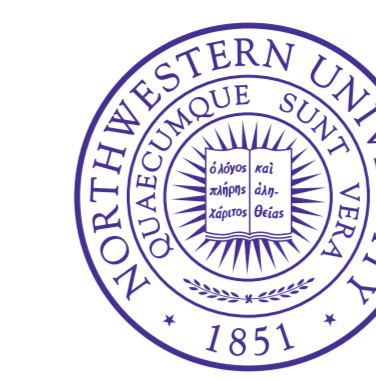


Crowdsourced Pairwise Comparison for Source Separation Evaluation



NEW YORK UNIVERSITY



NORTHWESTERN UNIVERSITY



Adobe

Mark Cartwright¹, Bryan Pardo², Gautham Mysore³

1. Music and Audio Research Lab, New York University

2. Interactive Audio Lab, Northwestern University

3. Creative Intelligence Lab, Adobe Research

Please address correspondence to: mark.cartwright@nyu.edu

1. Introduction

Automated objective methods of audio source separation evaluation are fast, cheap, and require little effort by the investigator; but their output often correlates poorly with human quality assessments.

Subjective multi-stimulus listening tests are the gold standard for audio evaluation, but they are slow and onerous to run.

Our previous work showed that a crowdsourced multi-stimulus listening test can produce results comparable to lab-based multi-stimulus tests [1], but they are limited to evaluating 12 or fewer stimuli and require ground-truth stimuli for reference.

We present a web-based pairwise-comparison listening test for source separation evaluation that addresses these limitations while still promising to speed and facilitate conducting listening tests. We compare to multi-stimulus lab- and web-based tests (referred to as lab-MS and web-MS)

2. Baseline Dataset

PEASS Dataset [2]

10 mixtures (5 music, 5 speech)

5 sec long w/ 2 - 7 sources each

8 test stimuli per mixture:

- Reference
- 3 anchors
- 4 source separation algorithm outputs

MUSHRA multi-stimulus evaluations from 20 experts on 4 quality scales

3. Listening Test Procedure

Participants were recruited from Amazon's Mechanical Turk

Each participant was limited to one quality scale and could perform up to 10 trials

We collected at least 30 trials per condition (mixture / quality pair)

Steps:

Participants completed a quick hearing evaluation

Participants completed a training phase

For each trial, participants compared all pairs in a set: $\binom{8}{2}$ i.e., 28 pairs

For each pair, participants choose which of two stimuli is higher on a quality scale

Payment: \$0.80 for first trial, \$0.50 for subsequent trials. Up to \$0.25 bonus per trial based on consistency.

Quality scales:

Overall quality

Preservation of the target source

Suppression of other sources

Absence of additional artificial noises (additive artifacts)

Preservation of the target source (subtractive artifacts)

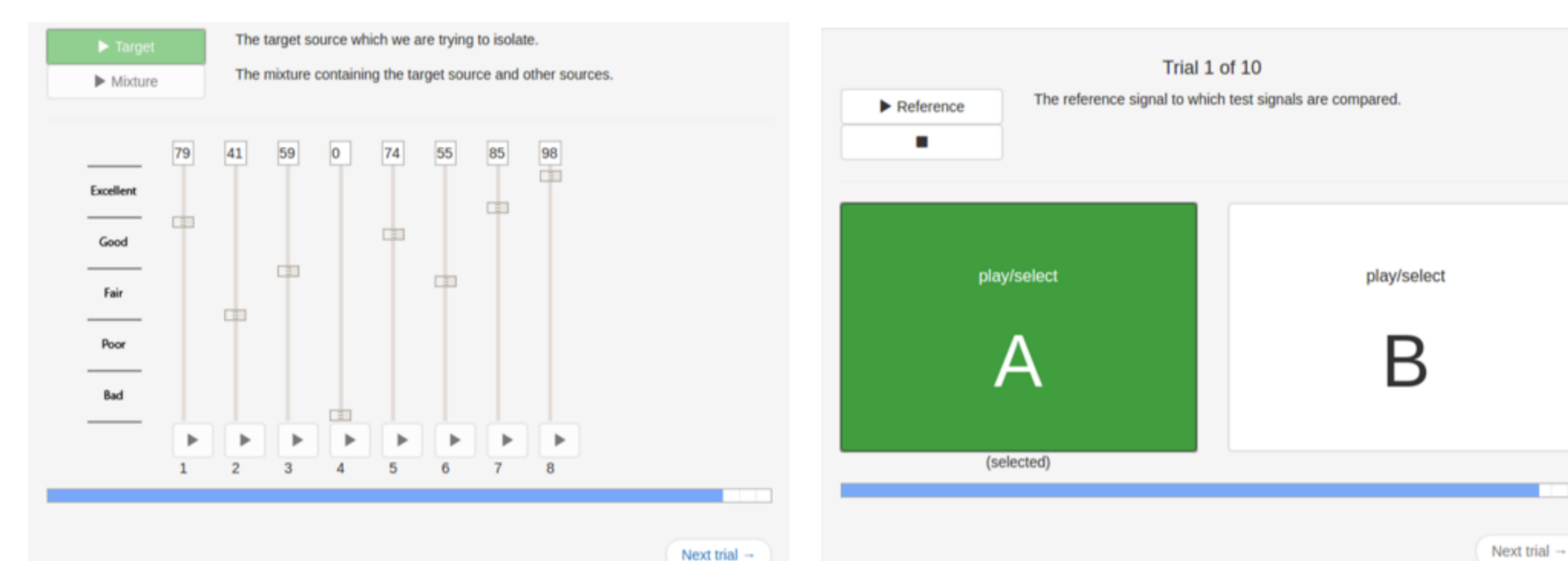
Lack of distortions to the target source (additive and subtractive artifacts)

* new quality scale added to address confusion between additive and subtractive artifacts.

Table 1. Listening tests details

	Web-based Multi-stimulus	Web-based Pairwise Comparison
# of Participants	530	458
# of participants that passed hearing screening	336	345
# of Trials	1763	1444
Mean trials per condition	34	30
Mean trials per participant	3.3	3.2

Figure 1. multi-stimulus (left) and pairwise (right) interfaces*



4. Quality Score Estimation

We used a Thurstone model to estimate quality scores from pairwise preferences. The basic Thurstone model is as follows:

$$S_n \sim \text{Normal}(\mu_n, \sigma^2), \text{ for } n \in 1 : N$$

$$\Pr(a_i \succ a_j) = \Pr(S_i > S_j), \text{ for } i, j \in 1 : N ; i \neq j$$

$$= \Pr(S_i - S_j > 0)$$

$$= \Phi\left(\frac{\mu_i - \mu_j}{\sigma\sqrt{2}}\right)$$

where for N items, S_n are the quality scale values with measurement error and μ_n are the latent quality scores. a_i and a_j are the two items in a paired comparison.

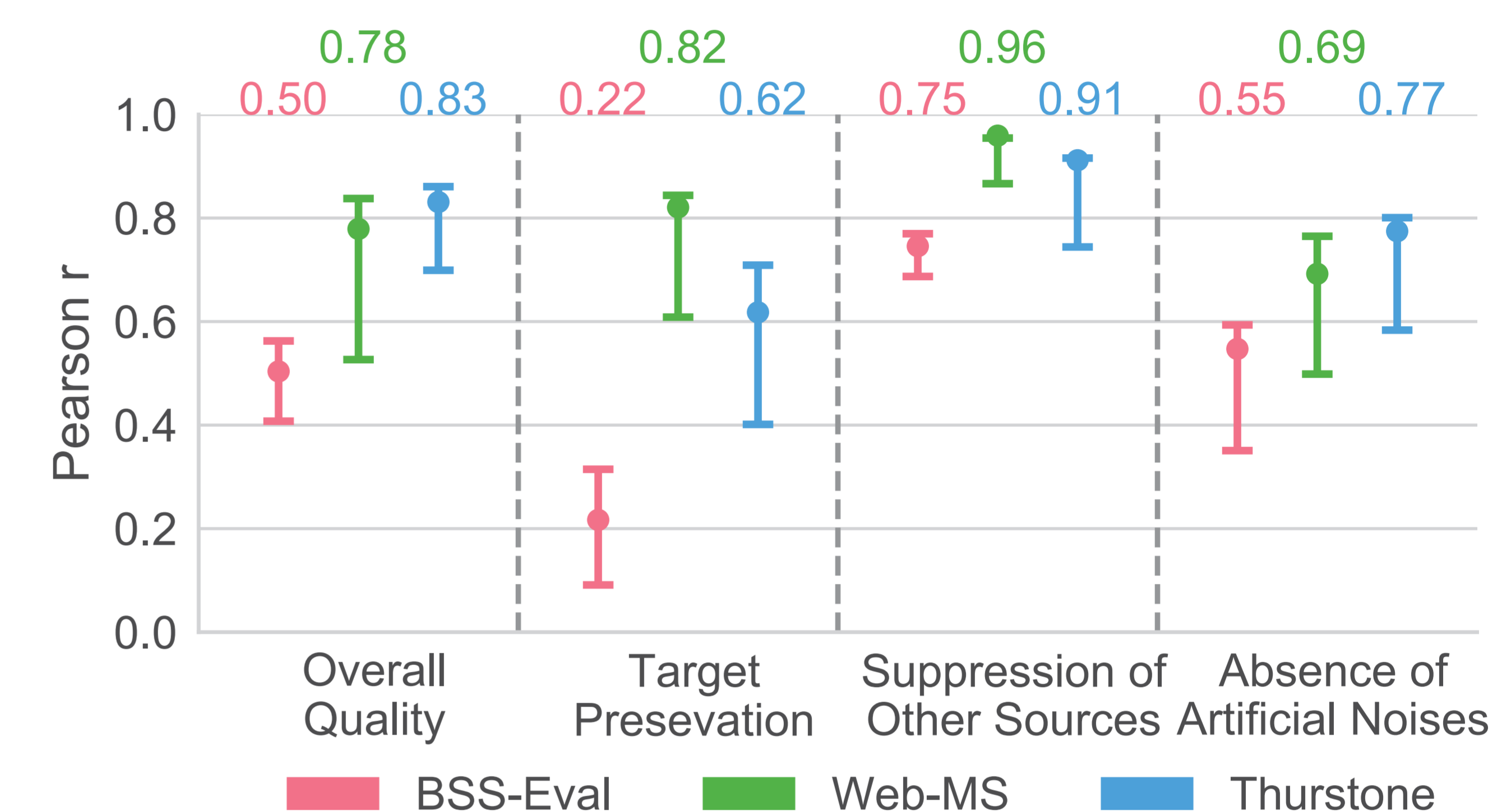
Using this model, we fit the likelihood of our data for each quality scale using MCMC sampling (NUTS) and with priors chosen so that the resulting scores are on the same scale as the multi-stimulus scores.

5. Results

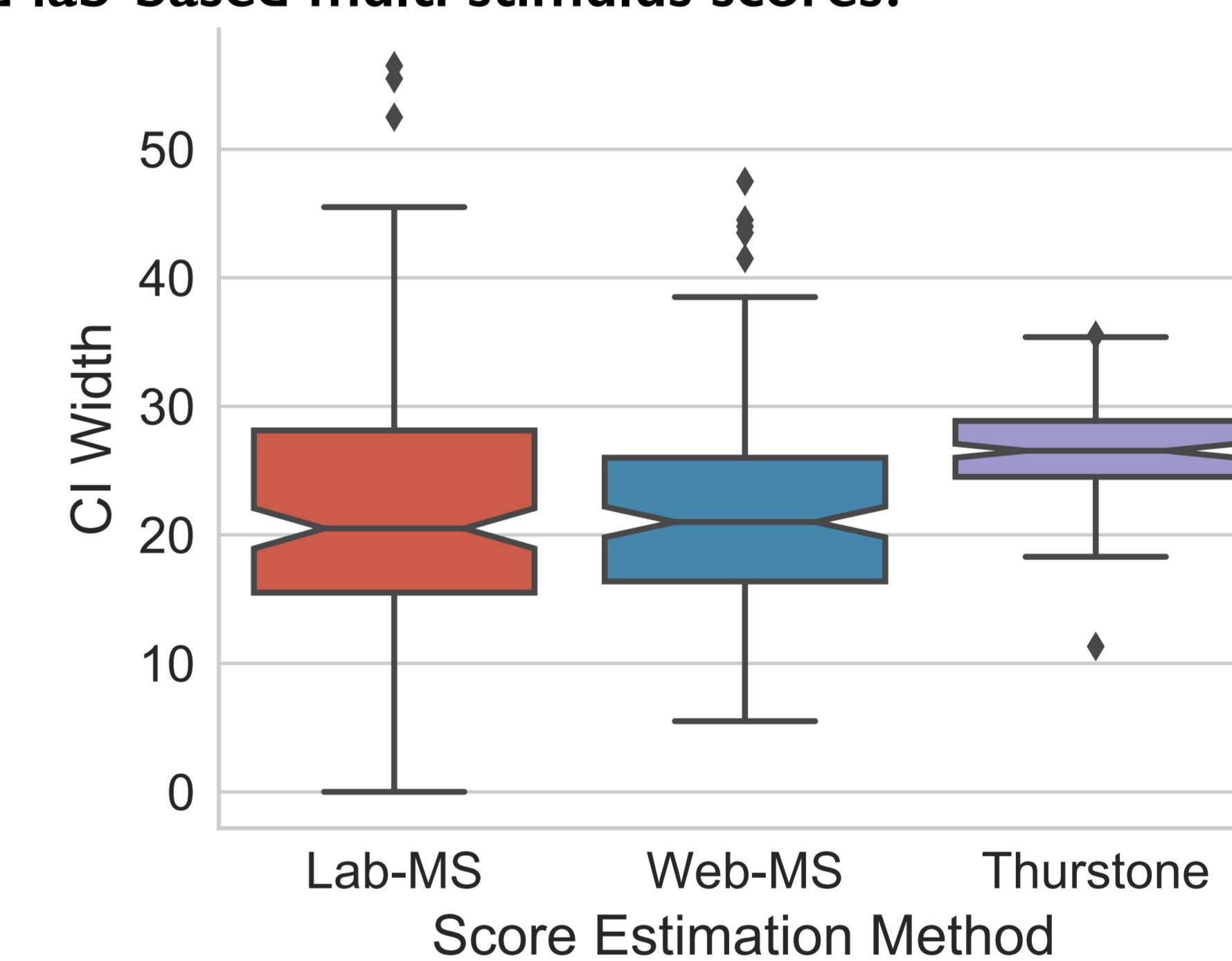
Table 2. Mean Pairwise Transitivity Statistics (N=10)

Quality Scale	Transitivity Satisfaction Rate	Weak Stochastic Transitivity	Medium Stoch. Transitivity	Strong Stoch. Transitivity
Overall Quality	0.91	0.97	0.93	0.61
Target Preservation	0.90	0.97	0.95	0.71
Suppression of Other Sources	0.92	0.99	0.94	0.60
Absence of Additional Artificial Noises	0.91	0.99	0.98	0.71
Lack of Distortion to the Target Source	0.93	1.00	0.99	0.73

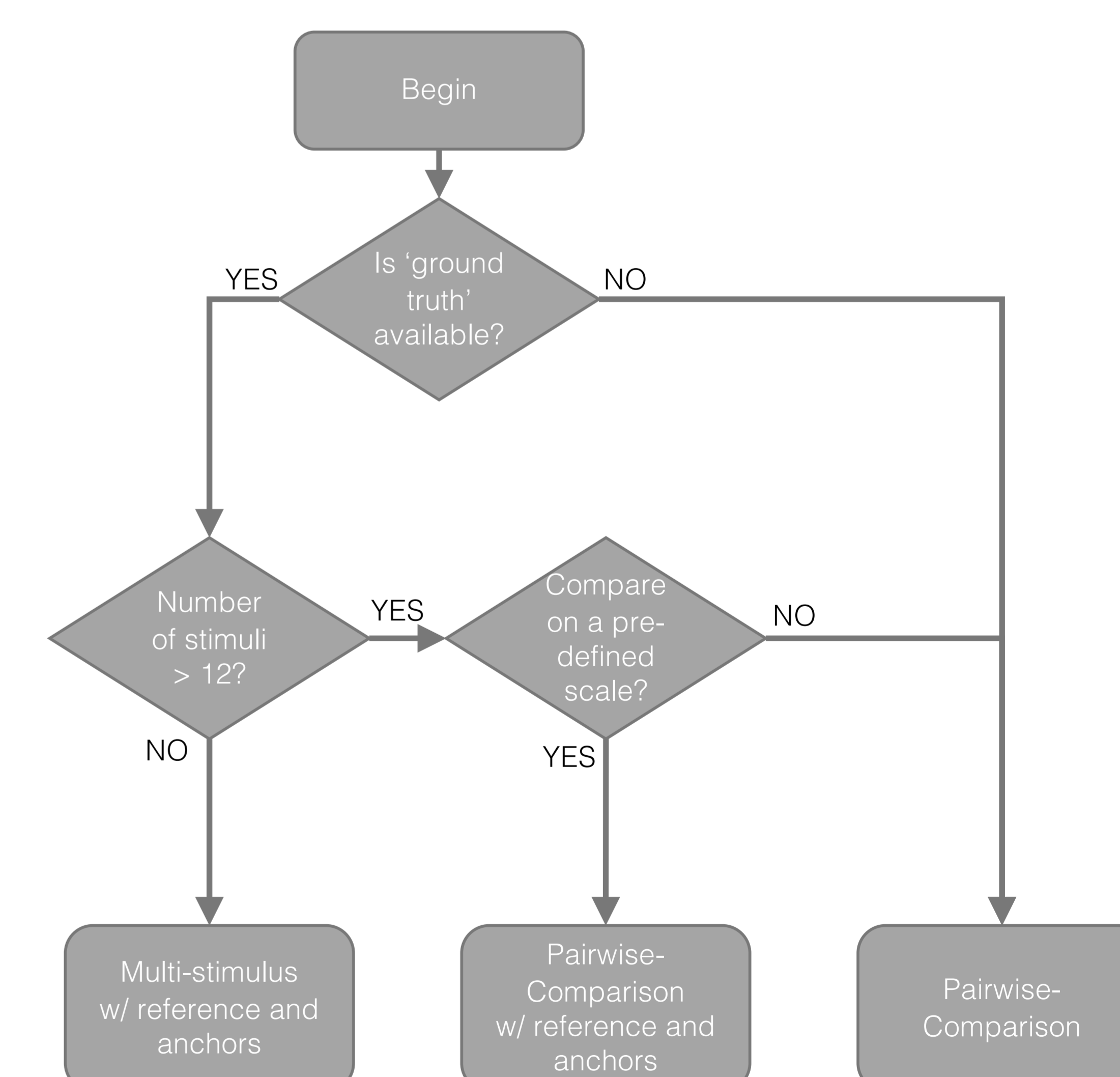
How do the scores from BSS-Eval, web-based pairwise comparison, and web-based multi stimulus tests correlate to lab-based multi stimulus?



Are the web-based pairwise-comparison scores noisier than web- and lab-based multi stimulus scores?



Which test should I use?



References

- [1] M. Cartwright, B. Pardo, G. Mysore, M. Hoffman. Fast and Easy Crowdsourced Perceptual Audio Evaluation. In Proc. of ICASSP, 2016.
- [2] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and Objective Quality Assessment of Audio Source Separation," IEEE TASLP, vol. 19, pp. 2046-2057, 2011.