Exploring the Correlations Between Measures of Listening Effort in Adults and Children: A Systematic Review with Narrative Synthesis

Callum Shields[1,2] (iD)**, Mark Sladen[1], Iain Alexander Bruce[1],
Karolina Kluk[2]** (iD) **and Jaya Nichani[1]**

Abstract

Listening effort (LE) describes the cognitive resources needed to process an auditory message. Our understanding of this notion remains in its infancy, hindering our ability to appreciate how it impacts individuals with hearing impairment effectively. Despite the myriad of proposed measurement tools, a validated method remains elusive. This is complicated by the seeming lack of association between tools demonstrated via correlational analyses. This review aims to systematically review the literature relating to the correlational analyses between different measures of LE. Five databases were used– PubMed, Cochrane, EMBASE, PsychINFO, and CINAHL. The quality of the evidence was assessed using the GRADE criteria and risk of bias with ROBINS-I/GRADE tools. Each statistically significant analysis was classified using an approved system for medical correlations. The final analyses included 48 papers, equating to 274 correlational analyses, of which 99 reached statistical significance (36.1%). Within these results, the most prevalent classifications were poor or fair. Moreover, when moderate or very strong correlations were observed, they tended to be dependent on experimental conditions. The quality of evidence was graded as very low. These results show that measures of LE are poorly correlated and supports the multi-dimensional concept of LE. The lack of association may be explained by considering where each measure operates along the effort perception pathway. Moreover, the fragility of significant correlations to specific conditions further diminishes the hope of finding an all-encompassing tool. Therefore, it may be prudent to focus on capturing the consequences of LE rather than the notion itself.

Introduction

The concept of listening effort (LE) is still yet to be fully understood, and as a result, it is not formally integrated into routine clinical practice. Many attempts have been made to define LE, but perhaps the most contemporary definition stems from the Framework of Understanding Effortful Listening (FUEL) workshop.

"The deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task, with listening effort applying more specifically when tasks involve listening" (Pichora-Fuller et al. 2016, page 10)

This definition and the resulting model poised by FUEL builds upon the effort perception theories first developed by Kahneman to better apply to the world of hearing (Kahneman 1973). It relays that the experience of LE is a complex interaction between factors such as working memory, attention, motivation, task difficulty, and cognitive capacity (Kahneman 1973; Pichora-Fuller et al. 2016; Shields et al. 2021). Despite these recent developments, our

[1]ENT department, Royal Manchester Children's Hospital, Manchester, UK
[2]University of Manchester, Manchester, UK

Corresponding author:
Callum Shields, ENT department, Royal Manchester Children's Hospital, Manchester, UK.
Email: callum.shields@mft.nhs.uk

understanding of the effort perception pathways involved with LE is limited.

The ability to accurately measure LE has wide-reaching potential ramifications for individuals with hearing impairment. Perhaps this is most acutely demonstrated by considering those with mild to moderate or asymmetrical hearing loss. These patients are often overlooked with regards to interventions, especially surgical devices, and may therefore represent an unmet disease burden. While this degree of hearing loss may not be severe enough to interfere with traditional outcomes such as speech and language development, it cannot be said to have no negative connotations for quality of life. Further, a workshop conducted via the Nottingham Hearing Biomedical Research Unit identified "What adverse effects are associated with not treating mild-to-moderate hearing loss in adults?" as the number one research priority for adults with mild to moderate hearing loss (Henshaw et al. 2015). In the paediatric population, there is growing research to suggest that overexposure to LE can negatively affect educational outcomes (Bess, Dodd-Murphey, and Parker 1998; Downs and Crum 1978; Hsu, Vanpoucke, and van Wieringen 2017). This is particularly concerning when 11.3% of the general school population is thought to have some degree of hearing-impairment (Bess et al. 1998). Since children with mild-to-moderate hearing loss are less frequently identified compared to those with more severe impairments, they may be more likely to endure LE without the benefit of interventions and thus further hindering their educational attainment (Archbold et al. 2015). With this considered, research must include the needs of the younger generation, especially as they are perhaps more susceptible to the sequelae of LE.

Outside of educational domains, other recognized consequences accrue following prolonged periods of effortful listening. Some of the most established outputs include fatigue and stress, which, in the long run, can have a deleterious impact on mental wellbeing (Bess et al. 2016; Bess and Hornsby 2014; Kramer, Kapteyn, and Houtgast 2006; Kumari et al. 2009; Mattys et al. 2012; Pichora-Fuller 2016; Pichora-Fuller et al. 2016; Schlotz et al. 2004; Schneiderman, Ironson, and Siegel 2005). Together these educational and psychological corollaries emphasise the importance of broadening our understanding of LE and the need to continue exploring this relatively uncharted territory to discern other possible consequences of LE. It also provides an alternative cache for developing outcome measures of LE by using these downstream effects as proxy measures for LE itself.

There are many barriers to incorporating LE as an outcome measure within clinical practice, paramount of which relates to the lack of consensus regarding the best way to measure LE (Pichora-Fuller et al. 2016). Three broad categories of LE measures are used across the literature: self-reported questionnaires, behavioural measures, and physiological measures (McGarrigle et al. 2014).

Within each category exists a cornucopia of different tools postulated to capture LE, each has its advocates and dissidents, yet the 'gold standard' measure remains frustratingly elusive. Moreover, it is not uncommon for tools that aim to represent the consequences of LE, such as fatigue questionnaires, to be used as a proxy measure of LE (Alhanbali et al. 2018; Dwyer et al. 2019). A brief outline of these measures and their advantages and disadvantages are presented below.

Self-reported measures: Questionnaires that assess the individual's perceived effort levels are subjective. These questionnaires often include a numerical rating scale to quantify the degree of effort. The advantages of self-reported measures include that they are quick to administer and can be used to trend LE over a period of time or assess for benefit before and after an intervention (McGarrigle et al. 2014). Potential pitfalls to questionnaires relate mainly to their subjective nature and the subsequent effect of user bias. Indeed, Kahneman (1973) distinguishes subjective and objective effort as two separate entities, inferring that mental effort is solely an objective construct, independent of perceived effort (Bruya and Tang 2018). Examples of self-reported measures include the NASA task load index or more superficial LE visual analogue scales (Alhanbali et al. 2019; Hart 2006). Alongside effort questionnaires, fatigue questionnaires operate similarly but focus on the tiredness evoked via effortful listening. This is an example of how the sequelae of LE can be utilised in the assessment of hearing-related effort (McGarrigle et al. 2014). They have the same advantages and disadvantages as effort questionnaires.

Behavioural measures: Centre around an individual's performance within a particular task and use their performance as an objective measure of effort. This is often through a single or dual-task paradigm and thus resonates with the capacity model of mental effort discussed above (Pichora-Fuller et al. 2016). A benefit to mapping performance is that it can resemble real-world challenges that listeners face, such as classroom or workplace scenarios, and thus has more ecological value (McGarrigle et al. 2014). The contrary stance is that laboratory-based testing frequently subverts the impact of motivation on capacity since it may be challenging to foster participant motivation for an arbitrary task (Shields et al. 2021). Additionally, their reliance on the capacity model to demonstrate effort requires the task to stretch the individual to their cognitive limit, which may not be feasible in all experiments (Pichora-Fuller et al. 2016). Behavioural measures include response times and reaction times within a dual-task (Rovetti et al. 2019; Visentin and Prodi 2021).

Physiological measures: This type of tool uses a myriad of instruments to determine effort, relaying the autonomic reaction of body systems to a stressful testing stimulus (McGarrigle et al. 2014). Their interpretation of neural and endocrine signals leaves little room for subjectivity, indicating that these measures reflect Kahneman's view of mental

effort more closely than the other measures (Bruya and Tang 2018; Kahneman 1973). Unfortunately, their objective nature does not eliminate all bias; they are prone to confounding bias in that many other factors could amplify or dampen somebody's physiological responses (Ayres et al. 2021). For instance, concurrent comorbidities (neurological disorders, endocrine disorders, ophthalmic conditions, or cardiovascular pathology), medication use, caffeine or alcohol intake, or even sleep hygiene (Ayres et al. 2021; Kemp et al. 2014; Pichora-Fuller et al. 2016). Therefore, physiological measures may represent the gold standard index of effort in ideal experimental conditions, but this is unlikely to translate into real-world environments. Examples of physiological measures include pupillometry, EEG oscillation, cortisol level, functional imaging, heart rate variability, and skin conductance (Francis et al. 2021; Kramer, Teunissen, and Zekveld 2016; Rovetti et al. 2019; Zekveld et al. 2019).

To add further complexity to the matter, individual studies have shown that these measures are poorly correlated, diminishing the hope of a validated clinical tool (Alhanbali et al. 2019). This observation has given rise to the multidimensional model of LE and suggests that no one tool can fully encompass LE. If multiple measures are required to assess LE accurately, the clinical acceptability of LE as an outcome measure will be seriously hindered (Mosadeghrad 2012). Given the vast array of tools and experimental conditions that can be applied to investigate LE, it remains problematic to generalise results from individual studies across the literature. Therefore, it is prudent to systematically review these individual studies to map out correlational data between measures across all relevant studies to understand this concern fully. If the results of the individual studies harmonise the lack of correlation between measures of LE, then it would be important to consider alternative strategies to embrace the concept as an outcome.

Aims

The primary aim of this paper is to perform a systematic review of the literature pertaining to the correlations between methods of measuring LE in adults and children to determine how related they are to one another. If the multidimensional model is accurate, this review hypothesises that weak and non-significant correlations will be observed across the literature.

A secondary aim is to compare the correlations between various LE instruments to measures of fatigue to investigate the association between the two concepts. If fatigue is a downstream consequence of LE, this paper predicts that the strength of correlations for effort-effort comparisons will be similar to effort-fatigue comparisons.

Methods

A protocol outlining the methodology of this review was published on PROSPERO prior to the commencement of the literature search. Access to the full PROSPERO protocol can be obtained via the following link: https://www.crd.york.ac.uk/prospero/display_record.php?RecordID=253010.

Literature Search

The literature search for this review was conducted using five databases – PubMed, Cochrane Central, EMBASE, PsycINFO, and CINAHL. An appropriate search strategy was used that utilised medical subject heading (MeSH) and operators was used. The two core themes of the search were – listening effort/fatigue and hearing impairment – a detailed overview of this strategy can be found in the PROSPERO registration. A date limit of 2000-present was applied, given the relative novelty of the term listening effort. The strategy was validated by a medical librarian, independent of the study. Once the final set of papers had been retrieved, they were uploaded to the systematic review software EPPI-reviewer to be screened (Thomas, Brunton, and Graziosi 2021).

Study Selection

The initial screening stage involved removing duplicates; this was done initially by the tool located within EPPI-reviewer and subsequently verified manually by one reviewer. Once duplicates had been removed, screening on title and abstract was conducted by two reviewers independently with blinding applied to the other's decisions to minimise any bias effects. Inclusion and Exclusion criteria were generated using a modified version of the Population, Intervention, Comparison, Outcome and Study design (PICOS) framework (Armstrong 1999; Ebell 1999). In this review, the domain 'intervention' was replaced with 'exposure' to resonate with the study objective. The following inclusion and exclusion criteria were applied:

Inclusion:
- Participants – The study population includes the following categories: Normal hearing, Hearing-impaired with a device, Hearing-impaired without a device, Adults and Children. Adults and children have been included in this review to allow for age-specific comparisons of the correlations between instruments. Children may be more prone to the negative consequences of LE in terms of educational attainment and therefore represent an important population to consider. However, there is only a small number of published papers on paediatric individuals; therefore, this review included adults to allow for a complete narrative.
- Exposure – The paper includes any tool used to capture LE or fatigue related to a listening task.
- Comparison – The paper uses statistical correlational methods to compare the results obtained for the different measures of LE

- Outcome – The paper produces correlation coefficients between measures and associated p-values
- Study design – Primary research, study must be available in English

Exclusion:
- Participants – Any documented comorbidities that may influence listening effort results or fatigue assessment. This may include chronic health conditions, fatigue-causing conditions, or dual sensory impairment.
- Exposure – Any tool which does not primarily measure listening effort or fatigue related to a listening task. This included general quality of life measures.
- Comparison – No analysis provided to compare different tools.
- Outcome – The paper does not comment on the relationship between two measures of LE derived from the results of the correlational analysis.
- Study design – Secondary research, editorials, interviews, book chapters or case reports, Studies not available in English

Any disagreement between reviewers during the coding stage was referred to a panel constituting an ENT consultant, audiologist and another ENT doctor for reconciliation. This process was then repeated similarly for full-text screening. After obtaining an initial set of included results, forward citation chaining was conducted to further search the literature for relevant papers. Forward citation was chosen as it had the added benefit of finding papers published after the initial search and papers contained within the grey literature and, therefore, may not have been identified (Lefebvre et al. 2021). An outline of this process is shown in the PRISMA flowchart contained within the results section of this paper.

Data Extraction and Data Analysis

Data extraction was performed via a google form document by two reviewers. Key information captured for each study included:

- Details of the study – Type of paper, date published, authors, country of publication
- Study design – Observational vs. interventional
- Study population – Cohorts, sample size, and whether a power analysis had been performed
- Setting of the study – Listening task and listening conditions
- Measure of LE – What exact tools were used, and which overarching category do they fall under (Behavioural/physiological/self-reported)
- Type of comparison method used for the analysis of LE measures (Spearman's/Pearson's)
- Strength of correlation.

Quality Assessment and Risk of Bias

To assess the quality of evidence and potential risk of bias, the Grading of Recommendations Assessment, Development and Evaluation (GRADE) criteria was applied to the pool of included studies (Schünemann et al. 2013). This approach has the benefit of assessing the quality of evidence in its entirety rather than focusing on individual studies. This was deemed particularly appropriate for this review, given the heterogenous designs of the studies included. The GRADE criteria contain several domains that assess the overall quality of evidence (Schünemann et al. 2013).

First, 'inconsistency' refers to how homologous the evidence is, where studies performed in similar conditions on comparable participants. 'Imprecision' describes the statistical power of the studies in relation to their sample sizes. 'Indirectness' is how applicable the included studies are when examining the research question of this review. Determinates of indirectness include the population under investigation and the outcome measures used within the included studies. Each item can be graded as 'not serious', 'serious', or 'very serious' depending on the proportion of studies that satisfy each criterion (Schünemann et al. 2013). The effect of publication bias was also investigated within the overall GRADE assessment. This was determined by examining whether studies were consistent between their methods and reported results.

In addition to the above markers for study quality, the risk of bias was also assessed by applying an appropriate checklist to each study which varied based on individual study design. The ROBINS-I checklist was used for interventional studies to determine study limitations (Sterne et al. 2016). The current review classified a study as interventional if it attempted to measure the change in LE within study participants upon application of an accepted therapy such as hearing aid settings. This tool was chosen as it was specifically designed for non-randomised interventional studies, which aligns well with the subset of included studies in this review. Additionally, it is endorsed by the Cochrane Handbook for systematic reviews (Sterne et al. 2021). ROBINS-I assesses several domains, including – Confounding bias due to differences in baseline characteristics; selection bias emanating from differences in participants between study groups; classification bias due to misclassifying participants based on intervention status; deviation bias when study participants deviate from intended interventions and outcome bias arising from differences in the methods used to measure outcomes between participants (Sterne et al. 2016). Following this assessment, the overall risk of bias is formulated as – 'low', 'moderate', 'serious', or 'critical' depending on how the study is assessed in the domains mentioned above (Sterne et al. 2016). An example of how the ROBINS-I checklist was applied within this study can be found in the supplementary information.

For observational studies, the GRADE checklist was utilised, which examines factors including - whether a control group was present, how the study confirmed the exposure of interest, how consistently was the outcome measure

applied to participants, did the study account for possible confounders, and finally did the study measure at multiple time points (Schünemann et al. 2013). For this review, studies were classed as observational if they measured LE within a cohort and did not attempt to discern the effect of a particular intervention such as hearing aid settings on the participant's LE. The overall risk of bias for the observational studies can be judged as 'low' whereby most information from studies is deemed to have a small risk of bias; 'unclear' if the information from studies indicates a small or uncertain risk of bias; or 'high' in which the information from studies suggests that the risk of bias is sufficient to affect the interpretation of the results (Schünemann et al., 2013). An example of how the GRADE checklist was considered in the current study can be found in the supplementary information.

Data Synthesis

It was not appropriate to conduct a meta-analysis of the included studies as the pooled sample of papers was heterogenous in design. Factors that contributed to the lack of congruency between papers included measurement tools, experimental design, and study population. Therefore, a narrative synthesis of the result was produced to answer the research question posed by this review.

To overcome the challenge associated with directly comparing correlations amongst a heterogeneous study pool, this paper classified each r-value of statistically significant results according to the taxonomy of correlations outlined by Chan (Chan 2003). This reporting system has been specifically designed for application to medical papers. Chan reported the following grading system, a perfect correlation (coefficient + 1 or −1), very strong (coefficient 0.8 to 1 or −0.8 to −1), moderate (coefficient 0.6 to 0.8 or −0.6 to −0.8), fair (coefficient 0.3 to 0.6 or −0.3 to −0.6), poor (coefficient 0 to 0.3 or 0 to −0.3) and none (coefficient 0) (Chan 2003). This review evaluated the performance of each measurement theme against these criteria to provide an overall understanding of the degree of association between effort questionnaires, behavioural measures, physiological measures, and fatigue questionnaires in relation to capturing LE. A Chi-squared test was used to determine if there was a statistically significant difference between the observed classifications across all significant results. Additionally, a Fisher's exact test was used to examine for observed differences between the results for the adult and child population groups. Fisher's exact test was chosen due to the relatively small number of children within the included papers.

Results

Upon completion of both title and abstract, full-text screening, and forward citation searching, 48 papers were included based on the above PICOS criteria. The PRISMA diagram (Figure 1) below summaries the results of this process (Page, McKenzie, et al. 2021). The full details of each included paper, with reference to the experimental design, can be found in the supplementary information.

The adult population across the papers comprised 775 hearing-impaired participants and 815 individuals with normal hearing, totalling 1687 participants. Of the 775 patients with hearing impairment, 124 were listed as hearing aid users, whereas 117 had cochlear implants. For the paediatric population, 34 had a hearing impairment, and 179 had normal hearing, totalling 213 paediatric participants. The total sample size across all studies, ages, and hearing level equated to 1900.

Summary of the Strength of Correlations

The total number of correlational analyses across all included papers totalled 274. These 274 were stratified across ten different 'subgroups' each of which provided a comparison between two categories of instruments (EQ-EQ, EQ-B, EQ-P, EQ-FQ, B-B, B-P, B-FQ, P-P, P-FQ, and FQ-FQ – EQ = Effort questionnaire, B = Behavioural measures, P = Physiological measure and FQ = Fatigue questionnaire).

This review defined the results as either statistically significant or non-significant based on the p-values reported by the individual papers. A reported p-value of less than 0.05 was considered statistically significant in line with the traditional cut-off (Andrade 2019). Across all subgroups, 99 analyses reached statistical significance representing 36.1% of the total analyses. The range of r-values has been plotted using a violin chart and stratified according to significance – Figure 2. Only those analyses that reported the r-value, despite non-statistically significant results were included in the figure (133/175).

Each statistically significant trial was classified using the criteria outlined by Chan, which provides an overview of how related each measure is to one another (Chan 2003). This is summarised in Table 1, which is overlaid with a heat map to illustrate the density of the results. The most common classification identified across all significant correlations was fair (n = 62). The observed difference between the frequency of each classification was shown to be statistically significant through Chi-squared testing (p < 0.01).

Comparison Between Adults and Children

A total of 253/274 correlational analyses exclusively contained an adult population, whereas only 21/274 correlational analyses contained children. The number of significant trials amongst the adult population was 89 (35.2%). In contrast, 10 (47.6%) of the analyses in children reached significance. Fisher's exact test showed no significant difference between the frequency of each type of classification between the adult and child population groups (p = 0.84).
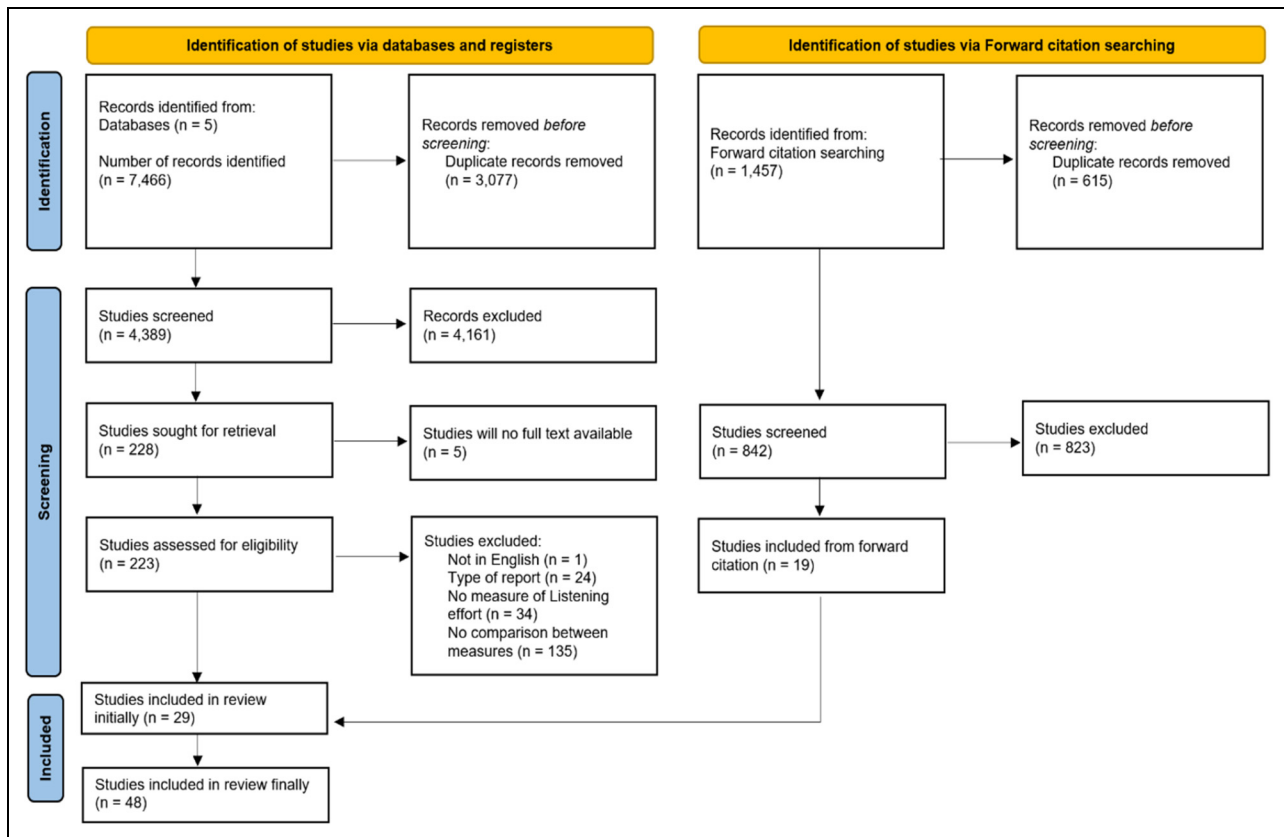
Figure 1. PRISMA Diagram Illustrating Literature Search and Screening Process (Page et al., 2021a).

Conditional and Unconditional Correlations

This review recognises two distinct types of correlations: conditional and unconditional. The latter refers to correlations in which the paper has provided a general comparison between two measures, collapsing for any changes to experimental conditions or study populations. Conditional correlations exist when the paper has published the strength of association between two measures of LE for a specific study condition. This review identified three categories of experimental conditions discussed below. The conditions for each analysis can be found in the supplementary information.

The first relates to the listening task used to evoke effort. This may include noise, light, signal quality, and task difficulty. This review noted that, in general, more challenging tasks tended to produce stronger correlations. For example, in one study, the r-value between a Visual analogue scale/Peak pupil diameter was 0.6 ($p < 0.01$) within the SRT50% condition. When this was altered to quiet conditions, the r-value fell to 0.008 ($p > 0.05$) (Kramer et al. 2016). In another study, an analysis comparing Peak pupil diameter/Need for recovery questionnaire revealed an r-value of - 0.35 ($p < 0.05$) within the light experimental conditions. When this was changed to dark conditions, the r-value was found to be - 0.18 ($p > 0.05$) (Wang, Kramer et al. 2018).

Next is the time point at which LE is measured. Examples of this condition include pre-test, during-test, post-test, pre-post-test difference (Gustafson et al. 2018; Key et al. 2017; Kramer et al. 2016; Zekveld et al. 2019), baseline period, retention period, speech period (Alhanbali et al. 2019; Russo et al. 2020), before response cue, and after response cue (Giuliani 2021). Again, similarly to the listening task, this review found that correlations were stronger when LE was sampled immediately following the listening task. When lapse of attention was compared to a fatigue scale, the pre-test r value changed from −0.5 ($p < 0.05$) to −0.702 ($p < 0.05$) in the post-test condition (Key et al. 2017). This was again shown during a comparison of the Speech, Spatial, and Qualities (SSQ) questionnaire, and pupillometry. Here, the baseline r value was not documented due to a lack of significance, but the retention interval findings reached significance with a coefficient of −0.78 (Russo et al. 2020).

The third is the hearing ability of the population used within the experiment. Normal hearing, hearing-impaired, and hearing device users constitute this condition. While it is accepted that those with a hearing impairment suffer from higher levels of effort, there was no apparent variation in the strength of correlations based on population conditions (Alhanbali et al. 2017; Holube et al. 2016).

It is rational to surmise that conditional correlations confer less evidence towards an association than
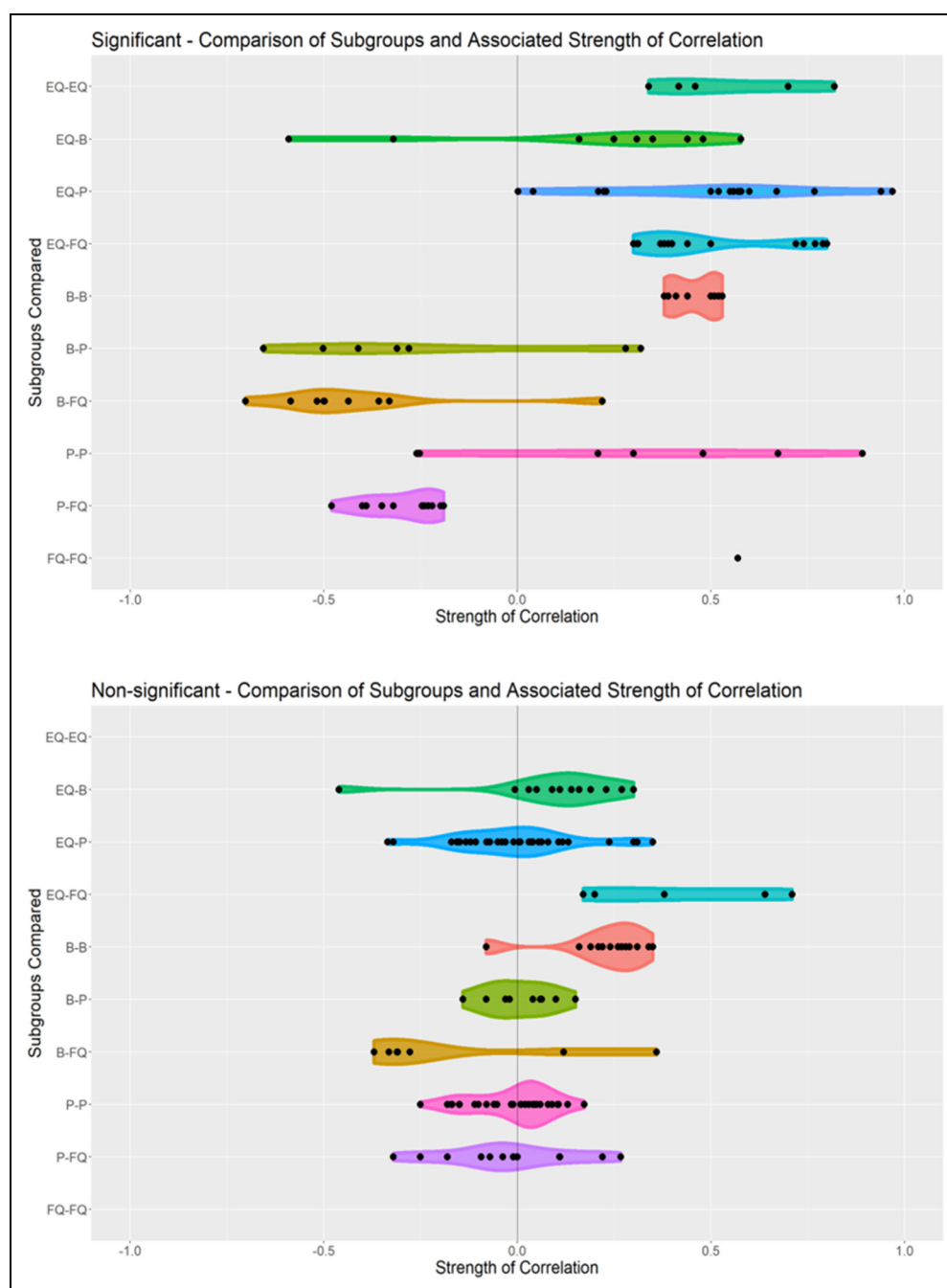
Figure 2. Violin Plots Showing the Range of r-values (x-axis) Against Each Subgroup Comparison (y-axis) for Both the Significant (top graph) and non-Significant (bottom graph) Correlational Analyses. Colour is Matched to Subgroup. (EQ = Effort Questionnaire, B = Behavioural Measures, P = Physiological Measures and FQ = Fatigue Questionnaire).

unconditional correlations as their reproducibility would be limited to the exact experimental conditions used to derive the finding.

Strength of Correlation Stratified by Subgroup

To further explore the correlations reported, this review presents both a within subgroup and between subgroup comparisons

below. Each section compares two subgroups of measures against one another to provide an insight into whether each theme captures similar or different components of LE. A summary of the number of significant analyses for each subgroup comparison is shown in Table 2.

Effort Questionnaires v Other Effort Questionnaires – A Within Subgroup Comparison. This review identified three papers

Table 1. Showing the Number (%) of Each Classification Type Against Each Subgroup Comparison.

Subgroup \ Strength	Poor (0 to 0.3 or 0 to -0.3)	Fair (0.3 to 0.6 or -0.3 to -0.6)	Moderate (0.6 to 0.8 or -0.6 to -0.8)	Very strong (0.8 to 1 or -0.8 to -1)	Total
EQ-EQ	0	*3 (60.0%)*	1 (20.0%)	1 (20.0%)	5
EQ-B	1 (11.1%)	*8 (88.9%)*	0	0	9
EQ-P	5 (23.8%)	*8 (38.0%)*	4 (19.1%)	4 (19.1%)	21
EQ-FQ	0	*14 (73.7%)*	4 (21.0%)	1 (5.3%)	19
B-B	0	*10 (100%)*	0	0	10
B-P	2 (28.6%)	*4 (57.1%)*	1 (14.3%)	0	7
B-FQ	1 (11.1%)	*7 (77.8%)*	1 (11.1%)	0	9
P-P	*3 (42.8%)*	2 (28.6%)	1 (14.3%)	1 (14.3%)	7
P-FQ	*6 (54.5%)*	5 (45.5%)	0	0	11
FQ-FQ	0	*1 (100%)*	0	0	1

Overlaid with a heat map to illustrate the concentration of results. The model value for each subgroup comparison is indicated by bold and italicised text. (EQ = Effort Questionnaire, B = Behavioural Measures, P = Physiological Measures and FQ = Fatigue Questionnaire).

Table 2. Summary of the Number of Significant Correlational Analyses Across Each Comparison of Subgroups.

Subgroup	Total number of trials	Number of significant results	% of trials reaching significance
EQ - EQ	5	5	100
EQ - B	28	9	32.1
EQ - P	73	21	28.8
EQ - FQ	24	19	79.2
B - B	15	10	66.7
B - P	26	7	26.9
B - FQ	19	9	47.4
P - P	43	7	16.3
P - FQ	30	11	36.7
FQ - FQ	1	1	100

that compared different effort-based questionnaires (Perreau et al., 2017; Picou et al., 2017; Picou & Ricketts, 2018). This resulted in five correlational analyses, all of which reached statistical significance. All five analyses were unconditional and therefore not dependent on changes to experimental conditions. Within this subgroup, the strongest association was observed between single entity questions that aimed to elicit 'loss of control' and questions that focused on the 'work required' to complete a task. This had an r-value of 0.82. The weakest correlation was noted between 'work required' questions and

questions which explored an individual's desire to 'give up' with an r-value of 0.34. Three of the correlations were classified as fair, one as moderate, and one as very strong.

Effort Questionnaires v Behavioural Measures – A Between Subgroup Comparison. Out of the 48 included papers, 13 of them compared effort questionnaires to behavioural measures of LE (Desjardins, 2016; Johnson et al., 2015; Megha & Maruthy, 2019; Perreau et al., 2017; Picou et al., 2017; Picou & Ricketts, 2018; Rovetti et al., 2019; Seeman & Sims, 2015; Strand et al., 2018; Visentin et al., 2019, 2021; Visentin & Prodi, 2021; White & Langdon, 2021). These papers produced 28 distinct correlational analyses. Only nine (32.1%) of these correlations reached statistical significance. For these nine correlations, six were unconditional, and three were conditional. The strongest association was demonstrated between control-themed questions and reaction times, with an r-value of - 0.59. This was a conditional correlation found only when the stimulus was presented in noise. The weakest correlation existed between a five-point LE scale/response times with an r-value of 0.16. Regarding the classification of correlations, one was poor, and eight were fair. The remaining 19 results were not statistically significant.

Effort Questionnaires v Physiological Measures – A Between Subgroup Comparison. The comparisons between effort

questionnaires and physiological measures of LE represent the most investigated subgroup within this review, with results stemming from 23 papers (Alhanbali et al., 2019, 2021; Bernarding et al., 2017; Dimitrijevic et al., 2019; Dwyer et al., 2019; Finke et al., 2016; Francis et al., 2021; Holube et al., 2016; Kramer et al., 2016; Lau et al., 2019; Mackersie et al., 2015; Mackersie & Cones, 2011; Magudilu Srishyla Kumar, 2020; McGarrigle et al., 2020; Rovetti et al., 2019; Russo et al., 2020; Schafer et al., 2015; Seeman & Sims, 2015; Strand et al., 2018; Visentin et al., 2021; White & Langdon, 2021; Wisniewski et al., 2015; A. A. Zekveld et al., 2011). This culminated in a total of 73 correlational analyses. Within these 73 analyses, 21 (28.8%) reached significance, and 52 failed to reach significance. Focusing on the 21 significant results, four were unconditional, and 17 were conditional. The joint strongest association compared National Aeronautics Space Administration (NASA) task load index/EEG oscillation and a five-point LE scale/EEG frontal midline theta power, both showing r values of 0.97. This represents a near-perfect correlation; similar results were also observed between a seven-point LE scale/EEG oscillation (r-value 0.94). These very strong analyses stemmed from three separate papers. The weakest association was between NASA task load index/Functional infra-red spectroscopy with an r-value of 0.13. In terms of classification, five were poor, eight were fair, four were moderate, and four were very strong.

Effort Questionnaires v Fatigue Questionnaires – A Between Subgroup Comparison. Within the category contrasting effort questionnaires to fatigue questionnaires, a total of eight papers published comparisons (Alhanbali et al., 2017, 2018, 2019, 2021; Dwyer et al., 2019; McGarrigle et al., 2020; Picou et al., 2017; Picou & Ricketts, 2018). This resulted in 24 separate correlation analyses; 19 (79.2%) reached significance, while five did not. Among the significant findings, 11 were unconditional, and eight were condition dependant. The largest effect size was demonstrated by Amsterdam Checklist for Hearing and Work/Profile of Mood States - frequency of tiredness questions with an r-value of 0.80. The weakest association was between an Effort assessment scale/Fatigue assessment scale resulting in an r-value of 0.30. The significant correlations obtained the following classifications of strength, 14 were fair, four were moderate, and one was very strong.

Behavioural Measures v Other Behavioural Measures – A Within Subgroup Comparison. Two papers compared different behavioural measures of LE and generated 25 distinct correlational analyses (Strand et al., 2018; Wu et al., 2014). Ten (66.7%) of these analyses were significant, and 15 were non-significant. Of the ten significant results, one was unconditional, and nine were conditional. The largest r value existed for listening span/cognitive spare capacity test, and this was a conditional correlation found within the high

predictability/updating conditions with an r-value of 0.53. The weakest correlation was observed between complex reaction times/cognitive spare capacity tests, again conditional for the 2-word recall condition; here, the r value was 0.38. All the results in this section were classified as fair.

Behavioural Measures v Physiological Measures – A Between Subgroup Comparison. A total of nine papers compared behavioural and physiological measures of LE (Bertoli & Bodmer, 2014; Gustafson et al., 2018; Key et al., 2017; Rovetti et al., 2019; Strand et al., 2018; Visentin et al., 2021; White & Langdon, 2021; A. Zekveld et al., 2020; Zhao et al., 2019). They conducted 26 correlational analyses, with only seven (26.9%) reaching statistical significance. The seven analyses which reached significance were all conditional. The largest effect size was between hit rate (correct response within a certain time frame)/pupillometry - pupil dilation response, r = - 0.65. This result was a conditional correlation found within the young cohort (age 18-35 years) within the 15–20 s time period after the stimulus. The joint weakest correlation was between reaction times/ pupillometry - mean pupil dilation in the compatible condition and reaction times/pupillometry - peak pupil dilation in the compatible/conflict condition with r values of 0.28 and - 0.28 respectively. Two analyses were poor, four were fair, and one was moderate.

Behavioural Measures v Fatigue Questionnaires – A Between Subgroup Comparison. Six papers reported correlations between behavioural measures of LE and questionnaires designed to measure fatigue, resulting in 19 correlational analyses (Athey, 2016; Gustafson et al., 2018; Hornsby, 2013; Key et al., 2017; Picou et al., 2017; Picou & Ricketts, 2018). Within this category, nine results reached significance; seven of these significant results were conditional, and all were dependent on the time point used to measure LE. The largest coefficient was shown between lapses of attention/fatigue scale within the post-test time condition with an r value - 0.70. The weakest correlation was between reaction times/tiredness questions, with an r-value of 0.22. These results are classified as follows, one poor, seven fair, and one moderate.

Physiological Measures v Other Physiological Measures – A Within Subgroup Comparison. Eight of the 48 papers included in this study compared physiological measures of LE with other physiological measures (Alhanbali et al., 2019; Francis et al., 2021; Giuliani, 2021; Kramer et al., 2016; McMahon et al., 2016; Miles et al., 2017; Seifi Ala et al., 2020; A. A. Zekveld et al., 2019). Altogether, this translated into 43 distinct correlational analyses. For these results, seven (16.3%) reached significance, and 36 failed to reach significance. When examining the significant correlations, this review found that two out of the seven were unconditional correlations, and the remaining five were conditional. The

strongest association existed between skin conductance response quantity/skin conductance response amplitude, with an r-value of 0.89. This was a conditional correlation demonstrated only in the before response cue period. The weakest association was between skin conductance/pupillometry - peak pupil diameter with an r-value of 0.20. These results can be stratified according to Chan's classification as follows, three poor, two fair, one moderate, and one very strong.

Physiological Measures v Fatigue Questionnaires – A Between Subgroup Comparison. Eight papers compare physiological measures against questionnaires related to fatigue (Alhanbali et al., 2019, 2021; Dwyer et al., 2019; Gustafson et al., 2018; Key et al., 2017; McGarrigle et al., 2020; Wang, Kramer, et al., 2018; Wang, Naylor, et al., 2018). These papers produced 30 correlational analyses, 11 (36.7%) reached significance. Eight of the 11 significant results were unconditional correlations. The strongest association existed between pupillometry - task-evoked pupillary response/fatigue scale 0–100. The r-value for this result was - 0.48. The weakest correlation was between pupillometry - peak pupil diameter/visual analogue scale - tiredness with an r-value of - 0.19. The classifications for the significant results were six poor and five fair.

Fatigue Questionnaires v Other Fatigue Questionnaires – A Within Subgroup Comparison. The final correlation comparison between fatigue questionnaires only yielded one trial (Wang, Naylor, et al., 2018). This result demonstrated a significant unconditional correlation between Need for Recovery questionnaire/Checklist Individual Strengths questionnaire. The r-value for this analysis was 0.57, resulting in a fair classification.

Quality Assessment and Risk of Bias

The GRADE approach was applied to the 48 included studies to provide an overview of the quality of the evidence contained within this review. This review utilised either the GRADE criteria or the ROBINS-I. In total, 38 papers were classified as observational and nine as interventional. Overall, the risk of bias was judged to be high/serious for both types of study design. At an individual study level, 38/48 were deemed to have a serious/high risk of bias, 9/48 a moderate risk of bias, and 5/48 a low risk of bias. The rationale behind this assessment was due to several factors. First, 31 out of the 48 studies only included one cohort; 17 of these studies only had a normal hearing group, and 14 only contained individuals with hearing impairment. Inconsistency was also noted between the exclusion criteria applied by each study with regard to possible confounding comorbidities. Some papers accounted for a range of possible pre-existing factors which may influence the measure of LE obtained. Examples include neurological disorders, learning disabilities, psychiatric conditions, chronic fatigue, medication use, smoking history, alcohol intake, and caffeine use. The antithesis was

found whereby some papers only partially attempted to adjust for this potential risk of bias, introducing a high degree of confounder bias. Finally, most papers, 44/48, only measured LE at a single time point and therefore had no follow-up. Generally, the literature mitigated the effects of exposure bias – by ensuring normal hearing participants were screened via audiological testing prior to inclusion - and outcome bias – by keeping missing data to a minimum.

The included studies displayed a large amount of heterogeneity in terms of study design; therefore, 'inconsistency' was judged to affect the quality of evidence seriously. Specifically, a large variation existed between studies regarding which tool was utilised to measure LE. Indeed, variations existed between specific types of tools, for example, differences in the numerical ranges contained within self-reported LE scales. The setting of the study experiments was also inconsistent in terms of listening task and listening conditions. Finally, the study population differed significantly between papers with reference to hearing level and age. 'Imprecision', most studies did not report a sample size calculation, 30/48, an additional 3/48 performed a power calculation but subsequently did not have sufficient sample size, one study sample fulfilled the power calculation for some but not all outcomes, and only 14/48 studies satisfied the required sample size. Therefore, the effect of 'Imprecision' on the quality of evidence was graded as serious. This review applied the criterion of 'indirectness' by considering three factors – Did the study contain hearing-impaired participants? Did the study contain both an adult and paediatric population? Did the study use an accepted measure for LE? Overall, no study satisfied all three factors; 17 met two, 30 met one, and one failed to meet any of the above factors. This finding resulted in 'indirectness' achieving a rating of moderate. Publication bias moderately impacted the quality of evidence due to the fact 42/274 correlational analysis did not report the r values of non-significant results (Page, Higgins, and Sterne 2021). Combining each of the GRADE criteria outlined above translates into an overall quality of evidence rating of very low, which has important implications for the conclusions drawn by this review. This aligns with previous systematic reviews investigating the literature pertaining to LE and reflects one of the inherent challenges to validating a clinical tool to measure this phenomenon (Holman, Drummond, and Naylor 2021; Ohlenforst et al. 2017). A summary of the quality of evidence assessment for the included papers is shown in Table 3 below.

Discussion

This review aimed to determine the association between different instruments used to capture LE by evaluating the correlational analyses performed across the literature. Additionally, the correlations between effort and fatigue instruments were contrasted to investigate the relationship between the two concepts of LE and fatigue.

Table 3. Summary of Risk of Bias Assessment.

Risk of bias tool	Risk of bias	Inconsistency	Imprecision	Indirectness	Publication bias	Quality	Importance
GRADE (n = 39)	High	Serious	Serious	Moderate	Moderate	Very low	Important
ROBINS-I (n = 9)	Serious						

Multi-Dimensional Model of Listening Effort

Spatial Stratification. This review notes that only 36.1% of the 274 correlation analyses reached statistical significance, and correlation strength was most often classified as fair. The distribution of classifications towards fair was found to be statistically significant. This lack of association lobbies the notion that different measures encompass different components of LE. Previous studies have explained this phenomenon by referring to a multi-dimensional model of LE; therefore, this idea is reaffirmed by the findings of this review. (Alhanbali et al. 2019; Francis and Love 2020).

This model outlines that behavioural, physiological, and self-reported measures capture different components of LE (Alhanbali et al. 2019; Francis and Love 2020; Peelle 2018). McMahon et al. (2016) hypothesize that this can be rationalised by considering that different measures align with the different cognitive processes involved with sound perception (McMahon et al. 2016). Only when we understand these cognitive processes will we be able to fully interpret information gathered from measures of LE. Hughes et al. (2018) have attempted to categorise some of these cognitive pathways to help clarify this topic. They outlined three distinct cognitive stages which can give rise to LE. These were described as the effort evoked by 'attending to', 'processing', and 'adapting to' an auditory message. The measures of LE, therefore, may resonate more closely with one particular part of the overall process (Hughes et al. 2018). From a physiological perspective, this theory is reflected by examining functional imaging studies highlighting the myriad of cortical areas involved with understanding degraded speech (Peelle 2018).

Further to these intrinsic processes involved in speech perception, individual factors may also contribute to the effort experienced. Francis and Love (2020) delineated the differences between exerted and assessed effort in that the tangible cognitive resources required to match the demands of a task may not be congruent to the perceived effort required as interpreted by the individual listener. This theory also accounts for the interrelated effects of motivation and pleasure on the perception of effort (Pichora-Fuller et al. 2016). Therefore, a combination of exerted effort and assessed effort contribute to the overall LE experienced (Francis and Love 2020). From this depiction, it is clear to see the complexities involved with LE mechanisms and the challenges associated with its measurement.

Together, the lack of association between subgroups demonstrated across the literature and the cognitive mechanisms discussed above suggest that measures of LE are spatially stratified according to which components of the cognitive pathways they encapsulate. This means that behavioural, physiological, and self-reported measures have the potential to encode different aspects of the neural and cognitive routes involved with effort perception.

Effect of Experimental Conditions. To further build upon the aforementioned model, it is important to consider the effect that experimental conditions have on the degree of association between measures of LE. As stated, significant correlations between measures can and do occur, a finding which would seemingly refute the premise of the multi-dimensional nature of LE. To help explain this incongruency, this review highlights the fragility of significant correlations by referencing the influence of experimental conditions on the strength of correlations. This is reflected by the fact that 59/99 (56%) of the significant correlation were related to a specific experimental condition. Changing either the listening task or the timeframe LE was measured has been shown to alter the strength of correlation between measures of LE. This inconsistency may reflect the sensitivity of different instruments and their reliance on the cognitive capacity model (Kahneman 1973; Pichora-Fuller et al. 2016). This model posits that more difficult scenarios require individuals to draw upon more cognitive reserves. The more of these limited resources the individual uses, the more effort they experience. The ability of a measure to capture LE is, therefore, dependent on the relationship between its sensitivity and the task difficulty (Hick and Tharpe 2002). Differences in the sensitivity of each measure make it challenging to demonstrate associations in easier listening conditions hence explaining why the strength of correlations appears to change based upon the timeframe LE was sampled and the task used to evoke the effort. This highlights how measures of LE are context dependant and thus are appropriately termed 'markers' rather than an 'invariant' as a one to one relationship between instrument and outcome cannot be demonstrated (Richter and Slade 2017). Additionally, cautions against generalising the results of one paper across the entire literature. One possible way to mitigate this effect is to conduct more ecological studies outside the laboratory to help encapsulate real-world conditions more representationally (Bruya and Tang 2018).

To analyse this finding, and in keeping with the astronomical theme of multi-dimensions, this paper denotes a model of LE which accounts for the conditional effect on the association between measures, termed the solar system model of LE. In this model, LE is depicted as the sun, and the different measurement tools take the role of the orbiting planets.

When the various planets align (Experiments with difficult listening situations), they absorb the same solar outputs (LE dimensions). This mirrors the notion that when the conditions are right, the measures capture the same dimensions of LE. As the experimental conditions become easier, the sensitivity of each measure pulls the planets out of alignment, and the solar outputs they capture are different. This suggests that LE measures depend on the experimental conditions within which they are applied. A diagrammatic representation of this model is shown in Figure 3 below. This model only applies to a minority of LE measures; the remainder, as shown by this review, encompass different dimensions of LE no matter what experimental conditions are enacted.

Temporal Stratification. The final observation noted from this review is the lack of association between different measures within the same subgroup. This can be evidenced by considering two factors:

1. How many 'within subgroup' comparisons reached statistical significance? (See Table 2)
2. What strength did the statistically significant results reach? (See Table 1)

The comparisons between EQ - EQ, B - B, and FQ - FQ subgroups satisfy the first point above, with the percentage of trials reaching significance equating to 100%, 66.7%, and 100%, respectively. Despite this, the modal classification for each remained as fair (coefficient 0.3 to 0.6 or −0.3 to −0.6). Moreover, the P - P subgroup failed to meet either

point. Indeed, the number of significant results totalled 16.3%, representing the lowest from any subgroup comparison, and the modal classification was poor (coefficient 0 to 0.3 or 0 to −0.3). Interestingly, six of the seven significant results within the P - P subgroup comparison involved measures that utilise the same underlying mechanism to sample effort. For instance, Skin conductance/EEG alpha were significantly correlated, and both measure the electrophysiological response to the effort. A significant correlation was also noted between Cortisol/Chromogranin A with both being hormonal measures. Given this disparity amongst physiological measures, it may be more accurate to subclassify them according to their mechanism of action. Logical groupings include functional imaging, electrophysiological measures, and hormonal measures.

Together, these results suggest that measures of LE may also be temporally stratified along a timeline which begins at the onset of an effortful stimulus. For example, functional imaging measures such as infra-red spectroscopy may reflect the more real-time cerebral processes that result from LE. Electrophysiological measures, including pupillometry, capture the subsequent neuronal response to these processes. Followed by the cascade of endocrine mechanisms recruited in response to effort, which would be sampled using hormonal measures of LE such as cortisol. Subsequently, behavioural measures would then encode for the immediate reactive response of the individual exposed to effort via their reaction times. Finally, both effort and fatigue questionnaires would capture the downstream consequences of LE, which manifests much further
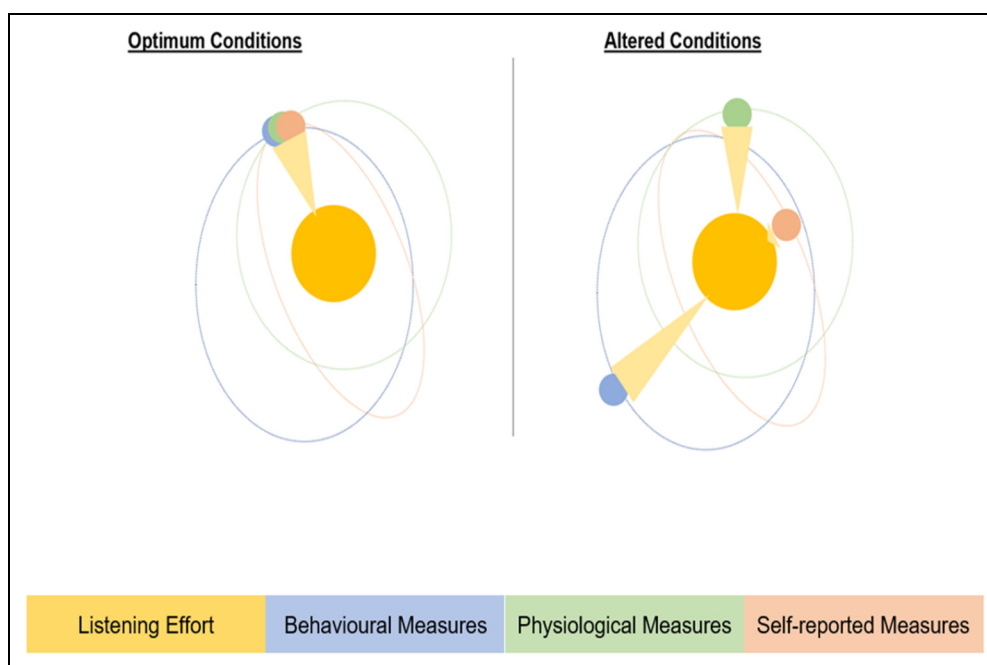

Figure 3. Illustration of the Effects Experimental Conditions Have When Measuring le. The Component of LE Captured is Represented by the Solar Rays.

along the effort perception pathway. The decision to group the subtypes of questionnaires together can be rationalised by examining that 79.2% of trials that compared EQ - FQ reached significance satisfying the first point above. A representation of the temporal stratification concept is found in Figure 4 below.

The spatial and temporal stratification of LE measures explains why the multi-dimensional model exists and why specific tools are weakly correlated. Moreover, the fragility of significant correlations has been explained by examining the effect of experimental conditions on the strength of association between two tools. Accepting these findings clarifies why the literature has not been able to pin down a single optimum measure for LE and advocates exploring different approaches to incorporate LE into clinical assessment.

An Alternative Approach to Measuring Listening Effort

Given the seemingly perpetual challenge of selecting one measure of LE to clinically assess individuals with hearing impairment, it is prudent to consider other possible alternatives to help document the abstract notion. As previously discussed, there are several recognised consequences associated with LE, which may represent a more suitable target for screening tools. Furthermore, this review has shown that fatigue questionnaires perform similarly to effort indices in the classification matrix of subgroups. Therefore, shifting attitudes towards capturing these downstream effects creates an opportunity to design a composite measure comprised of previously validated tools.

First, to start with fatigue, many of the papers included in this review have referred to fatigue, and some have already adopted the approach of measuring this metric alongside effort in hearing-impaired individuals (Alhanbali et al. 2018, 2019; Dwyer et al. 2019; Key et al. 2017; Picou, Moore, and Ricketts 2017; Wang et al. 2018). Many of the proposed measures for LE demonstrate considerable overlap with fatigue. Take, for instance, the use of cortisol, performance over time or effort/fatigue assessment scales (Bess and Hornsby 2014). The evidence for the connection
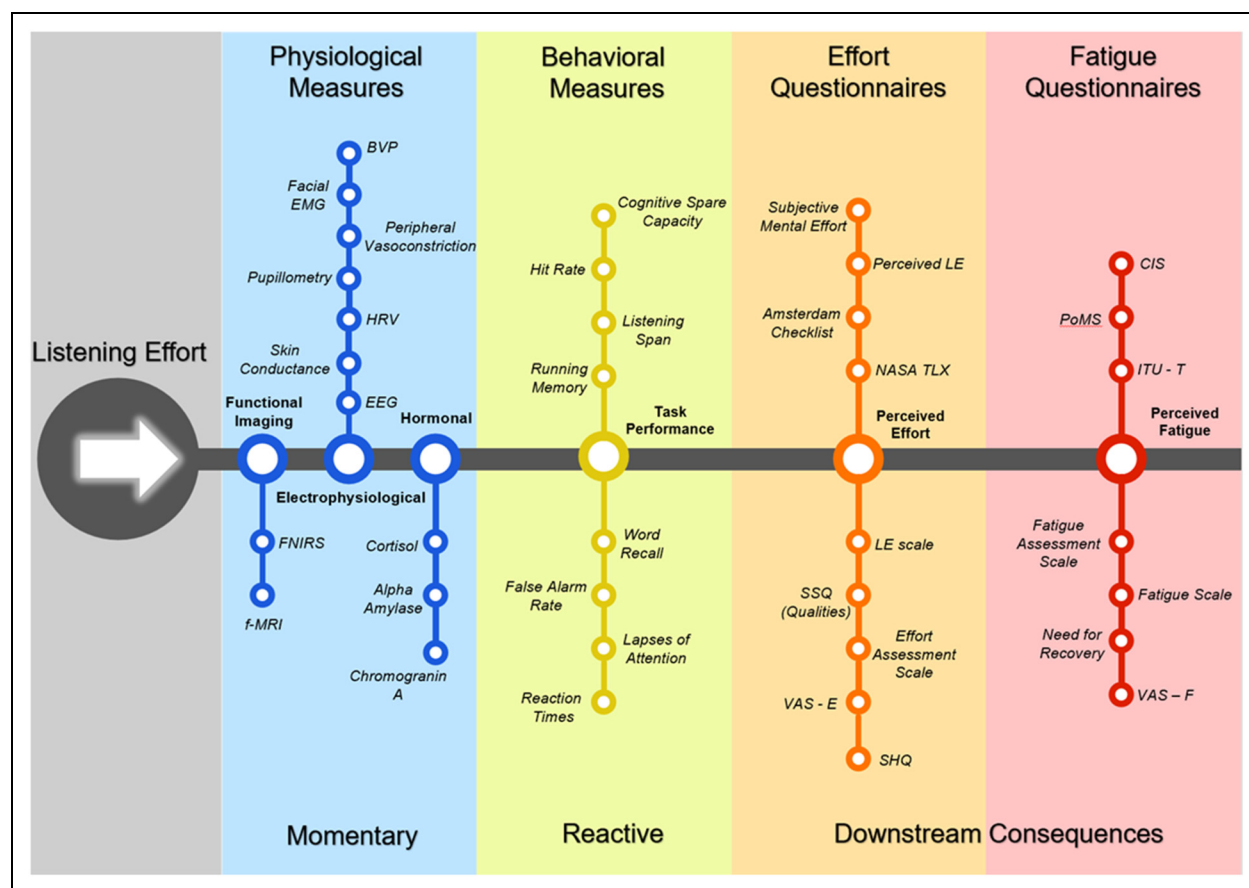

Figure 4. Metro Map Model Indicating the Time Point Each Tool Captures LE. (BVP = Blood Volume Pulse, EMG = Electromyogram, HRV = Heart Rate Variability, EEG = Electroencephalogram, FNIRS = Functional Near Infra-Red Spectroscopy, f-MRI = Functional Magnetic Resonance Imaging, NASA TLX = National Aeronautics and Space Administration Task Load Index, SSQ = Speech, Spatial and Qualities Questionnaire, VAS – E = Visual Analogue Scale – Effort, SHQ = Spatial Hearing Questionnaire, CIS = Checklist Individual Strengths, PoMS = Profile of Mood States, ITU-T = International Telecommunications Union – Tiredness, VAS – F = Visual Analogue Scale – Fatigue.

between effort and fatigue is becoming increasingly formidable and is now widely accepted as a consequence of prolonged intensive listening. Similarly to LE, fatigue displays multi-dimensional characteristics making it challenging to measure (Bess and Hornsby 2014; Gawron 2016). As a result, many self-reported tools have been proposed for the adult population. Gawron (2016) provides a detailed overview of these fatigue questionnaires and provides information on their possible optimal application (Gawron 2016). Examples that may be relevant to LE include the Checklist for individuals Strength's questionnaire due to its application to chronic fatigue and visual analogue scales for fatigue which have been widely used for hearing-impaired individuals already (Alhanbali et al. 2021; Gawron 2016). Regarding individuals with hearing loss, recent work by Hornsby et al. has validated a fatigue scale that aims to capture hearing-related fatigue specifically – The Vanderbilt Fatigue Scale. This instrument has now been validated in adults and children (Hornsby et al. 2021, 2022).

The next consequence to consider is the stress generated from effortful listening. Again, several studies included in this review have acknowledged stress as a corollary of LE and attempted to represent this within the testing strategy (Dwyer et al. 2019; Kramer et al. 2016; Mackersie, MacPhee, and Heldt 2015; Zekveld et al. 2019). Stress accumulates due to over-activation of autonomic processes and disruption of the endocrine system's cortisol axis (Sharma 2021). Given this, physiological measures (Cortisol, Skin conductance, Heart rate variability) which directly sample these changes may provide the most accurate insight. Their use in clinical practice is hindered by the ability to test repeatedly to establish a longitudinal trend of stress levels. An alternative to these objective measures is the Perceived Stress Scale (PSS), first proposed by Cohen in the 1980s (Cohen, Kamarck, and Mermelstein 1983). The PSS is the most widely used self-reported measure for stress and has been shown to apply to adults, children, and across ethnic groups (Baik et al. 2019; Huang et al. 2020; White 2014).

Outside of fatigue and stress, which are well-documented consequences of LE, there are more subtle effects that this paper will now explore. Individuals with hearing impairment have been shown to suffer lower levels of self-assurance and confidence compared with normal hearing counterparts (Bat-Chava 1993; Theunissen et al. 2014). Whilst this is generally accepted, an interesting observation was noted by Warner-Czyz et al., whereby children who had received a hearing intervention (hearing aid or cochlear implant) rated their self-esteem higher than normal-hearing children (Warner-Czyz et al. 2015). This highlights the reversibility of the detrimental impact hearing loss has on self-worth and the potential for confidence to be used as an indicator of benefit in before and after studies. The relationship between hearing ability and confidence is likely a multifactorial process influenced by factors such as the ability to communicate, social connectedness, physical appearance, and

personal circumstances. Notwithstanding these many facets, qualitative data from Hughes et al. (2018) did elicit a relationship between low confidence and the polarity between listening effort and reward through focus group interviews (Hughes et al. 2018). Further to these internal determinants of confidence, educational outcomes are closely linked to an individual's sense of esteem (Rubin, Dorle, and Sandidge 1977). As elaborated previously, many studies have observed a trend for poorer school-based outcomes amongst those with hearing impairment (Bess et al. 1998; Niedzielski et al. 2006; Purcell et al. 2016). It is therefore plausible to suggest that LE may adversely affect self-esteem levels through its connection to the aforementioned factors. The Rosenberg scale (1965) is perhaps the most widely utilised measure for self-esteem (Rosenberg 1965). It has undergone rigorous evaluation and consistently yielded impressive reliability indices across age groups, ethnicities, languages, and genders (Dukes and Martinez 1994). Within the UK population, the Rosenberg scale was demonstrated to have a Cronbach alpha level of 0.9, indicating strong internal consistency (Schmitt and Allik 2005). Moreover, it has already been used within the hearing-impaired population (Theunissen et al. 2014; Warner-Czyz et al. 2015). Therefore, the Rosenberg scale represents a strong candidate for a composite tool to measure the possible confidence sequalae emanating from LE.

The final downstream effects of LE identified through this review is an individual's desire to take control of a situation or to give up on a listening task. This was enshrined within the work of Picou and colleagues in the effect of hearing aid settings on subjective ratings of LE (Picou et al. 2017; Picou and Ricketts 2018). The idea is that at increasing levels of task difficulty, individuals would be more likely to act to change something about the situation (move to a quieter room) or give up on the task altogether (Picou and Ricketts 2018). As opposed to the other downstream effects discussed in this section, these reflect more acute consequences of LE and will depend heavily on the task's difficulty. The decision to take control and give up on a task represents reactive tipping points for action in response to effortful listening. They used specific questions to assess these phenomena – 'How likely would you be to try to do something else to improve the situation (e.g., move to a quiet room, ask the speaker to speak louder)?' and 'How likely would you be to give up or just stop trying?' (Picou and Ricketts 2018).

Justification for a Questionnaire-Based Approach

This review has provided evidence for the various downstream effects of LE in the form of – fatigue, stress, low self-assurance, and increased desire to take control and give up on a task. They represent some of the acute and chronic consequences faced by individuals exposed to prolonged periods of intense listening. Given the widespread use of fatigue

scales in LE studies and the difficulties associated with measuring LE directly, this paper advocates for a shift in approaches to focus on capturing these negative sequelae. As discussed, several self-reported tools are already validated for each domain, creating the possibility of amalgamating the most relevant parts of each tool into a single measure which provides a holistic insight into the quality-of-life detriments stemming from LE.

Additionally, questionnaire-based instruments are less resource-intensive than more complex physiological and behaviour measures of LE. This method allows for longitudinal sampling of LE, which helps to trend how the individual is affected over time (Farrington 1991). This approach reflects the dynamic nature of LE. It would also allow for 'real-world' data collection which may help alleviate some of the issues that stem from the influence that experimental conditions have on LE instruments discussed previously (Bruya and Tang 2018; Katkade, Sanders, and Zou 2018). Finally, since questionnaires can be administered outside the hospital/laboratory setting, this approach would align with the post-COVID shift to try and move towards a remote consultation model wherever possible (Gupta, Gkiousias, and Bhutta 2021; Hutchings 2020; Samarrai et al. 2021).

Review Limitations

Several factors limit the results obtained through this systematic review. The most important relates to the quality of included studies. Many papers were graded as having a serious risk of bias and low overall quality, which has overt connotations for the above inferences. The low sample size of many of the included papers may have limited ability to demonstrate statistical significance, creating the possibility for the results to appear poorly correlated when in fact, this may be due to poor study design. Linked to study design, publication bias my have artificially influenced the observed results due to a tendency for those articles demonstrating significant results being looked upon more favourably in publication process (Nair 2019). Thus, even weaker correlations may have been noted from unpublished research. Furthermore, the high degree of heterogeneity between included studies prevented the use of meta-analytical methods for data synthesis. Another inherent challenge with reviewing correlational data from a heterogenous study pool, is the difficulty in using quantitative methods to compare individual analyses. Although this review did not find a difference in the strength of correlations observed for adults and children, only three papers out of the 48 included studies contained a paediatric population. This small sample limits the ability to draw inferences and should be addressed in future work. Finally, a date limit of 2000 was placed on the search strategy of this review. While the specific terminology of "listening effort" may not have been used commonly prior to this date, there is still potential for some relevant articles which explore the effort related to

hearing being missed. For example, Feuerstein correlated "perceived ease of listening" to word recognition in 1992, and Kramer correlated "Self-rated handicap" to pupil dilation in 1996 (Feuerstein 1992; Kramer et al. 1997).

Future Work

Future work should focus on more robust studies to compare different measures of LE, including hearing-impaired individuals alongside normal hearing controls, measuring LE at multiple timepoints, ensuring sample size meets power requirements and the inclusion of paediatric participants. This will help address some of the fundamental limitations of this review and provide further clarification regarding the multi-dimensional model of LE. This paper has probed into some of the downstream effects of LE; however, the list created above is not exhaustive, and further research should aim to identify other potential consequences. Finally, exploratory research should be conducted to assess whether a composite tool, proposed within the discussion section, may be feasible as a proxy measure for LE.

Conclusion

The results of this review have shown that the LE measures are poorly correlated, further supporting the notion of the multi-dimensional model of LE. Taken further, this review has identified some of the potential dimensions that may impact LE, namely the cognitive processes involved with LE (Spatial dimension), the experimental conditions used to elicit LE and the time lapse between stimulus and dimension along the effort pathway measured (temporal dimension). This may indicate that no one measure will be adequate to capture LE in its entirety. To overcome this problematic finding, this review discusses the possibility of shifting focus from measuring LE directly to measuring the consequences of prolonged exposure to effortful listening. These include, but are not limited to, fatigue, stress, low confidence, and increased desire to take control of a situation or give up on a task. A composite tool that comprises components of previously validated measures for these effects may provide a better insight into the burden of LE.

ORCID iDs

Callum Shields https://orcid.org/0000-0002-6232-3898
Karolina Kluk https://orcid.org/0000-0003-3638-2787

Supplemental Material

Supplemental material for this article is available online.

References

Included studies

Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. J. (2017). Self-Reported listening-related effort and fatigue in hearing-impaired adults. *Ear and Hearing*, *38*(1), e39. https://doi.org/10.1097/AUD.0000000000000361

Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. J. (2018). Hearing handicap and speech recognition correlate with self-reported listening effort and fatigue. *Ear and Hearing*, *39*(3), 470–474. https://doi.org/10.1097/AUD.0000000000000515

Alhanbali, S., Dawes, P., Millman, R. E., & Munro, K. J. (2019). Measures of listening effort are multi-dimensional. *Ear and Hearing*, *40*(5), 1084–1097. https://doi.org/10.1097/AUD.0000000000000697

Alhanbali, S., Munro, K. J., Dawes, P., Carolan, P. J., & Millman, R. E. (2021). Dimensions of self-reported listening effort and fatigue on a digits-in-noise task, and association with baseline pupil size and performance accuracy. *International Journal of Audiology*, *60*(10), 762–772. https://doi.org/10.1080/14992027.2020.1853262

Athey, H. (2016). '*The effect of auditory fatigue on reaction time in normal hearing list' by Haley Athey*. Retrieved 19 December 2021, from https://commons.lib.jmu.edu/diss201019/108/

Bernarding, C., Strauss, D. J., Hannemann, R., Seidler, H., & Corona-Strauss, F. I. (2017). Neurodynamic evaluation of hearing aid features using EEG correlates of listening effort. *Cognitive Neurodynamics*, *11*(3), 203–215. https://doi.org/10.1007/s11571-017-9425-5

Bertoli, S., & Bodmer, D. (2014). Novel sounds as a psychophysiological measure of listening effort in older listeners with and without hearing loss. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *125*(5), 1030–1041. https://doi.org/10.1016/j.clinph.2013.09.045

Desjardins, J. L. (2016). The effects of hearing aid directional microphone and noise reduction processing on listening effort in older adults with hearing loss. *Journal of the American Academy of Audiology*, *27*(1), 29–41. https://doi.org/10.3766/jaaa.15030

Dimitrijevic, A., Smith, M. L., Kadis, D. S., & Moore, D. R. (2019). Neural indices of listening effort in noisy environments. *Scientific Reports*, *9*(1), 11278. https://doi.org/10.1038/s41598-019-47643-1

Dwyer, R. T., Gifford, R. H., Bess, F. H., Dorman, M., Spahr, A., & Hornsby, B. W. Y. (2019). Diurnal cortisol levels and subjective ratings of effort and fatigue in adult cochlear implant users: A pilot study. *American Journal of Audiology*, *28*(3), 686–696. https://doi.org/10.1044/2019_AJA-19-0009

Finke, M., Büchner, A., Ruigendijk, E., Meyer, M., & Sandmann, P. (2016). On the relationship between auditory cognition and speech intelligibility in cochlear implant users: An ERP study. *Neuropsychologia*, *87*, 169–181. https://doi.org/10.1016/j.neuropsychologia.2016.05.019

Francis, A. L., Bent, T., Schumaker, J., Love, J., & Silbert, N. (2021). Listener characteristics differentially affect self-reported and physiological measures of effort associated with two challenging listening conditions. *Attention, Perception & Psychophysics*, *83*(4), 1818–1841. https://doi.org/10.3758/s13414-020-02195-9

Giuliani, N. P. (n.d.). *Comparisons of physiologic and psychophysical measures of listening effort in normal-hearing adults* [University of Iowa]. Retrieved 20 December 2021, from https://iro.uiowa.edu/esploro/outputs/doctoral/Comparisons-of-physiologic-and-psychophysical-measures/9983776858302771

Gustafson, S. J., Key, A. P., Hornsby, B. W. Y., & Bess, F. H. (2018). Fatigue related to speech processing in children with hearing loss: Behavioral, subjective, and electrophysiological measures. *Journal of Speech, Language, and Hearing Research: JSLHR*, *61*(4), 1000–1011. https://doi.org/10.1044/2018_JSLHR-H-17-0314

Holube, I., Haeder, K., Imbery, C., & Weber, R. (2016). Subjective listening effort and electrodermal activity in listening situations with reverberation and noise. *Trends in Hearing*, *20*, 2331216516667734. https://doi.org/10.1177/2331216516667734

Hornsby, B. W. Y. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and Hearing*, *34*(5), 523–534. https://doi.org/10.1097/AUD.0b013e31828003d8

Johnson, J., Xu, J., Cox, R., & Pendergraft, P. (2015). A comparison of two methods for measuring listening effort as part of an audiologic test battery. *American Journal of Audiology*, *24*(3), 419–431. https://doi.org/10.1044/2015_AJA-14-0058

Key, A. P., Gustafson, S. J., Rentmeester, L., Hornsby, B. W. Y., & Bess, F. H. (2017). Speech-processing fatigue in children: Auditory event-related potential and behavioral measures. *Journal of Speech, Language, and Hearing Research : JSLHR*, *60*(7), 2090–2104. https://doi.org/10.1044/2016_JSLHR-H-16-0052

Kramer, S. E., Teunissen, C. E., & Zekveld, A. A. (2016). Cortisol, chromogranin A, and pupillary responses evoked by speech recognition tasks in normally hearing and hard-of-hearing listeners: A pilot study. *Ear and Hearing*, *37*(Suppl 1), 126S–135S. https://doi.org/10.1097/AUD.0000000000000311

Lau, M. K., Hicks, C., Kroll, T., & Zupancic, S. (2019). Effect of auditory task type on physiological and subjective measures of listening effort in individuals with Normal hearing. *Journal of Speech, Language, and Hearing Research: JSLHR*, *62*(5), 1549–1560. https://doi.org/10.1044/2018_JSLHR-H-17-0473

Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indices of listening effort in a competing-talker task. *Journal of the American Academy of Audiology*, *22*(2), 113–122. https://doi.org/10.3766/jaaa.22.2.6

Mackersie, C. L., MacPhee, I. X., & Heldt, E. W. (2015). Effects of hearing loss on heart-rate variability and skin conductance measured during sentence recognition in noise. *Ear and Hearing*, *36*(1), 145–154. https://doi.org/10.1097/AUD.0000000000000091

Magudilu Srishyla Kumar, L. (2020). *Measuring listening effort using physiological, behavioral and subjective methods in normal hearing subjects: Effect of signal to noise ratio and presentation level* [James Madison University]. https://commons.lib.jmu.edu/diss202029/7

McGarrigle, R., Rakusen, L., & Mattys, S. (2020). Effortful listening under the microscope: Examining relations between

pupillometric and subjective markers of effort and tiredness from listening. *Psychophysiology* 58(1).https://doi.org/10.31234/osf.io/kych5

McMahon, C. M., Boisvert, I., de Lissa, P., Granger, L., Ibrahim, R., Lo, C. Y., Miles, K., & Graham, P. L. (2016). Monitoring alpha oscillations and pupil dilation across a performance-intensity function. *Frontiers in Psychology*, 7, 745. https://doi.org/10.3389/fpsyg.2016.00745

Megha, null, & Maruthy, S. (2019). Auditory and cognitive attributes of hearing aid acclimatization in individuals with sensorineural hearing loss. *American Journal of Audiology*, 28(2S), 460–470. https://doi.org/10.1044/2018_AJA-IND50-18-0100

Miles, K., McMahon, C., Boisvert, I., Ibrahim, R., de Lissa, P., Graham, P., & Lyxell, B. (2017). Objective assessment of listening effort: Coregistration of pupillometry and EEG. *Trends in Hearing*, 21, 2331216517706396. https://doi.org/10.1177/2331216517706396

Perreau, A. E., Wu, Y.-H., Tatge, B., Irwin, D., & Corts, D. (2017). Listening effort measured in adults with Normal hearing and cochlear implants. *Journal of the American Academy of Audiology*, 28(8), 685–697. https://doi.org/10.3766/jaaa.16014

Picou, E. M., Moore, T. M., & Ricketts, T. A. (2017). The effects of directional processing on objective and subjective listening effort. *Journal of Speech, Language, and Hearing Research: JSLHR*, 60(1), 199–211. https://doi.org/10.1044/2016_JSLHR-H-15-0416

Picou, E. M., & Ricketts, T. A. (2018). The relationship between speech recognition, behavioural listening effort, and subjective ratings. *International Journal of Audiology*, 57(6), 457–467. https://doi.org/10.1080/14992027.2018.1431696

Rovetti, J., Goy, H., Pichora-Fuller, M. K., & Russo, F. A. (2019). Functional near-infrared spectroscopy as a measure of listening effort in older adults who use hearing aids. *Trends in Hearing*, 23, 2331216519886722. https://doi.org/10.1177/2331216519886722

Russo, F. Y., Hoen, M., Karoui, C., Demarcy, T., Ardoint, M., Tuset, M.-P., De Seta, D., Sterkers, O., Lahlou, G., & Mosnier, I. (2020). Pupillometry assessment of speech recognition and listening experience in adult cochlear implant patients. *Frontiers in Neuroscience*, 14, 556675. https://doi.org/10.3389/fnins.2020.556675

Schafer, P. J., Serman, M., Arnold, M., Corona-Strauss, F. I., Strauss, D. J., Seidler-Fallbohmer, B., & Seidler, H. (2015). Evaluation of an objective listening effort measure in a selective, multi-speaker listening task using different hearing aid settings. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, 2015, 4647–4650. https://doi.org/10.1109/EMBC.2015.7319430

Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech, Language, and Hearing Research: JSLHR*, 58(6), 1781–1792. https://doi.org/10.1044/2015_JSLHR-H-14-0180

Seifi Ala, T., Graversen, C., Wendt, D., Alickovic, E., Whitmer, W. M., & Lunner, T. (2020). An exploratory study of EEG alpha oscillation and pupil dilation in hearing-aid users during effortful listening to continuous speech. *PloS One*, 15(7), e0235782. https://doi.org/10.1371/journal.pone.0235782

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research: JSLHR*, 61(6), 1463–1486. https://doi.org/10.1044/2018_JSLHR-H-17-0257

Visentin, C., & Prodi, N. (2021). How reliable are 11- to 13-year-olds' self-ratings of effort in noisy conditions? *Frontiers in Built Environment*, 7, 108. https://doi.org/10.3389/fbuil.2021.688016

Visentin, C., Prodi, N., Cappelletti, F., Torresin, S., & Gasparella, A. (2019). Speech intelligibility and listening effort in university classrooms for native and non-native Italian listeners. *Building Acoustics*, 26, 1351010X1988231. https://doi.org/10.1177/1351010X19882314

Visentin, C., Valzolgher, C., Pellegatti, M., Potente, P., Pavani, F., & Prodi, N. (2021). A comparison of simultaneously-obtained measures of listening effort: Pupil dilation, verbal response time and self-rating. *International Journal of Audiology*, 61(7), 1–13. https://doi.org/10.1080/14992027.2021.1921290

Wang, Y., Kramer, S. E., Wendt, D., Naylor, G., Lunner, T., & Zekveld, A. A. (2018). The pupil dilation response during speech perception in dark and light: The involvement of the parasympathetic nervous system in listening effort. *Trends in Hearing*, 22, 2331216518816603. https://doi.org/10.1177/2331216518816603

Wang, Y., Naylor, G., Kramer, S. E., Zekveld, A. A., Wendt, D., Ohlenforst, B., & Lunner, T. (2018). Relations between self-reported daily-life fatigue, hearing Status, and pupil dilation during a speech perception in noise task. *Ear and Hearing*, 39(3), 573–582. https://doi.org/10.1097/AUD.0000000000000512

White, B. E., & Langdon, C. (2021). The cortical organization of listening effort: New insight from functional near-infrared spectroscopy. *NeuroImage*, 240, 118324. https://doi.org/10.1016/j.neuroimage.2021.118324

Wisniewski, M. G., Thompson, E. R., Iyer, N., Estepp, J. R., Goder-Reiser, M. N., & Sullivan, S. C. (2015). Frontal midline θ power as an index of listening effort. *Neuroreport*, 26(2), 94–99. https://doi.org/10.1097/WNR.0000000000000306

Wu, Y.-H., Aksan, N., Rizzo, M., Stangl, E., Zhang, X., & Bentler, R. (2014). Measuring listening effort: Driving simulator versus simple dual-task paradigm. *Ear and Hearing*, 35(6), 623–632. https://doi.org/10.1097/AUD.0000000000000079

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, 32(4), 498–510. https://doi.org/10.1097/AUD.0b013e31820512bb

Zekveld, A., Scheepen, J., Versfeld, N., Kramer, S., & Steenbergen, H. (2020). The influence of hearing loss on cognitive control in an auditory conflict task: Behavioral and pupillometry findings. *Journal of Speech, Language, and Hearing Research*, 63, 1–10. https://doi.org/10.1044/2020_JSLHR-20-00107

Zekveld, A. A., van Scheepen, J. A. M., Versfeld, N. J., Veerman, E. C. I., & Kramer, S. E. (2019). 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. *Hearing Research*, 381, 107768. https://doi.org/10.1016/j.heares.2019.07.005

Zhao, S., Bury, G., Milne, A., & Chait, M. (2019). Pupillometry as an objective measure of sustained attention in young and older listeners. *Trends in Hearing*, 23, 2331216519887815. https://doi.org/10.1177/2331216519887815

General references

Andrade, C. (2019). The P value and statistical significance: Misunderstandings, explanations, challenges, and alternatives. *Indian Journal of Psychological Medicine*, 41(3), 210–215. https://doi.org/10.4103/IJPSYM.IJPSYM_193_19

Archbold, S., Ng, Z. Y., Harrigan, S., Gregory, S., Holland, L., & Mulla, I. (2015). 'Views of Parents and Teachers'. 48.

Armstrong, E. C. (1999). The well-built clinical question: The key to finding the best evidence efficiently. *WMJ: Official Publication of the State Medical Society of Wisconsin*, 98(2), 25–28.

Ayres, P., Lee, J. Y., Paas, F., & van Merriënboer, J. J. G. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology*, 12, 3905. https://doi.org/10.3389/fpsyg.2021.702538

Baik, S. H., Fox, R. S., Mills, S. D., Roesch, S. C., Sadler, G. R., Klonoff, E. A., & Malcarne, V. L. (2019). Reliability and validity of the perceived stress scale-10 in hispanic Americans with English or spanish language preference. *Journal of Health Psychology*, 24(5), 628–639. https://doi.org/10.1177/1359105316684938

Bat-Chava, Y. (1993). Antecedents of self-esteem in deaf people: A meta-analytic review. *Rehabilitation Psychology*, 38(4), 221–234. https://doi.org/10.1037/h0080303

Bess, F. H., Dodd-Murphey, J., & Parker, R. (1998). 'Children with Minimal Sensorineural Hearing Loss: Prevalence, Educational Performance, and Functional Status. *Ear and Hearing*, 19(5), 339–354. https://doi.org/10.1097/00003446-199810000-00001

Bess, F. H., Gustafson, S. J., Corbett, B. A., Warren Lambert, E., Camarata, S. M., & Hornsby, B. W. Y. (2016). Salivary cortisol profiles of children with hearing loss. *Ear and Hearing*, 37(3), 334–344. https://doi.org/10.1097/AUD.0000000000000256

Bess, F. H., & Hornsby, B. W. Y. (2014). Commentary: Listening can be exhausting–fatigue in children and adults with hearing loss. *Ear and Hearing*, 35(6), 592–599. https://doi.org/10.1097/AUD.0000000000000099

Bruya, B., & Tang, Y.-Y. (2018). 'Is attention really effort? Revisiting Daniel Kahneman's influential 1973 book attention and effort'. *Frontiers in Psychology*, 9, 1133. https://doi.org/10.3389/fpsyg.2018.01133

Chan, Y. H. (2003). Biostatistics 104: Correlational analysis. *Singapore Medical Journal*, 44(12), 614–619.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, 24(4), 385–396. https://doi.org/10.2307/2136404

Downs, D. W., & Crum, M. A. (1978). Processing demands during auditory learning under degraded listening conditions. *Journal of Speech and Hearing Research*, 21(4), 702–714. https://doi.org/10.1044/jshr.2104.702

Dukes, R. L., & Martinez, R. (1994). The impact of ethgender on Self-Esteem among adolescents. *Adolescence*, 29(113), 105–115.

Ebell, M. (1999). Information at the point of care: Answering clinical questions. *The Journal of the American Board of Family Practice*, 12(3), 225–235. https://doi.org/10.3122/jabfm.12.3.225

Farrington, D. P. (1991). Longitudinal research strategies: Advantages, problems, and prospects. *Journal of the American Academy of Child & Adolescent Psychiatry*, 30(3), 369–374. https://doi.org/10.1097/00004583-199105000-00003

Feuerstein, J. F. (1992). Monaural versus binaural hearing: Ease of listening, word recognition, and attentional effort. *Ear and Hearing*, 13(2), 80–86. https://doi.org/10.1097/00003446-199204000-00003

Francis, A. L., & Love, J. (2020). 'Listening effort: Are we measuring cognition or affect, or both?'. *WIRES Cognitive Science*, 11(1), e1514. https://doi.org/10.1002/wcs.1514

Gawron, V. J. (2016). Overview of self-reported measures of fatigue. *The International Journal of Aviation Psychology*, 26(3–4), 120–131. https://doi.org/10.1080/10508414.2017.1329627

Giuliani, N. P. (2021). 'Comparisons of Physiologic and Psychophysical Measures of Listening Effort in Normal-Hearing Adults'. University of Iowa.

Gupta, T., Gkiousias, V., & Bhutta, M. F. (2021). A systematic review of outcomes of remote consultation in ENT. *Clinical Otolaryngology*, 46(4), 699–719. https://doi.org/10.1111/coa.13768

Hart, S. G. (2006). Nasa-Task load Index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9), 904–908. https://doi.org/10.1177/154193120605000909

Henshaw, H., Sharkey, L., Crowe, D., & Ferguson, M. (2015). Research priorities for mild-to-moderate hearing loss in adults. *The Lancet*, 386(10009), 2140–2141. https://doi.org/10.1016/S0140-6736(15)01048-X

Hick, C. B., & Tharpe, A. M. (2002). 'Listening effort and fatigue in school-age children with and without hearing loss'. *Journal of Speech, Language, and Hearing Research: JSLHR*, 45(3), 573–584. https://doi.org/10.1044/1092-4388(2002/046)

Holman, J. A., Drummond, A., & Naylor, G. (2021). The effect of hearing loss and hearing device fitting on fatigue in adults: A systematic review. *Ear and Hearing*, 42(1), 1–11. https://doi.org/10.1097/AUD.0000000000000909

Hornsby, B. W. Y., Camarata, S., Cho, S.-J., Davis, H., McGarrigle, R., & Bess, F. H. (2021). Development and validation of the vanderbilt fatigue scale for adults (VFS-A). *Psychological Assessment*, 33(8), 777–788. https://doi.org/10.1037/pas0001021

Hornsby, B. W. Y., Camarata, S., Cho, S.-J., Davis, H., McGarrigle, R., & Bess, F. H. (2022). 'Development and evaluation of pediatric versions of the vanderbilt fatigue scale for children with hearing loss'. *Journal of Speech, Language, and Hearing Research*, 65(6), 2343–2363. https://doi.org/10.1044/2022_JSLHR-22-00051

Hsu, B. C.-L., Vanpoucke, F., & van Wieringen, A. (2017). Listening effort through depth of processing in school-age children. *Ear and Hearing*, 38(5), 568–576. https://doi.org/10.1097/AUD.0000000000000436

Huang, F., Wang, H., Wang, Z., Zhang, J., Du, W., Su, C., Jia, X., Ouyang, Y., Wang, Y., Li, L., Jiang, H., & Zhang, B. (2020). Psychometric properties of the perceived stress scale in a community sample of Chinese. *BMC Psychiatry*, 20(1), 130. https://doi.org/10.1186/s12888-020-02520-4

Hughes, S. E., Hutchings, H. A., Rapport, F. L., McMahon, C. M., & Boisvert, I. (2018). Social connectedness and perceived listening effort in adult cochlear implant users: A grounded theory to establish content validity for a new patient-reported outcome measure. *Ear and Hearing*, 39(5), 922–934. https://doi.org/10.1097/AUD.0000000000000553

Hutchings, R. (2020). The Impact of COVID-19 on the Use of Digital Technology in the NHS. *Nuffieldtrust*, 23.

Kahneman, D. (1973). *Attention and Effort* (1st ed). Prentice-Hall.

Katkade, V. B., Sanders, K. N., & Zou, K. H. (2018). Real world data: An opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. *Journal of Multidisciplinary Healthcare*, 11, 295–304. https://doi.org/10.2147/JMDH.S160029

Kemp, A. H., Brunoni, A. R., Santos, I. S., Nunes, M. A., Dantas, E. M., de Figueiredo, R. C., Pereira, A. C., Ribeiro, A. L. P., Mill, J. G., Andreão, R. V., Thayer, J. F., Benseñor, I. M., & Lotufo, P. A. (2014). Effects of depression, anxiety, comorbidity, and antidepressants on resting-state heart rate and its variability: An ELSA-brasil cohort baseline study. *American Journal of Psychiatry*, 171(12), 1328–1334. https://doi.org/10.1176/appi.ajp.2014.13121605

Kramer, S. E., Kapteyn, T. S., Festen, J. M., & Kuik, D. J. (1997). Assessing aspects of auditory handicap by means of pupil dilatation. *Audiology*, 36(3), 155–164. https://doi.org/10.3109/00206099709071969

Kramer, S. E., Kapteyn, T. S., & Houtgast, T. (2006). Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam checklist for hearing and work. *International Journal of Audiology*, 45(9), 503–512. https://doi.org/10.1080/14992020600754583

Kumari, M., Badrick, E., Chandola, T., Adam, E. K., Stafford, M., Marmot, M. G., Kirschbaum, C., & Kivimaki, M. (2009). Cortisol secretion and fatigue: Associations in a community based cohort. *Psychoneuroendocrinology*, 34(10), 1476–1485. https://doi.org/10.1016/j.psyneuen.2009.05.001

Lefebvre, C., Glanville, J., Briscoe, S., Littlewood, A., Marshall, C., Metzendorf, M.-I., Noel-Storr, A., Rader, T., Shokraneh, F., Thomas, J., & Wieland, S. (2021). 'Chapter 4: Searching for and Selecting Studies'. Retrieved 22 November 2021 (https://training.cochrane.org/handbook/current/chapter-04)

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, 27(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British society of audiology cognition in hearing special interest group "white paper". *International Journal of Audiology*, 53(7), 433–440. https://doi.org/10.3109/14992027.2014.890296

Mosadeghrad, A. M. (2012). A conceptual framework for quality of care. *Materia Socio-Medica*, 24(4), 251–261. https://doi.org/10.5455/msm.2012.24.251-261

Nair, A. S. (2019). 'Publication Bias - importance of studies with negative results!'. *Indian Journal of Anaesthesia*, 63(6), 505–507. https://doi.org/10.4103/ija.IJA_142_19

Niedzielski, A., Humeniuk, E., Błaziak, P., & Gwizda, G. (2006). Intellectual efficiency of children with unilateral hearing loss. *International Journal of Pediatric Otorhinolaryngology*, 70(9), 1529–1532. https://doi.org/10.1016/j.ijporl.2006.02.011

Ohlenforst, B., Zekveld, A. A., Jansma, E. P., Wang, Y., Naylor, G., Lorens, A., Lunner, T., & Kramer, S. E. (2017). Effects of hearing impairment and hearing aid amplification on listening effort: A systematic review. *Ear and Hearing*, 38(3), 267–281. https://doi.org/10.1097/AUD.0000000000000396

Page, M. J., Higgins, J., & Sterne, J. (2021a). 'Chapter 13: Assessing Risk of Bias Due to Missing Results in a Synthesis'. Retrieved 6 February 2022 (https://training.cochrane.org/handbook/current/chapter-13)

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021b). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Systematic Reviews*, 10(1), 89. https://doi.org/10.1186/s13643-021-01626-4

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, 39(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Pichora-Fuller, M. K. (2016). How social psychological factors may modulate auditory and cognitive functioning during listening. *Ear and Hearing*, 37(Suppl 1), 92S–100S. https://doi.org/10.1097/AUD.0000000000000323

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, 37(Suppl 1), 5S–27S. https://doi.org/10.1097/AUD.0000000000000312

Purcell, P. L., Shinn, J. R., Davis, G. E., & Sie, K. C. Y. (2016). Children with unilateral hearing loss may have lower intelligence quotient scores: A meta-analysis. *The Laryngoscope*, 126(3), 746–754. https://doi.org/10.1002/lary.25524

Richter, M., & Slade, K. (2017). Interpretation of physiological indicators of motivation: Caveats and recommendations. *International Journal of Psychophysiology*, 119, 4–10. https://doi.org/10.1016/j.ijpsycho.2017.04.007

Rosenberg, M. (1965). *Society and the Adolescent Child*.

Rubin, R. A., Dorle, J., & Sandidge, S. (1977). Self-Esteem and school performance. *Psychology in the Schools*, 14(4), 503–507. https://doi.org/10.1002/1520-6807(197710)14:4<503::AID-PITS2310140423>3.0.CO;2-Z

Samarrai, R., Riccardi, A. C., Tessema, B., Setzen, M., & Brown, S. M. (2021). Continuation of telemedicine in otolaryngology post-COVID-19: Applications by subspecialty. *American Journal of Otolaryngology*, 42(3), 102928. https://doi.org/10.1016/j.amjoto.2021.102928

Schlotz, W., Hellhammer, J., Schulz, P., & Stone, A. A. (2004). Perceived work overload and chronic worrying predict weekend-weekday differences in the cortisol awakening response. *Psychosomatic Medicine*, 66(2), 207–214. https://doi.org/10.1097/01.psy.0000116715.78238.56

Schmitt, D. P., & Allik, J. (2005). Simultaneous administration of the Rosenberg self-esteem scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, 89(4), 623–642. https://doi.org/10.1037/0022-3514.89.4.623

Schneiderman, N., Ironson, G., & Siegel, S. D. (2005). STRESS AND HEALTH: Psychological, behavioral, and biological determinants. *Annual Review of Clinical Psychology*, 1, 607–628. https://doi.org/10.1146/annurev.clinpsy.1.102803.144141

Schünemann, H., Brożek, J., Guyatt, G., & Oxman, A. 2013. 'GRADE Handbook for Grading Quality of Evidence and

Strength of Recommendations'. *BMJ*. https://doi.org/10.1136/bmj.39500.677199.AE

Sharma, D. (2021). Physiology of stress and its management by Dushyant Kumar Sharma. 2018, *1*(001), 01–05.

Shields, C., Willis, H., Nichani, J., Sladen, M., & Kluk-de Kort, K. (2021). 'Listening effort: WHAT is it, HOW is it measured and WHY is it important?'. *Cochlear Implants International*, *0*(0), 1–4. https://doi.org/10.1080/14670100.2021.1992941

Sterne, J., Hernán, M., McAleenan, A., Reeves, B., & Higgins, J. (2021). 'Chapter 25: Assessing Risk of Bias in a Non-Randomized Study'. Retrieved 22 November 2021 (https://training.cochrane.org/handbook/current/chapter-25)

Sterne, J. A., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., Ramsay, C. R., Regidor, D., Rothstein, H. R., Sandhu, L., Santaguida, P. L., Schünemann, H. J., Shea, B., Shrier, I., Tugwell, P., Turner, L., Valentine, J. C., Waddington, H., Waters, E., Wells, G. A., Whiting, P. F., & Higgins, J. P. (2016). ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. *BMJ*, *355*, i4919. https://doi.org/10.1136/bmj.i4919

Theunissen, S. C. P. M., Rieffe, C., Netten, A. P., Briaire, J. J., Soede, W., Kouwenberg, M., & Frijns, J. H. M. (2014). Self-Esteem in hearing-impaired children: The influence of communication, education, and audiological characteristics. *PLOS ONE*, *9*(4), e94521. https://doi.org/10.1371/journal.pone.0094521

Thomas, J., Brunton, J., & Graziosi, S. (2021). 'Eppi-Reviewer 4.0: Software for Research Synthesis'.

Visentin, C., & Prodi, N. (2021). 'How reliable are 11- to 13-year-olds' self-ratings of effort in noisy conditions?'. *Frontiers in Built Environment*, *7*, 108. https://doi.org/10.3389/fbuil.2021.688016

Warner-Czyz, A. D., Loy, B. A., Evans, C., Wetsel, A., & Tobey, E. A. (2015). Self-Esteem in children and adolescents with hearing loss. *Trends in Hearing*, *19*, 2331216515572615. https://doi.org/10.1177/2331216515572615

White, B. (2014). The perceived stress scale for children: A pilot study in a sample of 153 children. *International Journal of Pediatrics and Child Health*, *2*, 45–52. https://doi.org/10.12974/2311-8687.2014.02.02.4