



AKADEMIA GÓRNICZO-HUTNICZA
IM. STANISŁAWA STASZICA W KRAKOWIE
AGH UNIVERSITY OF KRAKOW

Laboratory and Crowdsourcing Environments in Video QoE Assessment

A comparative analysis of rating consistency and response times

DOI: 10.3390/electronics15081666

VQEG Meeting, 19 May 2026

Avrajyoti Dutta¹, Mohamedalfateh T. M. Saeed², Swapnil Arawade¹, Andreja Samčović³, Syed Uddin¹, Dawid Juszka¹, Michał Grega¹, Mikołaj Leszczuk¹

¹ AGH University of Krakow ² Ondokuz Mayıs University ³ University of Belgrade

Electronics 2026, 15, 1666 · MDPI

Outline

- Motivation and the question we wanted to answer
- Methodology: two environments, one protocol
- Stimuli: stable versus time varying quality
- The CrowdQoE platform
- Results on reliability, cognitive load, and equivalence
- Discussion and limitations
- Conclusions and future work

Why this study?

- Subjective video quality assessment matters because it shapes user satisfaction and the way providers optimize streaming.
- Laboratory testing is reliable but slow and expensive.
- Crowdsourcing platforms like Prolific give us scale and demographic reach, but the environment is no longer controlled.
- The open question: when does the crowd give us the same answer as the lab, and when does it not?

1,344

respondents in our earlier Prolific based crowdsourcing studies

ITU-T P.910

international standard followed in both environments

Response time

behavioural signal we record in milliseconds, in addition to ratings

Research gap and objectives

What is missing in the literature

- Most lab versus crowd comparisons look only at average ratings.
- Few studies measure the cognitive cost of remote assessment.
- The boundary where crowdsourcing stops being reliable is unclear.
- Temporally complex content has not been systematically tested.

What we set out to do

- Quantify reliability gaps between the AGH lab and the Prolific crowd.
- Use response time as an implicit indicator of cognitive load.
- Identify when remote crowdsourcing is still statistically valid.
- Give practical guidance for scalable QoE pipelines.

Where this work sits

Crowdsourcing for QoE

- Hoßfeld et al., best practices for QoE crowd testing (TMM 2014).
- Naderi and Cutler, crowdsourced P.910 extension (ICASSP 2024).
- Saupe et al., crowd MOS correlates above 0.96 with the lab.

Validity and reliability

- Keimel et al., crowd designs need to be adapted.
- Anegekuh et al., about a third of workers do not watch the video.
- Behrend et al., crowd samples behave like university subject pools.

Cognitive and behavioural

- Lebreton et al., gaze tracking reveals a gap between what users do and say.
- Wanat et al., behavioural QoE closer to realistic settings.
- Jakubiec and Janowski, fNIRS based neuro QoE.

Two environments, one protocol

CONTROLLED LABORATORY

AGH University of Krakow

- 32 bachelor students, aged 19 to 25.
- Identical desktops, Full HD displays, 1 Gbps wired network.
- Standardised lighting, sessions supervised in person.
- No multitasking, sessions of about 55 minutes.

UNCONTROLLED CROWDSOURCING

Prolific platform, global pool

- Recruited internationally: UK, US, Ireland, Germany, France, others.
- Own PC, at least 40 Mbps download, Full HD enforced.
- No supervision, participants in their own homes.
- Automated checks on browser, resolution, and connection speed.

Same protocol on both sides: ITU-T P.910, 5 point ACR scale, response time in ms, same CrowdQoE platform

Two databases by design

DATABASE A

Time varying quality

- 80 processed sequences, made from 26 short and 20 long source clips.
- Quality fluctuates within each clip.
- Rating it requires temporal pooling, the viewer keeps integrating quality over time.
- High cognitive load on purpose.

DATABASE B

Stable quality

- 40 clips, each shown three times, so 120 stimuli in total.
- Constant quality through each clip.
- A viewer can make an immediate snapshot judgement.
- Low cognitive load, serves as a baseline.

Five quality levels (Excellent to Bad), set via bitrate, quantization and compression. Audio removed. Order randomized. Datasets are public: CrowdQoE-2025 on GitHub.

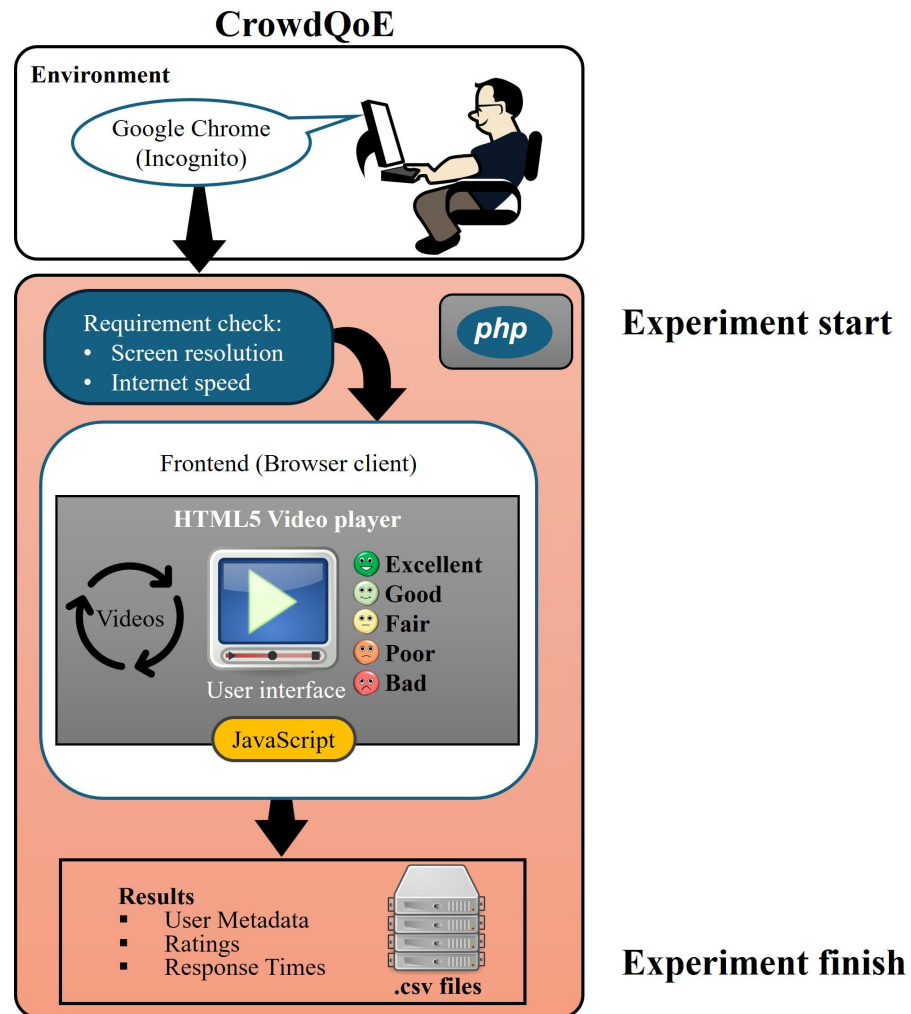
SETUP, TABLE 1

Environmental control summary

Environment	Device	Network	Supervision	Ambience	Validation
Laboratory	Identical AGH desktops, Full HD	1 Gbps LAN	Supervised	Standardised	Yes
Prolific	Own PC, Full HD required	At least 40 Mbps	None	Personal setting	Yes

Same experimental logic on both sides, so environmental control is the only thing that varies between arms.

The CrowdQoE platform



Same pipeline for both arms

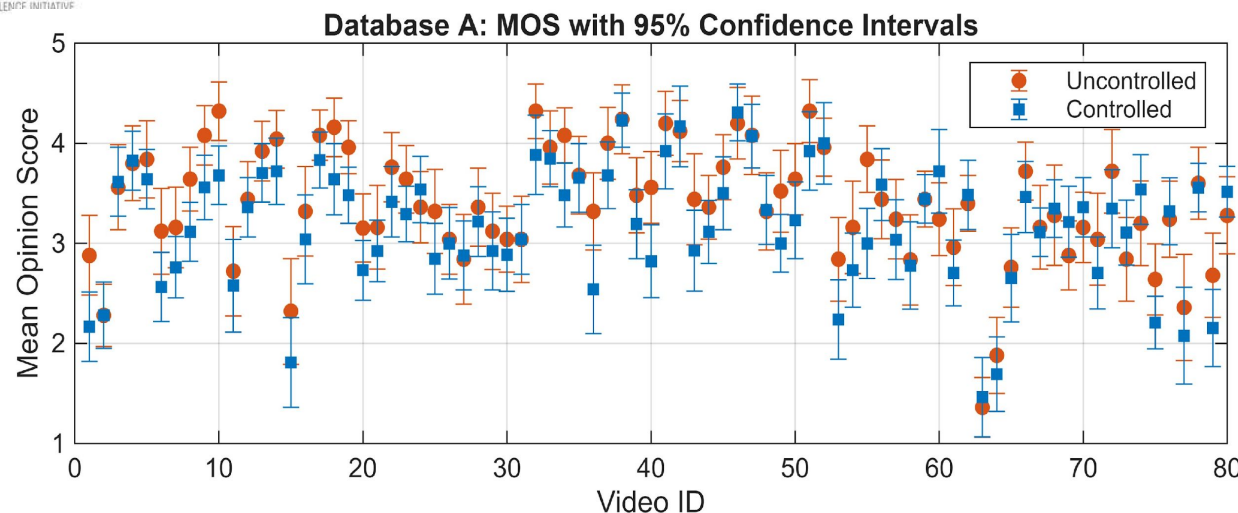
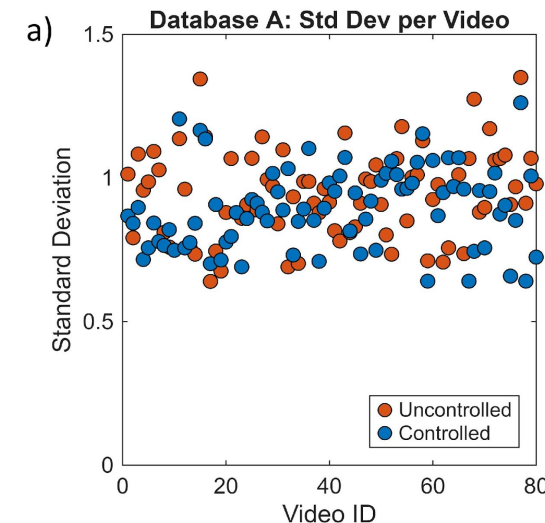
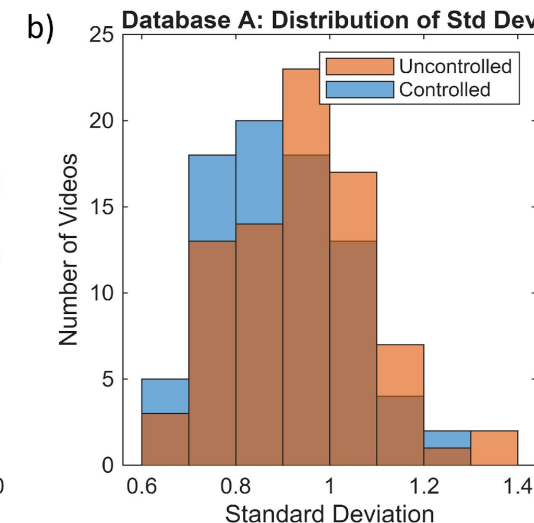
- Built in PHP, JavaScript, and an HTML5 video player.
- Real time checks on browser, screen resolution and connection speed before entry.
- Incognito mode reduces cache effects and keeps sessions anonymous.
- Playback order randomized to remove ordering bias.
- ACR rating and millisecond response time recorded for every clip.
- All raw data and metadata logged under a Session ID.

RESULTS I

Raw MOS and rating variability

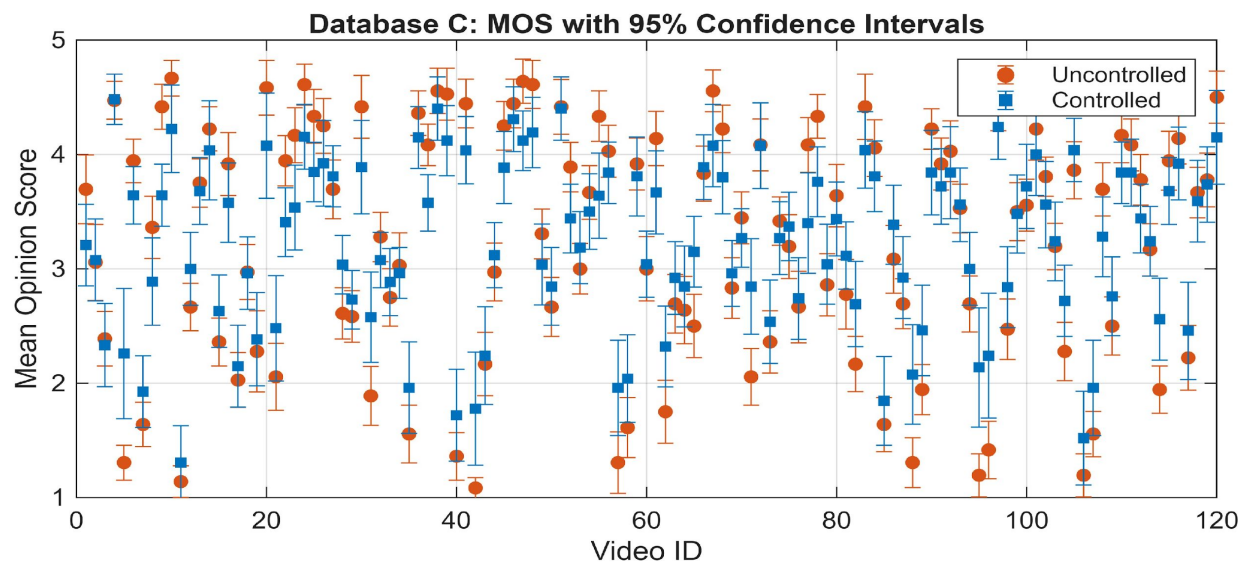
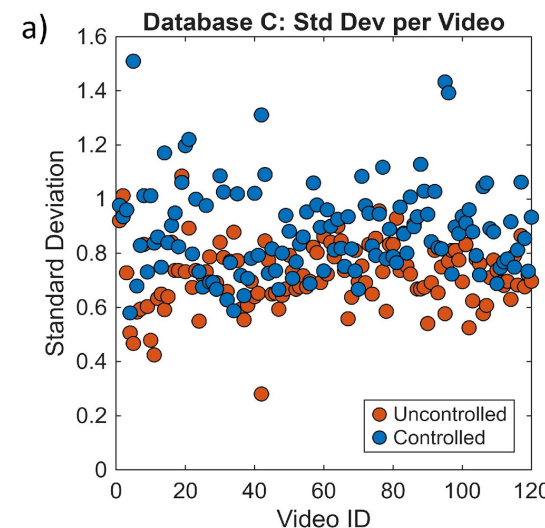
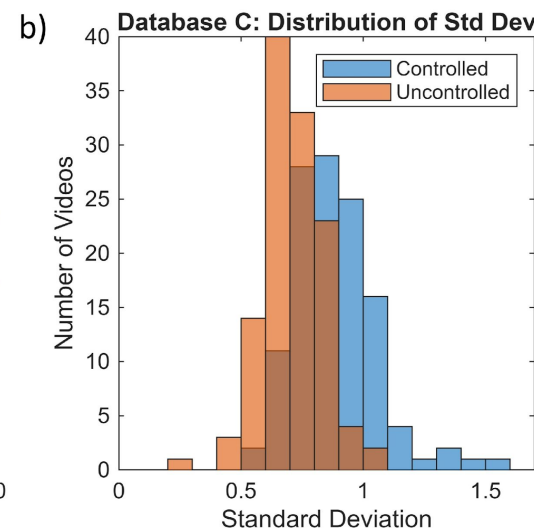
DATABASE A

Time varying



DATABASE B

Stable



Central tendency is similar across environments; rating spread is wider for time varying content.

Reliability screening exposes the gap

Two stage ITU-R BT.500 screening

- Session completion of at least 95 percent.
- Leave one out Pearson correlation, r at least 0.75.

92.0 %

rejection rate, Database A, crowd (only 2 valid subjects)

11.1 %

rejection rate, Database B, crowd (32 valid subjects)

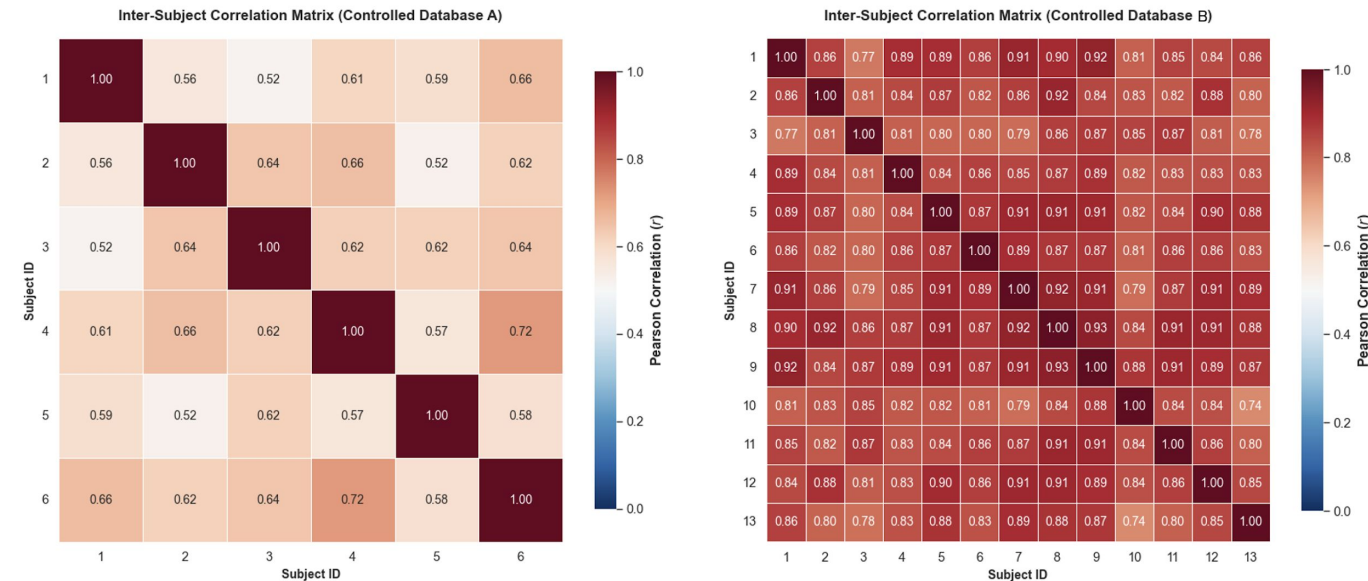
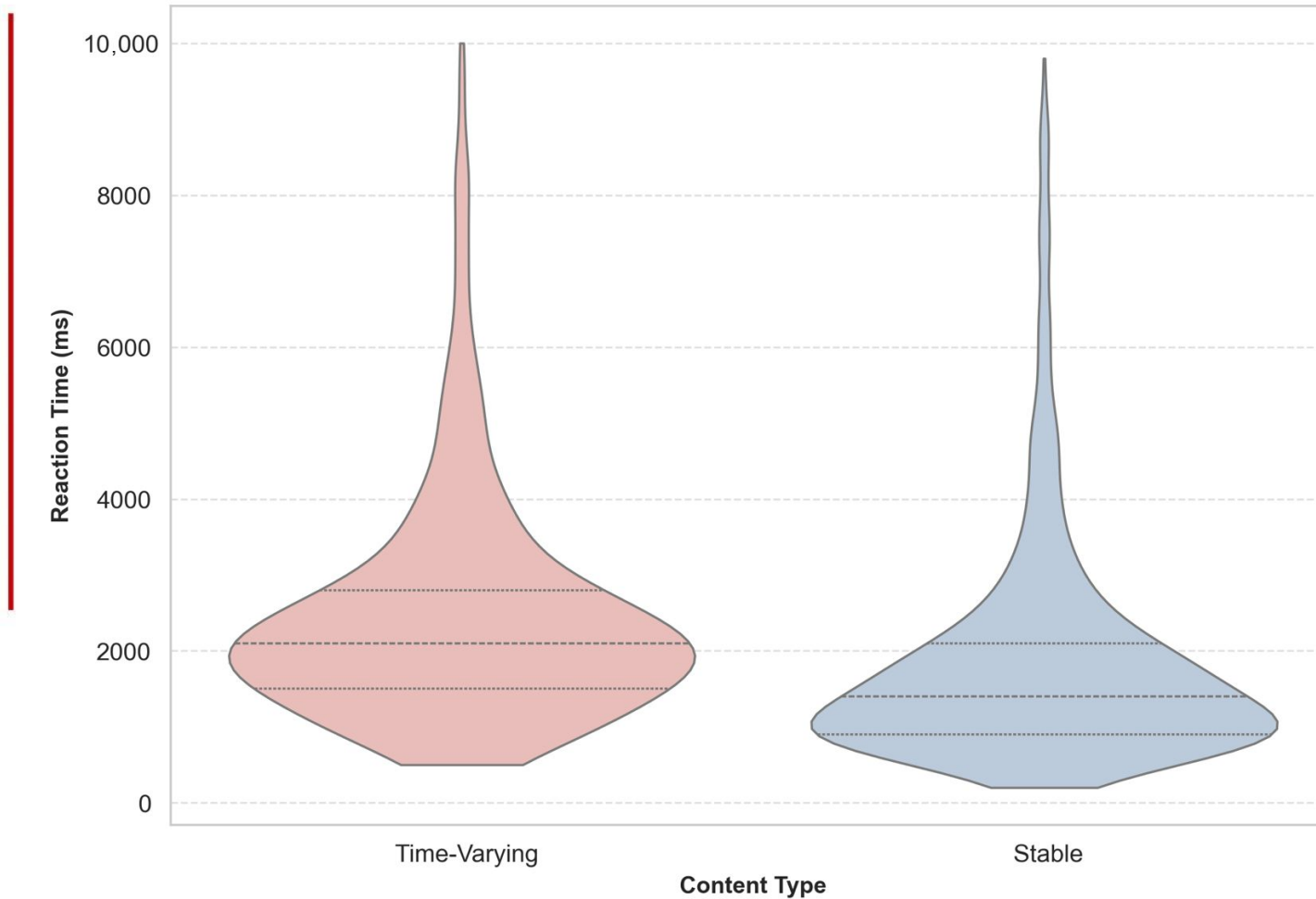


Figure 6. Inter-subject correlation for retained valid subjects.

Cronbach α = 0.81 (DB A) and 0.94 (DB B)

Hidden cost: cognitive load

Cognitive Load Analysis (Voting Reaction Time)



Same MOS, different effort

Lab subjects, Mann–Whitney U test, $p < 0.001$

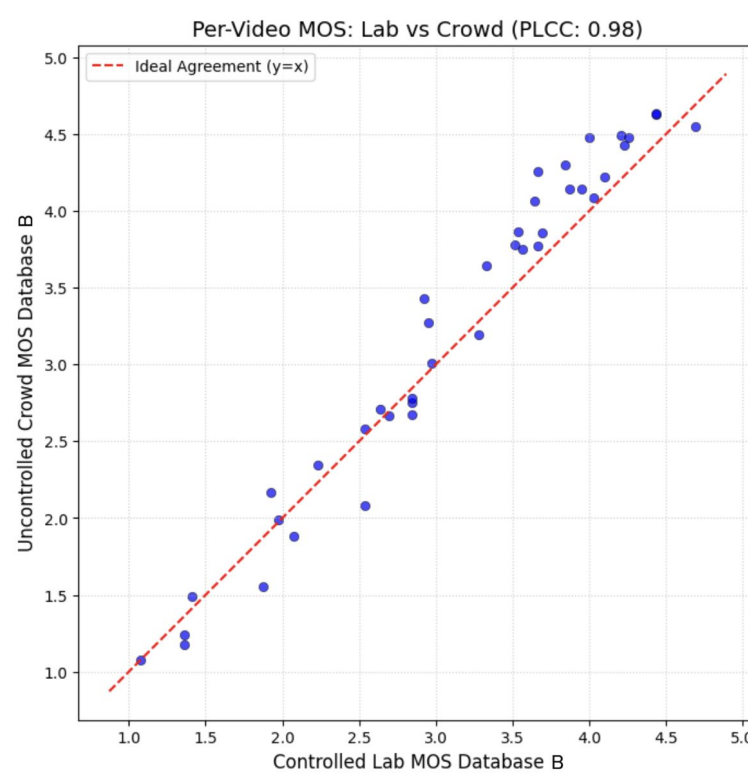
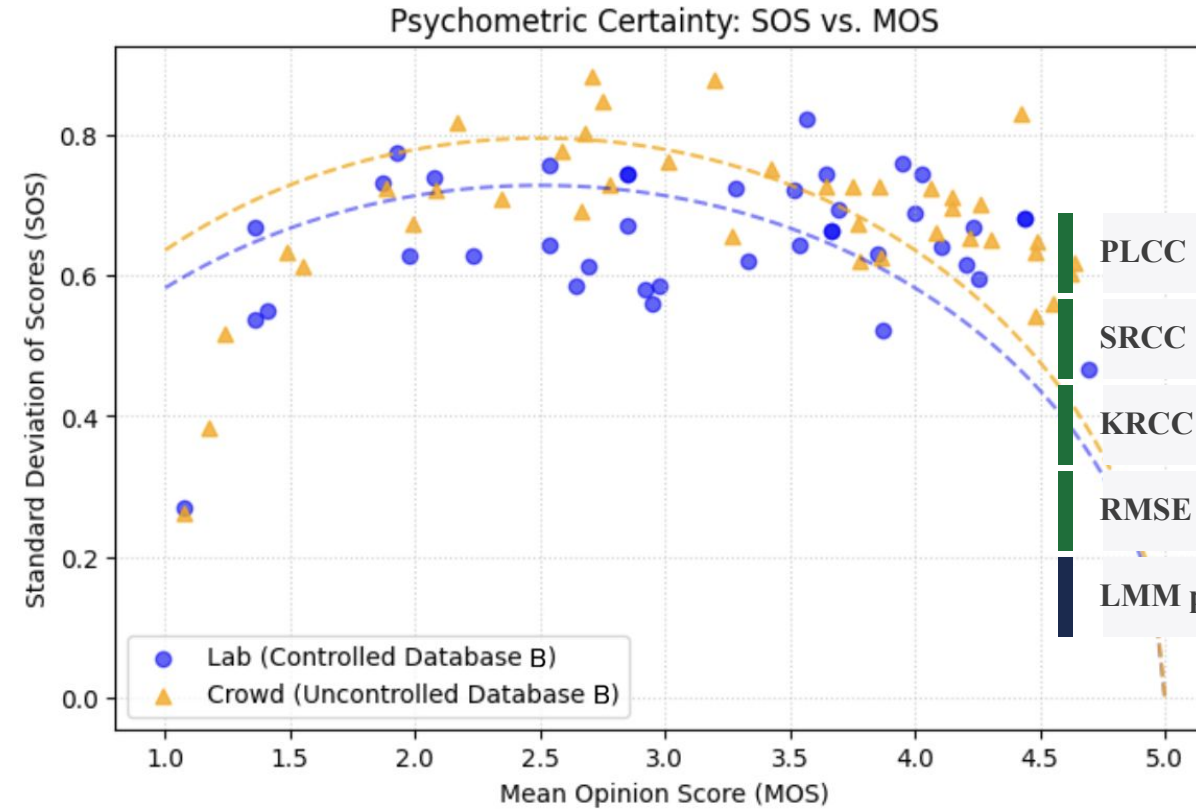
+506 ms

longer to rate time varying than stable content

Cognitive compensation

- MOS values match across content types ($p = 0.353$).
- Viewers pay in processing time, not in the score they give.
- Temporal pooling has a measurable mental cost.

Lab versus crowd: equivalence on stable content



For stable content (DB B):

PLCC	0.9829
SRCC	0.9809
KRCC	0.8988
RMSE	0.255 MOS pt
LMM p	0.258 (n.s.)

A video engineer using crowd data would reach the same conclusions as one using lab data, for stable content.

What this means

Time varying content

Crowdsourcing is currently not valid

- Temporal pooling demands sustained attention.
- Without supervision, attention does not hold up (92 percent rejection).
- The compensation we see in the lab does not transfer to the crowd.
- MOS by itself does not show this cost.

Stable content

Crowdsourcing matches the lab

- Strict LOO screening keeps the data clean.
- Ratings, rankings, and rating variance all line up.
- Practical, scalable ground truth at lower cost.
- Real world diversity becomes a feature.

Limitations: lab cohort skewed to engineering students (95 percent male), $N = 19$ valid lab subjects, display hardware varies in the crowd.

Conclusions and future work

1

Crowdsourcing is a bounded method

Its validity depends on the cognitive load of the task, not on the platform.

2

Stable content: crowd matches lab

PLCC 0.98, SRCC 0.98, LMM $p = 0.258$. Same rating certainty as the lab.

3

Time varying content: crowd fails

92 percent rejection. Attention collapses without supervision. MOS alone hides this.

4

Future QoE designs should track cognitive load

Response time today. fNIRS, gaze, and behavioural signals next.



CONTACT

About me

Avrajyoti Dutta

PhD Researcher, AGH University of Krakow

Member, AGH Quality of Experience (QoE) Team

Email	avrajyoti.dutta@agh.edu.pl
LinkedIn	linkedin.com/in/avrajyotidutta
ORCID	orcid.org/0000-0001-5749-0256
AGH QoE	qoe.agh.edu.pl/members/
Code/Data	github.com/dutta-agh/CrowdQoE-2025



Scan to connect on LinkedIn

Thank you

Questions, comments, collaborations welcome.

avrajyoti.dutta@agh.edu.pl

Paper DOI: [10.3390/electronics15081666](https://doi.org/10.3390/electronics15081666)

Code and data: github.com/dutta-agh/CrowdQoE-2025

Presented at the VQEG Meeting, 19 May 2026.