

# Combining Multiple Psychophysiological Measures of Listening Effort: Challenges and Recommendations

**Michael Richter, Ph.D.,<sup>1</sup> Tanveer Buihyan, Ph.D.,<sup>2</sup>  
Lars Bramsløw, Ph.D.,<sup>3</sup> Hamish Innes-Brown, Ph.D.,<sup>3,8</sup>  
Lorenz Fiedler, Ph.D.,<sup>3</sup> Lauren V. Hadley, Ph.D.,<sup>4</sup> Graham Naylor, Ph.D.,<sup>4</sup>  
Gabrielle H. Saunders, Ph.D.,<sup>5</sup> Dorothea Wendt, Ph.D.,<sup>3,6</sup>  
William M. Whitmer, Ph.D.,<sup>4</sup> Adriana A. Zekveld, Ph.D.,<sup>7</sup> and  
Sophia E. Kramer, Ph.D.<sup>7</sup>**

## 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 article discusses in detail the various obstacles for combining multiple physiological measures in listening effort research and provides recommendations on how to overcome them.

<sup>1</sup>School of Psychology, Liverpool John Moores University, Liverpool, United Kingdom; <sup>2</sup>R&D, Demant A/S, Kongebakken, Denmark; <sup>3</sup>Eriksholm Research Centre, Oticon A/S, Snekkersten, Denmark; <sup>4</sup>Hearing Sciences, School of Medicine, University of Nottingham, Nottingham, United Kingdom; <sup>5</sup>Manchester Centre for Audiology and Deafness, University of Manchester, Manchester, United Kingdom; <sup>6</sup>Department of Health Technology, Technical University of Denmark, Lyngby, Denmark; <sup>7</sup>Section of Ear and Hearing, Department of Otolaryngology – Head and Neck Surgery, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam Public Health Research Institute, Amsterdam, The Netherlands; <sup>8</sup>Hearing Systems Section,

Department of Health Technology, Technical University of Denmark, Kongens Lyngby, Denmark.

Address for correspondence: Michael Richter, Ph.D., Byrom Street, Liverpool, L3 3AF, United Kingdom (e-mail: m.richter@ljmu.ac.uk).

Listening Effort Methodologies: Challenges and Future Directions; Guest Editor, Erin M. Picou, Au.D., Ph.D.

Semin Hear 2023;44:95–105. © 2023. Thieme. All rights reserved. Thieme Medical Publishers, Inc., 333 Seventh Avenue, 18th Floor, New York, NY 10001, USA DOI: <https://doi.org/10.1055/s-0043-1767669>. ISSN 0734-0451.

**KEYWORDS:** listening effort, psychophysiological measure, listening demand

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. Among 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,<sup>1,2</sup> EEG-evoked potential components,<sup>3,4</sup> functional near-infrared spectroscopy (fNIRS),<sup>5,6</sup> and peripheral measures, such as skin conductance,<sup>7,8</sup> pupil diameter,<sup>9,10</sup> heart rate variability,<sup>11,12</sup> and cardiovascular prejection period.<sup>12,13</sup> 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 conceptual 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.<sup>14</sup> 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.<sup>11,13,15-18</sup> 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.<sup>7,19-21</sup> 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<sup>22</sup>—for instance, to

decide whether a listening situation is more effortful or more arousing<sup>11,16</sup>—to find (psychophysiological) measures that appropriately match our concept,<sup>22,23</sup> and to build a refined theory of listening effort.<sup>24</sup> 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.<sup>25</sup> 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.<sup>21</sup> 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.<sup>26–32</sup> This implies a definition of listening effort as a state that changes in a predictable way when the level of the criterion variable (e.g., 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<sup>33</sup> 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 descrip-

tion of it. For instance, McGarrigle and colleagues<sup>20</sup> defined listening effort as “the mental exertion required to attend to, and understand, an auditory message,” Picou and colleagues<sup>34</sup> conceptualized it as “cognitive resources allocated for speech recognition,” and Pichora-Fuller and colleagues<sup>35</sup> 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<sup>21</sup>—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<sup>34</sup> 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<sup>13</sup> 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.<sup>36</sup> 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 using only 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.<sup>37,38</sup> 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.<sup>39,40</sup> Another example is the interference due to the magnetic field of magnetic resonance imaging (MRI) systems that can influence the ECG signal.<sup>41,42</sup>

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.<sup>43,44</sup> 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 (e.g., systems that only require electrodes on the thorax and neck<sup>45</sup>). Eye tracking is compatible with head-mounted sensors if a screen-based (remote) eye tracker is used. The problem of MRI artifacts 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 artifacts.<sup>41,42</sup> 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.<sup>46</sup> 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.<sup>47,48</sup>

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.<sup>39,40</sup> 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 standalone 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**

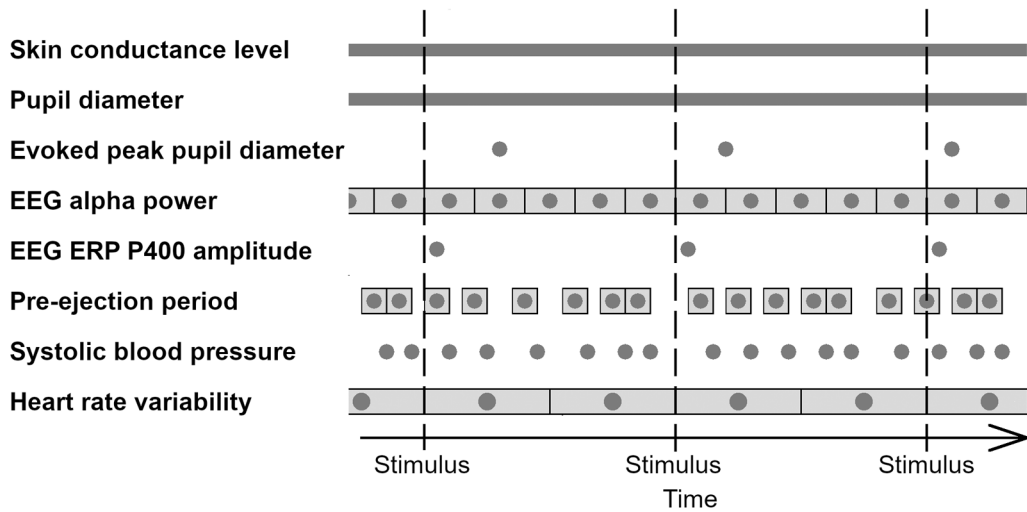
Determine how and whether the selected biosignals will interfere with one another and acquire appropriate specialized equipment accordingly to mitigate any problems caused by the physical and electromagnetic interference of the measurement devices and to attenuate the data synchronization issue. Consider these issues already at the planning stage of projects to ensure that the required financial, logistical, and knowledge-related resources (e.g., 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<sup>10</sup> or skin conductance level<sup>29</sup> has one particular value at any given moment and all such values provide meaningful information. In contrast, noncontinuous 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.<sup>49</sup> 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<sup>50</sup> or systolic blood pressure,<sup>12</sup> the maximum blood pressure between two consecutive heart beats.

In addition to noncontinuous measures that exist only at specific points in time, there are noncontinuous measures that refer to specific time periods and can therefore not be associated with a specific point in time. For instance, preejection period<sup>12</sup> 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<sup>51</sup> and is not associated with one single, specific point in time. Another example is heart period,<sup>52</sup> which refers to the time interval between two consecutive heart beats. There are also listening effort measures that are noncontinuous because of how they are calculated. For instance, the determination of EEG alpha<sup>53</sup> and theta power<sup>54</sup> requires the use of epochs to extract the frequency components of interest (e.g., an epoch of 1,250 ms would be required for the quantification of theta power<sup>55</sup>). Another example is heart rate variability,<sup>29</sup> which can also be determined only by quantifying variability over a certain time period (e.g., 1-minute intervals if a Fast Fourier Transform is used to quantify high-frequency heart rate variability<sup>56</sup>). Fig. 1 provides an



**Figure 1** Time characteristics of selected physiological measures. Dark gray lines indicate continuous measures; dark gray dots indicate noncontinuous measures. Dark gray dots with surrounding light gray boxes indicate noncontinuous 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.

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., 1,000–200 ms) prior to stimulus onset.<sup>57</sup> In some studies, this baseline is corrected for the individual dynamic range in the pupil size.<sup>58</sup> 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).<sup>12</sup>

It should be evident then that aggregating noncontinuous and continuous measures is complex. While it is technically possible to treat the noncontinuous measure as a continuous one by resampling to obtain one data point of the noncontinuous measure for each data point of the continuous measure,<sup>44,59</sup> this leads to a bias given that data points are created where the measure does not exist or that a noncontinuous 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 noncontinuous 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 (e.g., changes in pupil response evoked by auditory stimuli<sup>49,60</sup>), 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.<sup>61,62</sup> 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.<sup>63</sup> EEG evoked potentials rely on even faster mechanisms, and can be observed after a few milliseconds.<sup>64</sup>

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,<sup>12,13,29</sup> whereas paradigms using pupil-related measures tend to present different stimulus conditions in intervals of a few seconds.<sup>30,65,66</sup> 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 (MANOVA) is used as gatekeeper for follow-up univariate tests.<sup>67</sup> For instance, Plain and colleagues<sup>11</sup> analyzed seven different physiological listening effort measures by first conducting a 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.<sup>67</sup> Avoiding multivariate procedures also mitigates the problem of multicollinearity between the dependent variables, which can influence the interpretability of the results.<sup>68</sup> 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.<sup>69</sup> 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.<sup>70,71</sup> For instance, prejection period and pupil diameter in the dark—when the parasympathetic contribution is minimal<sup>72</sup>—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.<sup>73–75</sup> 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 each measure to have the same influence on the aggregated index. However, the standardization—for instance a z-standardization<sup>76</sup>—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 (mm Hg) and a standard deviation of 10 mm Hg—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 mm Hg 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.<sup>77</sup> 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.

#### **FUNDING**

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

#### **CONFLICT OF INTEREST**

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



## 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* 2021;16(07):e0254162
2. Wisniewski MG, Zakrzewski AC, Bell DR, Wheeler M. EEG power spectral dynamics associated with listening in adverse conditions. *Psychophysiology* 2021;58(09):e13877
3. Hunter CR. Tracking cognitive spare capacity during speech perception with EEG/ERP: effects of cognitive load and sentence predictability. *Ear Hear* 2020;41(05):1144–1157
4. Silcox JW, Payne BR. The costs (and benefits) of effortful listening on context processing: a simultaneous electrophysiology, pupillometry, and behavioral study. *Cortex* 2021;142:296–316
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* 2019;23:2331216519886722
6. White BE, Langdon C. The cortical organization of listening effort: new insight from functional near-infrared spectroscopy. *Neuroimage* 2021;240:118324
7. Alhanbali S, Dawes P, Millman RE, Munro KJ. Measures of listening effort are multidimensional. *Ear Hear* 2019;40(05):1084–1097
8. Giuliani NP, Brown CJ, Wu YH. Comparisons of the sensitivity and reliability of multiple measures of listening effort. *Ear Hear* 2021;42(02):465–474
9. Koelewijn T, Zekveld AA, Lunner T, Kramer SE. The effect of monetary reward on listening effort and sentence recognition. *Hear Res* 2021;406:108255
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* 2019;381:107768
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* 2021;410:108334
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* 2021;410:108348
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* 2021;42(03):718–731
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(04):169–188
15. Książek 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* 2021;25:23312165211009351
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* 2021;42(06):1577–1589
17. Seifi Ala T, Graverson 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(07):e0235782
18. Fiedler L, Seifi Ala T, Graverson 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* 2021;42(06):1590–1601
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* 2022;23(02):114–117
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* 2014;53(07):433–440
21. Francis AL, Love J. Listening effort: Are we measuring cognition or affect, or both? *Wiley Interdiscip Rev Cogn Sci* 2020;11(01):e1514
22. Podsakoff PM, MacKenzie SB, Podsakoff NP. Recommendations for creating better concept definitions in the organizational, behavioral, and social sciences. *Organ Res Methods* 2016;19(02):159–203
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* 2017;26(3S):378–392
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* 2012;32(40):14010–14021
27. Piquado T, Isaacowitz D, Wingfield A. Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology* 2010;47(03):560–569
28. Obleser J, Wöstmann M, Hellbernd N, Wilsch A, Maess B. Adverse listening conditions and memory load drive a common  $\alpha$  oscillatory network. *J Neurosci* 2012;32(36):12376–12383

29. Mackersie CL, Calderon-Moultrie N. Autonomic nervous system reactivity during speech repetition tasks: heart rate variability and skin conductance. *Ear Hear* 2016;37(Suppl 1):118S–125S
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* 2017;351:68–79
31. Wendt D, Koelewijn T, Książek 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* 2018;369:67–78
32. Winn MB, Teece KH. Listening effort is not the same as speech intelligibility score. *Trends Hear* 2021;25:23312165211027688
33. Richter M, Slade K. Interpretation of physiological indicators of motivation: caveats and recommendations. *Int J Psychophysiol* 2017;119:4–10
34. Picou EM, Ricketts TA, Hornsby BWY. Visual cues and listening effort: individual variability. *J Speech Lang Hear Res* 2011;54(05):1416–1430
35. Pichora-Fuller MK, Kramer SE, Eckert MA, et al. Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). *Ear Hear* 2016;37(Suppl 1):5S–27S
36. Richter M. The moderating effect of success importance on the relationship between listening demand and listening effort. *Ear Hear* 2016;37(Suppl 1):111S–117S
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. *J Electr Bioimpedance* 2011;2(01):98–101
38. Sherwood A, McFetridge J, Hutcheson JS. Ambulatory impedance cardiography: a feasibility study. *J Appl Physiol* (1985) 1998;85(06):2365–2369
39. Uchitel J, Vidal-Rosas EE, Cooper RJ, Zhao H. Wearable, integrated EEG-fNIRS technologies: a review. *Sensors (Basel)* 2021;21(18):6106
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* 2017;2017:4475–4478
41. Cieslak M, Ryan WS, Macy A, et al. Simultaneous acquisition of functional magnetic resonance images and impedance cardiography. *Psychophysiology* 2015;52(04):481–488
42. Oster J, Clifford GD. Acquisition of electrocardiogram signals during magnetic resonance imaging. *Physiol Meas* 2017;38(07):R119–R142
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
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* 2016;7(10):3882–3898
45. Dalton KM, Davidson RJ. The concurrent recording of electroencephalography and impedance cardiography: effects on EEG. *Psychophysiology* 1997;34(04):488–493
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* 2020;41 Suppl 1(Suppl 1):5S–19S
47. van der Mee DJ, Gevonden MJ, Westerink JHDM, de Geus EJC. Validity of electrodermal activity-based measures of sympathetic nervous system activity from a wrist-worn device. *Int J Psychophysiol* 2021;168:52–64
48. Calamia M. Practical considerations for evaluating reliability in ambulatory assessment studies. *Psychol Assess* 2019;31(03):285–291
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* 2018;22:2331216518800869
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* 2020;24(03):119–126
51. Boudoulas H. Systolic time intervals. *Eur Heart J* 1990;11(Suppl 1):93–104
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* 2021;83(04):1818–1841
53. Miles K, McMahon C, Boisvert I, et al. Objective assessment of listening effort: coregistration of pupillometry and EEG. *Trends Hear* 2017;21:2331216517706396
54. Wisniewski MG, Thompson ER, Iyer N, Estépp JR, Goder-Reiser MN, Sullivan SC. Frontal midline  $\theta$  power as an index of listening effort. *Neuroreport* 2015;26(02):94–99
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* 2017;58, Suppl 4(Suppl 4):53–67
56. Berntson GG, Bigger JT Jr, Eckberg DL, et al. Heart rate variability: origins, methods, and interpretive caveats. *Psychophysiology* 1997;34(06):623–648
57. Mathôt S, Fabius J, Van Heusden E, Van der Stigchel S. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behav Res Methods* 2018;50(01):94–106

58. Ayasse ND, Wingfield A. Anticipatory baseline pupil diameter is sensitive to differences in hearing thresholds. *Front Psychol* 2020;10:2947
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* 2017;119:87–92
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* 2015;323:81–90
61. Mokrane A, Nadeau R. Dynamics of heart rate response to sympathetic nerve stimulation. *Am J Physiol* 1998;275(03):H995–H1001
62. Warner HR, Cox A. A mathematical model of heart rate control by sympathetic and vagus efferent information. *J Appl Physiol* 1962;17:349–355
63. Mathôt S. Pupillometry: psychology, physiology, and function. *J Cogn* 2018;1(01):16
64. Picton TW, Hillyard SA, Krausz HI, Galambos R. Human auditory evoked potentials. I. Evaluation of components. *Electroencephalogr Clin Neurophysiol* 1974;36(02):179–190
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* 2018;365:90–99
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* 2013;50(01):23–34
67. Bird KD, Hadzi-Pavlovic D. Controlling the maximum familywise Type I error rate in analyses of multivariate experiments. *Psychol Methods* 2014;19(02):265–280
68. Schroeder MA. Diagnosing and dealing with multicollinearity. *West J Nurs Res* 1990;12(02):175–184, discussion 184–187
69. Slinker BK, Glantz SA. Multiple regression for physiological data analysis: the problem of multicollinearity. *Am J Physiol* 1985;249(1, Pt 2):R1–R12
70. Behnke M, Kaczmarek LD. Successful performance and cardiovascular markers of challenge and threat: a meta-analysis. *Int J Psychophysiol* 2018;130:73–79
71. Miyake S. Multivariate workload evaluation combining physiological and subjective measures. *Int J Psychophysiol* 2001;40(03):233–238
72. Lowenstein O, Loewenfeld IE. Role of sympathetic and parasympathetic systems in reflex dilation of the pupil; pupillographic studies. *Arch Neurol Psychiatry* 1950;64(03):313–340
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* 2016;21(01):112–120
76. Moisl H. Variable scaling in cluster analysis of linguistic data. *Corpus Linguist Linguist Theor* 2010;6(01):. Doi: 10.1515/clt.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