Listening Test Interface Design

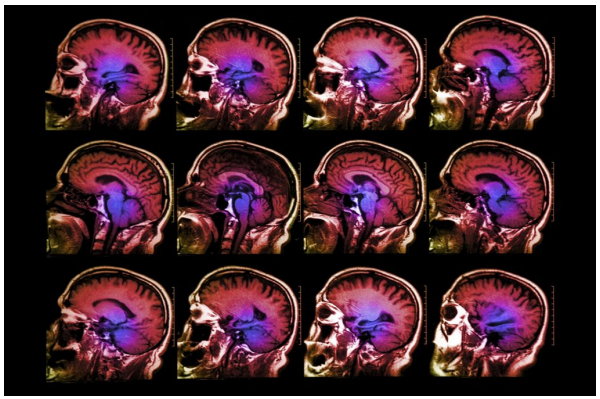
Brecht De Man
Correspondence: <b.deman@qmul.ac.uk>

Centre for Digital Music
School of Electronic Engineering and Computer Science
Queen Mary University of London

July 13, 2016

In the news

Bug in fMRI software calls 15 years of research into question (Wired UK)
Software faults raise questions about the validity of brain studies (Ars Technica)



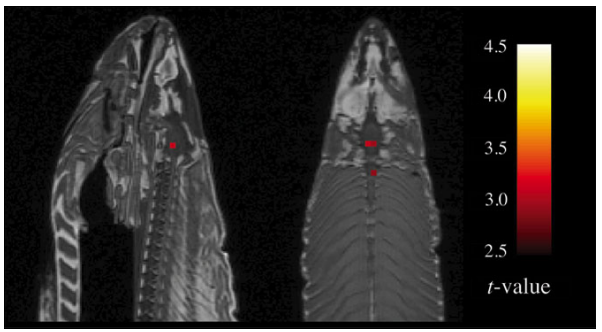
Source: iStock

"In theory, we should find 5% false positives (for a significance threshold of 5%), but instead we found that the most common software packages for fMRI analysis can result in **false-positive rates of up to 70%**. These results question the validity of **some 40,000 fMRI studies** and may have a large impact on the interpretation of neuroimaging results."

[1] A. Eklund, T. E. Nichols, and H. Knutsson, "Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates," Proceedings of the National Academy of Sciences, 2016.

In the news

Bug in fMRI software calls 15 years of research into question (Wired UK)
Software faults raise questions about the validity of brain studies (Ars Technica)



"In theory, we should find 5% false positives (for a significance threshold of 5%), but instead we found that the most common software packages for fMRI analysis can result in **false-positive rates of up to 70%**. These results question the validity of **some 40,000 fMRI studies** and may have a large impact on the interpretation of neuroimaging results."

[1] A. Eklund, T. E. Nichols, and H. Knutsson, "Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates," Proceedings of the National Academy of Sciences, 2016.
[2] Bennett et al., "Neural Correlates of Interspecies Perspective Taking in the **Post-Mortem Atlantic Salmon**: An Argument For Proper Multiple Comparisons Correction," Journal of Serendipitous and Unexpected Results, 2010.

Applications

- Audio codecs
- Electroacoustic design (loudspeakers, headphones, ...)
- Human perception
- Standards
- ...

Perceptual evaluation

- **Interface design**
- Subject selection
- Subject training
- Stimulus selection
- Listening environment
- Playback system
- Subject exclusion
- Research questions
- ...

Goals

- Accurate, reproducible, representative
- High discrimination
- Low time and effort

Basic reading

[3] Søren Bech and Nick Zacharov, "Perceptual Audio Evaluation - Theory, Method and Application," Wiley, 2006.

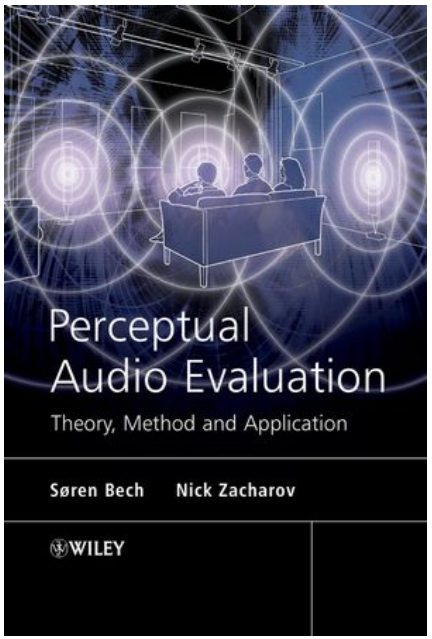


Table of Contents

Table of Contents

Available listening test tools

Name	Supported interfaces	Language
APE	multi-stimulus multi-axis, pairwise	MATLAB
BeaqleJS	MUSHRA + ABX	JavaScript
HULTI-GEN	*many*	Max
MUSHRAM	MUSHRA	MATLAB
Scale	semantic differential, n-AFC	MATLAB
WhisPER	*many (7)*	MATLAB

APE

- MATLAB-based
- Pairwise (AB) mode
- Multiple stimulus mode
  - All stimuli on one axis
  - Multiple axes for multiple attributes
- Made available to public
  - Others can spot bugs and make improvements
  - code.soundsoftware.ac.uk/projects/APE
  - AES 136th Convention [4]

Web Audio Evaluation Tool

- Browser-based: OS-independent, no 3rd party software
- Wide range of highly customisable interfaces
- Remote or local tests
- Easy quick test setup: no programming required
- High compatibility: configuration and results in XML
- Immediate diagnostics and result in browser
- Intuitive, flexible, works out of the box

Web Audio Evaluation Tool

Developers

- Nicholas Jillings *Web Audio*
- Brecht De Man *perception of music production*
- David Moffat *realism of sound synthesis*

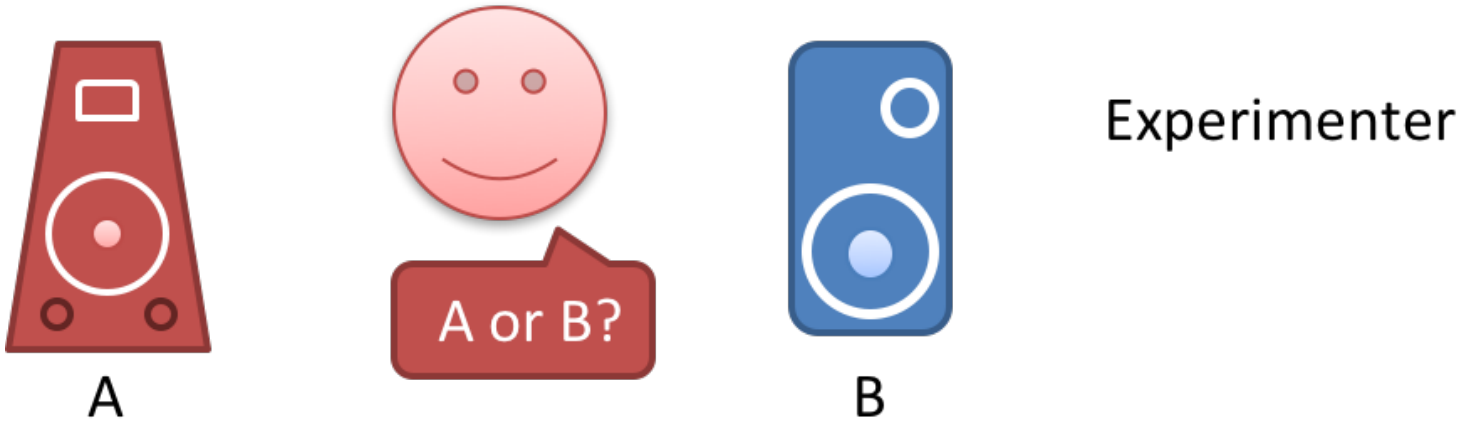
Contributors

- Giulio Moro *Hammond organs*
- Adib Mehrabi *vocal imitation of musical sounds*
- Alex Wilson *audio quality in music*
- Alejandro Osses Vecchi *psychoacoustics*

Table of Contents

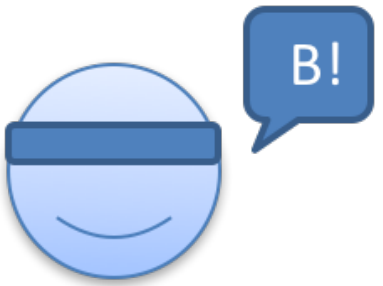
Double blind

No blind:



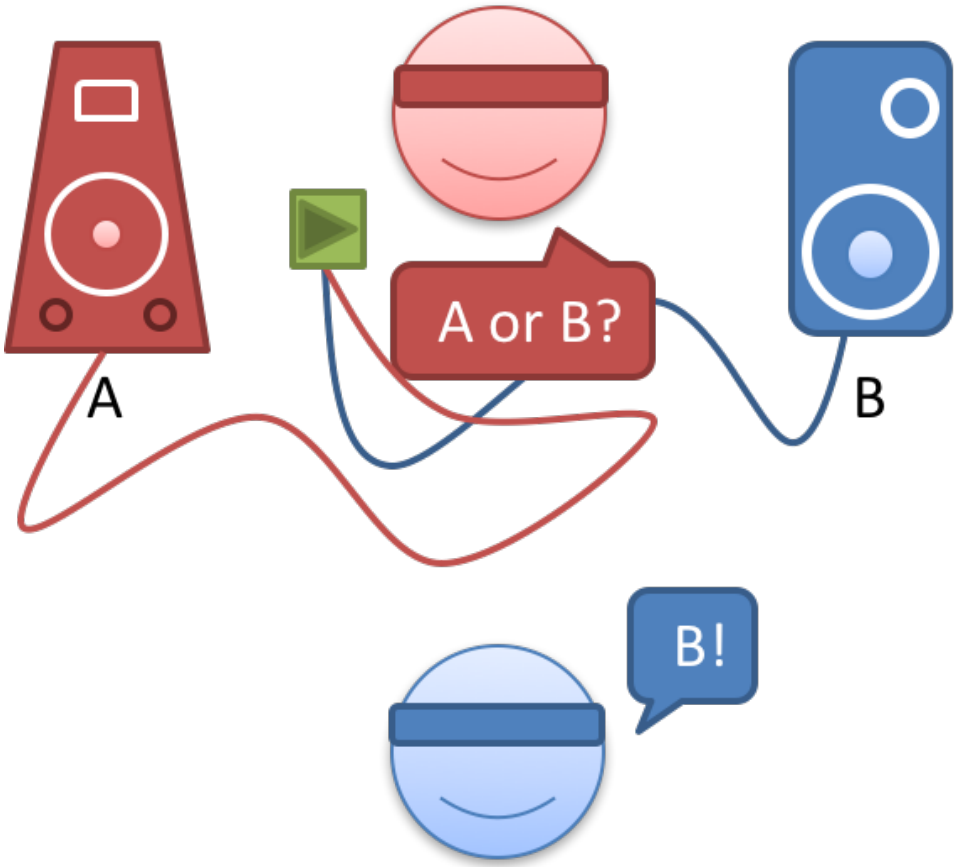
Double blind

Single blind:

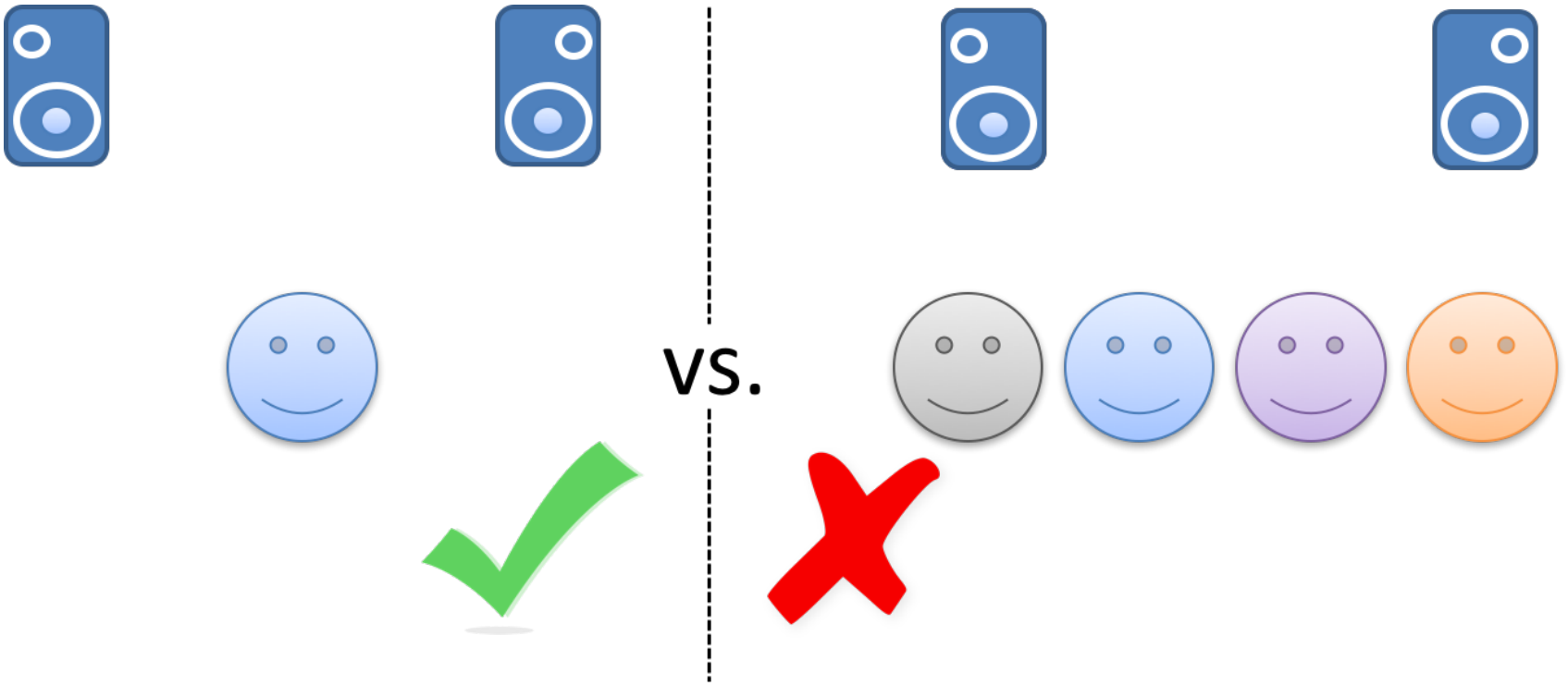


Double blind

Double blind:



Individual tests

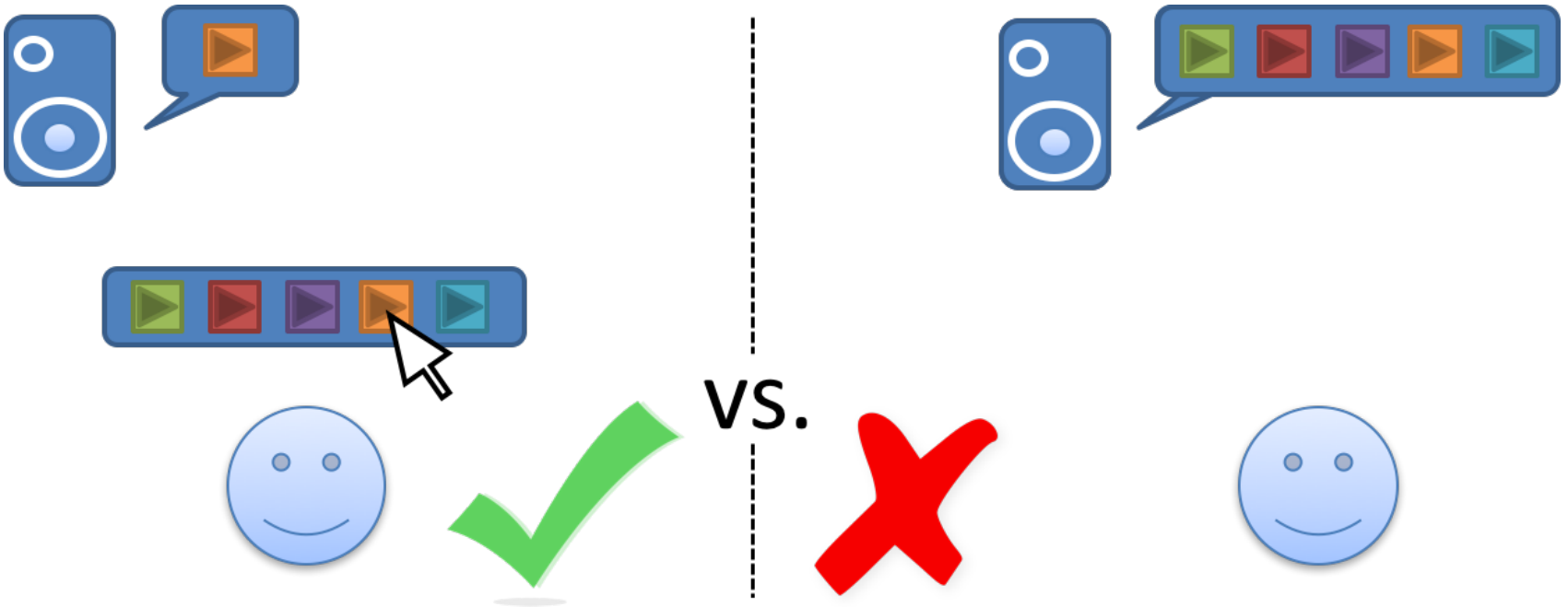


Randomisation

Remove bias due to order of

- (Presentation and labelling of) stimuli
- Parts of the experiment
- Tests
- ...

Free switching



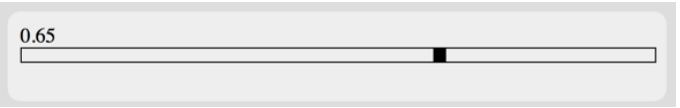
"Increases ability to perceive more delicate differences"

[6] J. Berg and F. Rumsey, "Spatial attribute identification and scaling by repertory grid technique and other methods," 16th AES Conference: Spatial Sound Reproduction, March 1999.

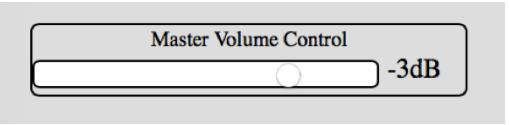
Time-aligned stimuli

- Switch to corresponding position in next stimulus
- More balanced over full length of stimuli

Reduce visual distractions



Page 1 of 2



- "Visual information processing [induces] *Inattentional Deafness*"
- "Requiring use of scroll bar: significant negative effect on critical listening reaction times"

[7] Joshua Mycroft, Joshua D. Reiss and Tony Stockman, "The Influence of Graphical User Interface Design on Critical Listening Skills," Sound and Music Computing (SMC), July 2013.

Automated, computerised tests

- Double blind
- Individual tests
- Randomisation
- Free switching
- Time-aligned stimuli
- Reduce visual distractions

Automated, computerised tests

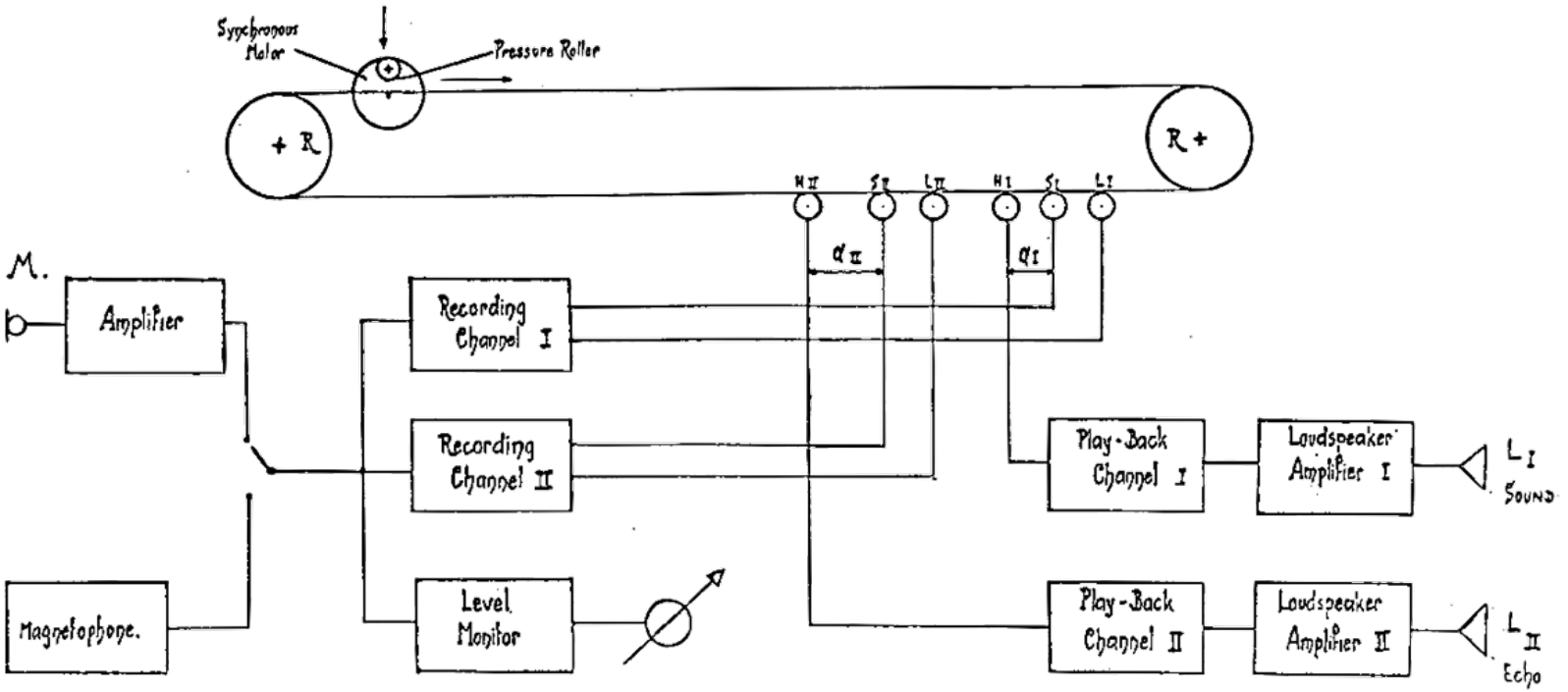


Fig. 1. Block diagram of echo apparatus.

[8] Helmut Haas, "The influence of a single echo on the audibility of speech," JAES Vol. 20 No. 2, pp. 146-159, 1972.

Automated, computerised tests

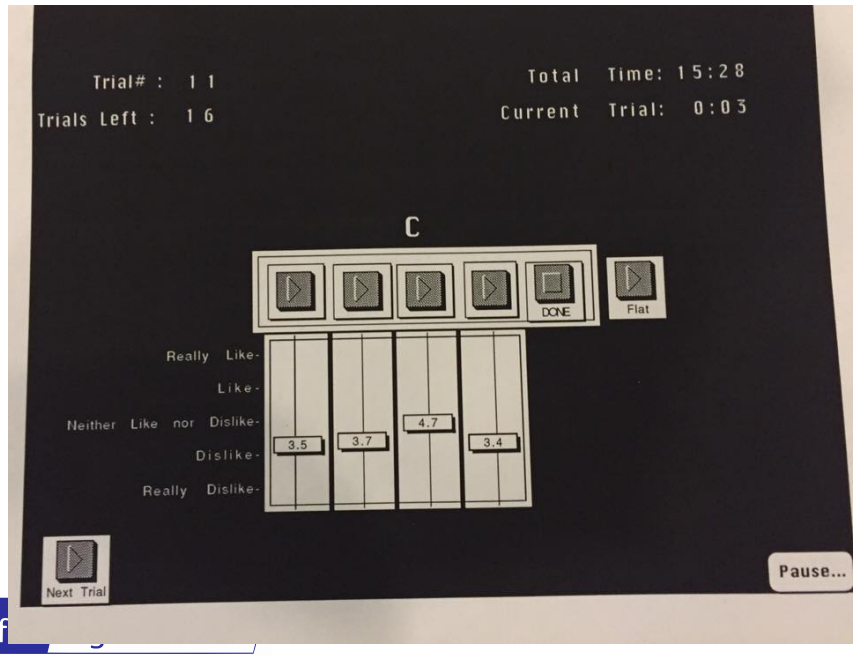
Dr Sean Olive (HARMAN)

- Pencil and paper

Automated, computerised tests

Dr Sean Olive (HARMAN)

- Pencil and paper
- HyperCard (late 1980s)



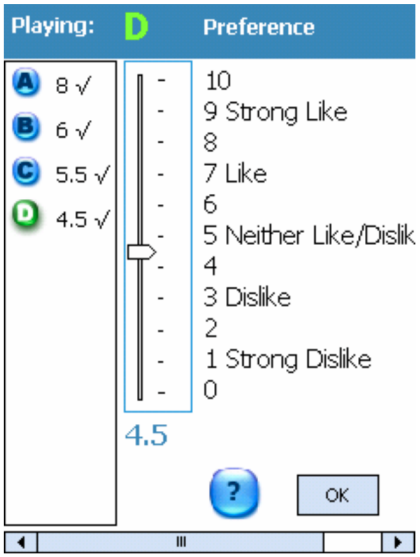
centre f

Queen Mary University of London

Automated, computerised tests

Dr Sean Olive (HARMAN)

- Pencil and paper
- HyperCard (late 1980s)
- Pocket PC / PDA (early 2000s)



centre for digital music

Queen Mary University of London

Automated, computerised tests

Dr Sean Olive (HARMAN)

- Pencil and paper
- HyperCard (late 1980s)
- Pocket PC / PDA (early 2000s)
- Apple iPad

Pilot test and dummy analysis

Source: Muppet Show Wikia

- Feedback on interface and stimuli
- Know what type of data to expect

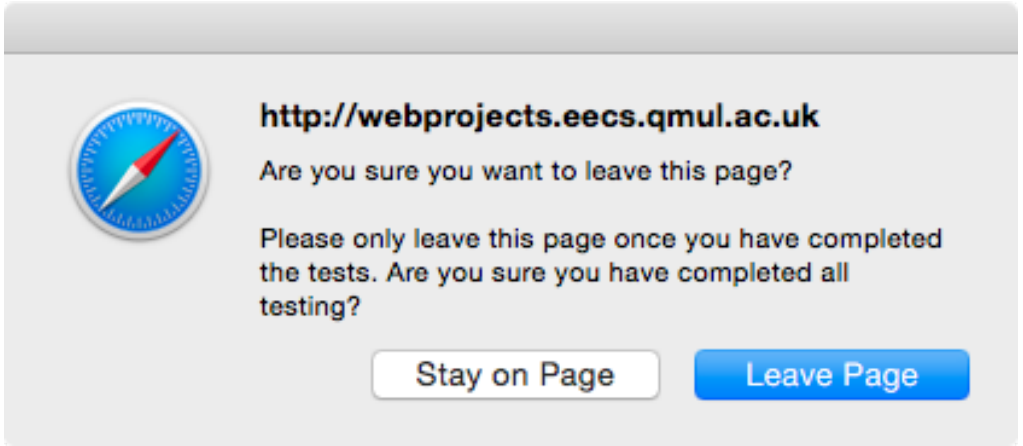
Language

Native tongue: "Minimise (risk of) semantic constraints on subjects"

[9] J. Berg and F. Rumsey, "Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction," AES Convention 109, September 2000.

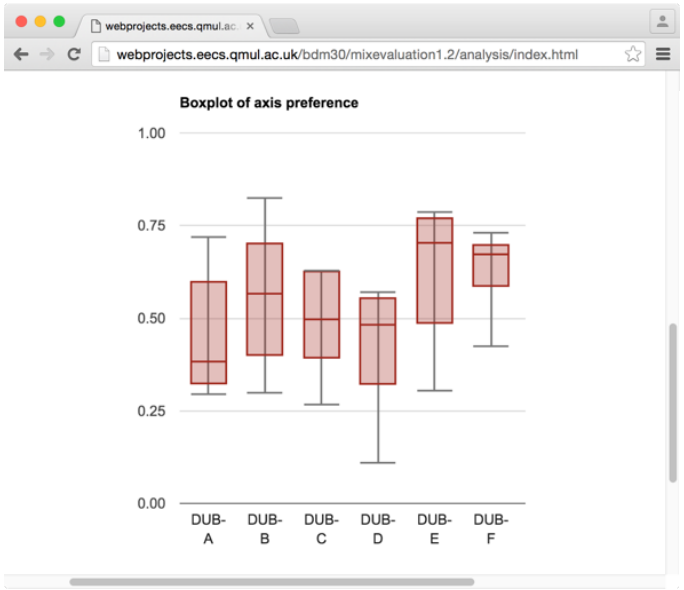
Safety

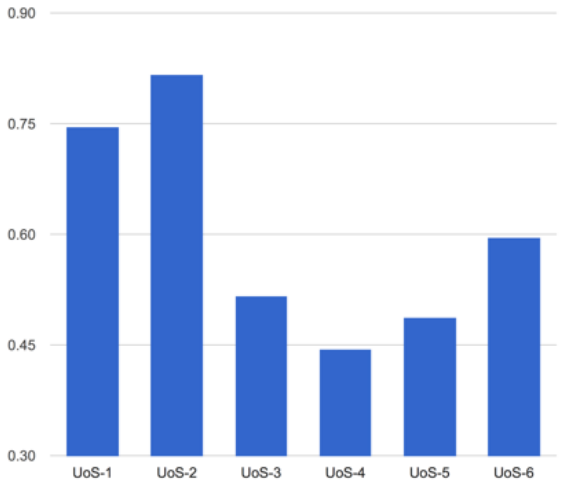
Safety



- Don't allow accidental closing of window
- Intermediate session saves

Reward



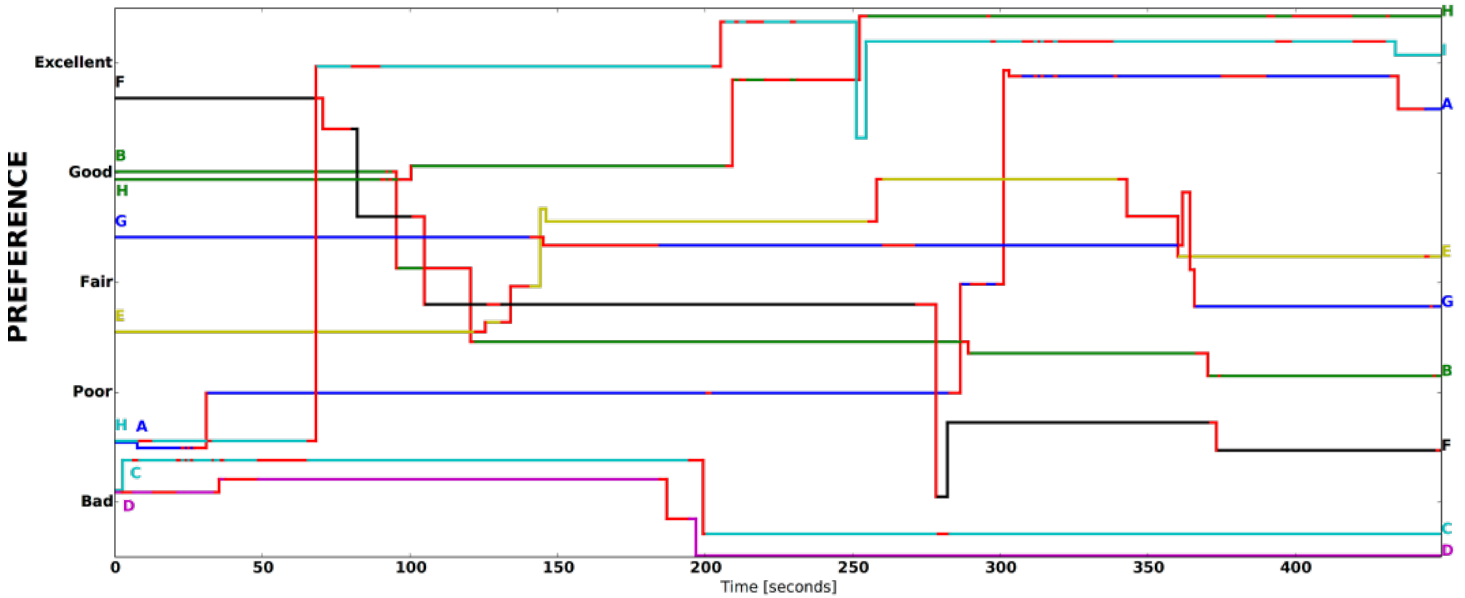


- Immediately display all / own results
- Candy

centre for digital music

Queen Mary University of London

Diagnostics

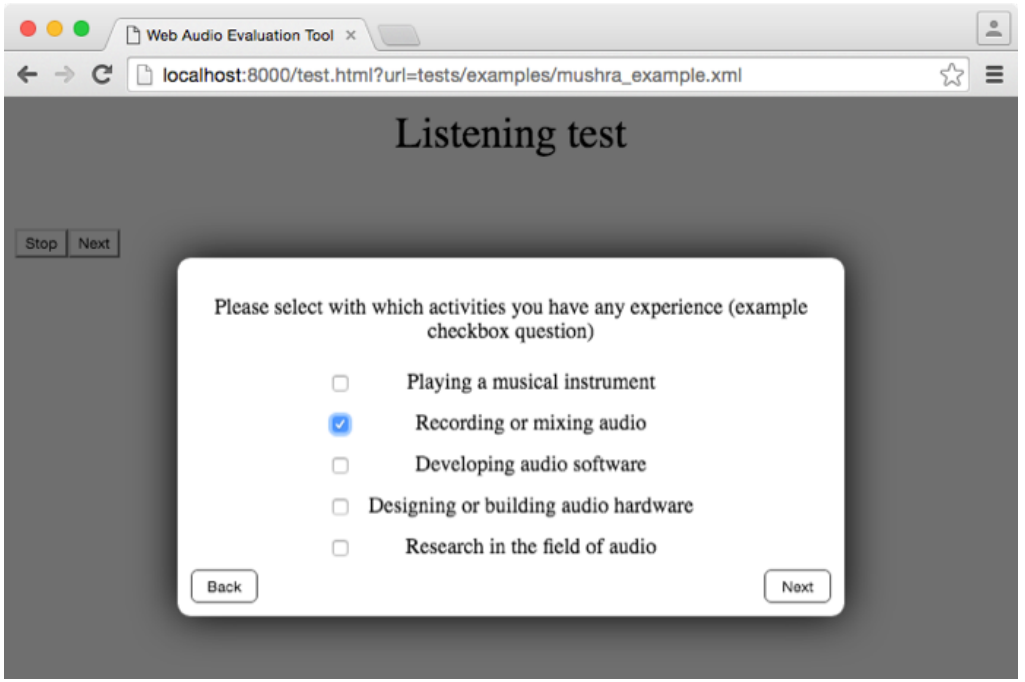


- Spot errors early
- Exclude subjects
- Insights to better design tests
- Statistics for experiment report

centre for digital music

Queen Mary University of London

Surveys



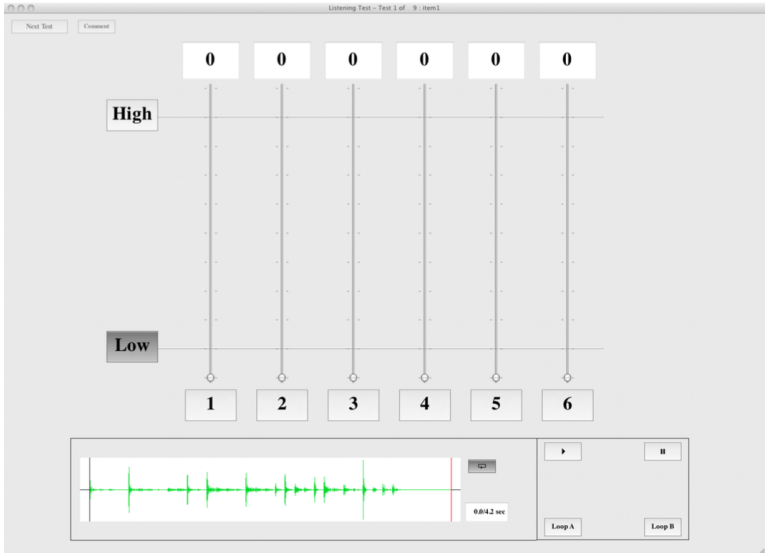
- Immediately before/after test/page
  - Avoid chasing subjects afterwards
  - Some questions might relate to the test

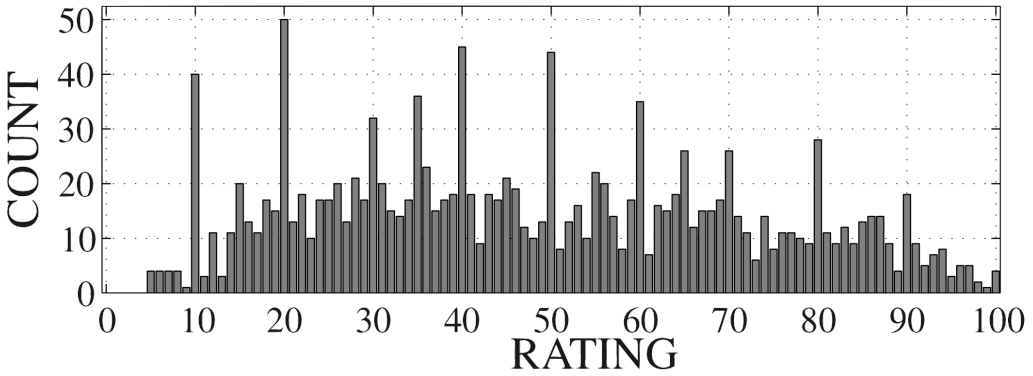
Comments

- Learn more than simple, single-attribute ratings
- Understand how subjects used the scale
- "People get frustrated when they don't have comment boxes" (Sean Olive, AES140 Workshop)
- Separate comment boxes (one for each stimulus) instead of one
  - Proportion of stimuli commented on: 96.5% instead of 82.1%
  - Comments 47% longer

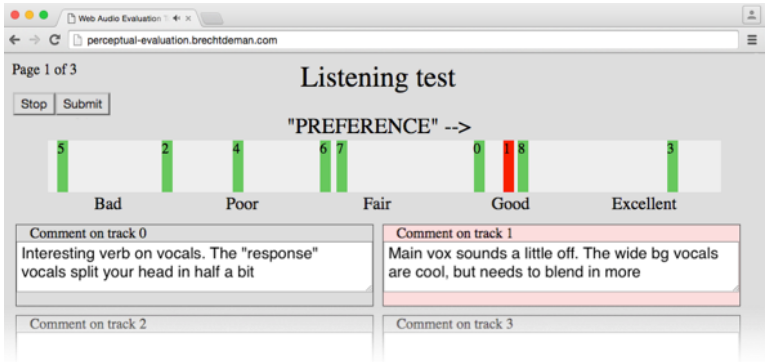
[4] Brecht De Man and Joshua D. Reiss, "APE: Audio perceptual evaluation toolbox for MATLAB," 136th Convention of the Audio Engineering Society, April 2014.

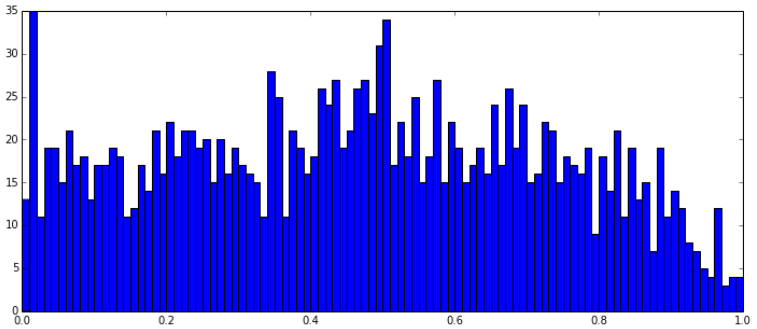
Axis ticks





[10] Jouni Paulus, Christian Uhle and Jürgen Herre, "Perceived Level of Late Reverberation in Speech and Music," AES Convention 130, 2011.





centre for digital music

Queen Mary University of London

Table of Contents

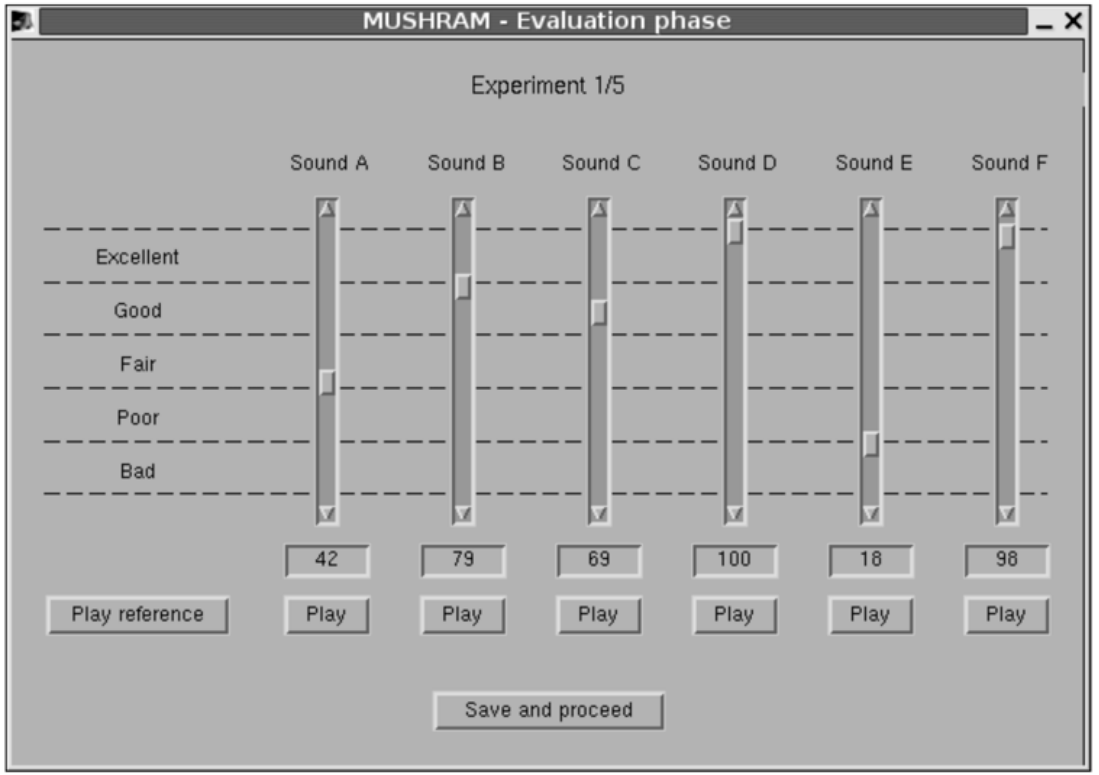
Multiple stimulus or pairwise?

"Multiple comparison methods produce more reliable and discriminating judgments of sound quality compared to single stimulus or paired comparison methods."
[11] S. Olive and T. Welti, "The relationship between perception and measurement of headphone sound quality," in Audio Engineering Society Convention 133, October 2012.

"Multiple stimulus tests are substantially less time-consuming."
[12] B. De Man and J. D. Reiss, "A pairwise and multiple stimuli approach to perceptual evaluation of microphone types," in Audio Engineering Society Convention 134, May 2013.

Multiple stimulus only viable for limited number of stimuli to be compared against each other.

To MUSHRA or not to MUSHRA?



[13] Method for the subjective assessment of intermediate quality level of coding systems. Recommendation ITU-R BS.1534-1, 2003.
[14] E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary guidelines for subjective evalutation of audio source separation algorithms," in UK ICA Research Network Workshop, 2006.

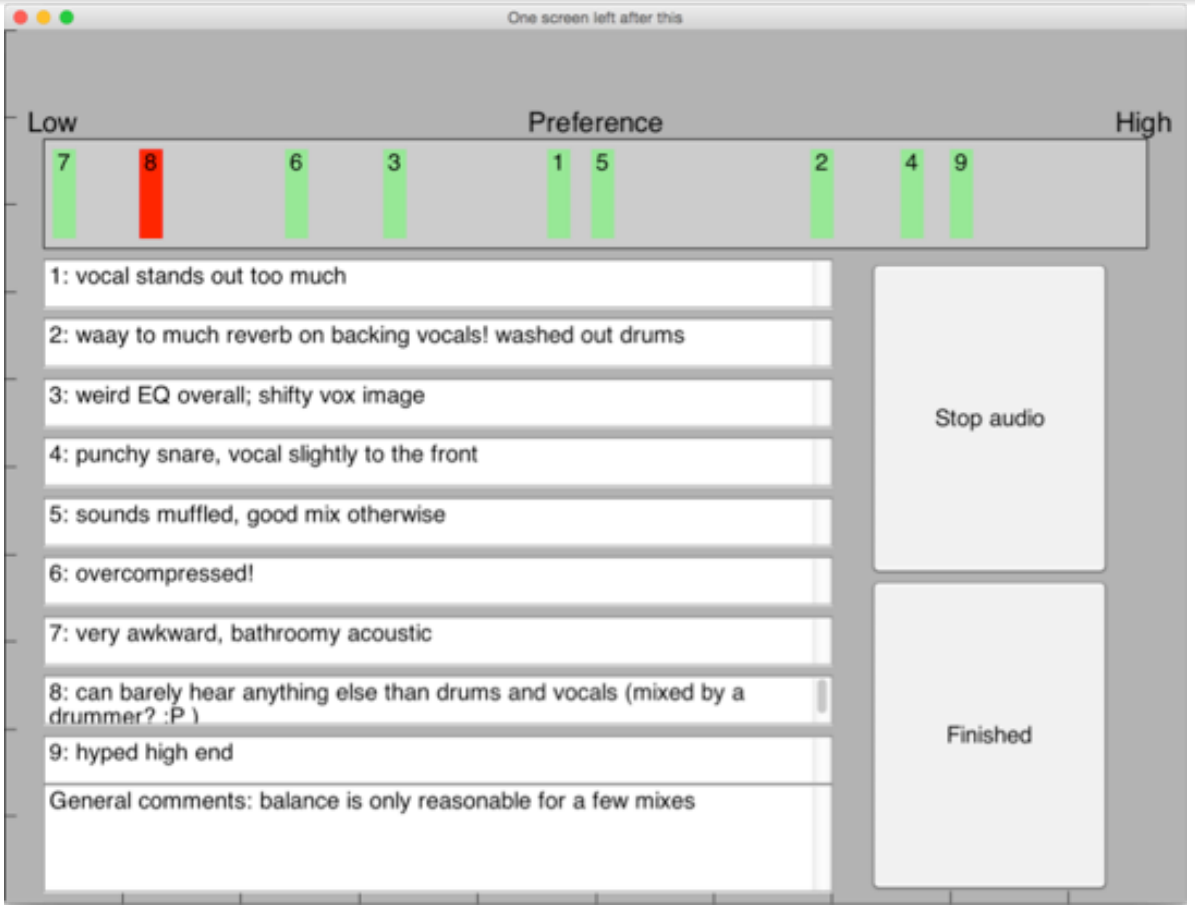
To MUSHRA or not to MUSHRA?

- "[MUSHRA aims] to establish the detectability, magnitude of impairment or basic audio quality of the codec relative to a reference. [It does] not attempt to measure the listeners' preference for the audio codec when compared to the hidden lossless reference."

[16] Sean E. Olive, "Some new evidence that teenagers and college students may prefer accurate sound reproduction," in Audio Engineering Society Convention 132, April 2012.

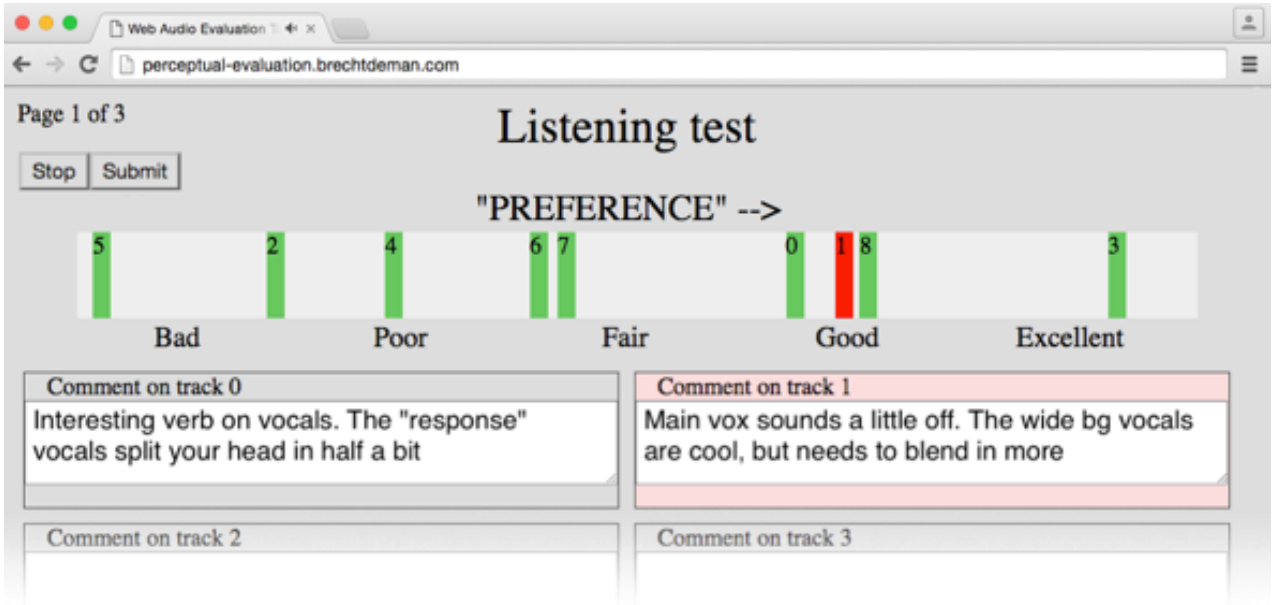
- What if there is no reference to compare to?

To MUSHRA or not to MUSHRA?



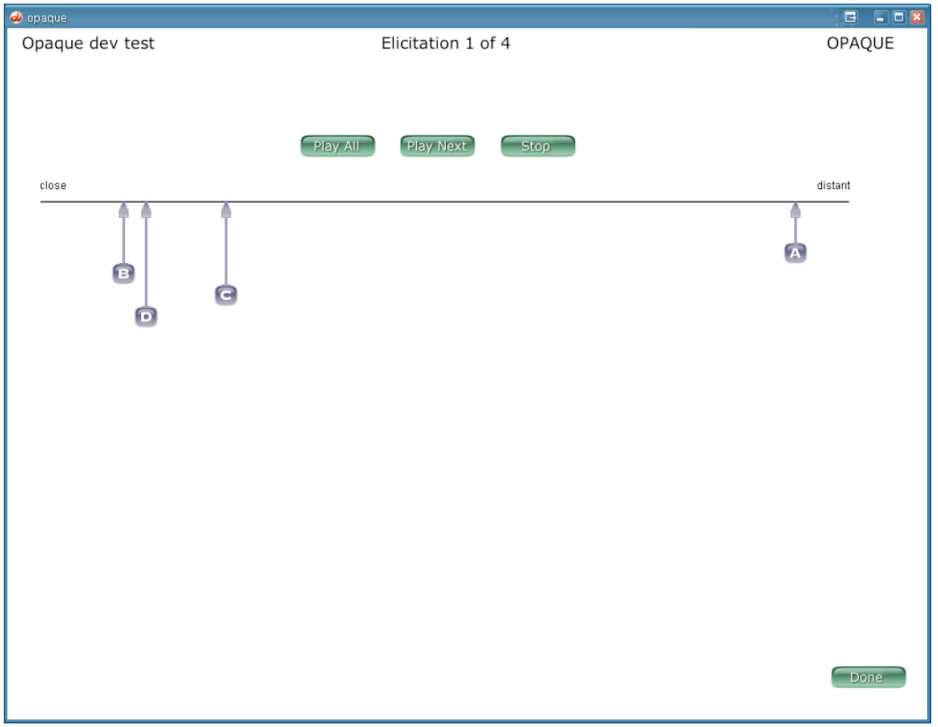
[4] Brecht De Man and Joshua D. Reiss, "APE: Audio perceptual evaluation toolbox for MATLAB," 136th Convention of the Audio Engineering Society, April 2014.

To MUSHRA or not to MUSHRA?



[17] Nicholas Jillings, Brecht De Man, David Moffat and Joshua D. Reiss, "Web Audio Evaluation Tool: A browser-based listening test environment," 12th Sound and Music Computing Conference, July 2015.

To MUSHRA or not to MUSHRA?



[18] Jan Berg, "OPAQUE - A tool for the elicitation and grading of audio quality attributes," AES Convention 118, 2005.

Constraints on use of scale

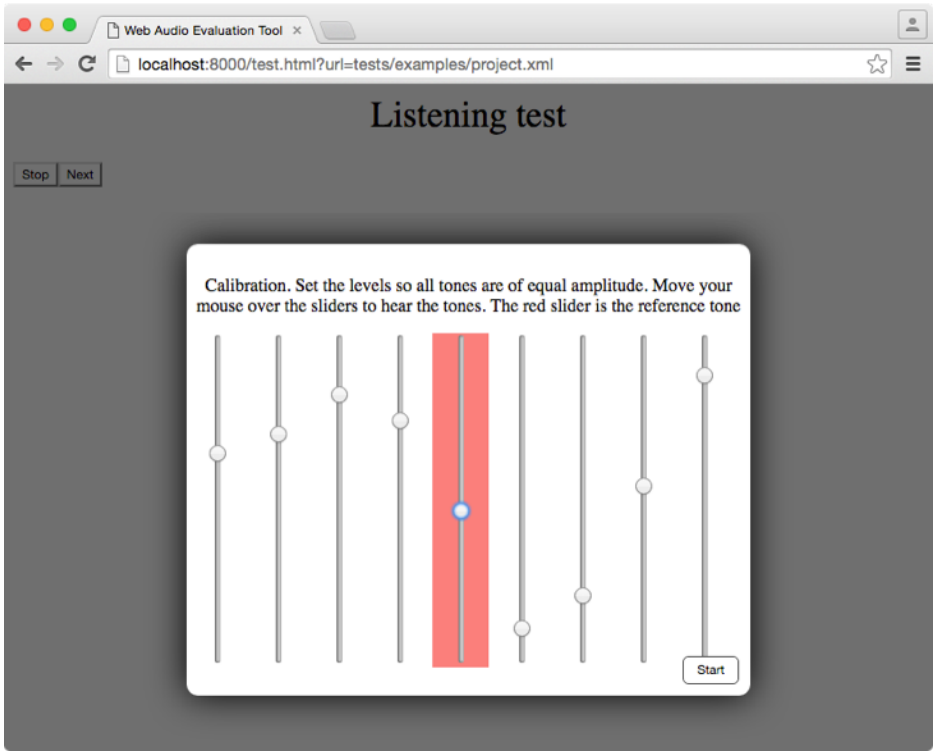
Examples

- "Put at least one slider at 100%"
- "The anchor needs to be below 20%"
- ...

Options

- Instructions
- Enforced (not possible to proceed)
- Remove subject from test
- Normalise ratings after the fact

Calibration

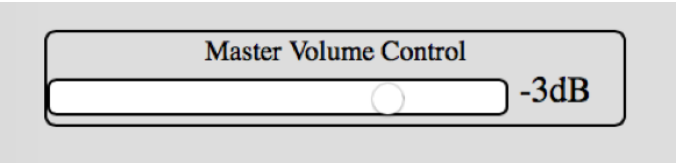


- Audiometric test
- Test of playback system

centre for digital music

Queen Mary University of London

Listening level



Options

- Everyone at same level
- Choose most comfortable/familiar level
  - Set once at start
  - Change throughout test

Remote tests

- Pro: Low effort, scales easily, different locations/cultures/languages within reach
- Con: Loss of control **... or is there?**

[19] M. Cartwright, B. Pardo, G. Mysore and M. Hoffman "Fast and Easy Crowdsourced Perceptual Audio Evaluation," ICASSP, 2016.
[20] M. Schoeffler, F.-R. Stöter, H. Bayerlein, B. Edler and J. Herre, "An Experiment about Estimating the Number of Instruments in Polyphonic Music: A Comparison Between Internet and Laboratory Results," ISMIR, 2013.
[21] F.-R. Stöter, M. Schoeffler, B. Edler and J. Herre, "Human Ability of Counting the Number of Instruments in Polyphonic Music," Meetings on Acoustics Vol. 19, 2013.

Online tests

Why online?

- Cross-platform
- Centralised results collection (over web or local server)
- Multiple machines at once
- Leveraging other web technologies

Source: sewellsupport.com

Online tests

Why online?

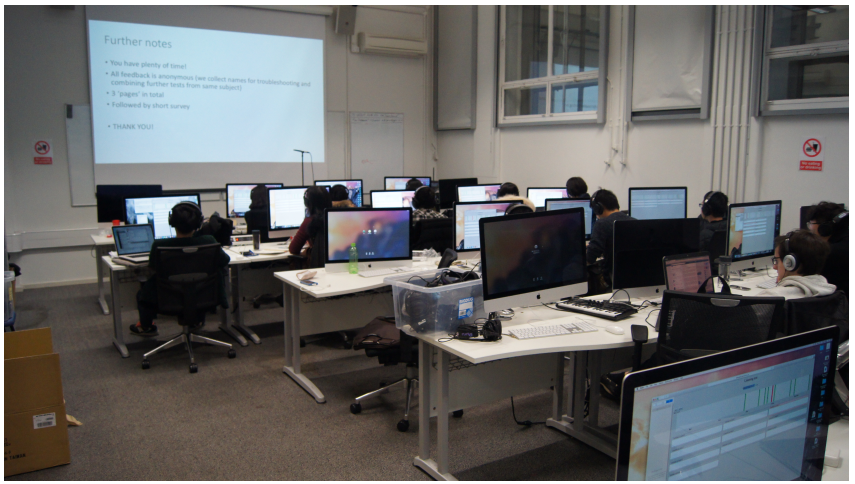
- Cross-platform
- Centralised results collection (over web or local server)
- Multiple machines at once
- Leveraging other web technologies

Source: jaxonraye.com

Online tests

Why online?

- Cross-platform
- Centralised results collection (over web or local server)
- Multiple machines at once
- Leveraging other web technologies



Online tests

Why online?

- Cross-platform
- Centralised results collection (over web or local server)
- Multiple machines at once
- Leveraging other web technologies

Source: LinkedIn.com

Conclusion

Try it!

- Conduct an experiment
- Contribute (`github.com/BrechtDeMan/WebAudioEvaluationTool`)
- Any problems? Feedback?
  - Report issue
  - `b.deman@qmul.ac.uk`

AES 2nd Workshop on Intelligent Music Production

- Tuesday 13 September 2016
- Charterhouse Square, Queen Mary University of London
- Short (<2 pages) paper submission deadline: 1 August 2016
- Confirmed speakers:
  - Bryan Pardo (Northwestern University)
  - François Pachet (Sony CSL Music Paris)
- More info and registration: `aes-uk.org/wimp`

References

A. Eklund, T. E. Nichols, and H. Knutsson, "Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates," *Proceedings of the National Academy of Sciences*, 2016.

B. et al., "Neural correlates of interspecies perspective taking in the post-mortem atlantic salmon: An argument for proper multiple comparisons correction," *Journal of Serendipitous and Unexpected Results*, 2010.

S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons, 2007.

B. De Man and J. D. Reiss, "APE: Audio Perceptual Evaluation toolbox for MATLAB," in *Audio Engineering Society Convention 136*, April 2014.

F. E. Toole and S. Olive, "Hearing is believing vs. believing is hearing: Blind vs. sighted listening tests, and other interesting things," in *Audio Engineering Society Convention 97*, November 1994.

J. Berg and F. Rumsey, "Spatial attribute identification and scaling by repertory grid technique and other methods," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, March 1999.

J. Mycroft, J. D. Reiss, and T. Stockman, "The influence of graphical user interface design on critical listening skills," *Sound and Music Computing (SMC), Stockholm*, 2013.

H. Haas, "The influence of a single echo on the audibility of speech," *J. Audio Eng. Soc*, vol. 20, no. 2, pp. 146–159, 1972.

J. Berg and F. Rumsey, "Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction," in *Audio Engineering Society Convention 109*, September 2000.

J. Paulus, C. Uhle, and J. Herre, "Perceived level of late reverberation in speech and music," in *Audio Engineering Society Convention 130*, May 2011.

References (cont.)

S. Olive and T. Welti, "The relationship between perception and measurement of headphone sound quality," in *Audio Engineering Society Convention 133*, October 2012.

B. De Man and J. D. Reiss, "A pairwise and multiple stimuli approach to perceptual evaluation of microphone types," in *Audio Engineering Society Convention 134*, May 2013.

Method for the subjective assessment of intermediate quality level of coding systems.
Recommendation ITU-R BS.1534-1, 2003.

E. Vincent, M. G. Jafari, and M. D. Plumbley, "Preliminary guidelines for subjective evalutation of audio source separation algorithms," in *UK ICA Research Network Workshop*, 2006.

C. Völker and R. Huber, "Adaptions for the MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA) for elder and technical unexperienced participants," in *DAGA 2015 Nürnberg*, 2015.

S. E. Olive, "Some new evidence that teenagers and college students may prefer accurate sound reproduction," in *Audio Engineering Society Convention 132*, April 2012.

N. Jillings, D. Moffat, B. De Man, and J. D. Reiss, "Web Audio Evaluation Tool: A browser-based listening test environment," in *12th Sound and Music Computing Conference*, July 2015.

J. Berg, "OPAQUE – A tool for the elicitation and grading of audio quality attributes," in *118th Convention of the Audio Engineering Society*, May 2005.

M. Cartwright, "Fast and easy crowdsourced perceptual audio evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

M. Schoeffler, F.-R. Stöter, H. Bayerlein, B. Edler, and J. Herre, "An experiment about estimating the number of instruments in polyphonic music: A comparison between internet and laboratory results," in *14th International Society for Music Information Retrieval Conference (ISMIR 2013)*, pp. 389–394, 2013.

References (cont.)

F.-R. Stöter, M. Schoeffler, B. Edler, and J. Herre, "Human ability of counting the number of instruments in polyphonic music," in *Proceedings of Meetings on Acoustics*, vol. 19, p. 035034, Acoustical Society of America, 2013.

Sound sources

- Fredy V: “In The Meantime” 
- ‘counting.wav’ by Corsica_S (Freesound) 

Q&A

b.deman@qmul.ac.uk

www.brechtdeman.com

@BrechtDeMan

github.com/BrechtDeMan/WebAudioEvaluationTool

linkedin.com/in/BrechtDeMan

aes.org/aes/BrechtDeMan

centre for digital music

Queen Mary University of London