Headphone and Loudspeaker Screening for Web-Based Auditory Experiments: Suggestions for a Reliable Estimation of Data Quality and Sample Size

Kilian Sander (1), Yves Wycisk (1), Reinhard Kopiez (1), Benedetto Manca (2), Friedrich Platz (3)

(1) Hanover University of Music, Drama and Media, Hanover, Germany
(2) University of Cagliary, Cagliary, Italy
(3) State University of Music and Performing Arts, Stuttgart, Germany

Background

Internet experiments on auditory perception often require specific playback devices, such as headphones. Although suggestions for screening methods already exist (Woods et al., 2017), their practical application does not consider the prevalence of playback devices. Unfortunately, the proportion of headphones to loudspeakers seems to be unknown. Additionally, in line with the standards of epidemiology, reliable information on sensitivity (true positive rate) and specificity (true negative rate) are required to evaluate screening procedures. In the current state of research, the assessment of correctly identified playback devices (data quality) based on screening methods is unclear.

Aims

Our primary aim was to develop a reliable screening method for detecting headphones and loudspeakers as playback devices. We wanted to provide an online tool to calculate application-oriented data quality and the required sample size for web-based surveys, which also considers both the prevalence of playback devices and the test procedure metrics.

Methods

In a laboratory study (*N* = 40), the headphone screening test suggested by Woods et al. (2017), two other self-developed tests based on interaural time differences (Bilsen & Raatgever, 2002), and the Franssen effect on sound localization (Franssen, 1960) were evaluated with three different types of playback devices: (1) headphones (circumaural/intra-aural), (2) loudspeakers, and (3) a laptop (built-in speakers). Each of the screening tests consisted of six items/tasks. Subsequently, the screening procedures were tested in an internet study (*N* = 211). To ensure trustworthy responses, we did not give any information about the study’s intention. Various control procedures (e.g., timeout, attention testing) were used to guarantee high data quality. Participants using tablets, smartphones, and built-in speakers of monitors/TVs were excluded from the survey.

Results

Considering all playback devices (*N* = 1194), the headphone prevalence was 17.6%. The prevalence in the trimmed data set (*N* = 211, excluding tablets, smartphones, monitors/TVs) was 37.9 %. As each test consisted of six items, the possible thresholds for the classification as a headphone within a test, that is, the minimum number of correct answers, ranged from 1 to 6. To combine the individual results of the three tests into one classification, we compiled different combination methods: for example, ensemble methods from machine learning. 2,178 values of sensitivity and specificity were calculated for all tests and combination methods for all possible thresholds from the trimmed data set using a k-fold cross-validation method (Stone, 1974). According to the decision theory, the best test or test combination for a certain application, that is, a given prevalence, has the highest overall utility, which is calculated from the test metrics and the prevalence (Treat & Viken, 2012). Our calculator computes the overall utility for the decision goal of *maximizing percent correct* for a given prevalence and returns the test or test combination with the highest value. As in a power analysis, the required sample size and the appropriate test to be used can be determined from (a) the minimum number of target devices, (b) their certainty and (c) the prevalence of playback devices. The online tool is part of the *Headphone and Loudspeaker Test* (HALT) R-package (https://github.com/klausfrieler/HALT).

Conclusion and Implications

The low prevalence of headphones used in web-based experiments indicates the central role of highly sensitive and specific screening methods. Considering the standards of epidemiology, it is insufficient to focus solely on sensitivity and specificity without obtaining information on device prevalence. Our findings can help to improve the data quality and efficiency of future studies.

References

Bilsen, F., & Raatgever, J. (2002). *Demonstrations of dichotic pitch* [CD].

Franssen, N. V. (1960). *Some considerations on the mechanism of directional hearing* [Doctoral dissertation]. Technische Hogeschool.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Treat, T. A., & Viken, R. J. (2012). Measuring test performance with signal detection theory techniques. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.* (pp. 723–744). American Psychological Association. https://doi.org/10.1037/13619-038

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7). https://doi.org/10.3758/s13414-017-1361-2