Subjective Audio Testing Cheat Sheet

Fora Soft Learn

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(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.