Effects of Hearing Impairment and Hearing Aid Amplification on Listening Effort: A Systematic Review

Barbara Ohlenforst, Adriana Zekveld, Elise P. Jansma, Yang Wang, Graham Naylor, Artur Lorens, Thomas Lunner and Sophia E. Kramer

Ohlenforst, B., Zekveld, 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

Effects of hearing impairment and hearing aid amplification on listening effort - a systematic review

Barbara Ohlenforst[1,4], Adriana A. Zekveld[1,2,3], Elise P. Jansma[6], Yang Wang[1,4], Graham Naylor[7], Artur Lorens[8], Thomas Lunner[3,4,5]and Sophia E. Kramer[1]

[1]Section Ear & Hearing, Dept. of Otolaryngology-Head and Neck Surgery, VU University Medical Center and EMGO Institute for Health Care Research, Amsterdam, The Netherlands;

[2]Department of Behavioral Sciences and Learning, Linköping University, Linköping, Sweden;

[3]Linnaeus Centre HEAD, The Swedish Institute for Disability Research, Linköping and Örebro Universities, Linköping, Sweden;

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

[5]Department of Clinical and Experimental Medicine, Linköping, University, Sweden;

[6]Medical Library, VU University Amsterdam, Amsterdam, The Netherlands;

[7]MRC/CSO Institute of Hearing Research, Scottish Section, Glasgow, United Kingdom

[8]Institute of Physiology and Pathology of Hearing, International Center of Hearing and Speech, Warsaw, Poland

Received October 28, 2015;

This work was supported by a grant from the European Commission (LISTEN607373).

Corresponding author: Barbara Ohlenforst, Rørtangvej 20, 3070 Snekkersten, Denmark
Email: baoh@eriksholm.com Phone: 0045-48298900 Fax: 0045-49223629

Abstract

Objectives: To undertake a systematic review of available evidence on the effect of hearing impairment and hearing-aid amplification on listening effort. Two research questions were addressed: Q1) does hearing impairment affect listening effort? and Q2) can hearing aid amplification affect listening effort during speech comprehension?

Design: English language articles were identified through systematic searches in PubMed, EMBASE, Cinahl, the Cochrane Library, and PsycINFO from inception to August 2014. References of eligible studies were checked. The Population, Intervention, Control, Outcomes and Study design (PICOS) strategy was used to create inclusion criteria for relevance. It was not feasible to apply a meta-analysis of the results from comparable studies. For the articles identified as relevant, a quality rating, based on the 2011 Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group guidelines, was carried out to judge the reliability and confidence of the estimated effects.

Results: The primary search produced 7017 unique hits using the key-words: hearing aids OR hearing impairment AND listening effort OR perceptual effort OR ease of listening. Of these, 41 articles fulfilled the PICOS selection criteria of: experimental work on hearing impairment OR hearing aid technologies AND listening effort OR fatigue during speech perception. The methods applied in those articles were categorized into subjective, behavioral and physiological assessment of listening effort. For each study, the statistical analysis addressing research question Q1 and/or Q2 was extracted. In 7 articles more than one measure of listening effort was provided. Evidence relating to Q1 was provided by 21 articles that reported 41 relevant findings. Evidence relating to Q2 was provided by 27 articles that reported 56 relevant findings. The quality of evidence on both research questions (Q1 and Q2) was very low, according to the GRADE Working Group guidelines. We tested the statistical evidence across studies with non-parametric tests. The testing revealed only one consistent effect across studies, namely that listening effort was higher for hearing-impaired listeners

compared to normal-hearing listeners (Q1) as measured by EEG measures. For all other studies the evidence across studies failed to reveal consistent effects on listening effort. **Conclusion:** In summary, we could only identify scientific evidence from physiological measurement methods, suggesting that hearing impairment increases listening effort during speech perception (Q1). There was no systematic finding across studies indicating that hearing-aid amplification decreases listening effort (Q2). In general, there were large differences in the study population, the control groups and conditions, and the outcome measures applied between the studies included in this review. The results of this review indicate that published listening effort studies lack consistency, lack standardization across studies, and have insufficient statistical power. The findings underline the need for a common conceptual framework for listening effort to address the current shortcomings.

Keywords: Listening effort, hearing impairment, hearing aid amplification, speech comprehension, subjective ratings, behavioral measures, physiologic measures, quality rating;

Introduction

Hearing impairment is one of the most common disabilities in the human population and presents a great risk in everyday life due to problems with speech recognition, communication, and language acquisition. Due to hearing impairment, the internal representation of the acoustic stimuli is degraded (Humes & Roberts, 1990). This causes difficulties that are experienced commonly by hearing-impaired listeners, as speech recognition requires that the acoustic signal is correctly decoded (McCoy et al. 2005). Additionally, in daily life, speech is often heard amongst a variety of sounds and noisy backgrounds that can make communication even more challenging (Hällgren et al. 2005). Previous research suggests that hearing-impaired listeners suffer more from such adverse conditions in terms of speech perception performance as compared to normal-hearing listeners (Hagerman, 1984; Plomp, 1986; Hopkins et al. 2005). It has been suggested that keeping up with the processing of ongoing auditory streams increases the cognitive load imposed by the listening task (Shinn-Chunningham & Best, 2008). As a result, hearing-impaired listeners expend extra effort to achieve successful speech perception (Rönnberg et al. 2013; McCoy et al. 2005). Increased listening effort due to impaired hearing can cause adverse psychosocial consequences, such as increased levels of mental distress and fatigue (Stephens & Hétu, 1991; Kramer Sophia et al. 1997; Kramer et al. 2006), lack of energy and stress-related sick leave from work (Edwards, 2007; Gatehouse & Gordon, 1990; Kramer et al. 2006; Hornsby, 2013a, 2013b). Nachtegaal and colleagues (2009) found a positive association between hearing thresholds and the need for recovery after a working day. Additionally, hearing impairment can dramatically alter peoples' social interactions and quality of life due to withdrawal from leisure and social roles (Weinstein, 1982; Demorest & Erdman, 1986; Strawbridge et al. 2000), and one reason for this may be the increased effort required for successful listening. There is growing interest amongst researchers and clinicians in the concept of listening effort and its relationship with hearing impairment (Gosselin &

Gagné, 2010; McGarrigle et al. 2014). The most common approaches to assess listening effort include subjective, behavioral and physiological methods (for details see Table 1). The concept of subjective measures is to estimate the amount of perceived effort, handicap reduction, acceptance, benefit and satisfaction with hearing aids (Humes & Humes, 2004). Subjective methods such as self-ratings or questionnaires provide immediate or retrospective judgment of how effortful speech perception and processing was perceived by the individual during a listening task. The ratings are typically made on a scale ranging between “no effort” and “maximum effort”. Questionnaires are often related to daily life experiences and typically offer a closed set of possible response opportunities (e.g. Speech, Spatial and Qualities of Hearing scale (SSQ), (Noble & Gatehouse, 2006). The most commonly used behavioral measure is the dual-task paradigm (Howard et al. 2010; Gosselin & Gagné, 2011; Desjardins & Doherty, 2013), where participants perform a primary and a secondary task simultaneously. The primary task typically involves word or sentence recognition. Secondary tasks may involve probe reaction time tasks (Downs, 1982; Desjardins & Doherty, 2014; Desjardins & Doherty, 2013), memory tasks (Feuerstein, 1992; Hornsby, 2013a), tactile pattern recognition tasks (Gosselin & Gagné, 2011) or even driving a vehicle in a simulator (Wu et al. 2014). The concept of dual-task paradigms is based on the theory of limited cognitive capacity (Kahneman, 1973). An increase in effort or cognitive load, related to performing the primary task, leads accordingly to a lower performance in the secondary task, which is typically interpreted as increased listening effort (Downs, 1982). The concept of physiological measures of listening effort is to illustrate changes in the central and/or autonomic nervous system activity during task performance (McGarrigle et al. 2014). The electroencephalographic (EEG) response to acoustic stimuli, which is measured by electrodes on the scalp, provides temporally-precise markers of mental processing (Obleser et al. 2012; Bernarding et al. 2012). Functional magnetic resonance imaging (fMRI) is another physiological method to assess listening effort. Metabolic consequences of neuronal activity

are reflected by changes in the blood oxygenation level. For example, increased brain activity in the left inferior frontal gyrus has been interpreted as reflecting compensatory effort required during a challenging listening task, such as the effect of attention during effortful listening (Wild et al. 2012). The measure of changes in the pupil diameter (in short 'pupillometry') has furthermore been used to assess the intensity of mental activity, for example in relation to changes in attention and perception (Laeng et al. 2012). The pupil dilates when a task evokes increased cognitive load, until the task demands exceed the processing resources (Granholm et al. 1996). Pupillometry has previously been used to assess how hearing impairment (Kramer et al. 1997; Zekveld et al. 2011), sentence intelligibility (Zekveld et al. 2010), lexical manipulation (Kuchinsky et al. 2013), different masker types (Koelewijn et al. 2012) and cognitive function (Zekveld et al. 2011) affect listening effort. Like the pupil response, skin conductance and heart rate variability also reflect parasympathetic and sympathetic activity of the autonomic nervous system. For example, an increase in mean skin conductance and heart rate has been observed when task demands during speech recognition tests increase (Mackersie & Cones, 2011). Finally, cortisol levels, extracted from saliva samples, have been associated with cognitive demands and fatigue as a response to stressors (Hicks & Tharpe, 2002).

Hearing aids are typically used to correct for the loss of audibility introduced by hearing impairment (Hicks & Tharpe, 2002). Modern hearing aids provide a range of signal processing algorithms such as amplitude compression, directional microphones, and noise reduction (Dillon, 2001). The purpose of such hearing aid algorithms is to improve speech intelligibility and listening comfort (Neher et al. 2013). If hearing impairment indeed increases listening effort, as suggested by previous research (Feuerstein, 1992; Hicks & Tharpe, 2002; Luts et al. 2010), then it is essential to investigate whether hearing aids can reverse this aspect of hearing loss too.

Given that the number of methods to assess listening effort is still increasing and the evidence emerging is not coherent, an exhaustive review of the existing evidence is needed to facilitate our understanding of state-of-the-art knowledge related to 1) the influence of hearing impairment on listening effort and 2) the effect of hearing aid amplification on listening effort. The findings should guide researchers in defining research priorities and designing future studies, and help clinicians in improving their practice related to hearing aid assessment and fitting. Therefore, this systematic review addressed the following research questions: Q1) does hearing impairment affect listening effort? and Q2) can hearing aid amplification affect listening effort during speech comprehension? We hypothesized that hearing impairment increases listening effort (HP1). On the other hand, the application of hearing aid amplification is hypothesized to reduce listening effort relative to the unaided condition (HP2).

Methods

Search strategy

We systematically searched the bibliographic databases PubMed, EMBASE, CINAHL, PsycINFO and the Cochrane Library. Search variables included controlled terms from MeSH in PubMed, EMtee in EMBASE, CINAHL Headings in CINAHL, and free text terms. Search terms expressing ‘hearing impairment’ or ‘hearing aid’ were used in combination with search terms comprising ‘listening effort’ or ‘fatigue’ (see appendix for detailed search terms). English language articles were identified from inception to August 2014.

Inclusion and exclusion

The PICOS strategy (Armstrong, 1999) was used to form criteria for inclusion and exclusion as precisely as possible. The formulation of a well-defined research question with well-articulated PICOS elements has been shown to provide an efficient tool to find high-quality

evidence and to make evidence-based decisions (Richardson et al. 1995; Ebell, 1999). To be included in the review, studies had to meet the following PICOS criteria:

I. **P**opulation: Hearing-impaired participants and/or normal-hearing listeners with a simulated hearing loss (for example by applying a low-pass filter to the auditory stimuli).

II. **I**ntervention: Hearing impairment or hearing aid amplification (including cochlear implant), such as the application of real hearing aids, laboratory simulations of hearing-aid amplification, comparisons between aided versus unaided conditions or different types of hearing aid processing technologies. Finally, we considered results of simulations of signal processing in cochlear implants (CIs), tested by applied vocoded stimuli. When a study was restricted to investigating the participant's cognitive status and/or when performance comparisons between groups of participants with different cognitive functioning and speech perception abilities were applied, but participants were only normal-hearing and/or no hearing aid amplification was applied, the study was not included. Furthermore, measures of cognition, such as memory tests for speech performance on stimulus recall, were not considered an intervention.

III. **C**ontrol: Group comparisons (e.g. normal-hearing vs. hearing-impaired) or a within-subjects repeated measures design (subjects are their own controls). We included studies that compared listeners with normal-hearing versus impaired hearing, monaural versus binaural testing or simulations of hearing impairment, or with different degrees of hearing impairment, and studies that applied noise maskers to simulate hearing impairment.

IV. **O**utcomes: Listening effort, as assessed by (i) subjective measures of daily life experiences, handicap reduction, benefit or satisfaction, (ii) behavioral measures of changes in auditory tasks performance, or (iii) physiological measures corresponding

to higher cognitive processing load, such as N2 and/or P3 EEG responses, pupillometry, fMRI or cortisol measures. Subjective assessments that were not directly related to listening effort or fatigue (e. g. quality-of-life ratings, preference ratings) were not categorized as measure of listening effort. Furthermore, physiological measures of early-stage auditory processing, such as ERP components N1, mismatch negativity (MMN), and P2a were not considered as reflecting measures of listening effort.

V. **S**tudy design: Experimental studies with repeated measures design or randomized control trials, published in peer-reviewed journals of English language were included. Studies describing case reports, systematic reviews, editorial letters, legal cases, interviews, discussion papers, clinical protocols or presentations were not included.

The identified articles were screened for relevance by examining titles and abstracts. Differences between the authors in their judgment of relevance were resolved through discussion. The reference lists of the relevant articles were also checked to identify potential additional relevant articles. The articles were categorized as ‘relevant’ when they were clearly eligible, ‘maybe’ when it was not possible to assess the relevance of the paper based on the title and abstract, and ‘not relevant’ when further assessment was not necessary. An independent assessment of the relevance of all the articles categorized as ‘relevant’ or ‘maybe’ was carried out on the full texts by three authors (BO, AZ and SK).

Data extraction and management

For each relevant study, the outcome measures applied to assess listening effort were extracted and categorized into subjective, objective or physiological indicators of listening effort. We identified and extracted the findings addressing Q1 and/or Q2 from all relevant studies. The results of each study were evaluated with respect to the two hypotheses (HP1

and/or HP2) based on Q1 and Q2. When HP1 was supported (i.e. hearing impairment was associated with increased listening effort during speech understanding relative to normal hearing), statistical results were reported in the category 'more effort' (+). Results that did not show significant effects of hearing impairment on listening effort were categorized as 'equal effort' (=). If hearing impairment was associated with a reduction in listening effort, the results were reported as 'less effort' (-). HP2 stated decreased listening effort due to hearing aid amplification. Results supporting, refuting and equivocal with respect to HP2 were respectively reported as 'less effort' (+), 'more effort' (-) and 'equal effort' (=). Any given study could provide more than one finding relating to Q1 and/or Q2. General information related to PICOS was additionally extracted, such as on population (number and mean age of participants), intervention (type of hearing loss and configurations and processing), outcomes (methods to measure listening effort and test stimulus), and control and study design (test parameters).

An outright meta-analysis across studies with comparable outcomes was not feasible, because the studies were too heterogeneous with respect to characteristics of the participants, controls, outcome measures used, and study designs. However, we made across-studies comparisons based on the categorized signs (+, =, -) of evidence from each study, to get some insight into the consistency of the reported outcomes. Study findings and study quality were incorporated within a descriptive synthesis and by numerical comparisons across studies, to aid interpretation of findings and to summarize the findings.

Quality of evidence

The evaluation of the level of evidence, provided by all included studies, was adapted from the GRADE Working Group guidelines (Guyatt et al. 2011). The quality of evidence is rated for each measurement type (see Table 3, 5) corresponding to the research questions, as a body of evidence across studies, rather than for each study as a single unit. The quality of evidence

is rated by explicit criteria including “study limitations”, “inconsistency”, “indirectness”, “imprecision” and the “risk of publication bias”. How well the quality criteria were fulfilled across all studies on each measurement type was judged by rating how restricted those criteria were (“undetected”, “not serious”, “serious” or “very serious”). The quality criteria “inconsistency”, “indirectness”, “imprecision” and “publication bias” were judged by the same approach, as follows. If all the studies fulfilled the given criterion, restrictions on that criterion were judged as “undetected”, whereas “not serious” restrictions applied, when more than half of the studies from a measurement type fulfilled the criterion, a “serious” rating was given if less than half of the studies from a measurement type fulfilled the criterion, and “very serious” if none of the studies fulfill the criterion. The quality criterion “study limitations” was based on five sub-criteria (lack of allocation concealment, lack of blinding, incomplete accounting of patients and outcome events, selective outcome reporting, and early stop of trials for benefit) and rated as “undetected” if all the studies fulfilled the given criterion, “not serious” if more than half of the sub-criteria were fulfilled across studies, “serious” if less than half of the sub-criteria were fulfilled, and “very serious” if none of the sub-criteria was fulfilled. For example, with studies using Visual Analog Scales (VAS), the criterion “study limitations” was rated as “not seriously” restricted as none of the studies on VAS showed lack of allocation concealment, some studies lacked blinding and some had incomplete accountancy of patients but no selective outcome reporting and no early stop for benefit were identified across studies. The quality criterion called “inconsistency” was evaluated based on the experimental setup across studies, including the choice of stimulus, stimulus presentation and the measurement type for listening effort within each outcome. When findings across studies were not based on consistent target populations, consistent interventions and/or consistent factors of interest with respect to Q1 and/or Q2, “serious inconsistency” was judged for evidence on that measurement type. The quality criterion “indirectness” was related to differences between tested populations and/or differences in comparators to the

intervention. The criterion “indirectness” was seriously affected when findings across studies were based on comparing young normal-hearing listeners with elderly hearing-impaired listeners and/or when normal-hearing listeners were compared to listeners with simulated, conductive hearing impairment or sensorineural hearing-impairment. The quality criterion “imprecision” was evaluated based on statistical power sufficiency or provided power calculations across studies for each measurement type. We did not detect selective publication of studies in terms of study design (experimental versus observational), study size (small versus large studies) or lag bias (early publication of positive results), and thus “publication bias” was judged as “undetected”. The overall quality of evidence is a combined rating of the quality of evidence across all quality criteria on each measurement type. The quality is down rated, if the five quality criteria (limitations, inconsistency, indirectness, imprecision and publication bias) are not fulfilled by the evidence provided by the studies on a measurement type (Table 3, 5). When large effects were shown for a measurement type, dose response relations (e.g. between different levels of hearing impairment or hearing-aid usage and listening effort) and plausible confounders are taken into account, an uprating in quality of evidence is possible (Table 6). There are four possible levels of quality ratings, including high, moderate, low and very low quality. We created a separate evidence profile for each research questions (Table 3 on Q1, Table 5 on Q2) to sum up the key information on each measurement type. For each of our two research questions, evidence was provided by studies with diverse methods, which made it problematic to compute confidence intervals on absolute and relative effects of all findings on each individual measurement type. Therefore a binomial test (Sign test) was applied as alternative statistical method. We counted the signs (+, =, - in Table 1) corresponding to each measurement type for findings addressing HP1 and/or HP2 (more, equal or less effort). Our hypotheses were that listening effort is greater for hearing-impaired listeners than for those of normal hearing (HP1) and that aided listening helps to reduce effort compared to unaided listening (HP2), i.e. one-sided in both cases. Therefore we

applied a one-sided (directional) Sign test. The standard binomial test was used to calculate significance, as the test statistics were expected to follow a binomial distribution (Baguley, 2012). Overall, evidence across all measurement types on Q1 was judged as important to health and life quality of hearing-impaired listeners, as hearing impairment affects people in their daily lives. However, no life threatening impact, myocardial infarction, fractures or physical pain are expected from hearing impairment and the importance was not characterized as critical (see Table 3, 4 “Importance”) (Schünemann et al. 2013).

Two authors (BO and TL), were mainly involved in the design of the evidence profiles and the scoring of quality of evidence. Uncertainties or disagreement were discussed and solved according to the GRADE handbook (Guyatt et al. 2011).

Results

Results of the search

The PRISMA (Moher et al. 2009) flow-chart in Figure 1 illustrates details of the search and selection procedure including the number of removed duplicates, the number of articles that were excluded and the reasons for their exclusion. The main electronic database search produced a total of 12210 references: 4430 in PubMed, 3521 in EMBASE.com, 2390 in Cinahl, 1639 in PsycINFO and 230 in the Cochrane Library. After removing duplicates, 7017 references remained. After screening the abstracts and titles of those 7017 articles, further 6910 articles were excluded. The most common reasons for exclusion were that measures of listening effort- as outlined above - were not applied (n=4234 articles), hearing aid amplification was not provided (n=564) or studies focused on the development of cochlear implants (CI) (n=746) or the treatment of diseases (n=359) and neither of the two research questions was addressed. We checked the full text for the remaining 107 articles for eligibility and excluded 68 articles. Finally, 39 articles fulfilled the search and selection criteria and

were included in the review process. The inspection of the reference lists of these relevant articles resulted in two additional articles that met the inclusion criteria. Thus in total, 41 articles were included in this systematic review.

Results of the selection process and criteria

Before examining the evidence arising from the 41 included studies, it is useful to consider the general characteristics of the sample, arranged according to the five elements of the PICOS strategy described earlier.

Population

In seven studies, only people with normal hearing thresholds <= 20 dB HL participated (mean n=22.4, SD = 12.8). In 18 studies, only people with hearing impairment (mean n=52.4, SD=72.1) were tested, without including normal-hearing controls. The remaining 16 studies assessed both normal-hearing and hearing-impaired participants (mean n=51.2, SD=27.3). Hearing-impaired participants had monaural and/or binaural hearing loss and the degree of hearing impairment varied. Some studies examined experienced hearing-aid users, whereas participants of other studies included non-users of hearing aids. In two studies CI users participated and monaural versus binaural implantation (Dwyer et al. 2014) or CI versus hearing-aid fitting (Noble et al. 2008) was compared. Other studies compared hearing abilities between different age-groups (Desjardins & Doherty, 2013; Hedley-Williams et al. 1997; Tun et al. 2009). Overall, there was great variety in the tested populations in terms of hearing status and hearing aid experience.

Intervention

The intervention or exposure of interest was either hearing impairment (Q1) or hearing-aid amplification (Q2). In a number of studies, a certain type of hearing aid was chosen and

binaurally fitted in hearing-impaired participants (Bentler et al. 2008; Ahlstrom et al. 2014; Desjardins & Doherty, 2014). Other studies compared different hearing-aid types, such as analogue versus digital hearing-aids (Bentler & Duve, 2000) or hearing-aids versus CIs (Noble et al. 2008; Dwyer et al. 2014) which were tested in a variety of environments. Seven studies simulated hearing aid algorithms or processing, for example by using implementations of a 'master hearing aid' (Luts et al. 2010).

Comparators

The most commonly applied approach to assess the effect of hearing impairment on listening effort was to compare subjective perception or behavioral performance between normal-hearing and hearing-impaired listeners (Q1) (Feuerstein, 1992; Rakerd et al. 1996; Humes et al. 1997; Kramer et al. 1997; Oates et al. 2002; Korczak et al. 2005; Martin & Stapells, 2005; Humes et al. 1997). When the effect of hearing-aid amplification was investigated, aided versus unaided conditions (Q2) (Downs, 1982; Gatehouse & Gordon, 1990; Humes et al. 1997; Humes, 1999; Hällgren et al. 2005; Korczak et al. 2005; Picou et al. 2013; Hornsby, 2013; Ahlstrom et al. 2014) or different types of processing (Humes et al. 1997; Bentler & Duve, 2000; Noble & Gatehouse, 2006; Noble et al. 2008; Harlander et al. 2012; Dwyer et al. 2014), different settings of the test parameters (Kulkarni et al. 2012; Bentler et al. 2008; Sarampalis et al. 2009; Luts et al. 2010; Kulkarni et al. 2012; Brons et al. 2013; Desjardins & Doherty, 2013; Pals et al. 2013; Desjardins & Doherty, 2014; Gustafson et al. 2014; Neher et al. 2014; Picou et al. 2014; Wu et al. 2014; Sarampalis et al. 2009), were compared.

Outcomes

There was no common outcome measure of listening effort that was applied in all of the studies. We identified 42 findings from subjective measures, 39 findings from behavioral measures and 16 findings from physiological measures (summed up across Table 2 and 4). Of

the 42 findings based on subjective assessment or rating of listening effort, 31 findings resulted from visual-analogue scales (VAS, see Table 1). Such effort rating scales ranged for example from 0 to 10, indicating conditions of “no effort” to “very high effort” (e. g. Zekveld et al. 2011; Hällgren et al. 2005). The remaining eleven findings based on subjective assessment of listening effort resulted from the SSQ (Noble & Gatehouse, 2006; Noble et al. 2008; Hornsby et al. 2013; Dwyer et al. 2014). Most findings from behavioral measures (n=32 of 39 in total) corresponded to Dual Task Paradigm (DTP) and seven findings resulted from reaction time measures. The sixteen findings from physiological assessment of listening effort, included 12 findings from EEG measures (Oates et al. 2002; Korczak et al. 2005; Martin & Stapells, 2005), two findings from task-evoked pupil dilation measures (Kramer et al. 1997; Zekveld et al. 2011), one finding from measures of diurnal saliva cortisol concentrations (Hicks & Tharpe, 2002) and one finding from fMRI was used (Wild et al. 2012).

Study design

In this systematic review, studies that used a repeated measures design and/or a randomized controlled design were included. A between-group design (normally-hearing vs hearing-impaired) was applied in 17 studies (Luts et al. 2010; Rakerd et al. 1996; Kramer et al. 1997; Humes et al. 1997; Humes, 1999; Hicks & Tharpe, 2002; Oates et al. 2002b; Korczak et al. 2005; Stelmachowicz et al. 2007; Noble et al. 2008; Tun et al. 2009; Luts et al. 2010; Zekveld et al. 2011; Kulkarni et al. 2012; Neher et al. 2013; Neher et al. 2014; Ahlstrom et al. 2014; Dwyer et al. 2014).

Results of the data extraction and management

We categorized the methods of assessing listening effort from all relevant articles, into subjective, behavioral and physiological measurement methods. In Table 1, first all studies

that applied subjective methods are listed in alphabetical order, followed by the studies using behavioral and finally physiological measurement methods of listening effort. In six studies, more than one method was used to measure listening effort. Those studies contributed multiple rows in Table 1. Evidence on HP1 was provided by 41 findings from 21 studies. The evidence on HP2 was based on 56 findings from 27 studies.

Evidence on the effect of hearing impairment on listening effort (Q1)

See Tables 1 and 2 respectively for detailed and summarized tabulations of the results described in this section.

Subjective measures, Q1

Six findings (out of n=9 in total) indicated that self-rated listening effort, for different fixed intelligibility conditions, was higher for hearing-impaired listeners than for normal-hearing listeners. The applied methods included VAS ratings (n=5 findings) and the SSQ (n=1 finding). However, different comparisons across studies were made. Some compared normal-hearing and hearing-impaired groups (n=4 findings). One finding concerned the difference in self-rated effort between monaural or binaural simulation of impaired hearing. Three findings, based on the comparison between normal-hearing and hearing-impaired listeners concluded that hearing impairment does not affect listening effort. Those three findings resulted from VAS ratings. None of the tests with subjective measures indicated less listening effort due to a hearing loss.

Behavioral measures, Q1

Ten findings (out of n=17 in total) indicated higher levels of listening effort for groups with hearing impairment compared to groups with normal hearing. Findings from DTPs were mainly (n=6 out of 7) based on comparing performance between hearing-impaired and normal-hearing listeners, while all findings from reaction time measures (n=3) were based on

simulations of hearing impairment on normal-hearing listeners. The remaining 7 findings (all related to DTP) did not demonstrate significant differences between normal-hearing and hearing-impaired listeners. So, roughly half of the tests showed higher effort (10 findings, +) in the hearing-impaired group, and slightly less than half showed no difference (7 findings, =). No clear evidence showed reduced listening effort due to hearing impairment.

Physiological measures, Q1

Most findings (n=13 of 15 in total) indicated higher levels of listening effort due to hearing impairment. The applied methods varied between measures of EEG (n=9 findings), pupil dilation (n=2 findings), diurnal Cortisol levels (n=1 finding), and fMRI (n=1 finding). Nine findings resulted from comparing normal-hearing and hearing-impaired listeners and six findings from simulations of hearing impairment. The two remaining findings both resulted from EEG measures; one indicated no effect of hearing impairment, and the other indicated less effort in the presence of hearing impairment.

Quality of evidence on Q1

The GRADE evidence profile on all findings on the effect of hearing impairment on listening effort (Q1) is shown in Table 3. We created a separate row for each measurement type: subjective assessment by VAS, behavioral assessment by DTP or reaction-time measures, and physiological assessment by pupillometry or EEG. All measurement types corresponded to studies of randomized controlled trials (RCT). For each measurement type, all findings across studies were evaluated with respect to the quality criteria (“limitations”, “inconsistency”, “indirectness”, “imprecision” and “publication bias”). Each row in Table 3, representing a separate measurement type, was based on at least two findings (across studies) to justify being listed in the evidence profile. In summary, five measurement types were identified for Q1 (1 subjective, 2 behavioral and 2 physiological methods). Most quality criteria (“inconsistency”,

"indirectness", "imprecision") across the five measurement types showed "serious" restrictions for the evidence rating. The quality criterion "study limitation" showed "not serious" restrictions across all five measurement types, as only lack of blinding and lack of information on missing data or excluded participants (incomplete accounting of patients and outcome events) were identified for some studies. But there was no lack of allocation concealment, no selective outcome reporting and no early stop for benefit across studies. Overall, "serious inconsistency", "serious indirectness" or "serious imprecision" caused down-rating in quality and consequently low or very low quality of evidence resulted for three out of five outcomes on Q1. The quality criteria "publication bias" was "undetected" for all five measurement types, as we did not detect selective publication of studies in terms of study design, study size or lag bias.

Quality of evidence for subjective measures, Q1

Subjective assessment of listening effort, assessed by VAS ratings, provided the first row within the evidence profile in Table 3, based on seven randomized controlled trials (RCT). We found the quality criterion "study limitations" (Table 3) "not seriously" affected, as across studies only a lack of blinding and lack of descriptions of missing data or exclusion of participants were identified. No lack of allocation concealment, no selective outcome reporting and no early stop for benefit was found across those seven studies. We rated the criterion "inconsistency" as "serious" due to a great variety of experimental setups across studies, including different stimuli (type of target and masker stimulus) and presentation methods (headphones versus sound field). We identified furthermore "serious indirectness" for VAS ratings, as the population across the seven studies varied in age and hearing ability (young normal-hearing versus elderly hearing-impaired, children versus adults). Only two studies provided sufficient power or information on power calculations, which resulted in "serious imprecision". Publication bias was not detected across the seven studies. We rated

the quality of evidence on VAS ratings as very low based on "serious inconsistency", "serious indirectness" and "serious imprecision". We counted the "+", "=" and "-" for all findings on VAS ratings for Q1 in Table 1 and we applied a binomial test (Sign test), which resulted in a p-value of p=0.25. This indicated that HP1 could not be rejected, and therefore we did not find evidence across studies that listener's effort assessed by VAS-scales show higher listening effort ratings for hearing impaired listeners compared to normal-hearing listeners.

Quality of evidence for behavioral measures, Q1

We identified two types of behavioral assessment of listening effort. The first measurement type corresponded to listening effort assessed by DTPs and was based on eight randomized control studies (see Table 3). The quality assessment for findings from DTPs indicated "not serious limitations" (lack of blinding and incomplete accounting of patients and outcome events), "serious inconsistency" (different stimulus and test setups between studies), "serious indirectness" (participant groups not consistent across studies) and "serious imprecision" (missing information on power analysis and sufficiency of study participants) across the eight studies, resulting in a low quality of evidence. The evidence across studies, showed that listening effort, as assessed by DTP, did not indicate higher listening effort for hearing-impaired listeners compared to normal-hearing listeners (Sign-test: p=0.61). The second behavioral measurement type was reaction time assessment. Only one randomized controlled study used this measurement type. "Study limitations" (lack of blinding and incomplete accounting of patients and outcome events), "inconsistency" and "indirectness" were "not serious". However, we found serious "imprecision", which caused a down-rating from high to moderate quality of evidence. Only 10 normal-hearing but no hearing-impaired listeners were included in the single study using reaction time measures. Thus it was not possible to answer Q1 for reaction time.

Quality of evidence for physiological measures, Q1

Two types of physiological measures were identified for studies addressing Q1 (see Table 3). The first was pupillometry. Two randomized controlled trials using pupillometry were found. We rated “not serious limitations” as no lack of allocation concealment, no selective outcome reporting and no early stop for benefit was found. Both studies lacked information on blinding but only one showed incomplete accounting of patients and outcome events. We identified “serious inconsistency” (different stimulus conditions and test setups across both studies), “serious indirectness” (young normal-hearing compared with elderly hearing-impaired listeners), “serious imprecision” (missing power analysis and sufficiency for both studies). Thus the quality assessment of studies using pupillometry was judged as very low due to “serious inconsistency”, “serious indirectness” and “serious imprecision” across studies. We counted two plus signs (+) from the two corresponding studies in Table 1 and the applied Sign test did not show a difference in listening effort (as indexed by pupillometry) between normal-hearing and hearing-impaired listeners ($p = 0.25$).

The second physiological measurement type was EEG. Three studies used EEG. We identified “not serious limitations” across studies as experimental blinding and information on missing data or excluded participants was not provided but no lack of allocation concealment, no selective outcome reporting or early stop for benefit were found. However, “not serious inconsistency” was found across studies. Similar stimuli were applied and only one study differed slightly in the experimental setup from the other two studies. We rated “indirectness” as “not serious”, as across studies, age-matched hearing-impaired and normal-hearing listeners were compared and only one study did not include hearing-impaired listeners. We found “serious imprecision”, as across studies neither information on power calculation nor power sufficiency was given. The results from the Sign test on the outcome of EEG measures indicated, that hearing-impaired listeners show higher listening effort than normal-hearing

listeners (p=0.03). The quality of evidence was moderate for the EEG data and very low for pupillometry studies.

Evidence on the effect of hearing aid amplification on listening effort (Q2)

See Tables 1 and 4 respectively for detailed and summarized tabulations of the results described in this section.

Subjective measures, Q2

Reduced listening effort associated with hearing aid amplification was found 17 times. The applied methods were VAS ratings (n=13 findings) and the SSQ (n=4 findings). Studies compared different types of signal processing (n=8 findings), unprocessed versus processed stimuli (n=4 findings), aided versus unaided listening (n=4 findings) and active versus inactive signal processing algorithms (n=1 finding).

We identified thirteen findings indicating no effect of hearing-aid amplification on listening effort based on comparing different signal-processing algorithms (n=7), aided versus unaided conditions (n=4) and signal processing algorithms in active versus inactive settings (n=2). Those findings resulted mainly from VAS ratings (n=9 findings) or from the application of the SSQ (n=4 findings).

Three findings from VAS ratings indicated increased listening effort with hearing aid amplification when active versus inactive signal processing algorithms (n=2 findings) or processed versus unprocessed stimuli (n=1 finding) were tested.

In sum, evidence from subjective assessment on Q2 was based on 33 findings in total. 17 findings indicated reduced listening effort, 13 findings equal effort and 3 findings increased listening effort associated with hearing-aid amplification.

Behavioral measures, Q2

Fourteen findings indicated reduced listening effort with hearing aid amplification: aided versus unaided listening (n=4 findings), active versus inactive signal processing algorithms (n=5 findings) and unprocessed versus processed stimuli (n=5 findings). These findings resulted from DTPs (n=10 findings) or reaction time measures (n=4 findings). Six findings, which resulted from DTPs, indicated that hearing aid amplification does not affect listening effort. Those findings resulted when unprocessed versus processed stimuli (n=3) or active versus inactive signal processing algorithms (n=2 tests) or aided versus unaided conditions (n=1 test) were compared.

Two findings from DTPs indicated that listening effort is actually increased with hearing aid amplification, from comparing active versus inactive hearing aid settings, such as aggressive DNR versus moderate DNR versus inactive DNR settings. So, 14 findings indicated a reduction of listening effort when using amplification, 6 failed to find a difference and 2 tests indicated an increase in listening effort in the group with amplification.

Physiological measures, Q2

Evidence from a single EEG finding that compared aided versus unaided listening, indicated reduced listening effort for the aided condition. We did not identify further findings from physiological measures of listening effort that provided evidence on Q2.

Quality of evidence on Q2

Four measurement types were identified on Q2, including VAS and the SSQ for subjective assessment and DTP and reaction time measures from behavioral assessment (see Table 5). We judged that evidence based on a single physiological finding provides too little information to create a separate row in Table 5. The quality criteria (“limitations”, “inconsistency”, “indirectness”, “imprecision” and “publication bias”) were checked for

restrictions and rated accordingly ("undetected", "not serious", "serious", or "very serious") across the studies on each measurement type, as done for Q1. The quality of evidence for each measurement type was then judged across all quality criteria.

Quality of evidence for subjective measures, Q2

We identified two measurement types, including sixteen studies using VAS ratings and four studies that applied SSQ (Table 5). We judged the quality of evidence from VAS as very low, based on "serious inconsistency", "serious indirectness" and "serious imprecision". We found a lack of experimental blinding and incomplete accounting of patients and outcome events (treatment of missing data or excluded participants) across studies but there was no lack of allocation concealment, no selective outcome reporting and no early stop for benefit, which caused "limitations" to be "not serious". We rated "inconsistency" as "serious" as target and masker material, hearing aid setting and algorithms and the applied scales for VAS were not consistent across studies. Furthermore, "indirectness" was at a "serious" level based on a large variety regarding the participant groups (young normal-hearing versus elderly hearing-impaired, experienced versus inexperienced hearing aid users, different degrees of hearing impairment). Finally, only six (out of n=16 in total) of the studies provided sufficient power, which caused "serious imprecision". We counted the "+", "=" and "-" signs in Table 1 for subjective findings for VAS on Q2 and applied the Sign test, which revealed a p-value of p=0.50, meaning that evidence from VAS across studies did not show higher listening effort ratings for hearing aid amplification compared to unaided listening.

The second measurement type on subjective assessment resulted from SSQ data. We found randomized controlled trials (RCT, Table 5) in three studies. One study (Dwyer et al. 2014) was an observational study where different groups of participants rated their daily life experience with either hearing impairment, cochlear implant or hearing aid fitting. As everyday scenarios were rated, randomization was not applicable for this study. We judged

the study limitations for observational studies (development and application of eligibility criteria such as inclusion of control population, flawed measurement of exposure and outcome, failure to adequately control confounding) as they differ from randomized controlled studies, according to GRADE (Guyatt et al. 2011). The quality criteria "limitations" for the observational study using SSQ was rated as "not seriously" restricted as we could not identify any limitations. Quality of evidence was very low, as the quality criteria across studies, were similar to VAS, barely fulfilled ("serious inconsistency", "serious indirectness", "serious imprecision"). Based on the Sign test (p=0.64), we did not find evidence across studies from SSQ showing higher listening effort ratings for aided versus unaided listening conditions.

Quality of evidence for behavioral measures, Q2

Two behavioral measurement types included evidence from the application of DTPs (n=10 studies) and reaction time measures (n=3 studies, Table 5). For DTPs, the quality criteria across studies showed "not serious limitations" (no lack of allocation concealment, no selective outcome reporting or early stop for benefit, but lack of experimental blinding and lack of description of treatment of missing data), "serious inconsistency" (no consistent stimulus, test setups and hearing aid settings), "serious indirectness" (young normal-hearing versus elderly hearing-impaired; experienced versus inexperienced hearing aid users) and "serious imprecision" (lack of power sufficiency), which resulted in very low quality of evidence. Based on the Sign test (p=0.41), evidence across studies did not show that listening effort assessed by DTPs was higher for aided versus unaided listening.

Evidence on Q2 from reaction time measures (n=3 studies) had very low quality, based on very similar findings on the quality criteria across studies as described for the DTP measures. The results from the Sign test (p=0.06) on findings from reaction time measures across studies, did not indicate that aided listeners show lower listening effort than unaided listeners.

Discussion

The aim of this systematic literature review was to provide an overview of available evidence on: Q1) does hearing impairment affect listening effort? and Q2) does hearing aid amplification affect listening effort during speech comprehension?

Outcome measures on Q1

Evidence and quality of evidence from subjective measures

Across studies using subjective measures, we did not find systematic evidence that listening effort assessed by subjective measures was higher for hearing-impaired compared to normal-hearing listeners. A possible explanation for the weakness of evidence could be the great diversity of subjective measurement methods. For example, we identified eleven different rating scales for VAS, with varying ranges, step sizes labels and different wordings. Even though a transformation of scales to the same range can provide more comparable findings, it may still be questionable whether labels and meanings, such as “effort”, “difficulty” or “ease of listening”, are actually comparable across studies. The great variety in VAS scales may arise as subjective ratings were sometimes applied as an additional test to behavioral (Feuerstein, 1992; Desjardins & Doherty, 2014; Bentler & Duve, 2000) or physiological measures of listening effort (Hicks & Tharpe, 2002; Zekveld et al. 2011), in studies with varying research questions and test modalities. The variety of subjective scales illustrates how immature the methods for subjective assessment of listening effort still are. Comparing subjective findings across studies requires greater agreement in terminology, standardized methods and comparable scales.

Evidence and quality of evidence from behavioral measures

Evidence from DTPs and reaction time measures did not support our first hypothesis (HP1; higher listening effort scores for hearing-impaired listeners compared to normal-hearing listeners). The barely fulfilled GRADE quality criteria on DTP are caused by the great diversity of test setups across DTPs. The primary tasks typically applied sentence or word recall, and varied mainly in the type of speech material. However, the variety across secondary tasks was much greater, including visual motor tracking, reaction time tasks, memory recall, digit memorization, or driving in a car simulator. The diversity of tasks within DTPs is probably related to the developmental stage of research on listening effort, aiming for the most applicable and realistic method and better understanding of the concept of listening effort. However, the applied tasks within the DTPs may actually tax different stages of cognitive processing, such as acquisition, storage and retrieval from working memory or selective and divided attention, which makes a direct comparison of the findings questionable. It is furthermore problematic to compare the results directly as they originate from studies with different motivations and research questions, such as the comparison of single- versus DTPs (Stelmachowicz et al. 2007), the effect of age (Stelmachowicz et al. 2007; Tun et al. 2009; Desjardins & Doherty, 2013), cognition (Neher et al. 2014) or different types of stimuli (Feuerstein, 1992; Desjardins & Doherty, 2013). Evidence on reaction time measures resulted from just one study and showed better quality according to the GRADE criteria compared to evidence from DTPs, mainly because findings within a single study (reaction times) are less diverse than findings across eight studies (DTP).

Evidence and quality of evidence from physiological measures

EEG measures indicated that certain brain areas, representing cognitive processing, were more active during the compensation for reduced afferent input to the auditory cortex (Oates et al. 2002; Korczak et al. 2005). It seems reasonable, that evidence from EEG measures supported HP1, as brain activity during auditory stimulus presentation was compared between

hearing-impaired and normal-hearing listeners or for simulations of hearing impairment. Brain activity increased in response to a reduced level of fidelity of auditory perception for listeners with impaired hearing compared to those with normal hearing. The findings on the outcome of EEG were consistent and directly comparable across studies, as the same deviant stimuli were presented at the same presentation levels. However, quality of evidence rating by GRADE (Table 3) was still moderate, and research with less "imprecision" is required to provide reliable findings and conclusions on the results.

Summary of evidence and quality of evidence on Q1

The quality of evidence across measurement methods was not consistent and we found evidence of moderate quality (reaction time and EEG), low quality (DTP) or very low quality (VAS, pupillometry). Overall, evidence from physiological assessment supported HP1, but the moderate quality of this evidence may not allow high confidence in this finding. However, this result raises the intriguing question of how it was possible to show a significant effect of hearing-impairment on listening effort when evidence was based on findings from EEG measures (physiological), but not for any subjective or behavioral measure. The time-locked EEG activity (especially N2, P3), which corresponds to neural activity related to cognitive processing, may more sensitively reflect changes in the auditory input (e. g. background noise or reduced hearing abilities) than measures corresponding to behavioral consequences (e. g. reaction time measures) or perceived experiences (e. g. subjective ratings) of listening effort. However, effects of hearing impairment may still cover unknown factors that may be difficult to capture as they depend on the degree of hearing impairment, the intensity of the stimulus and the level of cortical auditory processing that the response measure is assessing.

Outcome measures on Q2

Evidence and quality of evidence from subjective measures

We identified twice as many findings from subjective assessment for Q2 compared to Q1. However, great diversity across scales, great variety of applied comparisons (e. g. aided versus unaided, active versus inactive algorithms, processed versus unprocessed stimuli) together with a variety of tested hearing aid algorithms prevented comparisons across studies. Consequently quality criteria, such as “inconsistency” and “indirectness” were poorly fulfilled. We believe that self-report measures should be more uniform to increase comparability. Furthermore, information on applied stimulus, environmental factors and individual motivation should be taken into account to provide better understanding of the findings.

Evidence and quality of evidence from behavioral measures

The systematic evidence on behavioral measures is small due to the diversity of behavioral measurement methods across studies, as was also the case for Q1. It is very difficult to compare task evoked findings on varying levels of cognitive processing for a great diversity of tasks, factors of interest and compared settings and conditions. The quality of evidence suffers as a consequence.

Evidence and quality of evidence from physiological measures

We observed a general lack of evidence on the effect of hearing-aid amplification on listening effort assessed by physiological measures. The use of hearing aids or CIs may be incompatible with some physiological measures such as fMRI.

Summary of evidence and quality of evidence on Q2

Even though there was no consistent evidence showing increased listening effort due to hearing impairment (HP1), it was surprising to see that even the existing evidence for less listening effort due to hearing aid amplification (HP2) was not significant. The diversity of

tests within each measurement type (subjective, behavioral and physiological) seems to restrict the amount of comparable, systematic evidence and consequently the quality of evidence. It is for example still unclear which factors influence subjective ratings of perceived listening effort and what motivates listeners to stay engaged versus giving up on performance. This kind of information would support more clear interpretations of outcomes of self-ratings of listening effort.

Limitations of the body of the search

This review illustrates the great diversity in terms of methodology to assess listening effort between different studies, which makes a direct comparison of the data problematic. Furthermore, the comparability of those findings is questionable as the different measurement methods may not tax the same cognitive resources. For example, the subjective and behavioral measures may assess different aspects of listening effort (Larsby et al. 2005; Fraser et al. 2010). In addition, a study by Zekveld and colleagues (2011) failed to show a relation between a subjective and a physiological measure (the pupil dilation response). We recommend that interpretation differences need to be resolved, by determining which measurement types reflect which elements of cognitive processing and listening effort. As an important part of this resolution, a unifying conceptual framework for listening effort and its components is much needed.

Limitations of our review

The definition of listening effort and the strict inclusion and exclusion criteria created for the search could be one limitation of the outcome of this systematic review. Studies were only included when the wording “listening effort” was explicitly used and results were provided by an outcome measure reflecting the effects of hearing impairment or hearing-aid amplification. Meanwhile, there are potentially relevant studies which were not included, for example

focusing on the effect of adverse listening conditions on alpha oscillations (which are often interpreted as a measure for attention or memory load) (Obleser et al. 2012; Petersen et al. 2015), or studying the relationship between hearing impairment, hearing aid use and sentence processing delay by recording eye fixations (Wendt et al. 2015). Such studies often apply different terminologies or keywords, which prevents them passing our search filters. An alternative view of this situation might be that it reflects the current lack of definition of what is and is not 'Listening Effort'.

Only two additional articles were identified by checking the reference lists from the 39 articles deemed to be relevant from the initial search. This might indicate that the set of search terms was well defined, or alternatively, that researchers in this field tend not to look far afield for inspiration.

The search output was certainly limited by the fixed end date for the inclusion of articles. Furthermore, only English language articles were considered, which may limit the search output.

This review produced evaluations of evidence quality which were generally disappointing. This should not be interpreted as an indication that the measurement methods used in the many studies included are inherently inadequate, merely that they have been applied in ways which are inconsistent and imprecise across studies. According to GRADE, low or very low quality of evidence resulted mainly due to "inconsistency", "indirectness" and "imprecision" across studies. The applied experimental setups across studies were inconsistent as most presented target and masker stimuli differed and participants were tested in different listening environments. We identified "serious indirectness" across studies as findings across studies resulted from testing different populations, including young normal-hearing listeners, elderly hearing-impaired listeners, normal-hearing and hearing-impaired children, simulated, conductive impairment, unilateral or bilateral hearing-aid usage, unilateral and bilateral CI usage. This does not mean that applied measurement methods within each individual study

were flawed. However, “serious inconsistency and indirectness” within GRADE does indicate that different test methods across studies may influence the reliability of the results as the tasks and the tested populations, used to evoke those results, differ. Non-randomized observational studies were not considered flawed as compared to randomized control trial studies as GRADE accounts for the design of the assessed studies and different sub-criteria are applied to evaluate the criterion called “study limitations” (see Table 5). Within this review findings from only one non-randomized observational study were included.

Conclusions:

Reliable conclusions, which are much needed to support progress within research on listening effort, are currently elusive. The body of research so far is characterized by a great diversity regarding the experimental setups applied, stimuli used and participants included. This review revealed a generally low quality of evidence relating to the question Q1; does hearing impairment affect listening effort? and Q2; can hearing-aid amplification affect listening effort during speech comprehension? Amongst the subjective, behavioral and physiological studies included in the review, only the results from the Sign test on the outcome of EEG measures indicated, that hearing-impaired listeners show higher listening effort than normal-hearing listeners. No other measurement method provided statistical significant evidence indicating differences in listening effort between normal-hearing and hearing-impaired listeners. The quality of evidence was moderate for the EEG data as little variability across studies, including the test stimuli, the experimental setup and the participants, was identified. Only physiological studies generated moderately reliable evidence, indicating that hearing impairment increases listening effort, amongst the subjective, behavioral and physiological studies included in this review. It seems fair to say that research on listening effort is still at an early stage.

Future directions:

More research is needed to identify the components of listening effort, and how different types of measures tap into them. Less diversity across studies is needed to allow comparability and more reliable conclusions based on current findings. The community needs to develop more uniform measures for distinct components of listening effort, as well as clear definitions of different aspects of cognitive processing, in order to understand current findings and to apply further research resources efficiently.

Acknowledgments

The authors thank Dorothea Wendt for her intellectual input and the fruitful discussion of this study. This article presents independent research funded by the European Commission (grant LISTEN607373).

PubMed:

#1 Hearing aid (technology)
"Hearing Aids"[Mesh:NoExp] OR "background noise"[tiab] OR "noise reduction"[tiab] OR "hearing aid"[tiab] OR "hearing aids"[tiab]OR "hearing loss"[tiab] OR "hearing impaired"[tiab] OR "hearing impairment"[tiab]

#2 Listening effort
"Speech Perception"[Mesh] OR "Reflex, Pupillary"[Mesh] OR "listening effort"[tiab] OR "perceptual effort"[tiab] OR "speech perception"[tiab] OR "speech discrimination"[tiab] OR"speech understanding"[tiab] OR "auditory stress"[tiab] OR "auditory fatigue"[tiab] OR "listening fatigue"[tiab] OR "cognitive load"[tiab] OR "Speech Acoustics"[tiab]OR "Speech Intelligibility"[tiab] OR "Pupillary reflex"[tiab] OR "ease of listening"[tiab] OR"Memory"[Mesh:NoExp] OR memory[tiab]

EMBASE.com:

#1 Hearing aid (technology)
'hearing aid'/de OR'background noise':ti,abOR 'noise reduction':ti,ab OR 'hearing aid':ti,ab OR 'hearing aids':ti,abOR 'hearing loss':ti,abOR 'hearing impaired':ti,ab OR 'hearing impairment':ti,ab

#2 Listening effort
'speech perception'/exp OR 'pupil reflex'/exp OR'listening effort':ti,ab OR 'perceptual effort':ti,ab OR 'speech perception':ti,ab OR 'speech discrimination':ti,ab OR'speech understanding':ti,ab OR 'auditory stress':ti,ab OR 'auditory fatigue':ti,ab OR 'listening fatigue':ti,ab OR 'cognitive load':ti,ab OR 'Speech Acoustics':ti,abOR 'Speech Intelligibility':ti,ab OR 'Pupillary reflex':ti,ab OR 'ease of listening':ti,ab OR'memory'/de OR memory:ti,ab

Cinahl:

#1 Hearing aid (technology)
(MH "Hearing Aids+") OR (MH "Auditory Brain Stem Implants") OR TI ("background noise" OR "noise reduction" OR "hearing aid" OR "hearing aids" OR "hearing loss"OR "hearing impaired" OR "hearing impairment") OR AB ("background noise" OR "noise reduction" OR"hearing aid" OR "hearing aids" OR "hearing loss"OR "hearing impaired" OR "hearing impairment")

#2 Listening effort
(MH "Speech Perception") OR (MH "Reflex, Pupillary") OR (MH "Memory") OR TI ("listening effort" OR "perceptual effort" OR "speech perception" OR "speech discrimination" OR"speech understanding" OR "auditory stress" OR "auditory fatigue" OR "listening fatigue" OR "cognitive load" OR "Speech Acoustics"OR "Speech Intelligibility" OR "Pupillary reflex" OR "ease of listening" OR memory) OR AB ("listening effort" OR "perceptual effort" OR "speech perception" OR "speech discrimination" OR"speech understanding" OR "auditory stress" OR "auditory fatigue" OR "listening fatigue" OR "cognitive load" OR "Speech Acoustics"OR "Speech Intelligibility" OR "Pupillary reflex" OR "ease of listening" OR memory)

PsycINFO:

#1 Hearing aid (technology)
DE "Hearing Aids" OR TI ("background noise" OR "noise reduction" OR "hearing aid" OR "hearing aids" OR "hearing loss"OR "hearing impaired" OR "hearing impairment") OR AB

("background noise" OR "noise reduction" OR"hearing aid" OR "hearing aids" OR "hearing loss"OR "hearing impaired" OR "hearing impairment")

#2 Listening effort
(DE "Speech Perception") OR (DE "Memory") OR TI ("pupillary reflex" OR "listening effort" OR "perceptual effort" OR "speech perception" OR "speech discrimination" OR"speech understanding" OR "auditory stress" OR "auditory fatigue" OR "listening fatigue" OR "cognitive load" OR "Speech Acoustics"OR "Speech Intelligibility" OR "Pupillary reflex" OR "ease of listening" OR memory) OR AB ("pupillary reflex" OR "listening effort" OR "perceptual effort" OR "speech perception" OR "speech discrimination" OR"speech understanding" OR "auditory stress" OR "auditory fatigue" OR "listening fatigue" OR "cognitive load" OR "Speech Acoustics"OR "Speech Intelligibility" OR "Pupillary reflex" OR "ease of listening" OR memory)

Cochrane Library:

#1 Hearing aid (technology)
"background noise" OR "noise reduction" OR "hearing aid" OR "hearing aids"OR "hearing loss"OR "hearing impaired" OR "hearing impairment"

#2 Listening effort
"Speech Perception" OR "Pupillary reflex" OR "listening effort" OR "perceptual effort" OR "speech perception" OR "speech discrimination" OR"speech understanding" OR "auditory stress" OR "auditory fatigue" OR "listening fatigue" OR "cognitive load" OR "Speech Acoustics"OR "Speech Intelligibility" OR "ease of listening" OR memory

#4 excluded publication types
NOT ("addresses"[Publication Type] OR "biography"[Publication Type] OR "comment"[Publication Type] OR "directory"[Publication Type] OR "editorial"[Publication Type] OR "festschrift"[Publication Type] OR "interview"[Publication Type] OR "lectures"[Publication Type] OR "legal cases"[Publication Type] OR "legislation"[Publication Type] OR "letter"[Publication Type] OR "news"[Publication Type] OR "newspaper article"[Publication Type] OR "patient education handout"[Publication Type] OR "popular works"[Publication Type] OR "congresses"[Publication Type] OR "consensus development conference"[Publication Type] OR "consensus development conference, nih"[Publication Type] OR "practice guideline"[Publication Type]) NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms])

Figure legends

Figure 1: PRISMA flow diagram of the study identification, the screening, the eligibility and the inclusion process within the systematic search.

Reference List

Ahlstrom, J. B., Horwitz, A. R., & Dubno, J. R. (2014). Spatial separation benefit for unaided and aided listening. *Ear Hear, 35*(1), 72-85.

Armstrong, E. C. (1999). The well-built clinical question: the key to finding the best evidence efficiently. *Wis Med J 1999;98*(2), 25-8

Baguley, T. (2012). *Serious Stats: A Guide to Advanced Statistics for the Behavioral Science.* Palgrave Macmillan.

Bentler, R., Wu, Y. H., Kettel, J., & Hurtig, R. (2008). Digital noise reduction: outcomes from laboratory and field studies. *Int J Audiol, 47*(8), 447-460.

Bentler, R. A., & Duve, M. R. (2000). Comparison of hearing aids over the 20th century. *Ear Hear, 21*(6), 625-639.

Bernarding, C., Strauss, D. J., Seidler H., & Corona-Strauss, F. I. (2012). Neural correlates of listening effort related factors: Influence of age and hearing impairment. *Brain Res Bulletin, 91*, 21-30.

Bertoli, S., & Bodmer, D. (2016). Effects of age and task difficulty on ERP responses to novel sounds presented during a speech-perception-in-noise test. *Clin Neurophysiol, 127(1)*, 360-368.

Brons, I., Houben, R., & Dreschler, W. A. (2013). Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort. *Ear Hear, 34*(1), 29-41.

Brons, I., Houben, R., & Dreschler, W. A. (2014). Effects of noise reduction on speech intelligibility, perceived listening effort, and personal preference in hearing-impaired listeners. *Trends hear, 18*.

Demorest, M. E., & Erdman, S. A. (1986). Scale composition and item analysis of the Communication Profile for the Hearing Impaired. *J Speech Lang Hear Res, 29*(4), 515-535.

Desjardins, J. L., & Doherty, K. A. (2013). Age-Related Changes in Listening Effort for Various Types of Masker Noises. *Ear Hear, 34*, 261-272.

Desjardins, J. L., & Doherty, K. A. (2014). The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. *Ear Hear, 35*(6), 600-610.

Dillon H. (2001). Hearing Aids (1st ed.). Turramurra, Australia: Boomerang Press.

Downs, D. W. (1982). Effects of hearing aid use on speech discrimination and listening effort. *J Speech Hear Dis, 47*(2), 189-193.

Dwyer, N. Y., Firszt, J. B., & Reeder, R. M. (2014). Effects of unilateral input and mode of hearing in the better ear: self-reported performance using the speech, spatial and qualities of hearing scale. *Ear Hear, 35*(1), 126-136.

Ebell, M. H. (1999). Information at the point of care: answering clinical questions. *J Am Board Fam Med, 12*(3), 225-235.

Edwards, B. (2007). The future or hearing aid technology. *Trends Amplif, 11*, 31-45.

Feuerstein, J. F. (1992). Monaural versus binaural hearing: ease of listening, word recognition, and attentional effort. *Ear Hear, 13*(2), 80-86.

Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the Effort Expended to Understand Speech in Noise Using a Dual-Task Paradigm: The Effects of Providing Visual Speech Cues. *J Speech Lang Hear Res, 53*, 18-33.

Gatehouse, S., & Gordon, J. (1990). Response times to speech stimuli as measures of benefit from amplification. *Br J Audiol, 24*(1), 63-68.

Gosselin, P. A., & Gagne, J. P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *J Speech Lang Hear Res, 54*(3), 944-958.

Gosselin, P. A., & Gagné, J. P. (2010). Use of a Dual-Task Paradigm to Measure Listening Effort. *Can J Speech Lang Pathol Audiol, 34*, 43-51

Granholm, E., Asarnow, R. F., Sarkin, A. J., & et al. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology, 33*, 457-461.

Gustafson, S., McCreery, R., Hoover, B., Kopun, J. G., & Stelmachowicz, P. (2014). Listening effort and perceived clarity for normal-hearing children with the use of digital noise reduction. *Ear Hear, 35*(2), 183-194.

Guyatt, G. H., Oxman, A. D., Gunn, E., Kunz, R., Falck-Ytter, Y., & Schünemann, H. J. (2008). GRADE: what is 'quality of evidence' and why is it important to clinicians? *Analysis, 336*, 995-998.

Guyatt, G. H., Oxman, D., Schünemann, H. J., Tugwell, P., & Knottnerus, A. (2011). GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol, 64*(4), 380-382.

Hagerman, B. (1984). Clinical Measurements of Speech Reception Threshold in Noise. *Scand Audiol, 13*(1).

Hällgren, M., Larsby, B., Lyxell, B., & Arlinger, S. (2005). Speech understanding in quiet and noise, with and without hearing aids. *Int J Audiol, 44*(10), 574-583.

Harlander, N., Rosenkranz, T., & Hohmann, V. (2012). Evaluation of model-based versus non-parametric monaural noise-reduction approaches for hearing aids. *Int J Audiol, 51*(8), 627-639.

Hedley-Williams, A., Humes, L. E., Christensen, L. A., Bess, F. H., & Hedley-Williams, A. (1997). A Comparison of the Benefit Provided by Well-Fit Linear Hearing Aids and Instruments With Automatic Reductions of Low-Frequency Gain. *J Speech Lang Hear Res, 40*, 666-685.

Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *J Speech Lang Hear Res, 45*(3), 573-584.

Hopkins Katheryn, Moore Brian C.J., & Stone Michael A. (2005). Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. *J Acoust Soc Am, 123*(2), 1140-1153.

Hornsby, B. W. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear Hear, 34*(5), 523-534.

Howard Clara S., Munro Kevin J., & Plack Christopher J. (2010). Listening effort at signal-to-noise ratios that are typical of the school classroom. *Int J Audiol, 49*, 1708-8186.

Humes Larry E., & Roberts Lisa. (1990). Speech-Recognition Difficulties of the Hearingimpaired Elderly: The Contributions of Audibility. *J Speech Hear Res, 33,* 726-735.

Humes, L. E. (1999). Dimensions of hearing aid outcome. *J Am Acad Audiol, 10*(1), 26-39.

Humes, L. E., Christensen, L. A., Bess, F. H., & Hedley-Williams, A. (1997). A comparison of the benefit provided by well-fit linear hearing aids and instruments with automatic reductions of low-frequency gain. *J Speech Hearing Res, 40*(3), 666-685.

Humes, L. E., & Humes, L. E. (2004). Factors affecting long-term hearing aid success. *In Seminars in Hearing* (Vol. 25, No. 01, pp. 63-72). Copyright© 2004 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA.

Kahneman, D. (1973). *Attention and effort.* New Jersey: Prentice-Hall.

Koelewijn Thomas, Zekveld, A., Festen Joost M., & Kramer Sophia E. (2012). Pupil Dilation Uncovers Extra Listening Effort in the Presence of a Single-Talker Masker. *Ear Hear, 33,* 291-300.

Korczak, P. A., Kurtzberg, D., & Stapells, D. R. (2005). Effects of sensorineural hearing loss and personal hearing AIDS on cortical event-related potential and behavioral measures of speech-sound processing. *Ear Hear, 26*(2), 165-185.

Kramer Sophia E., Kapteyn Theo S., Festen Joost M., & Kuik Dirk J. (1997). Assessing Aspects of Auditory Handicap by Means of Pupil Dilatation. *Audiology, 36,* 155-164.

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. *Int J Audiol, 45*(9), 503-512.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology, 50*(1), 23-34.

Kulkarni, P. N., Pandey, P. C., & Jangamashetti, D. S. (2012). Multi-band frequency compression for improving speech perception by listeners with moderate sensorineural hearing loss. *Speech Communication, 54*(3), 341-350.

Laeng, B., Sirois, S., & Gredebaeck, G. (2012). Pupillometry: A Window to the Preconscious? *Perspectives on Psychological Science, 7*(1), 18-27.

Larsby, B., Hallgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *Int J Audiol, 44*(3), 131-143.

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Buechler, M. et al. (2010). Multicenter evaluation of signal enhancement algorithms for hearing aids. *J Acoust Soc Am, 127*(3), 1491-1505.

Mackersie, C. L., MacPhee, I. X., & Heldt, E. W. (2015). Effects of Hearing Loss and Heart Rate Variability and Skin Conductance Measured During Sentence Recognition in Noise. *Ear Hear, 36*(1), 145-154.

Mackersie, C. L., & Cones Heather. (2011). Subjective and psychophysiological indices of listening effort in a competing-talker task. *J Am Acad Audiol, 22*(2), 113-122.

Martin, B. A., & Stapells, D. R. (2005). Effects of low-pass noise masking on auditory event-related potentials to speech. *Ear Hear, 26*(2), 195-213.

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *Quarterly J Exp Psychol, 58*(1), 22–33.

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'. *Int J Audiol, 53,* 433–445.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine, 151*(4), 264-269.

Neher, T., Grimm Giso, Hohmann Volker, & Kollmeier Birger. (2013). Do Hearing Loss and Cognitive Function Modulate Benefit From Different Binaural Noise-Reduction Settings? *Ear Hear, 35,* e52-e62.

Neher, T., Grimm, G., & Hohmann, V. (2014). Perceptual consequences of different signal changes due to binaural noise reduction: do hearing loss and working memory capacity play a role? *Ear Hear, 35*(5), e213-e227.

Noble, W., & Gatehouse, S. (2006). Effects of bilateral versus unilateral hearing aid fitting on abilities measured by the Speech, Spatial, and Qualities of Hearing scale (SSQ). *Int J Audiol, 45*(3), 172-181.

Noble, W., Tyler, R., Dunn, C., & Bhullar, N. (2008). Unilateral and bilateral cochlear implants and the implant-plus-hearing-aid profile: comparing self-assessed and measured abilities. *Int J Audiol, 47*(8), 505-514.

Oates, P. A., Kurtzberg, D., & Stapells, D. R. (2002). Effects of sensorineural hearing loss on cortical event-related potential and behavioral measures of speech-sound processing. *Ear Hear, 23*(5), 399-415.

Obleser Jonas, Wöstmann Malte, Hellbernd Nele, Wilsch Anna, & Maess Burkhard. (2012). Adverse Listening Conditions and Memory Load Drive a Common Alpha Oscillatory Network. *J Neurosci, 32*(36), 12376-12383.

Pals, C., Sarampalis, A., & Baskent, D. (2013). Listening Effort With Cochlear Implant Simulations. *J Speech, Lang Hear Res, 56*(4), 1075-1084.

Petersen Eline B., Wöstmann Malte, Obleser Jonas, Stenfelt Stefan, & Lunner, T. (2015). Hearing loss impacts neural alpha oscillations under adverse listening conditions. *Front Psychol, 6*(177).

Picou, E. M., Aspell, E., & Ricketts, T. A. (2014). Potential benefits and limitations of three types of directional processing in hearing aids. *Ear Hear, 35*(3), 339-352.

Picou, E. M., Ricketts, T. A., & Hornsby, B. W. Y. (2013). How hearing aids, background noise, and visual cues influence objective listening effort. *Ear Hear, 34*(5), 52-64.

Plomp Reinier. (1986). A Signal-to-Noise Ratio Model for the Speech-Reception Threshold of the Hearing Impaired. *J Speech, Lang Hear Res, 29*, 146-154.

Rakerd, B., Seitz, P. F., & Whearty, M. (1996). Assessing the cognitive demands of speech listening for people with hearing losses. *Ear Hear, 17*(2), 97-106.

Richardson WS, Wilson MC, Nishikawa J, & Hayward RS. (1995). The well-built clinical question: a key to evidence based descriptions. *ACP Journal Club, 123*(3).

Rönnberg J., Lunner, T., Zekveld, A., Sörqvist P., Danielsson H., Lyxell, B. et al. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Front Syst Neurosci, 7*(13).

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: effects of background noise and noise reduction. *J Speech, Lang Hear Res, 52*(5), 1230-1240.

Schünemann, H., Brozek, J., Guyatt, G., & Oxman, A. (2013). *GRADE handbook: Handbook for grading the quality of evidence and the strength of recommendations using the GRADE approach.*

Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends Amplif.*

Steel, M. M., Papsin, B. C., & Gordon, K. A. (2015). Binaural fusion and listening effort in children who use bilateral cochlear implants: A psychoacoustic and pupillometric study. *PloS one, 10*(2), e0117611.

Stelmachowicz, P. G., Lewis, D. E., Choi, S., & Hoover, B. (2007). Effect of stimulus bandwidth on auditory skills in normal-hearing and hearing-impaired children. *Ear Hear, 28*(4), 483-494.

Stephens Dafydd, & Hétu Raymond. (1991). Impairment, Disability and Handicap in Audiology: Towards a Consensus. *Audiology, 30,* 185-200.

Strawbridge William, Wallhagen Margaret I., Shema Sarah J., & Kaplan George A. (2000). Negative Consequences of Hearing Impairment in Old Age: A Longitudinal Analysis. *The Gerontologist, 40*(3), 320-326.

Tun Patricia A., McCoy Sandra L., & Wingfield Arthur. (2009). Aging, Hearing Acuity, and the Attentional Costs of Effortful Listening. *Psychology and Aging, 24*(3), 761-766.

Weinstein, B. E., & Ventry, I. M. (1982). Hearing impairment and social isolation in the elderly. *J Speech, Lang, Hear Res, 25*(4), 593-599.

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful listening: The processing of degraded speech depends critically on attention. *J Neurosci, 32*(40), 14010-14021.

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear Hear, 36*(4), e153-e165.

Wu, Y. H., Aksan, N., Rizzo, M., Stangl, E., Zhang, X., & Bentler, R. (2013). Measuring listening effort: driving simulator versus simple dual-task paradigm. *Ear Hear, 35*(6), 623-632.

Xia, J., Nooraei, N., Kalluri, S., & Edwards, B. (2015). Spatial release of cognitive load measured in a dual-task paradigm in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am, 137*(4), 1888-1898.

Xia, J., Nooraei, N., Kalluri, S., & Edwards, B. (2016). Spatial release of cognitive load measured in a dual-task paradigm in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am, 137*(4), 1888-1898.

Zekveld, 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 Hear, 32*, 498-510.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear Hear, 31*, 480-490.

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2-3
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	7 (4-6)
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	7
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Not available
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	7-9
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	7,9
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	34-35
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Figure 1, and page 7-10
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	9-10
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	8-9, Tables 1, 3, 5
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	Tables 3,5 10-13
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	12-13

PRISMA 2009 Checklist

Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	9-10

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	GRADE, page 10-13, Table 3, 5
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	Sign-test page 12
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Figure 1, 13-16
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Q1: 16-18: Q2: 22-23
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Table 3, 5
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Summary: Table 1; p-value for binomial distribution in Table 3, 5
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	No meta-analysis carried out
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	GRADE, Table 3, 5
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	Sign-test based on binomial distribution, p-values in Table 3, 5
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	25-29

Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	30-31
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	32
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	1, 33

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

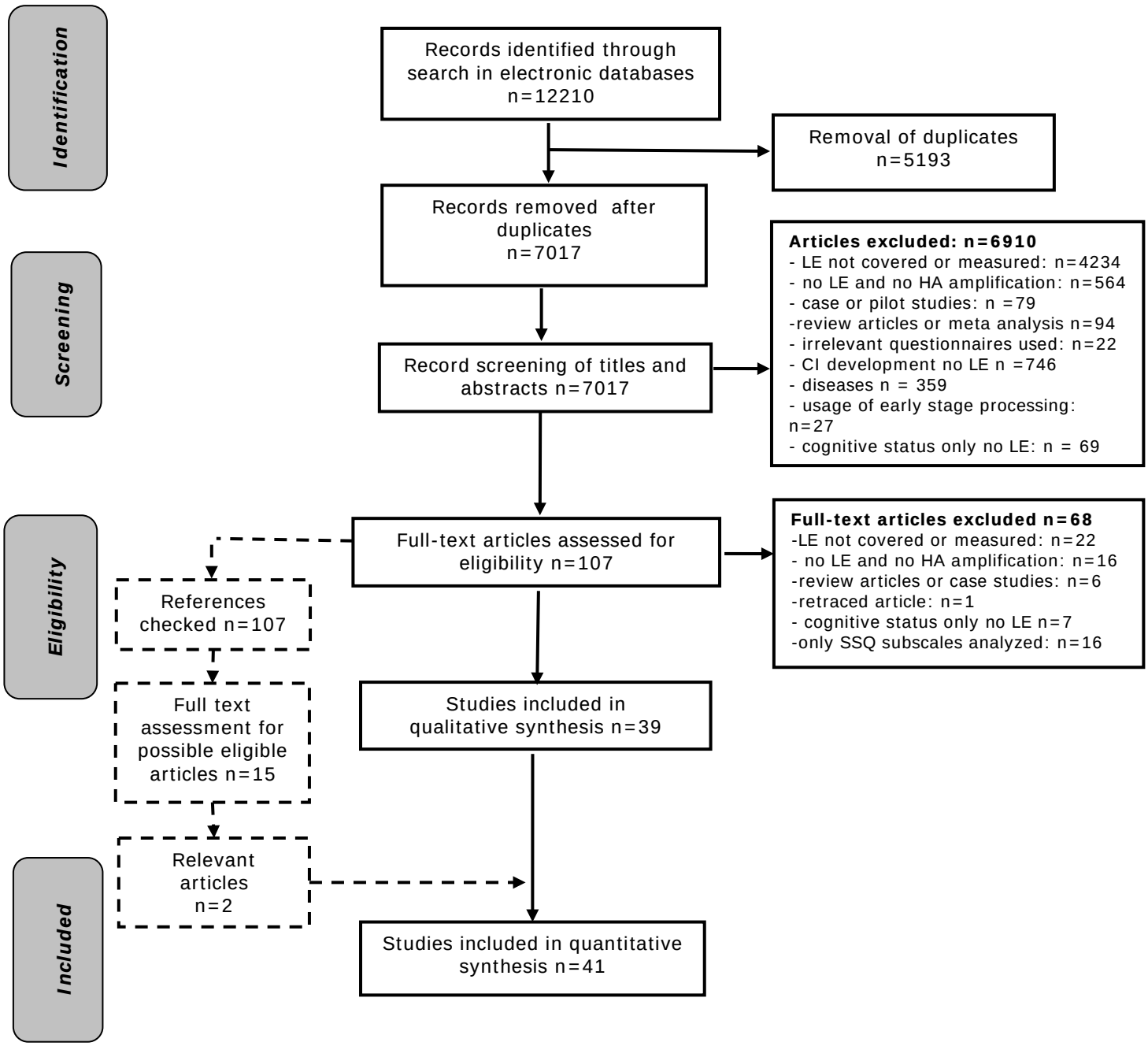
Identification
Screening
Eligibility
Included
Records identified through search in electronic databases n=12210
Removal of duplicates n=5193
Records removed after duplicates n=7017
Articles excluded: n=6910
- LE not covered or measured: n=4234
- no LE and no HA amplification: n=564
- case or pilot studies: n =79
-review articles or meta analysis n=94
- irrelevant questionnaires used: n=22
- CI development no LE n =746
- diseases n = 359
- usage of early stage processing: n=27
- cognitive status only no LE: n = 69
Record screening of titles and abstracts n=7017
Full-text articles assessed for eligibility n=107
Full-text articles excluded n=68
-LE not covered or measured: n=22
- no LE and no HA amplification: n=16
-review articles or case studies: n=6
-retraced article: n=1
- cognitive status only no LE n=7
-only SSQ subscales analyzed: n=16
References checked n=107
Full text assessment for possible eligible articles n=15
Relevant articles n=2
Studies included in qualitative synthesis n=39
Studies included in quantitative synthesis n=41

TABLE 1: Descriptive summary of data extracted from the 41 included articles.

	Study	Number of subjects	Mean age (years)	Type of hearing loss Mod.: moderate Bilat.: bilateral Mon.: monaural	Stimuli	Configurations and processing	Method to measure listening effort	Test parameters A: HA vs. none B: HA 1 vs. HA 2 C: NR; D: DIR; E: NAL; F: DSL; G: other	Author hypothesis (HP) **1) LE: HI > NH** **+: HP supported** **-: HP not supported** **=: no effect**	Author hypothesis (HP) **2) LE: aided < unaided** **+: HP supported** **-: HP not supported** **=: no effect**
	Subjective measures									
1.	Ahlstrom et al. (2014)	a)n=14 NH b)n=10 NH c)n=12 HI	a)23.4 b)67.7 c)69.8	c) Mild to mod.	consonant recognition (0° ,70 dB SPL) in SSN (0° or 90°, 66 dB SPL), stimuli were low-pass filtered at fcut= 1.7, 3.4, 7.1 kHz	BTEs, bilateral, quasi-DSL v4.0	VAS (0-15)	**A:** HA vs. none **Hearing:** NH vs. HI	**1+)** LE unaided decreased with increasing bandwidth for group a, b and c **Age effect:** LE: a < b for fcut = 1.7 and 3.4 kHz, no benefit of spatial separation	**2+)** group c: LE: aided < unaided only for spatially separated speech and noise at fcut = 1.7 kHz
2.	Bentler & Duve (2000)	n=14 HI	60.9	Mild to mod. SHL	Real life (church and restaurant) recordings	5 HAs tested: **a)** linear HA, NAL-R **b)** 1-channel compression HA, FIG6 prescription **c)** digital HA**:** 2-channel WDRC, loudness fitting **d)** digital HA	VAS (1-10) for all 5 HAs, random order	**B:** analogue vs. digital amplification **B:** b) vs. c) vs. d)		**2=)** effect of HAs on LE in both real life conditions
3.	Bentler et al., (2008)	n=25 HI	65.1	Mild to mod. bilateral SHL	AVE™ created virtual sound environment of 3 personally problematic situations, 68-92 dBA	Bilateral BTE fitting: 4-channels, directional microphone, feedback cancellation, DNR, NAL-NL1	VAS (1-10) effect of DNR	**B(C)** DNR-off vs. DNR-on (4, 8, 16s onset)		**2+)** LE rating: DNR-on < DNR-off, but no effect of DNR onset
4.	Brons et al., (2013)	n=10 NH	20.8	<= 15 dB HL	sentences in multi-talker babble noise, SNRs -4, 0, + 4dB; HA output recordings presented via head phones	**B:** 4 DNR conditions from 4 HAs **E**: NAL-RP	VAS (9-1), 1: no effort, 9: extreme effort as #15	**C:** DNR-off vs. DNR-on **B(C):** NR1 to NR4		**2=)** LE: DNR-on = DNR-off; **2=)** LE: DNR1 and DNR4 < DNR2 and DNR3
5.	Brons et al., (2014)	n=20 HI	61.3	Mod. HL	Sentences in multi-talker babble, SRT 50% and fixed SNRs: -4, +4, +10 dB, monaural via head-phones	Hearing aid output recorded, with NAL-RP, for a) DNR mild b) DNR moderate c) DNR strong d) unprocessed	VAS (1-9)	**A:** processed vs. unprocessed **B(C)**: a vs. b vs. c vs. d **Stimulus:** -4 dB vs. +4 dB vs. +10 dB		**2-)** LE at fixed SNRs: c > a and d **Stimuli**: -4 dB > +4 dB > +10 dB for a, b, c, d
6.	Desjardins & Doherty (2013), *see # 25)	a)n=15 YNH b)n=15 ONH c)n=16 OHI	a: 21.66, b: 66.86, c: 68.18	a and b: <= 25 dB HL; c: mild-mod., sensorineural HL	Sentences in ii) 1-talker babble, iii) 6-talker babble or iiii) SSN in free field, SNRs for 76% intelligibility	Only for c: personal HAs, bilateral, DSL, no DNR or DIR processing	VAS (100-0); 100: very easy, 0: very difficult	**Hearing**: a and b vs. c **Age:** a vs. b and c **Stimulus**: ii vs. iii vs. iiii	**1+)** LE: b < c for ii, iii and iiii **1=)** LE: iiii < iii < ii: a = b = c	

7.	Desjardins & Doherty (2014), *see # 26)	n=12 HI	66.0	Mild to mod., bilateral SHL	sentences in quiet and 2-talker babble for a) 50 % and b) 76% performance	bilateral BTEs: DSL-v.5, DNR (such as #39)	VAS (100-0), 100: very easy, 0: very difficult	**B(F)**: DNR-on vs. DNR-off **Stimulus:** a) vs. b)		**2=)** LE: DNR-on = DNR-off for a) and b) **stimuli)** LE: b<a
8.	Dwyer et al. (2014)	a)n=21 NH b) n=30 UHI c) n=20 UCI d) n=16 UHA e)c) and d) got CI→ CICI or CIHA	a)50 b) 50.5 c) 53.4 d) 60.3	b, c, d): asymmetrical, severe to profound HL	daily life environment → UCI, UHA, UHI, NH	Testing with own HAs/ CIs	SSQ (1-10) for e) pre- and post-CI (3 and 6 months)	**A:** NH (a) vs. HI (b, c, d) **Group effects:** UHI, UCI, UHA **Pre-vs. post CI:** UCI vs. CICI and UHA vs. CIHA	**1+)** LE: HI (b, c ,d) > NH	**2=)** LE: c = d **2=)** c and d = b **2=)** LE: post-CI = pre-CI **2+)** LE: CICI and CIHA < b
9.	Feuerstein (1992), *see # 28)	n=48 NH	19	ear plug: mon., conductive HL ~30 dB HL	Sound field: sentences (68°) in multi-talker background noise (68° and 65°), SNR: -5dB	a) monaural near (MN): good ear towards target b) monaural far (MF) good ear toward masker c) binaural listening (BIN)	Perceived LE: VAS (100-0)	**Stimuli:** c vs. a vs. b	**1+)** LE: c < a < b	
10.	Hällgren et al. (2005)	n=24 HI	G1:36.8 (n=12) G2:71.8 (n=12)	mild-to-mod. SHL	Sentences in: a) quiet b) modulated noise c) competing talker in two modalities: i)auditory (loudspeaker) ii) audio-visual (loudspeaker+ visual cues)	Own bilateral HAs	VAS (0-10)	**A:** HA vs. none **Background:** a vs.. b vs.. c		**2+)** LE: aided < unaided Greater HA benefit in a) than in b) and c); **LE(background):** c > b> a
11.	Harlander et al. (2012)	n=14 HI	66	Asymmetrical, mod. SHL	Sentences in: a)multi-talker babble, b)street noise, c)printer noise d)quasi-stationary noise SNRs: -5 to 15 dB, in 4dB-steps	MHA, 3 DNR systems: i)CBB ii) AMS iii) MS	VAS (1-13)	**B(C):** unprocessed vs. i vs. ii vs. iii **Stimuli:** a vs. b vs. c vs. d for all SNRs		**2-)** LE: unprocessed < iii especially for c and d **2+)** LE: unprocessed > ii only for b **2+)** LE: unprocessed > i for a, b, c; LE: iii > ii > i Stimuli: positive (better) SNRs reduced LE
12.	Hicks & Tharpe (2002), *see # 31) and # 46), **only exp. 2**	n=14 HI n=14 NH	NH and HI: 8.8	mild to mod. SHL	a) words (PBK) in quiet b) speech babble: +20, +15, +10 dB SNRs	Own HAs during testing N=12: own HA binaural N=1: HA monaural N=1: no HA	VAS (1-5)		**1=)** NH = HI	
13.	Hornsby (2013), *see # 32)	n=16 HI	65.8	Mild-severe SHL	Words, 70% intelligibility (0°) in cafeteria babble noise from 60°, 120°, 180° , 240°	Multi-channel HAs: a) OMNI and feedback management; b) DIR, DNR, wind noise and reverberation reduction,	pre-DTP and post-DTP LE ratings; SSQ (0-10) questions # 14, # 18, # 19	**A:** HA vs. none		**2=)** LE: unaided = aided (for a and b); LE: pre-DTP < post-DTP
14.	Humes et al. (1997)	a) n=41 HI b) n=53 HI c) n=16 HI	Age range a) 27-85 b) 34-87 c) 47-90	Symmetrical a) mild SHL b) mod. SHL c) severe SHL	sentence recognition in quiet, multi-talker babble and cafeteria noise, free field, +5, +10 dB SNRs	binaural HA fitting, NAL-R	VAS (100-0), 100: extremely easy, 0: very difficult	**A:** HA vs. none **B:** BILL vs. linear	**1+)** LE interaction: HL x noise x HA setting: c > a for BILL in babble noise	**2=)** LE: BILL = linear

15.	Humes et al. (1999)	n=55 HI	Not given	Symmetrical, mild-severe HL	sentence recognition in quiet (50, 60, 75 dB SPL) and babble noise at +5, +10 dB SNR	2 binaural sets of ITC-HAs HA1: BILL with NAL-R HA2: 2-channel WDRC, DSL[i/o]	VAS (0-100)	**A:** HA vs. none **B**: HA1 vs. HA2		**2+)** LE: unaided > HA1 > HA2 for n=55 in quiet (50 and 60 dB SPL)
16.	Luts et al. (2010)	n=38 NH n=34 HI-F n=37 HI-S (F: flat, S: steep slope)	NH: 30 HIF: 62 HIS: 68	NH: <= 20 dB HL HI: mod. SHL	Sentences (0° azimuth) in background noise (90°, 180°, 270°), in office like room; SNRs: -10, -5, 0, +5, +10 dB	Binaural MHA: 3 microphones, NAL-RP	VAS (0-6), 0: no effort, 6: extreme effort	**B1:** unprocessed vs. processed **B2:** SC1, SC2, BSS, WMF, COH **Hearing:** NH vs. HI **Stimulus:** 5 SNRs	**1+)** LE: HI > NH	**2+)** LE: MWT < unprocessed, largest effect at -10,-5, 0 dB SNRs **2+)** LE: SCI, SC2, COH < unprocessed at 0 dB SNR only **2-)** LE: BSS > unprocessed for all SNRs
17.	Mackersie et al. (2009)	n=14 HI	76	mild to mod. bilateral SHL	Sentences (75 dB SPL) in speech babble, +5dB SNR, free field	Simulated cellular phone encoding: wide and narrow band; HA functions: multi-channel, DIR, feedback cancellation, NAL-NL1	VAS (9-1)	**Stimulus:** quiet vs. noise **B1:** wide-band vs. narrow band encoding **B2:** individualized-amplification vs. standard setting		**Stimuli:** LE: quiet < noise **2+)** LE: individual amplification < standard setting **2=)** LE: wide band = narrow band encoding
18.	Neher et al. (2014), *see # 36)	n=10:G_{H+C+}, n=10: G_{H-C+}, n=10: G_{H+C-}, n=10: G_{H-C-} H: hearing C: cognition +:good, -: poor	G_{H+C+}: 73 G_{H-C+}: 75 G_{H+C-}: 76,6 G_{H-C-}: 75.5	bilateral SHL	Sentences in cafeteria noise, at -4, 0, +4 dB SNR	MHA:NAL-RP; OFF: NR inactive; ON: active NR; ON+g: active NR with restored stimuli long term spectrum Noise (N) or speech (S) processing	VAS(1-9)	**B:** $S_{off}N_{off}$, $S_{on}N_{on}$, $S_{on+g}N_{on+g}$, $S_{off}N_{on+g}$, $S_{on+g}N_{off}$ **Stimulus:** -4 vs. 0 vs. +4 dB SNR **Group effect**: H+ vs. H- and C+ vs. C-		**2+)** LE: $S_{on}N_{on}$ < $S_{on+g}N_{off}$ and $S_{off}N_{off}$ **2+)** LE: $S_{on}N_{on}$ and $S_{off}N_{on+g}$ < $S_{off}N_{off}$ and $S_{on+g}N_{off}$ at 0dB and +4dB SNR **2+)** LE (interaction: processing condition x group): $S_{on}N_{on}$ < $S_{off}N_{off}$ for all groups but smallest benefit for G_{H-C-} compared to other groups **2=)** LE: at -4 dB SNR similar across processing conditions **Stimuli**: LE: -4 dB > 0 dB > +4 dB
19.	Noble & Gatehouse (2006)	n=144: no HA n=118: 1 HA n=42: 2 HAs	no HA: 68.3 1 HA: 66.2 2 HAs: 66.4	Mild to mod. SHL	Rating of listening in everyday world, conversations	HA fitting: volume control, NAL-RP	SSQ (1-10)	**A**: no HA vs. 1 HA vs. 2 HAs		**2+)** LE: 2HA < 1HA < no HA
20.	Noble et al. (2008)	n=36: CICI n=70 CI n=39 CIHA	CICI: 64.8 CI: 60.7 CIHA: 61.4	severe SHL	Rating of daily listening situations	Participants own CI or HA	SSQ-50	**B:**CI vs. CICI vs. CIHA **A:** pre vs. post implantation		**2+)** LE: CICI < CI < CIHA **2+)** LE: pre implant > post-implant for CI and CICI but not for CIHA

21.	Palmer et al. (2006)	49 HI	62.1	Mild-mod. SHL	daily listening environment	BTEs fitted: NAL-NL1, DNR, adaptive feedback management; P1: fixed OMNI; P2: adaptive DIR P3: fixed DIR, P4: telecoil	VAS (completely agree - disagree) and diary on HA use	**B:** P1 vs. P2 vs. P3		**2=)** tiredness: P1 = P2 = P3
22.	Pals et al. (2013), *see # 37)	19 NH	22	< 20dB HL	noise-vocoded sentences	Noise vocoding for 2, 4, 6, 8, 12, 16, 24 channels	VAS(0-100)	**B:** unprocessed vs. vocoded speech (all channels) **Stimulus**: single task vs. dual tasks (judgment, mental-rotation) see 35)		**2+)** decreased LE for # channels < 6 compared to unprocessed speech; LE: both dual-tasks > single task
23.	Rudner et al. (2012) **only exp. 2**	30 HI	70	Mild to mod. SHL	sentences in SSN and modulated noise at -2, +4, +10 dB SNR	Participant's own BTEs	VAS(no effort to maximum possible effort)	**B**: fast vs. slow compression as between subject factor		**2=)** no main effect of compression on effort
24.	Zekveld et al. (2011), *see # 51)	n=28 NH n=36 HI n=38 YNH from 2010	NH: 55 HI: 61 NHY: 23	NH and YNH: <+20 dB HL HI: mild-mod. HL	Sentences in stationary noise, in free field, a)in quiet, b)50% intelligibility, c) 71%, d) 84%	Internal processing HI vs. NH	VAS(0-10), 0: no effort, 10: high effort	**Hearing:** NH vs. HI **Stimulus:** $SRT_{50\%}$ vs. $SRT_{71\%}$ vs. $SRT_{84\%}$ **Age:** 23 vs. 55,61 years	Stimuli: LE: b > c > d **1=)** LE: HI = NH	
	Behavioral measures									
25.	Desjardins & Doherty (2013), *see # 6)	a)n=15 YNH b)n=15 ONH c)n=16 OHI	a: 21.66, b: 66.86, c: 68.18	a and b: <= 25 dB HL; c: mild-mod., sensorineural HL	Sentences: high and low context in i) quiet (70dB SPL) or ii) 1-talker babble, iii) 6-talker babble or iiii) SSN in free field, SNRs for 76% intelligibility	See # 5)	DTP: Task 1: sentence recall Task 2: visual motor tracking	**hearing**: a and b vs. c **age**: a vs. b and c **Stimulus**: i vs. ii vs. iii vs. iiii **Stimulus:** high vs low context	LE: YNH: iii> iiii **1+)** LE: b and c > a in ii and iiii **1=)** LE: b = c, in all masker conditions **1=)** LE: high = low context across ii, iii and iiii for a, b and c	
26.	Desjardins & Doherty (2014), *see # 7)	n=12 HI	66.0	Mild to mod., bilateral SHL	sentences in quiet and 2-talker babble for a) 50 % and b) 76% performance	bilateral BTEs: DSL-v.5, DNR (such as #39)	DTP: Task 1: sentence recall Task 2: visual motor tracking	**B(C)**: DNR-on vs. DNR-off **Stimulus:** a) 50% vs. b) 76% performance		**2+)** LE: DNR-on < DNR-off for b **2=)** LE: DNR-off = DNR-on for a **2+)** LE: b > a for DNR-off but LE: a = b when DNR-on
27.	Downs (1982)	n=23 HI	51	N=16 bilat. SHL N=1 bilat. mixed N=6 mon. SHL, mon. mixed HL	CNC words in multi-talker babble, 0dB SNR, free field	Individual BTE fitting: -n=10: 3 fittings tested -n=12: 2 fittings tested -n=1: 1 fitting tested	DTP: task 1: word recall Task 2: ViRT	**A:** HA vs. none		**2+)** ViRTs: with HA < without HA
28.	Feuerstein (1992), *see # 9)	n=48 NH	19	ear plug: mon., conductive HL ~30 dB HL	SPIN sentences and light flash	a) monaural near (MN): good ear towards target b) monaural far (MF) good ear toward masker c) binaural listening (BIN)	attentional LE: DTP task 1: sentence recall task 2: ViRT	**Hearing:** c) vs. a) and b) a(MN) vs. b(MF)	**1=)** LE: c = a **1+)** LE: c < b and a < b Attentional LE not correlated with perceived LE (see #7)	
29.	Gatehouse & Gordon (1990)	n=44 HI	68	Symmetric, mild-mod. SHL	a) Tones: 60, 80 dB SPL b) SSN: 60, 70, 80 dB SPL	HA fitted	RT for response to all stimulus	**A:** HA vs. none		**2+)** RT: aided < unaided, effect larger for d) than c);

					c)words: 60, 70, 80dB SPL d) sentences, levels as c) c) and d) in quiet and +5 dB SNR with b); free field			**Stimulus:** 60 vs. 70 vs. 80 dB SPL		**Stimuli:** RT: 60 < 70 < 80 dB SPL for a)
30.	Gustafson et al. (2014)	n=24 NH	Age range: 7-12	thresholds <= 15 dB HL	CVC non-words in broadband noise; 60 and 65 dBSPL; SNRs: +5dB, 0dB, recorded HA output monaurally presented via headphones	2 different HAs fitted : DSL v5.0, amplitude modulation detection	Verbal RTs for non-word repetition	**B:** DNR-on vs. DNR-off **Stimulus:** 0dB vs. +5dB SNR **Hearing:** HA1 vs. HA2		**2+)** RTs: DNR-on < DNR-off, no variation across both HAs; **Stimuli:** RTs: +5 dB < 0 dB SNR
31.	Hicks & Tharpe (2002), *see #12) and #46), **only exp. 2**	n=14 HI n=14 NH	NH and HI: 8.8	mild to mod. SHL	a) words (PBK) in quiet b) speech babble: +20, +15, +10 dB SNRs	Own HAs during testing N=12: own HA binaural N=1: HA monaural N=1: no HA	DTP: Task 1: word recall Task 2: ViRT	**Hearing:** NH vs. HI **Stimulus:** a vs. b	**1=)** baseline RTs: NH = HI **1+**) ViRT: HI > NH in all conditions	
32.	Hornsby (2013), *see # 13)	n=16 HI	65.8	Mild-severe SHL	Words, 70% intelligibility (0°) in cafeteria babble noise from 60°, 120°, 180°, 240°	Multi-channel HAs: a) OMNI and feedback management; b) DIR, DNR, wind noise and reverberation reduction,	DTP: Task 1: word recall Task 2: memory recall and ViRTs	**A:** HA vs. none **B:** a vs. b		**2+)** RT: unaided > b **2=)** RT: unaided = a
33.	Kulkarni et al. (2012)	Exp.1: n=6 NH Exp.2. n= 8 HI	Exp.1: 35-45 Exp.2: 32-66	Exp.1: thresholds <20dB HL Exp.2: Mod.-severe SHL	Exp. 1 only NH: CVC recognition in broad band noise; SNR: 6, 3, 0, -3, -6, -12, -15 dB Exp. 2: only HI: CVC recognition in quiet Exp 1 and 2: monaural testing via headphones;	Exp.1 and 2 Multi-band frequency compression, CRs: uncompressed, 0.8, 0.6, 0.4, applied on speech	Exp. 1 and 2: RT for stimulus	Exp. 1 and 2 **B:** uncompressed vs. compressed speech (all CRs) **Stimulus:** Range of SNRs		Exp.1: **2+)**RTs for uncompressed: (SNR<0dB)>(SNR>0dB) **but** RT(all CRs): (SNR <0dB) < (SNR >0dB); Exp1: **stimuli:** increased RTs for decreasing SNRs Exp.2: **2+) RTs:** compressed<uncompressed
34.	Martin & Stapells (2005) *see # 49)	n=10 NH	33	<=20 dB HL	Deviant stimuli: /ba/, /da/, monaural, at i) 65 and ii) 80 dB ppSPL in a)quiet b)low pass filtered noise, fc = 250, 500, 1k, 2k, 4kHz, at 50% masking threshold, ipsilateral	HL simulation by masking noise	RTs during discrimination of deviant stimuli	**Noise:** fc effect **Stimulus:** a) vs. b)	**1+)** RTs: 65 dB > 80 dB ppSPL **1+)** RT: (fc<1kHz) > (fc>1Hz) only for i) **1+)** RTs: a) < b) for all fc but especially with increasing fc	
35.	Neher et al. (2013)	G_{H+C+}=10 NH G_{H-C+-}=10 HI G_{H+C-}=10 NH G_{H-C-}=10 HI H+/- NH / HI C+/- cognition good / poor	G_{H+C+}=72.1 G_{H-C+-}=74.7 G_{H+C-}=75.8 G_{H-C-}=75.0	G_{H+C+},PTA: 36.4 G_{H-C+-},PTA: 53.7 G_{H+C-},PTA: 38.1 G_{H-C-},PTA: 57.1	Sentences (convolved with head-related impulse responses from cafeteria babble) in quiet and SNRs: -4, 0, 4 dB	MHA with NAL-RP a) DNR_{off} b) $DNR_{moderate}$ c) DNR_{strong}	DTP (see # 39)) Task 1: sentence recall Task 2: ViRTs	**B:** a) vs. b) vs. c) **Stimulus:** -4 vs. 0 vs. +4 dB SNR **Hearing:** HI vs. NH		**2-)** ViRT: DNR_{strong} > $DNR_{moderate}$ and DNR_{off} for all groups **Stimuli:** LE: reduced ViRTs with increasing SNR

36.	Neher et al. (2014) *see # 18)	n=10:G_{H+C+}, n=10: G_{H-C+}, n=10: G_{H+C-}, n=10: G_{H-C-} H: hearing C: cognition +:good, -: poor	G_{H+C+}: 73 G_{H-C+}: 75 G_{H+C-}: 76,6 G_{H-C-}: 75.5	bilateral SHL	Sentences in cafeteria noise at 0 and -4dB SNR	MHA:NAL-RP; OFF: NR inactive; ON: active NR; ON+g: active NR with restored stimuli long term spectrum Noise (N) or speech (S) processing	DTP: task 1: sentence recall Task 2: ViRT	**B:** $S_{off}N_{off}$, $S_{on}N_{on}$, $S_{on+g}N_{on+g}$, $S_{off}N_{on+g}$, $S_{on+g}N_{off}$ **Stimuli** -4 vs. 0 dB SNR **Group effect**: H+ vs. H- and C+ vs. C-	Stimuli: -4 dB SNR > 0dB SNR **1=)** ViRT: no group effect	**2-)** ViRT: $S_{off}N_{off}$ and all other conditions < $S_{on}N_{on}$ **2+)** ViRT: $S_{on}N_{on}$ > $S_{on+g}N_{on+g}$,
37.	Pals et al. (2013) *see # 22)	n=19 NH	22	< 20dB HL	noise-vocoded sentences	Noise vocoding for 2, 4, 6, 8, 12, 16, 24 channels	DTP: task 1: sentence recall task 2: rhyme judgment and mental rotation	**B:** unprocessed vs. vocoded speech (2 to 24 channels)		**2+)** ViRT: (2-8 channels) > (9-24 channels)
38.	Picou et al. (2013)	n=27 HI	65.3	Bilateral, SHL	Words (65dBA) in quiet and 4-talker babble, 50% to 70% performance, AO and AV cues	bilateral HA fitting: NAL-NL1, digital feedback suppression	DTP: task 1: word recall task 2: ViRT	**A1:** HA vs. none **Stimulus1:** words in quiet vs. babble noise **Stimulus2**: AO vs. AV		**2+)** ViRT: HA < no HA ViRT: AO = AV and no interaction with HA use or noise ViRT: quiet < noise
39.	Picou et al. (2014)	n=18 HI	69.1	Mild to mod. SHL	Monosyllable words (62 dBA), in background noise (55dBA), audio-visual presentation	bilateral BTEs: NAL-NL1, feedback reduction DIR: a) mild, b) moderate, c) strong; DNR active only for b) and c)	DTP (as 35)) task 1: word recall task 2: ViRT	**B(D):** a vs. b vs. c **C:** DNR inactive for a vs. active b and c		**2=)** ViRT: DIR-off = DIR-on for n=17
40.	Rakerd et al. (1996)	Exp. 1 n=8 NH, n=9 HI Exp. 2: n=11 NH, n=11 HI	Exp.1:not given, young Exp.2: NH: 24; HI2: 62	Exp. 1: NH: < 15 dB HL; HI: congenital /early onset HL Exp.2: HL through presbyacusis	Exp.1 and 2: a)speech b) SSN	Exp. 1 and 2: monaural testing via headphones NH: speech at 65dB SPL HI: amplification for MCL	Exp. 1 and 2: DTP Task 1: listening to a) or b) with following questions for a) Task 2: digit memorization (9 digits for exp.1, 11 for exp.2	Exp. 1 and 2: **Hearing:** NH vs. HI **Stimulus:** a) vs. b)	Exp.1 and 2: **1+)** recall: HI < NH Exp1 and 2: recall: a < b **1+)** forgotten digits: HI (exp1) > HI (exp2) > NH	
41.	Sarampalis et al. (2009) **only exp. 2**	n=25 NH	21	<= 15 dB HL	sentences in a)quiet, 65 dB SPL b)4-speaker babble, SNRs: -6, -2, +2dB, headphone presentation	i)unprocessed ii) DNR only for b)	DTP: task 1: sentence recall Task 2:ViRT	**Stimulus:** a) vs. b) **B(C):** i) vs. ii)		**2=)** ViRTs: i) = ii) for SNRs: -2, +2dB **2+)** ViRTs: i) > ii) for SNR: -6dB only; ViRTs: a) < b) for all SNRs
42.	Stelmachowicz et al. (2007)	n=32 NH n=24 HI	Range: 7-14 n=14 in 4 age groups	NH: < 20dB HL HI: Mild to severe SHL	b) PBK words in SSN, SNR:+8 dB, bandwidth (BW): a)stimuli filtered at 5kHz, b) filtered at 10kHz	HI: DSL (v.5.0)	DTP: task 1: word recall Task 2: digit recall	**Stimulus**:5 vs. 10 kHz BW **Hearing:** NH vs. HI **Age:** 7-8 vs. 9-10 vs. 11-12 vs. 13-14 **Tasks:** dual vs. single	Stimuli: single-task < dual-task for HI and NH, for a) and b) **1=)** LE (digit recall): a) = b) for NH and HI Age: 9-10 HI < all other; 13-14 NH < all other	

43.	Tun et al., 2009	n=12: ONH n=12: OHI n=12: YNH n=12: YHI	YNH and YHI: 27.9; ONH and OHI: 73.9	Sensorineural, mild-mod. HL	Words: related and unrelated; insert ear phone, monaural at better ear, 70 dB HL	Internal processing HI vs. NH	DTP: task 1: word recall Task 2: mean % time on target for visual tracking	**Hearing**: NH vs. HI **Age**: young vs. old	**1+)** LE: OHI > ONH **1=)** LE: YHI = YNH	
44.	Wu et al. (2014)	Exp.1: n=29 HI Exp. 2: n=19 HI Exp.3: n=24 NH	Exp.1: 72.7 Exp. 2: 72.7 Exp. 3: 23.4	Exp. 1 and 2: SHL Exp. 3: <=25 dB HL	Exp. 1, 2, 3: Sentences in road-noise, -1 dB SNR;	Exp. 1 and 2: Recordings of HA output: compression, feedback suppression, NAL-NL1, DNR-off, stimuli presentation: ear-phones Exp. 3: MCL for directional conditions, no HL dependent amplification	Exp. 1: DTP Task 1: sentence recall Task 2: driving vehicle in simulator Exp. 2 and 3: DTP: Task 1: sentence recall Task 2: ViRT	Exp 1, 2: **A:** HA vs. none **B:** OMNI vs. DIR **Stimulus:** single-task vs. dual-task Exp. 3: **A:** unaided vs. aided **B:** OMNI vs. DIR	Exp. 1: Speech perception: dual-task < single-task Exp. 2: Speech perception: dual-task = single-task Exp. 3: **1+)** LE: young NH < elderly HI	Exp. 1: **2=)** no effect of HA conditions on driving performance Exp 2: **2=)** no effect of HA conditions on RTs Exp. 3: **2+)** RT: DIR < OMNI **2+)** RT: DIR < unaided
	Physiological measures									
45.	Kramer et al., 1997	n=14 NH n=14 HI	NH: 29 HI: 44	HI: mild- mod.	Sentences in fluctuating noise at: a) SRT, b) SRT+5, c) SRT+10, d) noise only	Internal processing HI vs. NH	Pupil during listening: i) peak amplitude, ii) mean dilation,	**Hearing:** NH vs. HI **Stimulus:** a vs. b vs. c vs. d	Stimulus: i: a > b for HI and NH **1+)** ii: HI < NH between a vs. b	
46.	Hicks & Tharpe (2002) *see # 12) and # 31), **only exp. 1**	n=10 HI n=10 NH	HI: 8.1 NH: 7.11	mild-to-mod. SHL	Classroom environment	Own HAs during testing n=8: binaural HAs n=1: monaural HA n=1: no HA	Saliva samples (beginning and end of school day, 2 days) for cortisol concentration	**Hearing:** NH vs. HI	**1+)** cortisol levels: HI > NH for morning and afternoon samples	
47.	Oates et al., 2002	n=20 NH, n=20 HI	NH: 30.3, HI: 31.2	NH <= 15 dB HL, HI: PTA1: 0-24 dB, PTA2: 25-49 dB, PTA3: 50-74 dB, PTA4: 75-120 dB	/ba and /da at i) 65dB ppe SPL ii) 80 dB ppe SPL via sound field	Internal processing HI vs. NH	EEG: N2 and P3: a) amplitude, b) latency, RT to deviant stimuli	**Hearing:** NH vs. HI **Stimulus:** i) vs. ii)	**1+)** b (N2,P3): PTA1 and PTA2 > NH for i **1=)** a(N2,P3): PTA1 and PTA2 = NH for i **1+)** a(N2,P3): PTA3 and PTA4 < NH for i and ii **1+)** b (N2,P3): PTA3 > NH for ii **1-)** a (P3): PTA3 > NH for ii	
48.	Korczak et al. (2005)	n=20 NH n=14 HI	30.3 29.3	NH: <= 15 dB HL HI: mild-severe HL	/ba and /da at i) 65dB ppe SPL ii) 80 dB ppe SPL via sound field	Only for HI: a) unaided b) personal HAs	EEG: N2 and P3 and RTs to stimuli	**A:** HA vs. no HA **Stimulus:** i) vs. ii)	**1+)** P3 and RT latencies for i): HI > NH	**2+)** detectability of N2, P3, aided > unaided, for HI in i)
49.	Martin & Stapells (2005) *see # 34)	n=10 NH	33	<=20 dB HL	/ba/, /da/, monaural, at i) 65 and ii) 80 dB ppe SPL in a)quiet b)low pass filtered noise, fc = 250, 500, 1k, 2k, 4kHz,	HL simulation by masking noise	EEG measures for a) RT for deviant stimuli perception b)ignoring deviant stimuli while reading	**Stimulus:** fc effect **Stimulus:** quiet vs. noise	**1+)** N1 amplitude decreases when fc increases only for i) **1+)** N1 latency: (fc<1kHz) >(fc>1kHz) **1+)** P3 only present for fc <4kHz	

					at 50% masking threshold, ipsilateral		i)N1: speech audibility ii)P3: stimulus discrimination		**1+)** P3 amplitude: (fc<1kHz) > (fc>2kHz) **1+)** P3 latency: (fc<1kHz) > (fc>1kHz)	
50.	Wild et al. (2012)	n=21 NH	21	Self-report of NH	3 attentional conditions: a) sentences b) auditory distracter c) visual distracter 4 intelligibility conditions i) clear speech ii) 6-band NV iii)compressed 6-band NV iiii) spectrally rotated NV	noise vocoding	fMRI while decision making for a, b or c by key press	**Stimulus**: a vs. b vs. c **Stimulus**: i, ii, iii, iiii	**1+)** attentional task: LE: i < ii and iii	
51.	Zekveld et al. (2011), *see # 24)	n=28 NH n=36 HI n=38 YNH from 2010	NH: 55 HI: 61 NHY: 23	NH and YNH: <+20 dB HL HI: mild-mod. HL	Sentences in stationary noise, in free field, a)in quiet, b)50% intelligibility, c) 71%, d) 84%	Internal processing HI vs. NH	Pupil during listening: i) peak amplitude, ii) mean dilation, iii) peak latency, iiii) response duration	**Hearing:** NH vs. HI **Stimulus**: $SRT_{50\%}$ vs. $SRT_{71\%}$ vs. $SRT_{84\%}$ **Age:** 23 vs. 55,61 years	**1+)** decline with increasing SNR in i), ii): HI < NH, less strong for NHY Age: iiii): NHY < NH, HI and baseline ii): NHY > NH, HI	

Extended data for 41 articles arranged by subjective, behavioral or physiological measurement types in alphabetical order. Articles describing studies using multiple types of measurement appear in multiple rows.
vs.: versus; **NH:** normal hearing**, HI:** hearing-impaired, **UHI:** unilateral HI; **(S)HL:** (sensorieneural) hearing loss, **CI:** Cochlear Implant; **UCI:** unilateral CI; **PTA**: pure tone average**; MCL:** most comfortable level; **HA:** hearing aid; **UHA:** unilateral HA; **VRT**: verbal response time; **DNR:** digital noise reduction; **NR:** noise reduction; **HA:** hearing aid**; BTE:** behind-the-ear HA; **ITC:** in the channel HA; **OMNI**: omnidirectional; **DIR:** directional; **ViRT:** visual response/reaction time; **RT:** reaction time; **SRT:** speech recognition test, **LE:** listening effort; **PR:** preference ratings; **Exp.:** experiment; **SSN**: speech-shaped noise; **ppeSPL**: peak-to-peak equivalent SPL; **DTP:** dual-task paradigm; **S+N:** speech + noise; **AO:** auditory only; **AV:** auditory visual; **SSQ:** Speech Spatial and Qualities of hearing scale SSQ (Gatehouse & Noble, 2004); **NAL:** National Acoustic Laboratories prescription (Byrne & Dillon, 1986), **CVC:** revised consonant-vowel-consonant words (Peterson and Lehiste, 1962); **CBB:** Codebook-Based (Rosenkranz, 2010), **AMS:** Amplitude Modulation System-Based (Tchorz & Kollmeier, 2003), **MS:** minimum statistics (Martin et al., 2004); **MHA:** Master Hearing Aid research platform (Grimm et al. 2006);

TABLE 2: Summary of extracted evidence from studies providing findings on the effect of hearing impairment on listening effort (Q1) (n=21 studies, 41 findings).

Q1	Type of effects	Methods	Participants
Less Effort	**1 tests in total:** • 1 NH vs. HI	Physiological: 1 findings	NH: 20 HI: 20
Equal Effort	**11 tests in total:** • 10 NH vs. HI • 1 monaural vs. binaural	Subjective: 3 findings Behavioral: 7 findings Physiological: 1 findings	NH: 278 HI: 164
More Effort	**29 tests in total**: • 14 NH vs. HI • 4 different degrees of hearing loss • 11 hearing loss simulations	Subjective: 6 findings Behavioral: 10 findings Physiological: 13 findings	NH: 450 HI: 481

Summary of evidence proposing more, equal or less effort (from top to bottom) due to hearing impairment with respect to the effect types, the applied methods and the corresponding number of participants. NH: normal-hearing; HI: hearing-impaired; vs: versus;

TABLE 3: GRADE evidence profile for findings on Q1

GRADE evidence profile: Q1: Does hearing-impairment affect listening effort?										
Quality assessment						Summary of findings				Importance
						No of participants		Effect (Sign test)	Quality	
No of studies (Design)	Study limitations	Inconsistency	Indirectness	Imprecision	Publication bias	Hearing-impaired	Normal-hearing	$HP_!$: LE: NH < HI		
Subjective assessment by visual-analogue scales (1-10)										
7 (RCT)	Not serious[2,3]	Serious[7,8]	Serious[9]	Serious[10]	undetected	259	220	$p_1 = 0.25$	Very low	Important
Behavioral assessment by dual-task paradigms										
8 (RCT)	Not serious[2,3]	Serious[7,8,11]	Serious[9,12]	Not serious[10]	undetected	187	147	$p_1 = 0.61$	Low	Important
Behavioral assessment by reaction time measures										
1 (RCT)	Not serious[2,3]	Not serious	Not serious	Serious[10]	undetected	0	10	$p_1 = 0.13$	Moderate	Important
Physiological assessment by pupillometry										
2 (RCT)	Not serious[2,3]	Serious[13]	Serious[9]	Serious[10]	undetected	50	80	$p_1 = 0.25$	Very low	Important
Physiological assessment by EEG measures										
3 (RCT)	Not serious[2,3]	not serious[14]	not serious	Serious[10]	undetected	50	34	$p_1 = 0.03$	Moderate	Important

Possible levels of quality criteria: not serious, serious, very serious and undetected; Possible range of quality of evidence: high, moderate, low or very low;
RCT: randomized controlled trial with corresponding limitations factors 1-5;
1)Lack of allocation concealment; 2) Lack of experimental blinding; 3) Incomplete accounting of patients and outcome events failure to adhere to an intention to treat analysis (excluded participants, missing data); 4) Selective outcome reporting; 5) Stopping trial earlier for benefit; 7) Differences in target stimulus: single sentences vs. sentence passages vs. words vs. consonants; 8) Differences in masker types: speech shaped noise vs. 1-talker babble vs. 6-talker babble cafeteria noise vs. stationary noise; 9) Differences between populations: young normal-hearing vs. elderly hearing-impaired participants; 10) Power sufficiency rarely provided; 11) Dual-task paradigm vs. single task paradigms; 12) Differences in comparators to the intervention: normal-hearing vs. sensorineural hearing-impaired vs. simulated, conductive hearing-impairment; 13) Differences in test setup: speech reception threshold at different levels; 14) Same stimulus and levels used for all three studies but in two studies presentation via sound field while in one via headphones;

TABLE 4: Summary of extracted evidence from studies providing findings on the effect of hearing aid amplification on listening effort (Q2) (n=27 studies, 56 findings).

Q2	Type of effects	Methods	Participants
Less Effort	**32 tests in total:** • 12 HA vs. none • 10 unprocessed vs. processed stimuli • 6 comparison of processing types • 4 algorithms on- vs. off	Subjective: 17 findings Behavioral: 14 findings Physiological: 1 findings	NH: 282 HI: 761
Equal Effort	**19 tests in total:** • 6 comparison of signal processing algorithms • 5 signal processing algorithms on vs. off • 5 HA vs. none • 3 unprocessed vs. processed stimuli	Subjective: 13 findings Behavioral: 6 findings	NH: 112 HI: 289
More Effort	**5 tests in total:** • 2 signal processing algorithm on vs. off • 3 comparison of signal processing algorithms	Subjective: 3 findings Behavioral: 2 findings	NH: 63 HI: 143

Summary of evidence proposing more, equal or less effort (from top to bottom) due to hearing-aid amplification with respect to the effect types, the applied methods and the corresponding number of participants. NH: normal-hearing; HI: hearing-impaired; HA: hearing-aid; vs.: versus;

TABLE 5: GRADE evidence profile for findings on Q2

GRADE evidence profile: Q2: Does hearing-aid amplification reduce listening effort?										
Quality assessment						Summary of findings				Importance
						No of participants		Effect (Sign test)	Quality	
No of studies (Design)	Study limitations	Inconsistency	Indirectness	Imprecision	Publication bias	Hearing-impaired	Normal-hearing	HP_2 : LE: HA < none		
Subjective assessment by visual-analogue scales (1-10)										
16 (RCT)	Not serious[2,3]	Serious[7,8,14]	Serious[9]	Serious[10]	undetected	419	127	$p_1 = 0.50$	Very low	important
Subjective assessment by the Speech, Spatial, and Qualities of Hearing Scale										
3 (RCT) 1 (OS)	Not serious[2,3] Not serious[6]	Serious[15]	Serious[12]	Serious[10]	undetected	638	21	$p_1 = 0.64$	Very low	important
Behavioral assessment by dual-task paradigms										
10 (RCT)	Not serious[2,3]	Serious[7,8,14]	Serious[9]	Serious[10]	undetected	184	108	$p_1 = 0.41$	Very low	important
Behavioral assessment by reaction time measures										
3 (RCT)	Not serious[2,3]	Serious[7,8]	Serious[9]	Serious[10]	undetected	52	30	$p_1 = 0.06$	Very low	important

Possible levels of quality criteria: not serious, serious, very serious and undetected; Possible range of quality of evidence: high, moderate, low or very low;
RCT: randomized controlled trial with corresponding limitations factors 1-5;
1)Lack of allocation concealment; 2) Lack of experimental blinding; 3) Incomplete accounting of patients and outcome events failure to adhere to an intention to treat analysis (excluded participants, missing data); 4) Selective outcome reporting; 5) Stopping trial earlier for benefit;
OS: non-randomized observational studies have the following limitation factors: 6a) failure to develop and apply appropriate eligibility criteria (inclusion of control population), 6b) flawed measurement of both exposure and outcome (differences in measured exposure),6c) failure to adequately control confounding (adjust analysis) and 6d) incomplete or inadequately short follow-up (follow all groups for same amount of time);
7) Differences in target stimulus: single sentences, sentence passages, words or consonants; 8) Differences in masker types: speech shaped noise, 1-talker babble, 6-talker babble, cafeteria noise, church environment or stationary noise; 9) Differences between populations: young normal-hearing vs. elderly hearing-impaired participants or experienced hearing-aid users vs. hearing-impaired listeners without hearing-aid experience; 10) Power sufficiency rarely provided; 11) Dual-task paradigm vs. single task paradigms; 12) Differences in comparators to the intervention: normal-hearing, sensorineural hearing-impaired or simulated, conductive hearing-impairment, unilateral vs. bilateral hearing-aid use, unilateral CI use vs. bilateral CI use or post-or pre-CI fitting, ; 13) Differences in test setup: speech reception threshold at different levels; 14) Differences in treatment: noise reduction, compression or linear amplification; 15) Differences of test environment: daily life vs. multi-talker babble or cafeteria noise; 16) Differences in test setup: verbal reaction times, button press reaction times, stimulus in sound-field vs. headphones;

TABLE 6: Factors determining the quality of evidence according to the GRADE handbook, chapter 5 (Schünemann et al. 2013).

Factors that can reduce the quality of the evidence	Consequence
Limitations in study design or execution (risk of bias)	lower 1 or 2 levels
Inconsistency of results	lower 1 or 2 levels
Indirectness of evidence	lower 1 or 2 levels
Imprecision	lower 1 or 2 levels
Publication bias	lower 1 or 2 levels
Factors that can increase the quality of the evidence	Consequence
Large magnitude of effect	increase 1 or 2 levels
All plausible confounding would reduce the demonstrated effect or increase the effect if no effect was observed	increase 1 level
Dose-response gradient	increase 1 level