Richter, M, Buhiyan, T, Bramslow, L, Innes Brown, H, Fiedler, L, Hadley, LV, Naylor, G, Saunders, GH, Wendt, D, Whitmer, W, Zekveld, AA and Kramer, SE

Combining Multiple Psychophysiological Measures of Listening Effort: Challenges and Recommendations

The published version of this article can be found at https://doi.org/10.1055/s-0043-1767669.

Combining Multiple Psychophysiological Measures of Listening Effort: Challenges and Recommendations

Michael Richter[1,a], Tanveer Buhiyan [2,b], Lars Bramsløw[3,c], Hamish Innes Brown [3,d], Lorenz Fiedler[3,e], Lauren V. Hadley[4,f], Graham Naylor [4,g], Gabrielle H. Saunders[5,h], Dorothea Wendt[3,6,i], William Whitmer [4,j], Adriana A. Zekveld[7,k], Sophia E. Kramer[7,l]

[1]School of Psychology, Liverpool John Moores University, Liverpool, UK

[2]R&D, Demant A/S, Kongebakken, Denmark

[3]Eriksholm Research Centre, Oticon A/S, Snekkersten, Denmark

[4]Hearing Sciences, School of Medicine, University of Nottingham, Nottingham, UK

[5]Manchester Centre for Audiology and Deafness, University of Manchester, Manchester, UK

[6]Department of Health Technology, Technical University of Denmark, Lyngby, Denmark

[7]Amsterdam UMC, Vrije Universiteit Amsterdam, Otolaryngology – Head and Neck Surgery, Ear & Hearing, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands

[a]Corresponding author. Dr.phil., Reader in Motivation Psychology. Mailing address: Byrom Street, Liverpool, L3 3AF, UK. E-mail: m.richter@ljmu.ac.uk. Telephone: +41 151 231 2220.

[b]PhD, tabh@demant.com.

[c]PhD, labw@eriksholm.com.

[d]PhD, hain@eriksholm.com.

[e]PhD, lfie@eriksholm.com.

[e]PhD, Lauren.Hadley1@nottingham.ac.uk.

[e]PhD, Professor of Hearing Sciences, Graham.Naylor@nottingham.ac.uk.

[e]PhD, gabrielle.saunders@manchester.ac.uk.

[e]PhD, dowe@eriksholm.com.

[e]PhD, bill.whitmer@nottingham.ac.uk.

[e]PhD, aa.zekveld@amsterdamumc.nl.

[e]PhD, Professor of Auditory Functioning and Participation,

se.kramer@amsterdamumc.nl.

Abstract

About one third of all recently published studies on listening effort have used at least one physiological measure, providing evidence of the popularity of such measures in listening effort research. However, the specific measures employed, as well as the rationales used to justify their inclusion, vary greatly between studies, leading to a literature that is fragmented and difficult to integrate. A unified approach that assesses multiple psychophysiological measures justified by a single rationale would be preferable because it would advance our understanding of listening effort. However, such an approach comes with a number of challenges, including the need to develop a clear definition of listening effort that links to specific physiological measures, customized equipment that enables the simultaneous assessment of multiple measures, awareness of problems caused by the different timescales on which the measures operate, and statistical approaches that minimize the risk of type-I error inflation. This paper discusses in detail the various obstacles for combining multiple physiological measures in listening effort research and provides recommendations on how to overcome them.

Keywords: Listening effort; psychophysiological measure;

Conflict of Interest

There are no conflicts of interest, financial, or otherwise.

Combining Multiple Psychophysiological Measures of Listening Effort: Challenges and Recommendations

Assessing physiological measures in listening effort research is common. Between 2019 and 2021, Clarivate's Web of Science database lists a total of 239 articles with the term "listening effort" in the title, abstract, or keywords. Amongst these articles, 35% (81) employed at least one physiological measure to examine listening effort; 7% (16) employed more than one physiological measure. The variety of measures used was large, and included measures directly indexing brain activity, such as electroencephalogram (EEG) alpha oscillations[1,2], EEG evoked potential components[3,4], and functional near-infrared spectroscopy (fNIRS)[5,6], and peripheral measures, such as skin conductance [7,8], pupil diameter [9,10], heart rate variability [11,12], and cardiovascular pre-ejection period [12,13]. The reason for the particular measures used seemed to be driven more by the researchers' interest and availability of measurement equipment than by a theoretical or conceptional rationale.

Given the lack of a unifying rationale, the heterogeneity in the employed measures constitutes a problem: It makes it difficult for listening effort researchers to decide which measure to use, to compare findings across studies involving different measures, and to draw straightforward conclusions from the existing literature[14]. Ultimately, a unifying rationale would boost theoretical progress and advance our understanding of the determinants, consequences, and mechanisms associated with listening effort. A more comprehensive approach that systematically integrates multiple physiological measures could be particularly useful when studying listening effort. However, there are a number of practical challenges to combining

more than a single physiological measure in a listening effort study. The purpose of this article is to highlight some of these challenges and to provide recommendations on how to address them. We hope that this will help listening effort researchers to develop a more integrative, unified approach to using physiological measures and thereby accelerate the advancement of our understanding of listening effort. Our discussion strongly draws on the experience that we have gained in the context of the HEAR-ECO project (http://hear-eco.eu/) in which we employed several physiological measures to examine listening effort[11,13,15-18]. The topics that we are going to discuss here are 1) the selection of appropriate physiological measures, 2) the simultaneous assessment of multiple physiological signals, 3) the aggregation and combination of simultaneously assessed physiological measures, and 4) the statistical analyses of multiple physiological measures.

Selection of Appropriate Physiological Measures

One of the most challenging aspects of a systematic, integrative approach that uses multiple physiological measures to examine listening effort is to find a good rationale for selecting the measures. For almost any common physiological measure, it is possible to find at least one publication where the authors associate the measure with listening effort or related constructs like effort, engagement, or resource allocation. Finding a published rationale that justifies the use of multiple physiological measures is, however, more difficult. Nonetheless, a unifying rationale seems to be desirable to facilitate the integration of results from different studies. Moreover, the lack of a unifying rationale increases the likelihood of a conflation of concept and measure, which is illustrated by the current discussion about the multiple dimensions of listening effort[7,19-21]. The lack of a unifying rationale linking the concept (listening

effort) to physiological measures makes it difficult to decide whether the discussion is about the dimensions of listening effort or about the dimensions of the measures employed in listening effort research.

The first step may thus be a clear and commonly accepted definition of the concept of listening effort. Without a clear definition of the concept, we will struggle to differentiate it from other phenomena[22]—for instance, to decide whether a listening situation is more effortful or more arousing[11,16]— to find (psychophysiological) measures that appropriately match our concept[22,23], and to build a refined theory of listening effort[24]. Psychophysiological measures can be viewed as proxies to self-report measures of subjectively perceived listening effort—a rating or other type of assessment of the individual's perception of how effortful listening is—which in common language may be viewed as the most meaningful definition of listening effort[25]. Whether or not this can be regarded as the 'ground truth' depends on the experimental setup and on the specific definition of listening effort. The same applies to objective behavioral measures of listening effort such as dual-task measures or delayed recall.

There are at least two approaches to developing a clear definition of a concept, and both have been used in listening effort literature[21]. The first is the empirical observation that a physiological measure responds to variations in an independent criterion variable—for instance, a listening demand-related variable like the signal-to-noise ratio of speech embedded in noise—as evidence that the measure constitutes a correlate of listening effort[26-32]. This implies a definition of listening effort as a state that changes in a predictable way when the level of the criterion variable (for instance, listening demand) changes. For example, it is usually assumed that a measure sensitive to listening effort should indicate relatively high effort in

moderately difficult listening demand conditions, and less effort in low listening demand conditions. Using such a concept definition, any physiological measure that has been demonstrated to respond to variations in the criterion variable would constitute a valid outcome of listening effort[33] and could be included in listening effort studies that employ physiological measures. Listening effort researchers favoring this approach should thus specify their criterion variable(s) and then review the literature to find out which psychophysiological measures respond to changes in it/them. These measures would then constitute the set of physiological measures that could legitimately be used to examine listening effort.

The second approach to define the concept of listening effort is to provide a verbal description of it. For instance, McGarrigle and colleagues[20] defined listening effort as "the mental exertion required to attend to, and understand, an auditory message", Picou and colleagues[34] conceptualized it as "cognitive resources allocated for speech recognition", and Pichora-Fuller and colleagues[35] defined it as "the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a [listening] task". The advantage of such a concept definition is that it avoids the risk of circularity of the criterion-variable approach[21]—the observed empirical relationship between a physiological measure and a listening-effort manipulation is considered to validate the measure as indicator of listening effort and, at the same time, hypotheses about whether the manipulation changes listening effort are tested using the physiological measure. If the concept definition refers to specific self-report or objective behavioral measures of listening effort, these measures provide an efficient way to resolve the circularity problem. For instance, a definition of listening effort as the subjective feeling of investing effort in a listening task would point to a self-report measure of listening

effort as criterion. However, the descriptive approach often requires additional concept definitions to allow the justification of the selection of physiological measures. For instance, it requires an additional operational definition of cognitive resource allocation as changes in pupil diameter to use Picou and colleagues' definition[34] to justify the use of pupil diameter in listening effort research. As far as we know, none of the current theoretical accounts of listening effort offer such a justification of specific physiological measures.

If these additional concept definitions refer to general physiological mechanisms (instead of referring to a specific measure), they offer the justification of multiple physiological measures that is needed for a unified approach to the use of physiological measures in listening effort research. For instance, using the operational definition of mental resource allocation as increased cardiac sympathetic activity in combination with Pichora-Fuller and colleagues' general definition of listening effort[13] would imply that all physiological measures that reflect cardiac sympathetic activity should be included in listening effort research. It is obviously not required to have two levels of concept definitions—a general one of listening effort and an operational one linking listening effort to a physiological mechanism. One could directly use an operational definition of listening effort that refers to physiological measures—for instance, a definition of listening effort as cardiac sympathetic activity in listening tasks[36]. However, including a broad, descriptive concept definition of listening effort probably offers a better integration of the listening effort literature that has not used physiological measures, such as those studies only using self-report or behavioral measures.

Recommendation 1

Use a clear definition of the concept of listening effort that creates an explicit link to the employed physiological measures. Make this definition salient. Other researchers will adopt your concept definition or present conflicting definitions, which will foster a discussion about the listening effort concept. This will hopefully lead over time to a commonly accepted definition of listening effort.

Simultaneous Collection of Multiple Physiological Biosignals

Once the physiological measures of interest have been selected, one needs to collect the biosignals that are required to calculate these measures. One of the most obvious challenges in the simultaneous collection of multiple biosignals is the parallel use of different measurement devices and sensors, which may interfere with one another and may result in discomfort and stress for study participants. For instance, EEG electrodes and fNIRS optodes often need to be placed at similar locations on the participant's head, which may be physically impossible if two separate sensor patches are necessary. EEG and fNIRS sensors may also interfere with the appropriate placement of the electrodes of impedance cardiograph systems (required for the determination of pre-ejection period) that use electrodes on the forehead or behind the ears[37,38]. Another example is the competition of eye tracking glasses and fNIRS and EEG sensors for space on the forehead. In addition to the competition for space, devices may also interfere with one another because of their electromagnetic properties. For instance, the simultaneous use of EEG and fNIRS can induce noise on the EEG signal caused by the electric activity of the fNIRS system[39,40]. Another example is the interference due to the magnetic field of magnetic resonance imaging (MRI) systems that can influence the ECG signal[41,42].

Many of these problems can be avoided by carefully selecting equipment. For instance, there are custom-made hybrid EEG-fNIRs systems that enable the simultaneous assessment of both signals[43,44]. Impedance cardiography and measures that require sensors mounted on the head can be made compatible by using impedance cardiographs with an electrode configuration that does not interfere with the other devices' sensors (for instance, systems that only require electrodes on the thorax and neck[45]). Eye tracking is compatible with head-mounted sensors if a screen-based (remote) eye tracker is used. The problem of MRI artefacts on the ECG signal can be mitigated by using carbon fiber electrodes and leads as well as by employing statistical methods to control for the induced artefacts[41,42]. However, careful planning, customization, and expertise are required for all involved biosignals.

Of course, in field research, the simultaneous measurement of multiple biosignals is highly limited by the need for equipment to be sufficiently unobtrusive and practical so as not to interfere with daily life, while also remaining reliable, valid, and sensitive. This of course is challenging especially when experiments take place over many days and thus require the participants to manage the fitting and charging of equipment at home[46]. However, the rapid development of commercially available mobile sensors might solve some of the issues once these systems have proven to be sufficiently reliable, valid and sensitive[47,48].

The second major problem related to the simultaneous assessment of multiple biosignals is the synchronization of the data. The most frequently used approach is to label the data during the data collection process with event markers and to use these recorded markers to align the different signals offline. However, given that the signals are digitized by separate devices with their own independent clocks, there will be some delays and misalignment

between the signals[39,40]. Moreover, if the signals were originally collected at different sampling rates, down-sampling of the raw signals to one and the same sampling rate may considerably distort the temporal aspects of the signals and introduce misalignment of the signals. A more sophisticated approach to data synchronization is to have one device that controls the sampling of all other devices. There are commercial solutions available, but the device (or software) would probably need to be customized to suit the needs of the specific, individual setup. Moreover, many stand-alone measurement devices are closed systems that do not allow a second device to control their data sampling process.

A researcher aiming to assess multiple physiological measures to examine listening effort thus needs to find a solution for the physical and electromagnetic interference of the employed devices and sensors as well as solve the problem of data synchronization. It may not always be possible to find an ideal solution, but awareness of these potential obstacles will allow for study designs to be optimized.

Recommendation 2

Determine how and whether the selected biosignals will interfere with one another and acquire appropriate specialized equipment accordingly to mitigate any problems caused by the physical and electromagnetic interference of the measurement devices and to attenuate the data synchronization issue. Consider these issues already at the planning stage of projects to ensure that the required financial, logistical, and knowledge-related resources (for instance, for the purchase of integrated measurement systems or for the recruitment of individuals with the expertise to provide custom-made solutions) are available.

Aggregation and Combination of Physiological Measures

Once one has managed to simultaneously sample the required biosignals and to synchronize them, the derived physiological measures must be aggregated and compared. One of the main challenges to this is caused by differences in the time characteristics of the physiological measures used in listening effort research. Continuous measures have a meaningful value at any given point in time and their time resolution is only limited by the quality of the measurement device. For instance, pupil diameter[10] or skin conductance level[29] have one particular value at any given moment and all such values provide meaningful information. In contrast, non-continuous measures either do not exist at some points in time or they cannot be related to one specific point in time in a meaningful manner. For instance, peak pupil diameter refers to a specific point in time when the pupil diameter attains its maximum value in a certain time interval[49]. At all other points in time, peak pupil diameter does not exist. The same applies to specific components of EEG event-related potentials like the P400 amplitude[50] or systolic blood pressure[12], the maximum blood pressure between two consecutive heart beats.

In addition to non-continuous measures that only exist at specific points in time, there are non-continuous measures that refer to specific time periods and can therefore not be associated with a specific point in time. For instance, pre-ejection period[12] refers to the time interval between the onset of the electrical excitation of the left heart ventricle and the opening of the aortic valve. Consequently, it does not exist during other periods of the cardiac cycle[51] and is not associated with one single, specific point in time. Another example is heart period [52], which refers to the time interval between two consecutive heart beats. There are also listening effort measures that are non-continuous because of how they are calculated. For

instance, the determination of EEG alpha[53] and theta power[54] requires the use of epochs to extract the frequency components of interest (for instance, an epoch of 1250 ms would be required for the quantification of theta power[55]). Another example is heart rate variability[29], which also can only be determined by quantifying variability over a certain time period (for instance, one-minute intervals if a Fast Fourier Transform is used to quantify high-frequency heart rate variability[56]). Figure 1 provides an illustration of the variability in the time characteristics of the discussed measures.

Associated with the various time-scales is the difference in baseline interval or nature of the baseline between various measures. For example, pupillometry measures often apply a trial-based baseline correction that is based on the mean pupil size in a relatively short period (e.g., 1000 – 200 msec) prior to stimulus onset[57]. In some studies, this baseline is corrected for the individual dynamic range in the pupil size[58]. On the other hand, the reactivity of cardiovascular measures like pre-ejection period or heart rate variability is often compared to a baseline measured before the onset of the task of interest (during rest)[12].

It should be evident then that aggregating non-continuous and continuous measures is complex. While it is technically possible to treat the non-continuous measure as a continuous one by resampling to obtain one data point of the non-continuous measure for each data point of the continuous measure [44,59], this leads to a bias given that data points are created where the measure does not exist or that a non-continuous measure is treated as a continuous one. The solution that probably introduces the least artificial information is to use averages across large time periods (e.g., over a block of stimulus response trials) for both continuous and non-continuous measures. One could still argue that the continuous measure is more reliable

because it depends on more measurement points and its values are not artificially introduced. However, averaging across longer time periods comes with a cost: a potential loss of sensitivity to shorter, phasic changes and only reflecting tonic changes in the measures. Given the popularity of paradigms in listening effort research that rely on the analysis of short stimulus evoked phasic changes (for instance, changes in pupil response evoked by auditive stimuli[49,60]), this constitutes a serious shortcoming.

In addition to the obstacles to the integration and comparison of multiple physiological measures created by the time characteristics of the measures themselves, differences in the time characteristics of the underlying physiological mechanisms must also be considered. Many of the physiological measures used in listening effort are driven by physiological mechanisms that operate on different time scales. For instance, it can take up to 20 seconds from the onset of nervous system activity to the maximum response of heart rate and blood pressure, and it also can take more than 10 seconds from the end of nervous system activity to the return of heart rate and blood pressure to their baseline values[61,62]. Pupil responses seem to be driven by faster physiological mechanisms given that they appear sooner (a few seconds after stimulus onset) and also disappear within seconds[63]. EEG evoked potentials rely on even faster mechanisms, and can be observed after a few milliseconds[64].

Given the differences in the time characteristics of the underlying physiological mechanisms, the paradigms used to optimize the assessment of the physiological measures of listening effort vary considerably. For instance, paradigms using cardiovascular measures normally present a single stimulus condition over a period of several minutes[12,13,29], whereas paradigms using pupil-related measures tend to present different stimulus conditions in

intervals of a few seconds[30,65,66]. Using multiple physiological measures that are driven by different physiological mechanisms consequently requires researchers to develop paradigms that are appropriate for the various time scales of their measures.

Recommendation 3

Take the individual time characteristics of the physiological measures and underlying physiological mechanisms into account when planning a study with multiple physiological measures. Develop paradigms that are appropriate for all involved measures.

Statistical Analysis of Multiple Physiological Measures

The final challenge to using multiple physiological measures in listening effort research is the selection of an appropriate statistical approach. The main concern here is the prevention of type-I error inflation due to the number of assessed physiological measures. One approach that is frequently adopted in listening effort research is to use an independent statistical test for each assessed measure. Unfortunately, this quickly increases type-I error. It is thus necessary to employ a type-I error control procedure. However, the big challenge is to find one that has a minimal impact on statistical power.

One option is to analyze all physiological measures in a two-step procedure where a first multivariate analysis of variance is used as gatekeeper for follow-up univariate tests[67]. For instance, Plain and colleagues[11] analyzed seven different physiological listening effort measures by first conducting a multivariate analysis of variance (MANOVA) that included all measures and then using univariate tests for those measures that were significant. If such a two-stage procedure is used with appropriately adapted critical *F*- and *t*-values for the follow-up tests, it can successfully control the maximum type-I error rate. However, in designs with more than

two groups, simple single-stage multiple-comparison procedures (like the Bonferroni procedure) perform as well as the more complex MANOVA-protected procedure and may thus be preferred[67]. Avoiding multivariate procedures also mitigates the problem of multicollinearity between the dependent variables, which can influence the interpretability of the results[68]. Multicollinearity—the correlation between the outcome variables in this case—is common in psychophysiological research given that the measures are often driven by the same or associated physiological mechanisms[69]. For instance, both pupil changes and heart rate changes are driven by sympathetic and parasympathetic nervous system activity and will highly correlate with one another if the autonomic outflow to the pupil and the heart does not differ.

An alternative approach is to aggregate the measures into a single index[70,71]. For instance, pre-ejection period and pupil diameter in the dark—when the parasympathetic contribution is minimal[72]—could be combined into a single index of sympathetic activity. A single aggregated index could be analyzed with a single statistical test and would thus prevent the problem of type-I error inflation discussed in the preceding paragraphs. Moreover, it would have higher statistical power because no type-I error inflation control would be needed and specific planned contrasts could be conducted[73-75]. Aggregating measures requires a decision on whether to standardize the individual measures before the aggregation. Standardizing the measures controls for the impact of the variability and magnitude of the responses of the individual measures. At first sight, this might seem to be a good idea because one would like to have each measure the same influence on the aggregated index. However, the standardization—for instance a z-standardization[76]—is often performed using the collected data, which introduces a bias. For instance, combining a z-standardized physiological measure

where participants showed originally almost no response variability—for instance, heart rate changes with a mean of 2 beats per minute (bpm) and a standard deviation of 1 bpm—with a z-standardized measure where participants showed strong response differences—for instance, systolic blood pressure responses with a mean of 20 millimeters of mercury (mmHg) and a standard deviation of 10 mmHg—leads to a huge bias because it treats a blood pressure change of 30 mmHg as being equivalent to a heart rate change of 3 bpm. A blood pressure change of 30 mmHg constitutes a much stronger physiological response than a heart rate response of 3 bpm, but this is neglected by the resulting index. This problem can be prevented by standardizing the individual physiological measures using their physiologically possible range as criterion (instead of their sample mean and variability). For fNIRS research, this approach has been taken recently by Zhang and colleagues who used a breath-holding task to scale the fNIRS response differences between conditions by the physiologically-plausible range of the fNIRS response before performing the statistical analysis[77]. Unfortunately, information about the absolute minimum and maximum response of many of the physiological measures employed in listening effort research is often not available. For instance, no information is known regarding the physiological maximum of a skin conductance response.

Recommendation 4

Plan your statistical analysis to account for the problems of assessing statistical significance (p-values) when running multiple tests (i.e., increased Type I-error when uncorrected or reduced statistical power when corrected for multiple testing) and of analyzing measures that are potentially highly correlated. If possible, use an aggregate index that represents the physiological mechanism that you are interested in.

Summary

Moving from using single physiological measures in listening effort research to combining multiple measures that are justified by a single, unifying rationale would help the field to overcome the fragmented approach that currently exists. The explicit presentation of researchers' concept definition of listening effort and its use to justify the employed physiological measures would promote a discussion about the core concept and hopefully lead to a commonly accepted definition of listening effort. Combining multiple measures does however require awareness of the problems that are caused by the simultaneous use of multiple measurement devices as well as sound knowledge about the time characteristics of the measures and the underlying physiological mechanisms. Moreover, awareness of the statistical issues associated with analyzing multiple measures is also required. The solutions to many of the challenges that we have outlined are still in their infancy or are yet to be developed. However, we are convinced that we should not leave it to future generations of researchers to integrate the fragmented field that we have created. Addressing these issues now is the only way forward to a more integrated approach to the use of physiological measures in listening effort research and to a comprehensive understanding of listening effort.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie-Sklodowska-Curie grant agreement No 765329; Medical Research Council (grant number MR/S003576/1). This work was supported by the NIHR Manchester Biomedical Research Centre.

References

1. Paul BT, Chen J, Le T, Lin V, Dimitrijevic A. Cortical alpha oscillations in cochlear implant users reflect subjective listening effort during speech-in-noise perception. *PLoS One*. Jul 2021;16(7):e0254162. e0254162. doi:10.1371/journal.pone.0254162

2. Wisniewski MG, Zakrzewski AC, Bell DR, Wheeler M. EEG power spectral dynamics associated with listening in adverse conditions. *Psychophysiology*. Sep 2021;58(9):e13877. e13877. doi:10.1111/psyp.13877

3. Hunter CR. Tracking Cognitive Spare Capacity During Speech Perception With EEG/ERP: Effects of Cognitive Load and Sentence Predictability. *Ear Hear*. Sep/Oct 2020;41(5):1144-1157. doi:10.1097/AUD.0000000000000856

4. Silcox JW, Payne BR. The costs (and benefits) of effortful listening on context processing: A simultaneous electrophysiology, pupillometry, and behavioral study. *Cortex*. Sep 2021;142:296-316. doi:10.1016/j.cortex.2021.06.007

5. Rovetti J, Goy H, Pichora-Fuller MK, Russo FA. Functional Near-Infrared Spectroscopy as a Measure of Listening Effort in Older Adults Who Use Hearing Aids. *Trends Hear*. Jan-Dec 2019;23:2331216519886722. 2331216519886722. doi:10.1177/2331216519886722

6. White BE, Langdon C. The cortical organization of listening effort: New insight from functional near-infrared spectroscopy. *Neuroimage*. Oct 15 2021;240:118324. 118324. doi:10.1016/j.neuroimage.2021.118324

7. Alhanbali S, Dawes P, Millman RE, Munro KJ. Measures of Listening Effort Are Multidimensional. *Ear and Hearing*. Sep-Oct 2019;40(5):1084-1097. doi:10.1097/aud.0000000000000697

8. Giuliani NP, Brown CJ, Wu YH. Comparisons of the Sensitivity and Reliability of Multiple Measures of Listening Effort. *Ear Hear*. Sep 10 2020;42(2):465-474. doi:10.1097/AUD.0000000000000950

9. Koelewijn T, Zekveld AA, Lunner T, Kramer SE. The effect of monetary reward on listening effort and sentence recognition. *Hear Res*. Jul 2021;406:108255. 108255. doi:10.1016/j.heares.2021.108255

10. Zekveld AA, van Scheepen JAM, Versfeld NJ, Veerman ECI, Kramer SE. Please try harder! The influence of hearing status and evaluative feedback during listening on the pupil dilation response, saliva-cortisol and saliva alpha-amylase levels. *Hear Res*. Sep 15 2019;381:107768. 107768. doi:10.1016/j.heares.2019.07.005

11. Plain B, Pielage H, Richter M, et al. Social observation increases the cardiovascular response of hearing-impaired listeners during a speech reception task. *Hear Res*. Oct 2021;410:108334. 108334. doi:10.1016/j.heares.2021.108334

12. Slade K, Kramer SE, Fairclough S, Richter M. Effortful listening: Sympathetic activity varies as a function of listening demand but parasympathetic activity does not. *Hear Res*. Oct 2021;410:108348. 108348. doi:10.1016/j.heares.2021.108348

13. Plain B, Richter M, Zekveld AA, Lunner T, Bhuiyan T, Kramer SE. Investigating the Influences of Task Demand and Reward on Cardiac Pre-Ejection Period Reactivity During a Speech-in-Noise Task. *Ear Hear*. Nov 16 2020;42(3):718-731. doi:10.1097/AUD.0000000000000971

14. Strand JF, Ray L, Dillman-Hasso NH, Villanueva J, Brown VA. Understanding Speech Amid the Jingle and Jangle: Recommendations for Improving Measurement Practices in Listening Effort Research. *Audit Percept Cogn*. 2020;3(4):169-188. doi:10.1080/25742442.2021.1903293

15. Ksiazek P, Zekveld AA, Wendt D, Fiedler L, Lunner T, Kramer SE. Effect of Speech-to-Noise Ratio and Luminance on a Range of Current and Potential Pupil Response Measures to Assess Listening Effort. *Trends Hear*. Jan-Dec 2021;25:23312165211009351. doi:10.1177/23312165211009351

16. Pielage H, Zekveld AA, Saunders GH, Versfeld NJ, Lunner T, Kramer SE. The Presence of Another Individual Influences Listening Effort, But Not Performance. *Ear Hear*. Nov-Dec 01 2021;42(6):1577-1589. doi:10.1097/AUD.0000000000001046

17. Seifi Ala T, Graversen C, Wendt D, Alickovic E, Whitmer WM, Lunner T. An exploratory Study of EEG Alpha Oscillation and Pupil Dilation in Hearing-Aid Users During Effortful listening to Continuous Speech. *PLoS One*. 2020;15(7):e0235782. doi:10.1371/journal.pone.0235782

18. Fiedler L, Seifi Ala T, Graversen C, Alickovic E, Lunner T, Wendt D. Hearing Aid Noise Reduction Lowers the Sustained Listening Effort During Continuous Speech in Noise-A Combined Pupillometry and EEG Study. *Ear Hear*. Nov-Dec 01 2021;42(6):1590-1601. doi:10.1097/AUD.0000000000001050

19. Shields C, Willis H, Nichani J, Sladen M, Kluk-de Kort K. Listening effort: WHAT is it, HOW is it measured and WHY is it important? *Cochlear Implants Int*. Nov 30 2021:1-4. doi:10.1080/14670100.2021.1992941

20. McGarrigle R, Munro KJ, Dawes P, et al. Listening effort and fatigue: what exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *Int J Audiol*. Jul 2014;53(7):433-40. doi:10.3109/14992027.2014.890296

21. Francis AL, Love J. Listening effort: Are we measuring cognition or affect, or both? *Wiley Interdiscip Rev Cogn Sci*. Jan 2020;11(1):e1514. doi:10.1002/wcs.1514

22. Podsakoff PM, MacKenzie SB, Podsakoff NP. Recommendations for Creating Better Concept Definitions in the Organizational, Behavioral, and Social Sciences. *Organizational Research Methods*. 2016;19(2):159-203. doi:10.1177/1094428115624965

23. Goertz G. *Social science concepts: A user's guide*. Princeton University Press; 2006.

24. DiRenzo GJ. *Concepts, theory, and explanation in the behavioral sciences*. Random House; 1966.

25. Krueger M, Schulte M, Zokoll MA, et al. Relation Between Listening Effort and Speech Intelligibility in Noise. *Am J Audiol*. Oct 12 2017;26(3S):378-392. doi:10.1044/2017_AJA-16-0136

26. Wild CJ, Yusuf A, Wilson DE, Peelle JE, Davis MH, Johnsrude IS. Effortful listening: the processing of degraded speech depends critically on attention. *J Neurosci*. Oct 3 2012;32(40):14010-21. doi:10.1523/JNEUROSCI.1528-12.2012

27. Piquado T, Isaacowitz D, Wingfield A. Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*. May 1 2010;47(3):560-9. doi:10.1111/j.1469-8986.2009.00947.x

28. Obleser J, Wostmann M, Hellbernd N, Wilsch A, Maess B. Adverse listening conditions and memory load drive a common alpha oscillatory network. *J Neurosci*. Sep 5 2012;32(36):12376-83. doi:10.1523/JNEUROSCI.4908-11.2012

29. Mackersie CL, Calderon-Moultrie N. Autonomic nervous system reactivity during speech repetition tasks: Heart rate variability and skin conductance. *Ear Hear*. Jul-Aug 2016;37 Suppl 1:118S-25S. doi:10.1097/AUD.0000000000000305

30. Ohlenforst B, Zekveld AA, Lunner T, et al. Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hear Res*. Aug 2017;351:68-79. doi:10.1016/j.heares.2017.05.012

31. Wendt D, Koelewijn T, Ksiazek P, Kramer SE, Lunner T. Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hear Res*. Nov 2018;369:67-78. doi:10.1016/j.heares.2018.05.006

32. Winn MB, Teece KH. Listening Effort Is Not the Same as Speech Intelligibility Score. *Trends Hear*. Jan-Dec 2021;25:23312165211027688. doi:10.1177/23312165211027688

33. Richter M, Slade K. Interpretation of physiological indicators of motivation: Caveats and recommendations. *Int J Psychophysiol*. Sep 2017;119:4-10. doi:10.1016/j.ijpsycho.2017.04.007

34. Picou EM, Ricketts TA, Hornsby BWY. Visual Cues and Listening Effort: Individual Variability. *Journal of Speech, Language, and Hearing Research*. 2011;54(5):1416-1430. doi:10.1044/1092-4388(2011/10-0154)

35. Pichora-Fuller MK, Kramer SE, Eckert MA, et al. Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL). *Ear Hear*. Jul-Aug 2016;37 Suppl 1:5S-27S. doi:10.1097/AUD.0000000000000312

36. Richter M. The Moderating Effect of Success Importance on the Relationship Between Listening Demand and Listening Effort. *Ear Hear*. Jul-Aug 2016;37 Suppl 1:111S-7S. doi:10.1097/AUD.0000000000000295

37. Meijer JH, Elbertse E, Boesveldt S, Berendse HW, Verdaasdonk RM. Using the Initial Systolic Time Interval to assess cardiac autonomic nervous function in Parkinson's disease. *Journal of Electrical Bioimpedance*. 2011;2(1):98-101. doi:10.5617/jeb.216

38. Sherwood A, McFetridge J, Hutcheson JS. Ambulatory impedance cardiography: a feasibility study. *J Appl Physiol (1985)*. Dec 1998;85(6):2365-9. doi:10.1152/jappl.1998.85.6.2365

39. Uchitel J, Vidal-Rosas EE, Cooper RJ, Zhao H. Wearable, Integrated EEG-fNIRS Technologies: A Review. *Sensors (Basel)*. Sep 12 2021;21(18)doi:10.3390/s21186106

40. von Luhmann A, Muller KR. Why build an integrated EEG-NIRS? About the advantages of hybrid bio-acquisition hardware. *Annu Int Conf IEEE Eng Med Biol Soc*. Jul 2017;2017:4475-4478. doi:10.1109/EMBC.2017.8037850

41. Cieslak M, Ryan WS, Macy A, et al. Simultaneous acquisition of functional magnetic resonance images and impedance cardiography. *Psychophysiology*. Apr 2015;52(4):481-8. doi:10.1111/psyp.12385

42. Oster J, Clifford GD. Acquisition of electrocardiogram signals during magnetic resonance imaging. *Physiol Meas*. Jun 22 2017;38(7):R119-R142. doi:10.1088/1361-6579/aa6e8c

43. Khan MJ, Ghafoor U, Hong KS. Early Detection of Hemodynamic Responses Using EEG: A Hybrid EEG-fNIRS Study. *Front Hum Neurosci*. 2018;12:479. doi:10.3389/fnhum.2018.00479

44. Al-Shargie F, Kiguchi M, Badruddin N, Dass SC, Hani AF, Tang TB. Mental stress assessment using simultaneous measurement of EEG and fNIRS. *Biomed Opt Express*. Oct 1 2016;7(10):3882-3898. doi:10.1364/BOE.7.003882

45. Dalton KM, Davidson RJ. The concurrent recording of electroencephalography and impedance cardiography: effects on EEG. *Psychophysiology*. Jul 1997;34(4):488-93. doi:10.1111/j.1469-8986.1997.tb02394.x

46. Keidser G, Naylor G, Brungart DS, et al. The Quest for Ecological Validity in Hearing Science: What It Is, Why It Matters, and How to Advance It. *Ear Hear*. Nov/Dec 2020;41 Suppl 1:5S-19S. doi:10.1097/AUD.0000000000000944

47. van der Mee DJ, Gevonden MJ, Westerink J, de Geus EJC. Validity of electrodermal activity-based measures of sympathetic nervous system activity from a wrist-worn device. *Int J Psychophysiol*. Oct 2021;168:52-64. doi:10.1016/j.ijpsycho.2021.08.003

48. Calamia M. Practical considerations for evaluating reliability in ambulatory assessment studies. *Psychol Assess*. Mar 2019;31(3):285-291. doi:10.1037/pas0000599

49. Winn MB, Wendt D, Koelewijn T, Kuchinsky SE. Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started. *Trends Hear*. Jan-Dec 2018;22:2331216518800869. doi:10.1177/2331216518800869

50. Kyong JS, Kwak C, Han W, Suh MW, Kim J. Effect of Speech Degradation and Listening Effort in Reverberating and Noisy Environments Given N400 Responses. *J Audiol Otol*. Jul 2020;24(3):119-126. doi:10.7874/jao.2019.00514

51. Boudoulas H. Systolic time intervals. *Eur Heart J*. Dec 1990;11 Suppl I:93-104. doi:10.1093/eurheartj/11.suppl_I.93

52. Francis AL, Bent T, Schumaker J, Love J, Silbert N. Listener characteristics differentially affect self-reported and physiological measures of effort associated with two challenging listening conditions. *Atten Percept Psychophys*. May 2021;83(4):1818-1841. doi:10.3758/s13414-020-02195-9

53. Miles K, McMahon C, Boisvert I, et al. Objective Assessment of Listening Effort: Coregistration of Pupillometry and EEG. *Trends Hear*. Jan-Dec 2017;21:2331216517706396. doi:10.1177/2331216517706396

54. Wisniewski MG, Thompson ER, Iyer N, Estepp JR, Goder-Reiser MN, Sullivan SC. Frontal midline theta power as an index of listening effort. *Neuroreport*. Jan 21 2015;26(2):94-9. doi:10.1097/WNR.0000000000000306

55. Moyer JT, Gnatkovsky V, Ono T, et al. Standards for data acquisition and software-based analysis of in vivo electroencephalography recordings from animals. A TASK1-WG5 report of the AES/ILAE Translational Task Force of the ILAE. *Epilepsia*. Nov 2017;58 Suppl 4:53-67. doi:10.1111/epi.13909

56. Berntson GG, Bigger JT, Jr., Eckberg DL, et al. Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology*. Nov 1997;34(6):623-48. doi:10.1111/j.1469-8986.1997.tb02140.x

57. Mathot S, Fabius J, Van Heusden E, Van der Stigchel S. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behav Res Methods*. Feb 2018;50(1):94-106. doi:10.3758/s13428-017-1007-2

58. Ayasse ND, Wingfield A. Anticipatory Baseline Pupil Diameter Is Sensitive to Differences in Hearing Thresholds. *Front Psychol*. 2019;10:2947. doi:10.3389/fpsyg.2019.02947

59. Kuipers M, Richter M, Scheepers D, Immink MA, Sjak-Shie E, van Steenbergen H. How effortful is cognitive control? Insights from a novel method measuring single-trial evoked beta-adrenergic cardiac reactivity. *Int J Psychophysiol*. Sep 2017;119:87-92. doi:10.1016/j.ijpsycho.2016.10.007

60. Koelewijn T, de Kluiver H, Shinn-Cunningham BG, Zekveld AA, Kramer SE. The pupil response reveals increased listening effort when it is difficult to focus attention. *Hear Res*. May 2015;323:81-90. doi:10.1016/j.heares.2015.02.004

61. Mokrane A, Nadeau R. Dynamics of heart rate response to sympathetic nerve stimulation. *Am J Physiol*. Sep 1998;275(3):H995-1001. doi:10.1152/ajpheart.1998.275.3.H995

62. Warner HR, Cox A. A mathematical model of heart rate control by sympathetic and vagus efferent information. *J Appl Physiol*. Mar 1962;17:349-55. doi:10.1152/jappl.1962.17.2.349

63. Mathot S. Pupillometry: Psychology, Physiology, and Function. *J Cogn*. Feb 21 2018;1(1):16. doi:10.5334/joc.18

64. Picton TW, Hillyard SA, Krausz HI, Galambos R. Human auditory evoked potentials. I: Evaluation of components. *Electroencephalography and Clinical Neurophysiology*. 1974;36:179-190. doi:10.1016/0013-4694(74)90155-2

65. Ohlenforst B, Wendt D, Kramer SE, Naylor G, Zekveld AA, Lunner T. Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. *Hear Res*. Aug 2018;365:90-99. doi:10.1016/j.heares.2018.05.003

66. Kuchinsky SE, Ahlstrom JB, Vaden KI, Jr., et al. Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*. Jan 2013;50(1):23-34. doi:10.1111/j.1469-8986.2012.01477.x

67. Bird KD, Hadzi-Pavlovic D. Controlling the maximum familywise Type I error rate in analyses of multivariate experiments. *Psychol Methods*. Jun 2014;19(2):265-80. doi:10.1037/a0033806

68. Schroeder MA. Diagnosing and dealing with multicollinearity. *West J Nurs Res*. Apr 1990;12(2):175-84; discussion 184-7. doi:10.1177/019394599001200204

69. Slinker BK, Glantz SA. Multiple regression for physiological data analysis: the problem of multicollinearity. *Am J Physiol*. Jul 1985;249(1 Pt 2):R1-12. doi:10.1152/ajpregu.1985.249.1.R1

70. Behnke M, Kaczmarek LD. Successful performance and cardiovascular markers of challenge and threat: A meta-analysis. *Int J Psychophysiol*. Aug 2018;130:73-79. doi:10.1016/j.ijpsycho.2018.04.007

71. Miyake S. Multivariate workload evaluation combining physiological and subjective measures. *International Journal of Psychophysiology*. 2001;40(3):233-238. doi:10.1016/S0167-8760(00)00191-4

72. Lowenstein O, Loewenfeld IE. Role of sympathetic and parasympathetic systems in reflex dilation of the pupil; pupillographic studies. *Arch Neurol Psychiatry*. Sep 1950;64(3):313-40. doi:10.1001/archneurpsyc.1950.02310270002001

73. Winer BJ, Brown DR, Michels KM. *Statistical principles in experimental design*. 3rd ed. McGraw-Hill; 1991.

74. Rosenthal R, Rosnow RL. *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge University Press; 1985.

75. Richter M. Residual tests in the analysis of planned contrasts: Problems and solutions. *Psychol Methods*. Mar 2016;21(1):112-20. doi:10.1037/met0000044

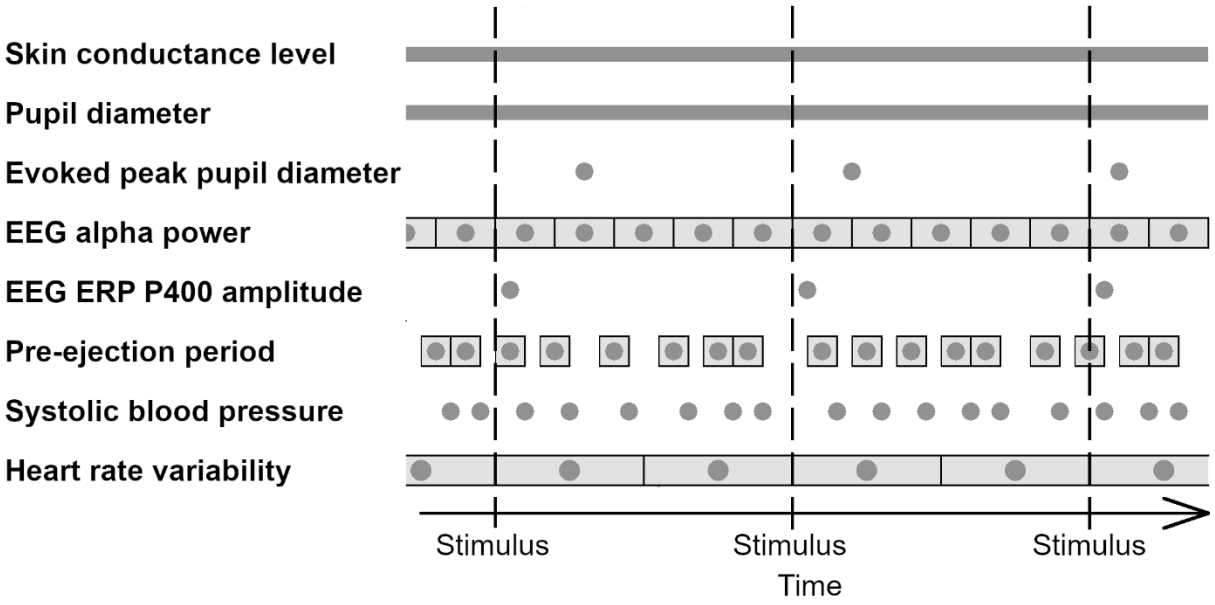
76. Moisl H. Variable scaling in cluster analysis of linguistic data. *Corpus Linguistics and Linguistic Theory*. 2010;6(1)doi:10.1515/cllt.2010.004

77. Zhang M, Alamatsaz N, Ihlefeld A. Hemodynamic Responses Link Individual Differences in Informational Masking to the Vicinity of Superior Temporal Gyrus. *Front Neurosci*. 2021;15:675326. doi:10.3389/fnins.2021.675326

Figure Captions

Figure 1. Time characteristics of selected physiological measures. Dark gray lines indicate continuous measures, dark gray dots non-continuous measures. Dark gray dots with surrounding light gray boxes indicate non-continuous measures that refer to time periods and not to specific points in time. The light gray boxes indicate the measurement epochs required to obtain the measure.

Figure 1


Skin conductance level
Pupil diameter
Evoked peak pupil diameter
EEG alpha power
EEG ERP P400 amplitude
Pre-ejection period
Systolic blood pressure
Heart rate variability
Stimulus
Stimulus
Stimulus
Time