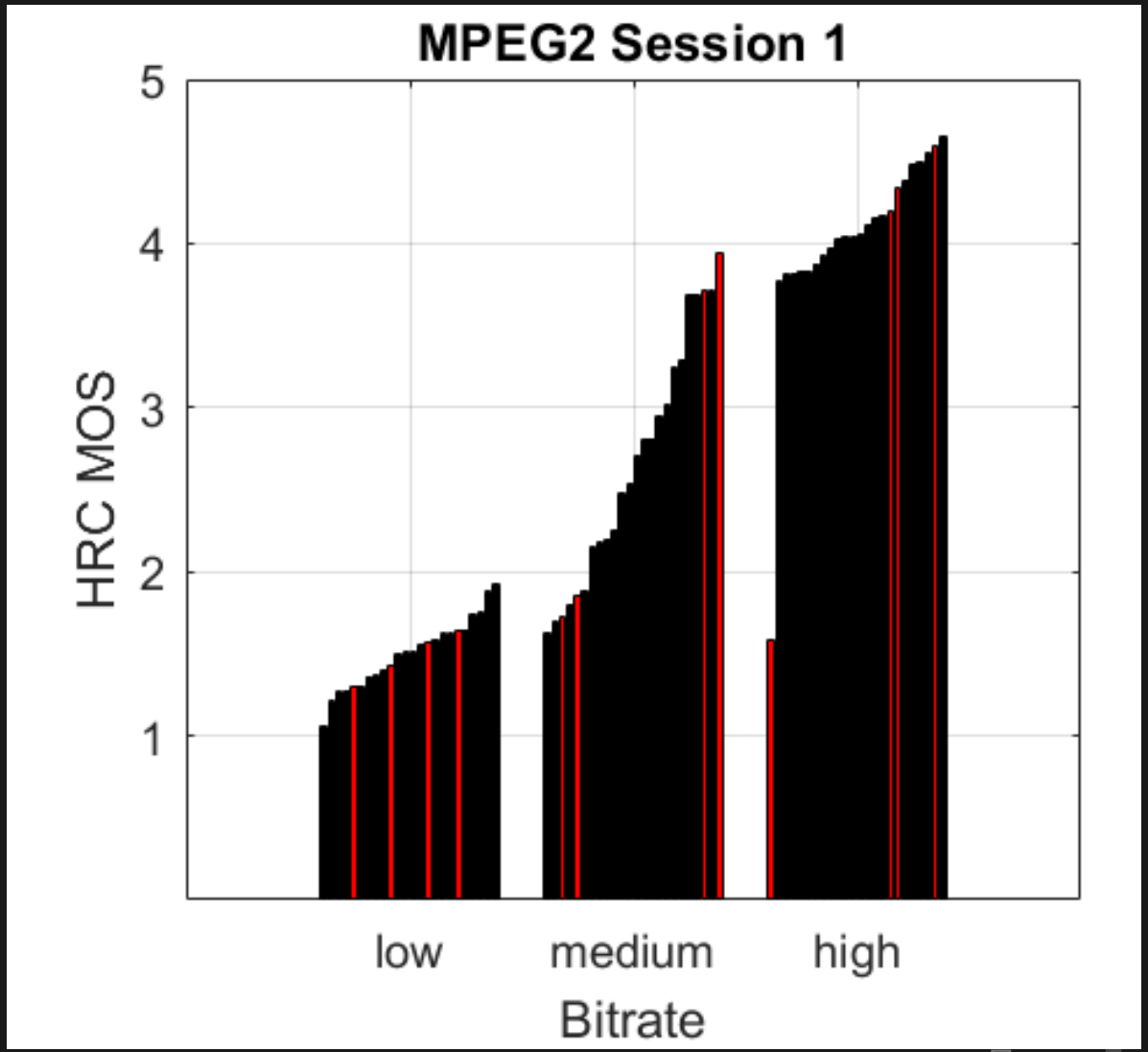
Relationship between Codecs & Bitrate Which Is Correct?



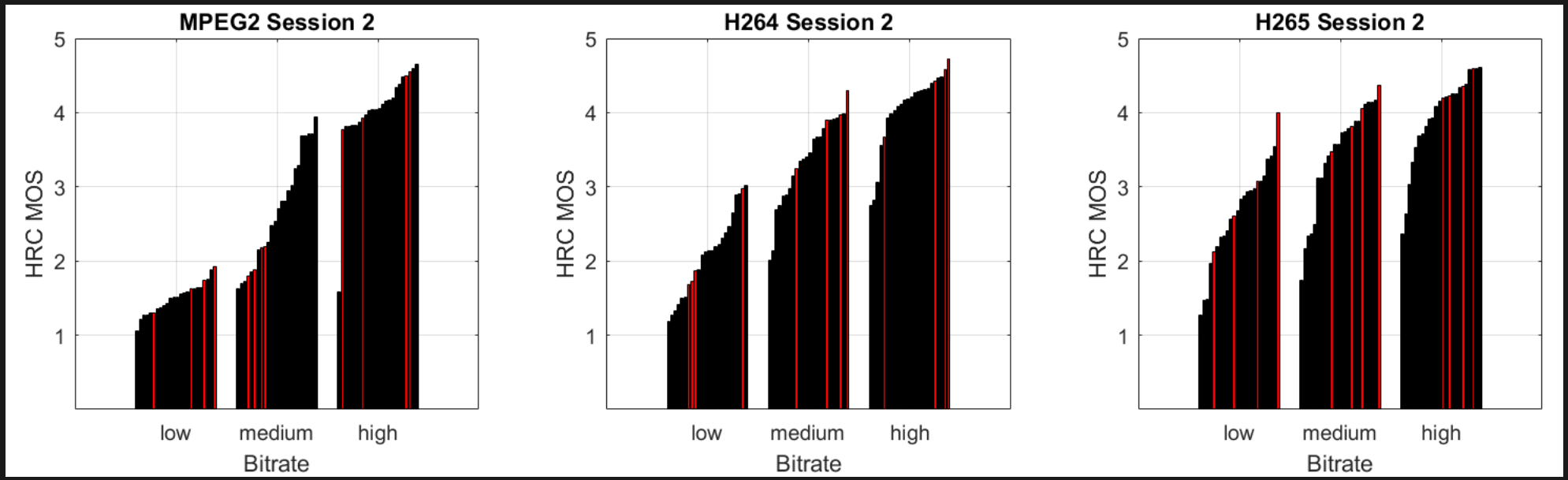
Spread of MOS for SRC



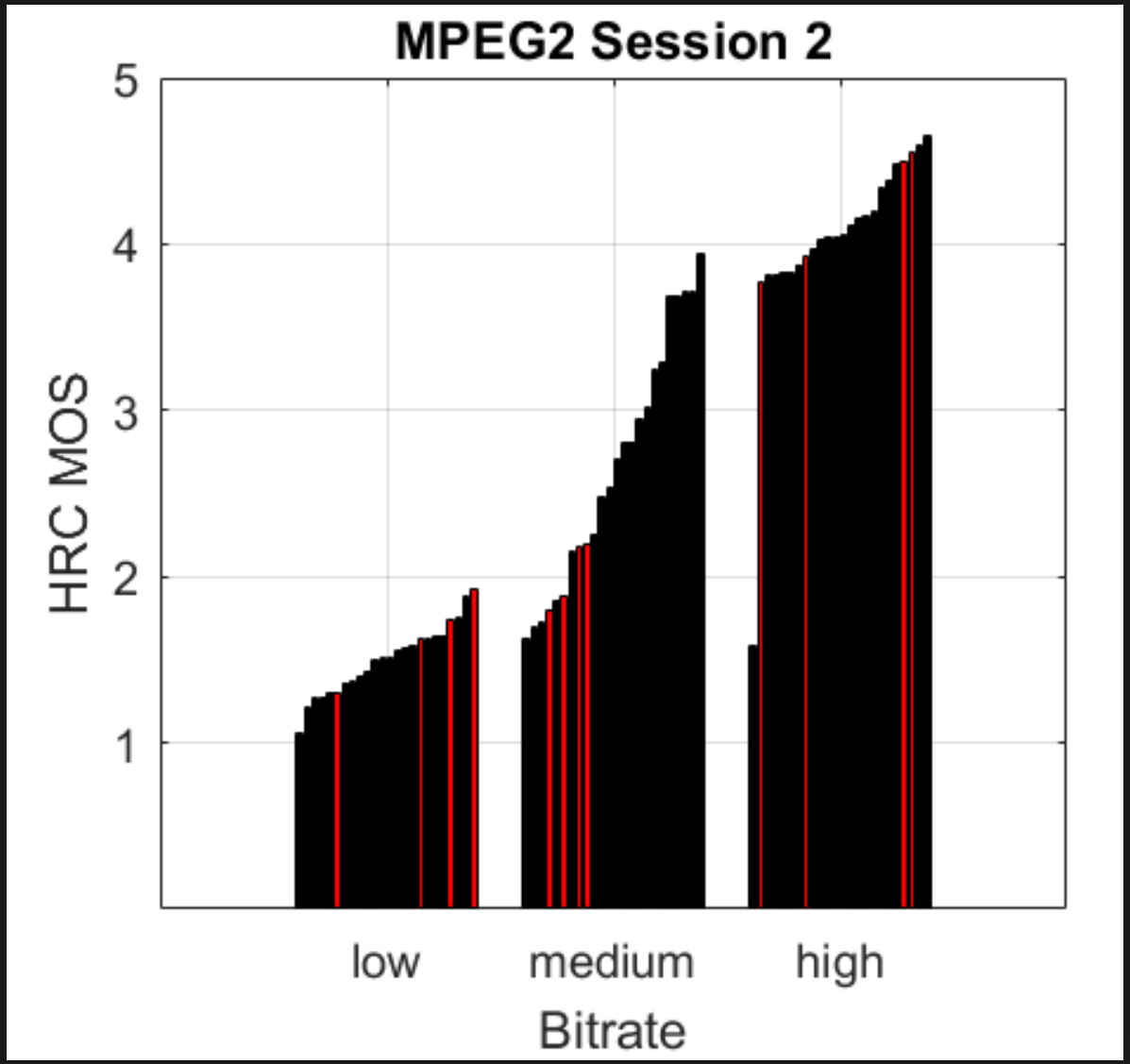
Spread of MOS for mpeg-2 SRC



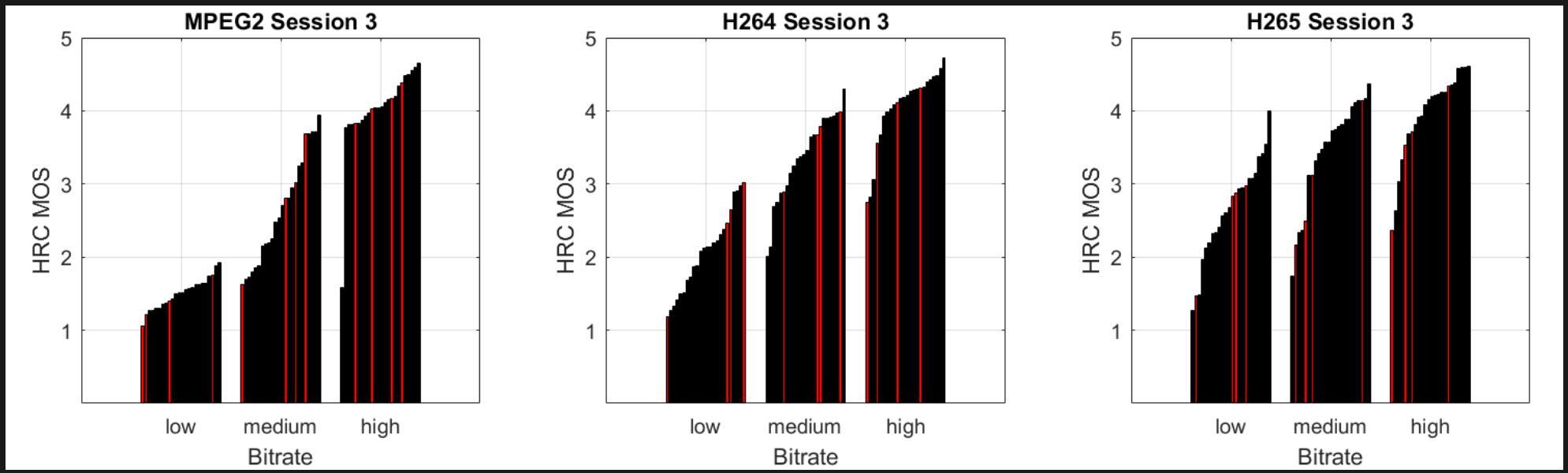
Spread of MOS for RSRC



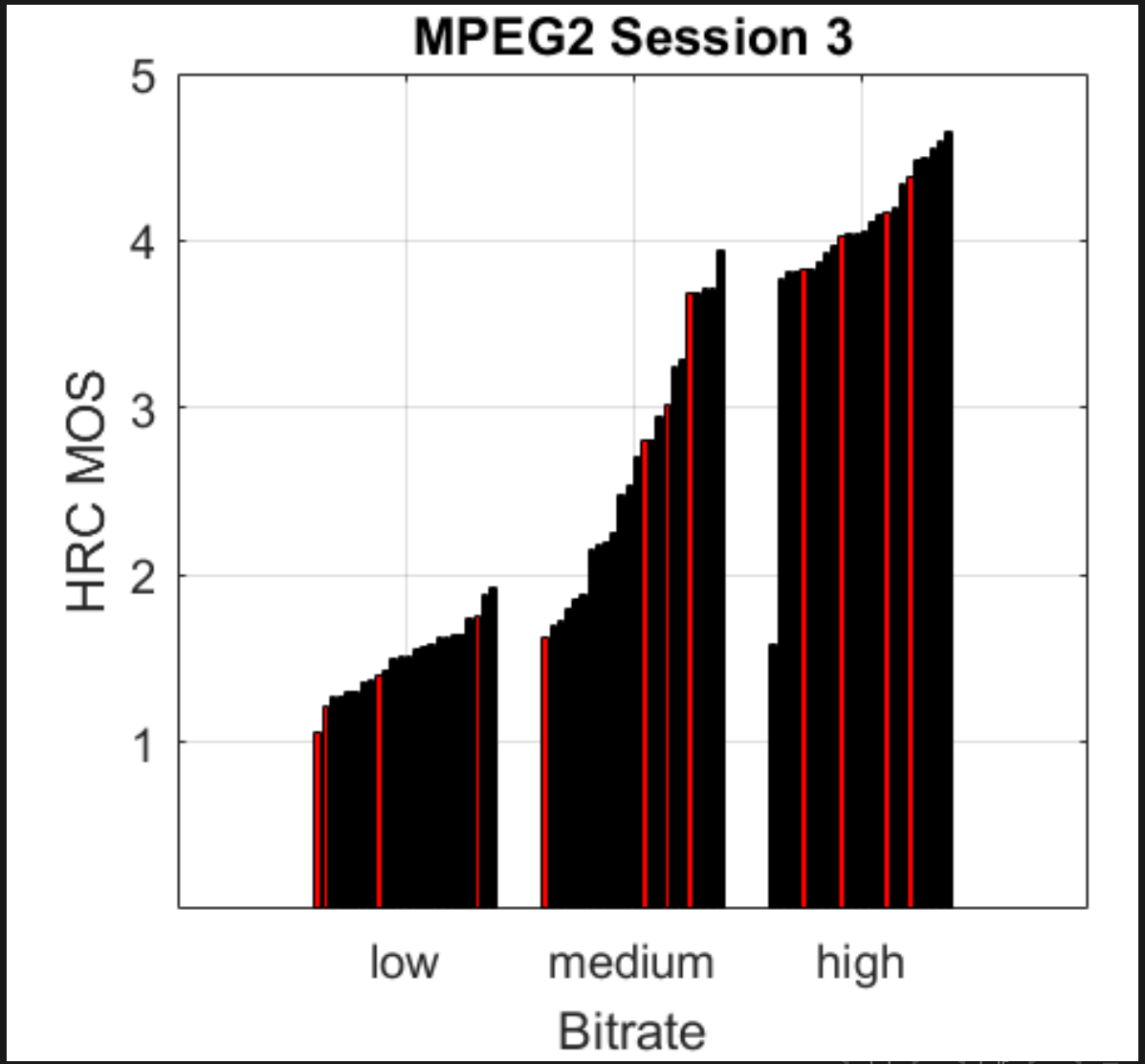
Spread of MOS for mpeg-2 RSRC



Spread of MOS for CD-SRC



Spread of MOS for mpeg-2 CD-SRC



Error Analysis

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

Error Analysis

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

Error Analysis

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

$$\epsilon \sim N(0, \sigma)$$

Error Analysis

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

$$\epsilon \sim N(0, \sigma)$$

We can ask for which experiment σ is the lowest?

Error Analysis

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

$$\epsilon \sim N(0, \sigma)$$

We can ask for which experiment σ is the lowest?

Design	σ	SRC	RSRC1	CD1
SRC	0.607		0.236	0.080
RSRC1	0.588	0.236		0.003
CD1	0.638	0.080	0.003	

Hypothesis Testing

	SRC	RSRC1	CD1	RSRC2	CD2	Rand
Same	27	16	27	14	24	22
Different	18	29	18	31	21	23

No sessions reached opposite conclusions
RSRC was most sensitive

Questionnaire AGH/NTIA

Does Scene Reuse Change Rating Behavior?

- Questionnaires say Yes
- Cannot see in data
  - Subject rating behaviors differ too much
  - Error terms are too large
  - No truth data to compare against

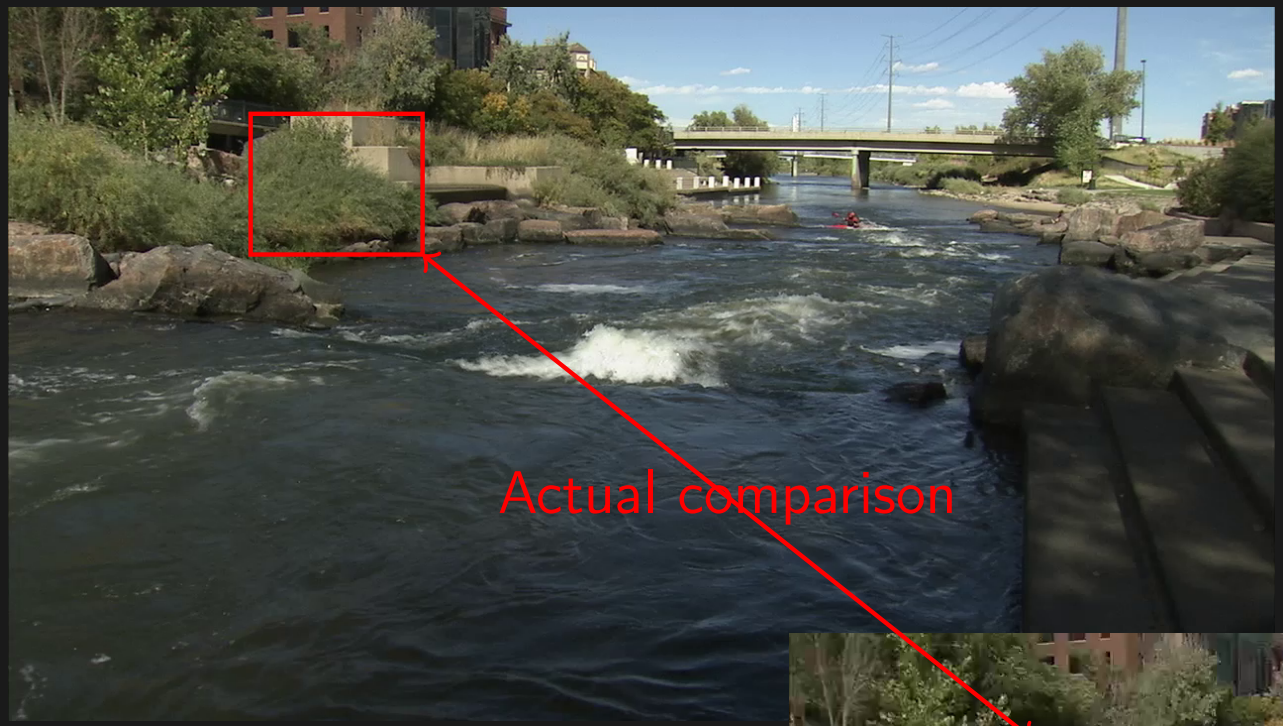
Outline

Can I Use Different Scenes?

Yes

What About Traditional Experiment Design?

What About Traditional Experiment Design?



What About Traditional Experiment Design?

Good reason to use

- Codec design, fine tune parameter
- Hypothesis testing for SRC comparisons

What About Traditional Experiment Design?

Good reason to use

- Codec design, fine tune parameter
- Hypothesis testing for SRC comparisons

Good reason to avoid

- Developing or testing a model
- Hypothesis testing for HRC comparisons
- New technology (HDR, VR, CR-2020)

What About Traditional Experiment Design?

Good reason to use

- Codec design, fine tune parameter
- Hypothesis testing for SRC comparisons

Good reason to avoid

- Developing or testing a model
- Hypothesis testing for HRC comparisons
- New technology (HDR, VR, CR-2020)

Compromise

- Repeat scenes only a few times
- Very careful design needed