

## Listening Test Starter Kit

How to design a fair A/B, ABX or MUSHRA-style test, and get answers about sound you can actually trust.

### Pick the right test

START WITH THE QUESTION

| YOUR QUESTION                                       | TEST                                | HOW IT WORKS                                                                                                                   | WHAT YOU GET                                        |
|-----------------------------------------------------|-------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------|
| Can anyone hear a difference at all?                | <b>ABX</b>                          | Listeners hear A, B and an unknown X, and say whether X is A or B. Repeat many trials.                                         | Percent correct and whether it beats guessing       |
| Which version do people prefer?                     | <b>Paired comparison (A/B)</b>      | Listeners hear two versions in random order and pick one, or rate the strength of preference.                                  | Preference share, ideally with confidence intervals |
| How do several versions compare to a reference?     | <b>MUSHRA-style (ITU-R BS.1534)</b> | Listeners rate several versions on a 0–100 scale against a known reference, with a hidden reference and a low anchor mixed in. | Mean scores per version                             |
| Are tiny impairments audible in high-quality audio? | <b>BS.1116-style</b>                | Double-blind comparison against a hidden reference, rated on an impairment scale, with experienced listeners.                  | Difference grades for small impairments             |

### Who should listen

**Experienced listeners** (engineers, musicians, trained panels) find small differences faster and more consistently.

**Everyday listeners** tell you what the public will actually notice. Pick the panel that matches the question.

### How many

More listeners and more trials make results more trustworthy. Small pilot tests are useful for finding problems; decisions deserve larger panels. Standards such as BS.1534 and BS.1116 give their own guidance on panel size.

1. **Level-match everything.** A slightly louder version almost always sounds better. Match by loudness (for example LUFS) and confirm by ear before testing.
2. **Keep it blind.** Listeners should not know which version is which. Double-blind, where the person running the test doesn't know either, is better still.
3. **Randomize.** Shuffle the order of trials and which button holds which version, so position and order don't drive answers.
4. **Control playback.** Same room, speakers or headphones, and level for every listener. Calibrate and write it down.
5. **Fight fatigue.** Keep sessions short, build in breaks, and give a short familiarization round before scored trials.
6. **Decide the analysis first.** Write the question, the hypothesis and how you'll judge the result before collecting data.

## ABX: how many correct is enough?

| TRIALS | CORRECT NEEDED | CHANCE OF GUESSING IT |
|--------|----------------|-----------------------|
| 10     | 9              | 1.1%                  |
| 12     | 10             | 1.9%                  |
| 16     | 12             | 3.8%                  |
| 20     | 15             | 2.1%                  |
| 24     | 17             | 3.2%                  |
| 30     | 20             | 4.9%                  |

One listener, one-sided binomial test, under 5% chance of reaching that score by guessing. Fewer correct answers means the result could be luck, not that there is no difference.

## Stimuli checklist

- ☐ Same excerpt, same start and end points
- ☐ Short clips (often 10–20 seconds) that show the difference
- ☐ Matched loudness, no clicks at the edges
- ☐ Neutral file names that don't reveal the version
- ☐ A hidden reference and, where useful, a low anchor

## People first

- ☐ Plain-language consent before anyone listens
- ☐ Safe, comfortable playback levels
- ☐ Responses stored without names
- ☐ An ethics review if results will be published

| ITEM                                                                | YOUR PLAN |
|---------------------------------------------------------------------|-----------|
| <b>Question</b><br>One sentence                                     |           |
| <b>Hypothesis</b><br>What you expect, and what would prove it wrong |           |
| <b>Test method</b><br>ABX, A/B, MUSHRA-style, other                 |           |
| <b>Versions</b><br>Including reference and anchor                   |           |
| <b>Program material</b><br>Excerpts, genres, length                 |           |
| <b>Listeners</b><br>Who, how many, how recruited                    |           |
| <b>Playback</b><br>Room, system, level, calibration                 |           |
| <b>Procedure</b><br>Training, trials, breaks, session length        |           |
| <b>Analysis</b><br>Decided before data comes in                     |           |

## What a good report includes

- ☐ The question and hypothesis
- ☐ Versions, excerpts and how levels were matched
- ☐ Listener panel and recruitment
- ☐ Playback setup and procedure
- ☐ Results with spread or confidence, not just averages
- ☐ What the result means, and what it doesn't

## Want it run for you?

HiNote Sound designs and runs listening studies for plugins, headphones, playback systems and mixes, and delivers a clear report. [courtney@hinotesound.com](mailto:courtney@hinotesound.com) · [hinotesound.com](http://hinotesound.com)