OPEN

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

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

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 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 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 keywords: hearing aids OR hearing impairment AND listening effort OR perceptual effort OR ease of listening. Of these, 41 articles fulfilled the Population, Intervention, Control, Outcomes, and Study design 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 seven 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 Grading of Recommendations Assessment, Development, and Evaluation Working Group guidelines. We tested the statistical evidence across studies with nonparametric tests. The testing revealed only one consistent effect across studies, namely that listening effort was higher for hearing-impaired listeners compared with normal-hearing listeners (Q1) as measured by electroencephalographic 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 scientific, 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.

Key words: Behavioral measures, Hearing aid amplification, Hearing impairment, Listening effort, Physiologic measures, Quality rating, Speech comprehension, Subjective ratings.

(Ear & Hearing 2017;38;267–281)

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). In addition, in daily life, speech is often heard among 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 with 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-Cunningham & Best 2008). As a result, hearing-impaired listeners expend extra effort to achieve successful speech perception (McCoy et al. 2005; Rönnberg et al. 2013). 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 et al. 1997, 2006), lack of energy and stress-related sick leave from work (Gatehouse & Gordon 1990; Kramer et al. 2006; Edwards 2007; Hornsby 2013a, b). Nachtegaal et al. (2009) found a positive association between hearing thresholds and the need for recovery after a working day. In addition, hearing impairment can dramatically alter peoples' social interactions and quality of life due to withdrawal from leisure and social roles (Weinstein & Ventry 1982; Demorest & Erdman 1986; Strawbridge et al. 2000), and one reason for this may

[1]Section Ear & Hearing, Department of Otolaryngology-Head and Neck Surgery, VU University Medical Center and Amsterdam Public Health Research Institute, Amsterdam, The Netherlands; [2]Eriksholm Research Centre, Oticon A/S, Snekkersten, Denmark; [3]Department of Behavioral Sciences and Learning, Linköping University, Linköping, Sweden; [4]Linnaeus Centre HEAD, The Swedish Institute for Disability Research, Linköping and Örebro Universities, Linköping, Sweden; [5]Medical Library, VU University Amsterdam, Amsterdam, The Netherlands; [6]MRC/CSO Institute of Hearing Research, Scottish Section, Glasgow, United Kingdom; [7]Institute of Physiology and Pathology of Hearing, International Center of Hearing and Speech, Warsaw, Poland; and [8]Department of Clinical and Experimental Medicine, Linköping, University, Sweden.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and text of this article on the journal's Web site (www.ear-hearing.com).

<zdoi; 10.1097/AUD.0000000000000396>

be the increased effort required for successful listening. There is growing interest among 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, and SDC Table 1, http://links.lww.com/EANDH/A335). 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 (DTP) (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 2013, 2014), memory tasks (Feuerstein 1992; Hornsby 2013), tactile pattern recognition tasks (Gosselin & Gagné 2011), or even driving a vehicle in a simulator (Wu et al. 2014). The concept of DTPs 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 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 (Bernarding et al. 2012; Obleser 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 (Hick & Tharpe 2002).

Hearing aids are typically used to correct for the loss of audibility introduced by hearing impairment (Hick & 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. 2014a). If hearing impairment indeed increases listening effort, as suggested by previous research (Feuerstein 1992; Hick & 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).

MATERIALS AND 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 the Supplemental Digital Content Appendix search terms, http://links.lww.com/EANDH/A323). English language articles were identified from inception to August 2014.

Inclusion and Exclusion

The Population, Intervention, Control, Outcomes, and Study design (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. Population: Hearing-impaired participants and normal-hearing listeners with a simulated hearing loss (for example by applying a low-pass filter to the auditory stimuli).
II. Intervention: Hearing impairment or hearing aid amplification (including cochlear implant [CI]), 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

TABLE 1. Summary of the 41 included articles

Publication	Method Used	Author Hypothesis (HP) 1) listening effort: HI > NH; +: HP Supported, –: HP Not Supported, =: No Effect	Author Hypothesis (HP) 2) listening effort: Aided < Unaided; +: HP Supported; –: HP Not Supported; =: No Effect
Subject Measures			
1. Ahlstrom et al. (2014)	VAS (0–15)	1+)	2+)
2. Bentler and Duve (2000)	VAS (1–10)		2=)
3. Bentler et al. (2008)	VAS (1–10)		2+)
4. Brons et al. (2013)	VAS (9-1)		2x 2=)
5. Brons et al. (2014)	VAS (1–9)		2–)
6. Desjardins and Doherty (2013), *see # 25)	VAS (100-0)	1+); 1=)	
7. Desjardins and Doherty (2014), *see # 26)	VAS (100-0)		2=)
8. Dwyer et al. (2014)	SSQ (1–10)	1+)	3 x 2=); 1x 2+)
9. Feuerstein (1992), *see # 28)	VAS (100-0)	1+)	
10. Hällgren et al. (2005)	VAS (0–10)		2+)
11. Harlander et al. (2012)	VAS (1–13)		2-); 2x 2+)
12. Hicks and Tharpe (2002), *see # 31) # 46), only exp. 2	VAS (1–5)	1=)	
13. Hornsby (2013),*see # 32)	SSQ (0–10) questions # 14, # 18, # 19		2=)
14. Humes et al. (1997)	VAS (100-0)	1+)	2=)
15. Humes et al. (1999)	VAS (0–100)		2+)
16. Luts et al. (2010)	VAS (0–6)	1+)	2x 2+); 1x 2-)
17. Mackersie et al. (2009)	VAS (9-1)		2+); 2=)
18. Neher et al. (2014), *see # 36)	VAS(1–9)		3x 2+); 1x2=)
19. Noble and Gatehouse (2006)	SSQ (1–10)		2+)
20. Noble et al. (2008)	SSQ-50		2x 2+)
21. Palmer et al. (2006)	VAS (completely agree - disagree)		2=)
22. Pals et al. (2013), *see # 37)	VAS(0–100)		2+)
23. Rudner et al. (2012); only exp. 2	VAS (no effort - maximum possible effort)		2=)
24. Zekveld et al. (2011), *see # 51)	VAS (0–10)	1=)	
Behavioral measures			
25. Desjardins and Doherty (2013), *see # 6)	DTP	1x 1+); 2x1=)	
26. Desjardins and Doherty (2014), *see # 7)	DTP		2x 2+); 1x 2=)
27. Downs (1982)	DTP		2+)
28. Feuerstein (1992), *see# 9)	DTP	1=);1+)	
29. Gatehouse and Gordon (1990)	RT for response to all stimulus		2+)
30. Gustafson et al. (2014)	Verbal RTs for nonword repetition		2+)
31. Hicks and Tharpe (2002), *see #12) #46), only exp. 2	DTP	1=); 1+)	
32. Hornsby (2013), *see # 13)	DTP		2+); 2=)
33. Kulkarni et al. (2012)	Exp. 1 and 2: RT for stimulus		Exp.1: 2+); Exp.2: 2+)
34. Martin and Stapells (2005); *see # 49)	RTs during discrimination of deviant stimuli	3x 1+)	
35. Neher et al. (2013)	DTP		2–)
36. Neher et al. (2014); *see # 18)	DTP	1=)	2–); 2+)
37. Pals et al. (2013); *see # 22)	DTP		2+)
38. Picou et al. (2013)	DTP		2+)
39. Picou et al. (2014)	DTP		2=)
40. Rakerd et al. (1996)	Exp. 1 and 2: DTP	Exp.1 and 2: 2x 1+)	
41. Sarampalis et al. (2009); only exp. 2	DTP		2=); 2+)
42. Stelmachowicz et al. (2007)	DTP	1=)	
43. Tun et al. (2009)	DTP	1+); 1=)	

(Continued)

TABLE 1. Continued.

Publication	Method Used	Author Hypothesis (HP) 1) listening effort: HI > NH; +: HP Supported, –: HP Not Supported, =: No Effect	Author Hypothesis (HP) 2) listening effort: Aided < Unaided; +: HP Supported; –: HP Not Supported; =: No Effect
44. Wu et al. (2014)	Exp. 1: DTP (sentence recall, driving vehicle in simulator); Exp. 2 and 3: DTP: (sentence recall, ViRT)	Exp. 3: 1+)	Exp. 1: 2=); Exp 2: 2=); Exp. 3: 2x 2+)
Physiological measures			
45. Kramer et al. (1997)	Pupil during listening: peak amplitude, mean dilation;	1+)	
46. Hicks and Tharpe (2002) *see # 12) # 31), only exp. 1	Saliva samples for cortisol concentration	1+)	
47. Oates et al. (2002)	EEG: N2 and P3	3x 1+); 1x 1=); 1x1-)	
48. Korczak et al. (2005)	EEG: N2 and P3 and RTs to stimuli	1+)	2+)
49. Martin and Stapells (2005); *see # 34)	EEG: RT for deviant stimuli; N1, P3	5x 1+)	
50. Wild et al. (2012)	fMRI while decision making	1+)	
51. Zekveld et al. (2011), *see # 24)	Pupil during listening: peak, mean amplitude and latency	1+)	

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.

NH, normal hearing; HI, hearing-impaired; HP, hypothesis; VAS, visual-analogue scale; ViRT, visual response/reaction time; RT, reaction time; SRT, speech recognition test; LE, listening effort; Exp., experiment; DTP, dual-task paradigm; SSQ, Speech Spatial and Qualities of hearing scale SSQ (Gatehouse & Noble, 2004).

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. Control: Group comparisons (e.g., normal hearing versus 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. Outcomes: Listening effort, as assessed by (1) subjective measures of daily life experiences, handicap reduction, benefit, or satisfaction, (2) behavioral measures of changes in auditory tasks performance, or (3) physiological measures corresponding to higher cognitive processing load, such as N2 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, and P2a were not considered as reflecting measures of listening effort.

V. Study 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 (B.O., A.Z., and S.K.).

Data Extraction and Management

For each relevant study, the outcome measures applied to assess listening effort were extracted and categorized into

TABLE 2. 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

GRADE, Grading of Recommendations Assessment, Development and Evaluation.

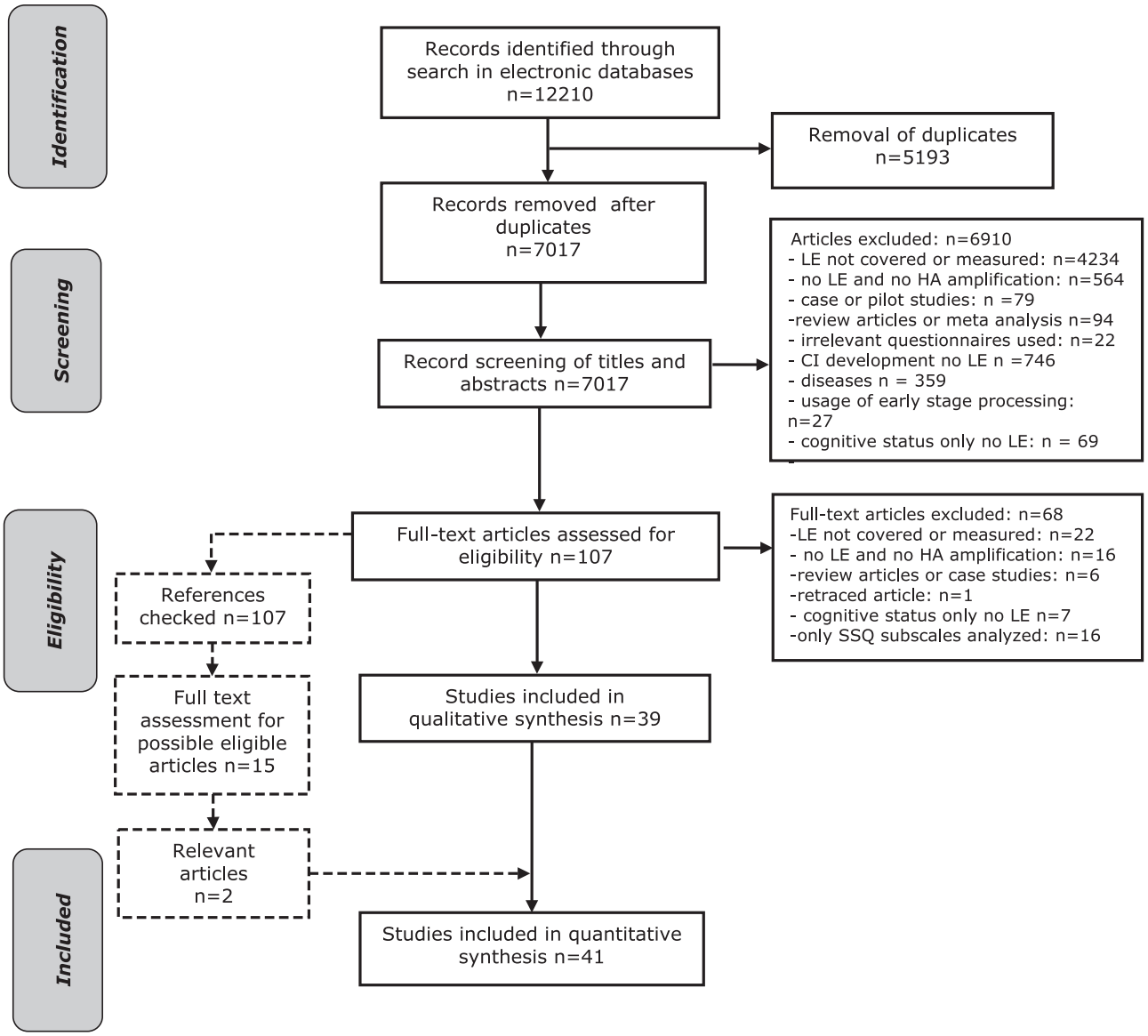

Fig. 1. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram of the study identification, the screening, the eligibility, and the inclusion process within the systematic search.

subjective, objective, or physiological indicators of listening effort. We identified and extracted the findings addressing Q1 or Q2 from all relevant studies. The results of each study were evaluated with respect to the two hypotheses (HP1 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 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 Grading of Recommendations Assessment, Development, and Evaluation (GRADE) Working

TABLE 3. 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	Number of 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 due to hearing impairment with respect to the effect types, the applied methods and the corresponding number of participants.
HI, hearing impaired; NH, normal hearing.

Group guidelines (Guyatt et al. 2011). The quality of evidence is rated for each measurement type (Tables 4, 6) 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 subcriteria (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 subcriteria were fulfilled across studies, "serious" if less than half of the subcriteria were fulfilled, and "very serious" if none of the subcriteria 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, or consistent factors of interest with respect to Q1 or Q2, "serious inconsistency" was judged for evidence on that measurement type. The quality criterion "indirectness" was related to differences between tested populations 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 when normal-hearing listeners were compared with 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 (Tables 4, 6). 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 2). 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 4 on Q1, Table 6 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, and SDC Table 1, http://links.lww.com/EANDH/A335) corresponding to each measurement type for findings addressing HP1 and 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 with unaided listening (HP2), that is, 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 "Importance" column in Tables 3, 5) (Schünemann et al. 2013).

Two authors (B.O. and T.L.) 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

TABLE 4. GRADE evidence profile for findings on Q1

GRADE Evidence Profile: Q1: Does Hearing Impairment Affect Listening Effort?										
						Summary of Findings				
Quality Assessment						No of Participants		Effect (Sign Test)		
No of Studies (Design)	Study Limitations	Inconsistency	Indirectness	Imprecision	Publication Bias	Hearing Impaired	Normal Hearing	HP_1: LE: NH < HI	Quality	Importance
Subjective assessment by visual analog scales (1–10)										
7 (RCT)	Not serious[2,3]	Serious[4,5]	Serious[6]	Serious[7]	Undetected	259	220	$p_1 = 0.25$	Very low	Important
Behavioral assessment by dual-task paradigms										
8 (RCT)	Not serious[2,3]	Serious[4,5,8]	Serious[6,9]	Not serious[7]	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[7]	Undetected	0	10	$p_1 = 0.13$	Moderate	Important
Physiological assessment by pupillometry										
2 (RCT)	Not serious[2,3]	Serious[10]	Serious[6]	Serious[7]	Undetected	50	80	$p_1 = 0.25$	Very low	Important
Physiological assessment by EEG measures										
3 (RCT)	Not serious[2,3]	Not serious[11]	Not serious	Serious[7]	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 with corresponding limitations factors: 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) Differences in target stimulus: single sentences vs. sentence passages vs. words vs. consonants; 5) Differences in masker types: speech shaped noise vs. 1-talker babble vs. 6-talker babble cafeteria noise vs. stationary noise; 6) Differences between populations: young normal-hearing vs. elderly hearing-impaired participants; 7) Power sufficiency rarely provided; 8) Dual-task paradigm vs. single task paradigms; 9) Differences in comparators to the intervention: normal-hearing vs. sensorineural hearing-impaired vs. simulated, conductive hearing-impairment; 10) Differences in test setup: speech reception threshold at different levels; 11) Same stimulus and levels used for all three studies but in two studies presentation via sound field while in one via headphones. GRADE, Grading of Recommendations Assessment, Development and Evaluation; HI, hearing impaired; NH, normal hearing; RCT, randomized controlled trial.

excluded, and the reasons for their exclusion. The main electronic database search produced a total of 12,210 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 CIs (n = 746) or the treatment of diseases (n = 359) and neither of the 2 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.

TABLE 5. 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	Number of 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 due to hearing aid amplification with respect to the effect types, the applied methods and the corresponding number of participants.
HA, hearing aid; HI, hearing impaired; NH, normal hearing.

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 7 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 or binaural hearing loss and the degree of hearing impairment varied. Some studies examined experienced hearing aid users, whereas participants of other studies included nonusers 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 (Hedley-Williams et al. 1997; Tun et al. 2009; Desjardins & Doherty 2013). 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 analog 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). 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; Hornsby 2013; Picou et al. 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 (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. 2014b; Picou et al. 2014; Wu et al. 2014), 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 Tables 3 and 5). Of the 42 findings based on subjective assessment or rating of listening effort, 31 findings resulted from VAS (Table 1, and SDC Table 1, http://links.lww.com/EANDH/A335). Such effort rating scales ranged for example from 0 to 10, indicating conditions of "no effort" to "very high effort" (Hällgren et al. 2005; Zekveld et al. 2011). The remaining 11 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 DTP and 7 findings resulted from reaction time measures. The 16 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 (Hick & 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 or a randomized controlled design were included. A between-group design (normally hearing versus hearing impaired) was applied in 17 studies (Rakerd et al. 1996; Kramer et al. 1997; Humes et al. 1997; Humes 1999; Hick & 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. 2014a, b; 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, and Supplemental Digital Content Table 1, http://links.lww.com/EANDH/A335, 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, and Supplemental Digital Content Table 1, http://links.lww.com/EANDH/A335. 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, Supplemental Digital Content Table 1, http://links.lww.com/EANDH/A335, and 3, respectively, for detailed and summarized tabulations of the results described in this section.

Subjective Measures, Q1 • Six findings (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 (see Table 3: more effort). 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 (of n = 17 in total) indicated higher levels of listening effort for groups with hearing impairment compared with groups with normal hearing (see Table 3: more effort). Findings from DTPs were mainly (n = 6 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 seven 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 (see Table 3: more effort). The applied methods varied between measures of EEG (n = 9 findings), pupil dilation (n = 2 findings), diurnal cortisol levels (n = 1 finding), and

TABLE 6. GRADE evidence profile for findings on Q2

GRADE Evidence Profile: Q2: Does Hearing Aid Amplification Reduce Listening Effort?										
						Summary of Findings				
Quality Assessment						No of Participants		Effect (Sign Test)		
No of Studies (Design)	Study Limitations	Inconsistency	Indirectness	Imprecision	Publication Bias	Hearing Impaired	Normal Hearing	HP_2: LE: HA < none	Quality	Importance
Subjective assessment by visual analog scales (1–10)										
16 (RCT)	Not serious[1,3]	Serious[5,6,10]	Serious[7]	Serious[8]	Undetected	419	127	$p_2 = 0.50$	Very low	Important
Subjective assessment by the speech, spatial, and qualities of hearing scale										
3 (RCT) 1 (OS)	Not serious[1,3] Not serious[4]	Serious[10]	Serious[9]	Serious[8]	Undetected	638	21	$p_2 = 0.64$	Very low	Important
Behavioral assessment by dual-task paradigms										
10 (RCT)	Not serious[1,3]	Serious[5,6,10]	Serious[7]	Serious[8]	Undetected	184	108	$p_2 = 0.41$	Very low	Important
Behavioral assessment by reaction time measures										
3 (RCT)	Not serious[1,3]	Serious[7,6]	Serious[7]	Serious[8]	Undetected	52	30	$p_2 = 0.06$	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. Randomized controlled trials (RTC) and nonrandomized observational studies (OS) with corresponding limitations factors: 1) 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) failure to develop and apply appropriate eligibility criteria (inclusion of control population), flawed measurement of both exposure and outcome (differences in measured exposure), failure to adequately control confounding (adjust analysis) and incomplete or inadequately short follow-up (follow all groups for same amount of time); 5) Differences in target stimulus: single sentences, sentence passages, words or consonants; 6) Differences in masker types: speech shaped noise, 1-talker babble, 6-talker babble, cafeteria noise, church environment or stationary noise; 7) Differences between populations: young normal-hearing vs. elderly hearing-impaired participants or experienced hearing-aid users vs. hearing-impaired listeners without hearing-aid experience; 8) Power sufficiency rarely provided; 9) 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; 10) Differences of test environment: daily life vs. multi-talker babble or cafeteria noise.

HA, hearing aid; GRADE, Grading of Recommendations Assessment, Development and Evaluation; RCT, randomized controlled trial.

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 4. 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 (RCTs). 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 4, 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 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 4, based on seven RCTs. We found the quality criterion "study limitations" (Table 4) "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 were 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 SDC Table 1, http://links.lww.com/EANDH/A335, and we applied a binomial test (Sign test), which resulted in a p value of 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 with 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 (Table 4). 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 with 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 (Table 4). The first was pupillometry. Two RCTs 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 SDC Table 1, http://links.lww.com/EANDH/A335, 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 5 as well as Supplemental Digital Content Table 1, http://links.lww.com/EANDH/A335 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 13 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. Seventeen 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 (Table 6). We judged that evidence based on a single physiological finding provides too little information to create a separate row in Table 6. 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 2 measurement types, including 16 studies using VAS ratings and four studies that applied SSQ (Table 6). 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 6 (out of n = 16 in total) of the studies provided sufficient power, which caused "serious imprecision." We counted the "+," "=," and "–" signs in Table 1, and SDC Table 1, http://links.lww.com/EANDH/A335, for subjective findings for VAS on Q2 and applied the Sign test, which revealed a p value of 0.50, meaning that evidence from VAS across studies did not show higher listening effort ratings for hearing aid amplification compared with unaided listening.

The second measurement type on subjective assessment resulted from SSQ data. We found RCTs (Table 6) in three studies. One study (Dwyer et al. 2014) was an observational study where different groups of participants rated their daily life experience with hearing impairment, CI, 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 6). 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 11 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; Bentler & Duve 2000; Desjardins & Doherty 2014) or physiological measures of listening effort (Hick & 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 with 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. 2014b), 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 with 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 with 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 4) 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 with 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 et al. (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 2 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 that 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, and 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. Nonrandomized observational studies were not considered flawed as compared with randomized control trial studies as GRADE accounts for the design of the assessed studies and different subcriteria are applied to evaluate the criterion called "study limitations" (Table 6). Within this review findings from only one nonrandomized 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? Among 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, among 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, 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 FP7-LISTEN607373).

The authors have no conflicts of interest to disclose.

Address for correspondence: Adriana A. Zekveld, Section Ear & Hearing, Dept. of Otolaryngology-Head and Neck Surgery, VU University Medical Center and Amsterdam Public Health Research Institute, Amsterdam, The Netherlands; P.O. Box 7057, 1007 MB, Phone: +31 20 4440952; Fax: +3120 444 2033.

Received October 28, 2015; accepted October 20, 2016.

REFERENCES

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

Armstrong, E. C. (1999). The well-built clinical question: The key to finding the best evidence efficiently. *WMJ*, *98*, 25–28.

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

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

Bentler, R., Wu, Y. H., Kettel, J., et al. (2008). Digital noise reduction: Outcomes from laboratory and field studies. *Int J Audiol*, *47*, 447–460.

Bernarding, C., Strauss, D. J., Hannemann, R., et al. (2013). Neural correlates of listening effort related factors: Influence of age and hearing impairment. *Brain Res Bull*, *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*, 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*, 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 Hear Res*, *29*, 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*, 600–610.

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

Downs, D. W. (1982). Effects of hearing and use on speech discrimination and listening effort. *J Speech Hear Disord*, *47*, 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*, 126–136.

Ebell, M. (1999). Information at the point of care: Answering clinical questions. *J Am Board Fam Pract*, *12*, 225–235.

Edwards, B. (2007). The future of 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*, 80–86.

Fraser, S., Gagné, J. P., Alepins, M., et al. (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*, 63–68.

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.

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

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., et al. (2014). Listening effort and perceived clarity for normal-hearing children with the use of digital noise reduction. *Ear Hear*, *35*, 183–194.

Guyatt, G. H., Oxman, A. D., Kunz, R., et al.; GRADE Working Group. (2008). What is "quality of evidence" and why is it important to clinicians? *BMJ*, *336*, 995–998.

Guyatt, G. H., Oxman, A. D., Schünemann, H. J., et al. (2011). GRADE guidelines: A new series of articles in the Journal of Clinical Epidemiology. *J Clin Epidemiol*, *64*, 380–382.

Hagerman, B. (1984). Clinical measurements of speech reception threshold in noise. *Scand Audiol*, *13*, 57–63.

Hällgren, M., Larsby, B., Lyxell, B., et al. (2005). Speech understanding in quiet and noise, with and without hearing aids. *Int J Audiol*, *44*, 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*, 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.

Hick, 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*, 573–584.

Hopkins, K., 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*, 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*, 523–534.

Howard C. 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, L. E. (1999). Dimensions of hearing aid outcome. *J Am Acad Audiol*, *10*, 26–39.

Humes, L. E., & Humes, L. E. (2004). Factors affecting long-term hearing aid success. In *Seminars in Hearing* (Vol. *25*, pp. 63–72). New York, NY: *Thieme Medical Publishers, Inc*.

Humes L. E.., & Roberts L. (1990). Speech-recognition difficulties of the hearing impaired elderly: The contributions of audibility. *J Speech Hear Res*, *33*, 726–735.

Humes, L. E., Christensen, L. A., Bess, F. H., et al. (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.

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

Koelewijn T, Zekveld, A., Festen J. M., Kramer S. 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*, 165–185.

Kramer S. E., Kapteyn T. S., Festen J. M., Kuik D. J. (1997). Assessing aspects of auditory handicap by means of pupil dilatation. *Audiology*, *36*, 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*, 503–512.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I. Jr, et al. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, *50*, 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 Commun*, *54*, 341–350.

Laeng, B., Sirois, S., Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspect Psychol Sci*, *7*, 18–27.

Larsby, B., Hällgren, M., Lyxell, B., et al. (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*, 131–143.

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

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

Mackersie, C. L., MacPhee, I. X., Heldt, E. W. (2015). Effects of hearing loss on heart rate variability and skin conductance measured during sentence recognition in noise. *Ear Hear*, *36*, 145–154.

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

McCoy, S. L., Tun, P. A., Cox, L. C., et al. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *Q J Exp Psychol A*, *58*, 22–33.

McGarrigle, R., Munro, K. J., Dawes, P., et al. (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*, 433–440.

Moher, D., Liberati, A., Tetzlaff, J., et al.; PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Ann Intern Med*, *151*, 264–9, W64.

Neher, T., Grimm, G., Hohmann, V., et al. (2014a). 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*, 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*, 172–181.

Noble, W., Tyler, R., Dunn, C., et al. (2008). Unilateral and bilateral cochlear implants and the implant-plus-hearing-aid profile: Comparing self-assessed and measured abilities. *Int J Audiol*, *47*, 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*, 399–415.

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., Maess, B. (2012). Adverse listening conditions and memory load drive a common alpha oscillatory network. *J Neurosci*, *32*, 12376–12383.

Pals, C., Sarampalis, A., Baskent, D. (2013). Listening effort with cochlear implant simulations. *J Speech Lang Hear Res*, *56*, 1075–1084.

Petersen E. B., Wöstmann, M., Obleser, J., Stenfelt, S., Lunner, T. (2015). Hearing loss impacts neural alpha oscillations under adverse listening conditions. *Front Psychol*, *6*.

Picou, E. M., Ricketts, T. A., Hornsby, B. W. (2013). How hearing aids, background noise, and visual cues influence objective listening effort. *Ear Hear*, *34*, e52–e64.

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*, 339–352.

Plomp, R. (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*, 97–106.

Richardson, W. S., Wilson, M. C., Nishikawa, J., et al. (1995). The well-built clinical question: A key to evidence-based decisions. *ACP J Club*, *123*, A12–A13.

Rönnberg, J., Lunner, T., Zekveld, A., et al. (2013). The ease of language understanding (ELU) model: Theoretical, empirical, and clinical advances. *Front Syst Neurosci*, *7*, 31.

Sarampalis, A., Kalluri, S., Edwards, B., et al. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *J Speech Lang Hear Res*, *52*, 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*. Available at: http://gdt.guidelinedevelopment.org/app/handbook/handbook.html.

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*, e0117611.

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

Stephens, D., & Hétu, R. (1991). Impairment, disability and handicap in audiology: Towards a consensus. *Audiology*, *30*, 185–200.

Strawbridge, W., Wallhagen, M. I., Shema, S. J., Kaplan, G. A. (2000). Negative consequences of hearing impairment in old age: A longitudinal analysis. *Gerontologist*, *40*, 320–326.

Tun, P. A., McCoy, S. L., Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful listening. *Psychol Aging*, *24*, 761–766.

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

Wild, C. J., Yusuf, A., Wilson, D. E., et al. (2012). Effortful listening: The processing of degraded speech depends critically on attention. *J Neurosci*, *32*, 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*, e153–e165.

Wu, Y. H., Aksan, N., Rizzo, M., et al. (2014). Measuring listening effort: Driving simulator versus simple dual-task paradigm. *Ear Hear*, *35*, 623–632.

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

Zekveld, A. A., Kramer, S. E., Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear 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.