

Subjective Audio Testing Cheat Sheet

MUSHRA / MOS / A/B / ABX - which method, how many listeners, on one page.

Step 1 - which method answers your question?

MOS	"How good is this clip, 1-5?" Absolute rating of speech (ITU-T P.800; P.808 for crowdsourcing). Score each clip on its own.
MUSHRA	"Rank these codecs, 0-100." Intermediate quality (ITU-R BS.1534-3). Many systems at once + hidden reference + low-pass anchor.
A/B	"Which is preferred?" Preference share (%). Cheap and intuitive - but loudness-match first or you measure volume.
ABX	"Is the difference audible at all?" X is secretly A or B; identify it. Significant only if a binomial test beats 50% guessing.

Step 2 - how many listeners, which statistic?

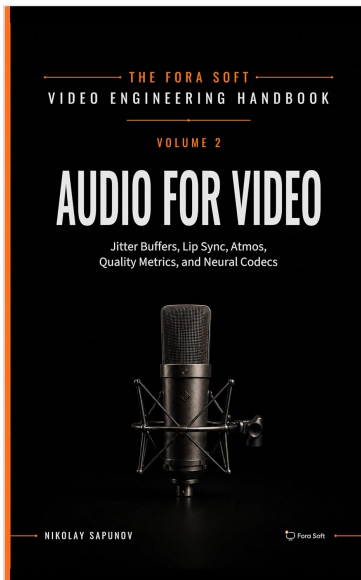
Method	Listeners (valid)	Statistical test	Output
MOS (P.800)	24+ lab / 100s crowd	mean + CI; ANOVA	1-5 score
MUSHRA (BS.1534-3)	~15+ after screening	mean + CI; RM-ANOVA	0-100 score
A/B preference	20-40+	binomial / sign test	preference %
ABX	1 trained + many trials	binomial vs 50%	p-value

Key facts that catch people

- MUSHRA = Multiple Stimuli with Hidden Reference and Anchor. The hidden reference pins the top (near 100); the 3.5 kHz low-pass anchor pins the bottom.
- MUSHRA post-screening drops any listener scoring the hidden reference below 90 on more than 15% of items - recruit above 15 so you keep 15 after screening.
- MOS is not a portable ruler: a 4.2 is only meaningful within one test. Never quote a bare MOS with no test context.
- ABX: a majority is not proof. 12 of 16 correct is significant ($p = 0.038$); 10 of 16 is not ($p = 0.23$). Precision grows with $\sqrt{\text{listeners}}$.
- BS.1116-3 (ABC/HR) is MUSHRA's sibling for near-transparent, audiophile-grade differences; MUSHRA is for the audible-but-acceptable range.

Running a defensible test - the discipline

- ☐ Loudness-match every clip to the same LUFS first - louder reliably beats better.
- ☐ Fix the method and the listener count before you collect a single rating.
- ☐ Randomise and balance clip order to kill order bias.
- ☐ Screen listeners (hidden reference / attention checks); drop the inattentive.
- ☐ Report confidence intervals and a significance test - a bare mean is a guess.



GO DEEPER · VOLUME 5 OF 7

Audio for Video: A Complete Guide to Sound in Video Products

From Jitter Buffers and Lip Sync to Atmos, Quality Metrics, and Neural Codecs

By Nikolay Sapunov, CEO at Fora Soft · 1,288 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0HH6XPY46

This sheet picks between MOS, MUSHRA, A/B and ABX and sets listener counts and statistics for each. Volume 5 places listening tests inside the wider quality toolkit: objective metrics such as PESQ, POLQA and ViSQOL, MUSHRA practice, and how neural codecs change what you need to measure.

WHAT YOU WILL FIND INSIDE

- MUSHRA test design and listener screening
- PESQ, POLQA and ViSQOL vs listening tests
- Evaluating neural audio codecs
- Lip sync and timing errors listeners notice

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI

Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF

Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery

Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing

Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs

Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video

Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Need to prove a codec or audio change sounds better before release?

We will design and run the listening test with objective metrics alongside, and give you a report with confidence intervals your team can act on. 250+ video products shipped since 2005.

[Book a 30-min call →](#)
[WhatsApp →](#)
[Email us →](#)