Published in final edited form as:

Psychophysiology. 2014 October ; 51(10): 1046–1057. doi:10.1111/psyp.12242.

Speech-perception training for older adults with hearing loss impacts word recognition and effort

Stefanie E. Kuchinsky*, Jayne B. Ahlstrom, Stephanie L. Cute, Larry E. Humes, Judy R. Dubno, and Mark A. Eckert*

Abstract

The current pupillometry study examined the impact of speech-perception training on word recognition and cognitive effort in older adults with hearing loss. Trainees identified more words at the follow-up than at the baseline session. Training also resulted in an overall larger and faster peaking pupillary response, even when controlling for performance and reaction time. Perceptual and cognitive capacities affected the peak amplitude of the pupil response across participants but did not diminish the impact of training on the other pupil metrics. Thus, we demonstrated that pupillometry can be used to characterize training-related and individual differences in effort during a challenging listening task. Importantly, the results indicate that speech-perception training not only affects overall word recognition, but also a physiological metric of cognitive effort, which has the potential to be a biomarker of hearing loss intervention outcome.

With increasing US life expectancy, the overwhelming majority of Americans may expect to experience hearing loss in their lifetime (Lin, Thorpe, Gordon-Salant, & Ferrucci, 2011). While the need for effective hearing loss interventions is clear, the optimal method for assessing their success is not (Saunders, Chisolm, & Abrams, 2005). Rather, research has shown that hearing aid benefit is better characterized through multiple outcome metrics, which aim to capture not only how accurately individuals perceive speech, but also how hard they had to work to achieve that level of performance (Humes, 1999). This evaluation of and response to the *demands* of a task (Navon & Gopher, 1979) is often termed *cognitive effort* (Kahneman, 1973). In the current study, we examined the extent to which metrics of word recognition, reaction time (RT), and a physiological indicator of effort uniquely contributed to assessments of benefit following speech-perception training.

Intervention outcomes typically include baseline and follow-up measures of speech recognition performance (Humes, Burk, Strauser, & Kinney, 2009; Metselaar et al., 2008; Miller et al., 2007). However, clinicians and researchers have long noted that the impact of hearing loss is not solely a decline in the number of words that can be correctly recognized. Individuals with hearing loss also report increased effort and fatigue, particularly when listening in noise (Héту, Riverin, Lalande, Getty, & St-Cyr, 1988; Hicks & Tharpe, 2002; McCoy et al., 2005; Nachttegaal et al., 2009; Pichora-Fuller, 2006), even when performance is equivalent (Bologna, Chatterjee, & Dubno, 2013; Hicks & Tharpe, 2002; McCoy et al.,

*Address correspondence to: Stefanie E. Kuchinsky (skuchins@umd.edu) or Mark A. Eckert (eckert@musc.edu), Department of Otolaryngology-Head and Neck Surgery, Medical University of South Carolina, 135 Rutledge Ave., MSC 550, Charleston, SC 29425-5500.

2005). Though not standards of clinical care, research evaluations of interventions have employed subjective measures of effort via self-assessment scales (Ahlstrom, Horwitz, & Dubno, 2014; Cox & Alexander, 1992; Humes, Garner, Wilson, & Barlow, 2001; Sweetow & Sabes, 2006). For example, Ahlstrom et al. found that even for conditions in which speech perception in noise did not improve, new hearing aid users reported experiencing less effort while listening with hearing aids that improve speech audibility compared to unaided listening.

Recently, researchers have sought to identify objective markers of the cognitive demands experienced in different listening conditions, in part to avoid biases that may arise from subjective reporting (Anderson Gosselin & Gagné, 2011; Fraser, Gagné, Alepins, & Dubois, 2010; Koelewijn, Zekveld, Festen, & Kramer, 2011; Zekveld, Kramer, & Festen, 2010, 2011). For example, aided listening has been shown to improve performance and reduce RTs on tasks that require individuals to divide their attention, suggesting that hearing aid use lessens the attentional demands associated with processing unaided speech, thus mitigating effort (Downs, 1982; Gatehouse & Gordon, 1990; Hällgren, Larsby, Lyxell, & Arlinger, 2005; Hornsby, 2013; Picou, Ricketts, & Hornsby, 2013).

Pupil dilation appears to be a particularly appropriate objective measure of effort because it can be collected throughout virtually any task of interest, is not under cognitive control, and is related to a well-studied neural system (the locus coeruleus-norepinephrine system; Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010; Rajkowski, Kubiak, & Aston-Jones, 1993). Indeed, increasing task-evoked pupil dilation has been shown to relate to increasing difficulty across a variety of studies (Cabestrero, Crespo, & Quirós, 2009; Engelhardt, Ferreira, & Patsenko, 2010; Kahneman & Beatty, 1966; Kahneman, Tursky, Shapiro, & Crider, 1969; Laeng, Ørbo, Holmlund, & Miozzo, 2011; Siegle, Steinhauer, Stenger, Konecky, & Carter, 2003), and especially for older adults (Piquado, Isaacowitz, & Wingfield, 2010; Zekveld et al., 2011), including older adults with hearing loss (Kramer, Kapteyn, Festen, & Kuik, 1997; Zekveld et al., 2011).

Although changes in effort with hearing loss interventions are of great interest, pupillometry has only been used to index changes within and between participants cross-sectionally. In the current study, we employed a speech-perception training protocol that has demonstrated consistent improvements in behavioral outcome measures of aided speech recognition by older adults (Burk, Humes, Amos, & Strauser, 2006; Burk & Humes, 2007, 2008; Humes et al., 2009). We examined the extent to which an objective physiological indicator of cognitive effort, pupillometry, provided an additive outcome metric of changes following training.

Pupil dilation has been shown to vary with the cognitive demands associated with speech perception, independently of behavioral metrics (Koelewijn et al., 2011; Kuchinsky et al., 2013; Zekveld et al., 2010). In the current study, we thus evaluated the effectiveness of a speech-perception training protocol using three outcome measurements: word identification, RT, and pupil response dynamics. We also examined the extent to which benefit is limited to the acoustic conditions experienced during training, or instead generalizes to new conditions, such as different signal-to-noise ratios (SNRs). Using a curve fitting technique,

we assessed changes in the shape of the pupil response (Kuchinsky et al., 2013), with the goal of characterizing both the magnitude and timing of cognitive demands (Beatty & Lucero-Wagoner, 2000; Nuthmann & van der Meer, 2005; van der Meer et al., 2010). By analyzing multiple components of the pupil response, we evaluated the impact of speech-perception training on both how and when cognitive effort is allocated during a words-in-noise identification task. We discuss the extent to which these outcomes may provide increasingly comprehensive assessments of hearing loss intervention benefits.

Method

Participants

Twenty-nine otherwise healthy older adults with mild-to-moderate hearing loss completed this speech-perception training study. An additional six individuals performed pupillometry at Baseline testing but did not complete the study due to scheduling conflicts (3 Training group individuals), illness (1 Control group individual), or meeting exclusion criteria for other study components that were communicated after initial testing (2 Control group individuals). All participants were right-handed, monolingual native speakers of American English and reported normal or corrected-to-normal visual acuity. The Control group involved 15 individuals (6 female) and the Training group involved 14 individuals (6 female). Participants' average age was 70.2 years [$SD = 8.5$], with no significant difference between the groups [$t(27) = -.48, p = .64, d = -.18$]. Participants provided written informed consent prior to participation in this University Institutional Review Board approved study.

Audiometry

Pure tone thresholds were measured at standard frequency intervals between .25 and 8.0 kHz with a Madsen OB922 clinical audiometer calibrated to ANSI standards (American National Standards Institute, 2004) and TDH-39 headphones. All participants had mild-to-moderate, sloping high-frequency sensorineural hearing loss (Figure 1) and could be considered candidates for hearing aids, though no participant had previous experience with hearing aids. Asymmetry between ears was limited to 15 dB at each frequency. For most participants, stimuli were presented to the right ear, except for the four participants (two in each group) for whom audibility could only be ensured in the left ear (discussed below).

The average threshold across all frequencies (.5 – 8.0 kHz) for the test ear was 36 dB HL, which did not differ significantly between the Training and Control groups [$t(27) = 1.48, p = .15, d = .57$]. Figure 1 displays the substantial overlap of thresholds across participants. However, if only the average threshold at .5, 1.0, and 2.0 kHz is considered (pure-tone average, PTA), the Training group had poorer hearing compared to the Control group [Training, $M = 17, SD = 6$ dB HL; Control, $M = 25, SD = 8$ dB HL; $t(27) = 2.95, p = .007, d = 1.14$]. The impact of this difference on outcome metrics was addressed by the following methods, described in detail below: 1) spectrally shaping of speech and noise for each participant's audiogram; 2) SNRs for the words-in-noise task were set for each participant based on word recognition scores from a separate task in order to achieve a moderate level of listening difficulty; and 3) PTA was included as a covariate in analyses for which significant group effects were observed.

Materials and Procedure

The materials, procedures, and equipment for the pupillometry task and speech-perception training are also described in previous studies (Kuchinsky et al., 2013, and Humes et al., 2009, respectively).

Auditory stimuli—To ensure audibility through at least 4.0 kHz and simulate a well-fit hearing aid, speech and background noise were spectrally shaped for each participant: levels at each one-third octave band frequency were generally 20 dB higher than threshold (Burk & Humes, 2008; Humes et al., 2009; Kuchinsky et al., 2013). A two-talker ICRA vocoded noise (Dreschler, Verschuure, Ludvigsen, & Westermann, 2001), which has the spectral properties of speech but is unintelligible, was used as background noise. The stimuli were presented at an SNR that was determined from the results of an initial open-set, word-in-noise recognition task for which participants listened to and repeated words without the benefit of selecting from a set of options. A default SNR of -2 dB was only adjusted for individuals for whom percent correct recognition fell outside a 30% – 70% range, thus providing all participants with a moderate level of task difficulty [word identification $M = 39.3\%$, $SD = 8.6\%$; no group difference, $t(27) = -.50$, $p = .62$, $d = -.19$]. An SNR of -2 dB was used for 22 of the 29 participants (11 in each group), -1 dB SNR was used for 6 participants (3 in each group), and 0 dB SNR was used for one person in the Control group.

Training—Participants were first recruited and assigned to the Training group. Control participants were recruited in parallel to match the Training group's demographic (age, gender) and hearing loss, as described above, and cognitive measures presented in Table 1. The Control group served as a comparison in order to ensure that observed changes were not driven by repeated experience with the experimental task and eyetracker setup. All participants, regardless of group assignment, completed audiometric and cognitive testing as well as Baseline and Follow-Up sessions of the pupillometry task.

Following the protocol developed by Humes and colleagues (Burk et al., 2006; Burk & Humes, 2007, 2008; Humes et al., 2009), training words were selected from the 600 most common words in the English language, which make up about 90% of the words in spoken conversation. Through pilot testing in studies by Humes and colleagues, the 400 most challenging of these words were selected for training. Each word was spoken by four talkers (2 male, 2 female; 1600 word stimuli), which were broken down into 32 blocks of 50 trials. Once a cycle of training on these 32 blocks was complete, the stimuli were repeated. Thus, a given stimulus (i.e., a word spoken by a particular talker) was never repeated within 1600 trials of one another. There were four cycles total and thus 6400 word training trials (32 blocks $\times$ 50 stimuli $\times$ 4 cycles of repetitions). Other separate training blocks consisted of either word recognition training trials for the remaining 200 easier to understand words or phrase recognition training trials 94 frequent phrases, each spoken by four talkers. These easier word and phrase items were not included in pupillometry testing. Participants typically finished 8 blocks per training session [$SD = 1$], lasting approximately 90 minutes, though because training was self-paced, participants could opt to complete more or less training on a particular day. Participants finished training in an average of 20 sessions [$SD = 4$] over 8.5 weeks [$SD = 3$].

For each training trial, participants listened to speech presented in ICRA background noise. A Tucker Davis Technologies (TDT) System III Real-Time Processor (RP2.1) presented auditory stimuli via a headphone buffer (TDT HB7) and one of a pair of ER-3A insert earphones (Etymotic). In word training blocks, participants then selected the word that they had just heard from a response set of 50 options (i.e., closed-set testing), which comprised the other words presented within that block. These options were presented on a 19-inch Elo touchscreen monitor (60 Hz) so that trainees received both auditory and visual feedback.

Control participants completed the study in a similar time frame as the Training participants; there was no significant difference between the number of days that passed between the two pupillometry sessions [$M = 81$, $SD = 28$] for the Control [$M = 90$, $SD = 27$] and Training groups [$M = 72$, $SD = 28$; $t(27) = 1.80$, $p = .08$, $d = .69$]. To increase the Control group's engagement in the study and provide interaction with the experimenters, Control participants were contacted weekly by phone by one of the study authors (SLC).

Pupillometry task—This task employed an orthographic version of the Visual World Paradigm (McQueen & Viebahn, 2007; Salverda & Tanenhaus, 2010). Participants sat 30 cm from the touchscreen monitor, allowing them to touch the screen to respond. The illumination of the room was kept constant and dimmed to 160 lux for all participants. Stimulus presentation was controlled by MATLAB 2009b running the Psychophysics Toolbox v.3.0.8 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997).

Each trial began with the presentation of a circle (0.80° diameter of visual angle) in the center of the touchscreen that the participant was instructed to fixate. After 1000 ms, the background noise began playing and continued through the rest of the trial. At 1800 ms, a response set display of four words appeared on the screen. Allowing for a 200 ms preview of the location of words on the screen, at 2000 ms, a target word stimulus (produced by one of the male talkers) was presented through the insert earphone as described during training. Words were presented in 24 point Courier font in all caps so that each letter subtended the same amount of visual space ($1.19^\circ \times 1.19^\circ$ of visual angle per letter). The four words on a given trial were the same length (two to seven letters) and were equally spaced on the screen such that the center of each word was 4 in. (18.93°) away from central fixation. Participants listened to each word stimulus and touched one of the four options to indicate the word they had heard. The trial ended when the participant made a response, triggering the start of the subsequent trial. The 256 words that made up these response sets were selected from the 400 words for which the trainees received speech-perception training and always included the target as one of the four options. Non-target alternatives never appeared as a target for a particular individual.

In each session (Baseline and Follow-Up), participants completed the same 192 word identification trials (plus four practice trials). For a third of these trials, the target word was presented in the SNR determined through open-set testing, to which the trainees were exposed throughout the training sessions. Two other SNRs were presented to participants that were 2 dB above (Easier) and 2 dB below (Harder) this baseline SNR and were included in analyses to examine the extent of training transfer across listening conditions. Thus, a given target word was presented once in each SNR, with the order of presentation

counterbalanced across words (i.e., one sixth of the words had each of the possible presentation orders: Hard-Train-Easy, Hard-Easy-Train, Easy-Train-Hard, etc.). Trials were spaced such that there were 60 trials between each item's repetitions. No individual word was presented at the same SNR or target location on the screen (upper left/right, lower left/right) as it had been previously in the experiment. SNR presentation occurred in blocks of 8 trials and the spatial location of the correct response was never the same for more than two trials in a row. For 16 of the participants, the Baseline data for only the Easier and Harder SNRs were presented in a previous paper (Kuchinsky et al., 2013). While this previous work had focused on the sensitivity of the pupil response to experimental manipulations of task difficulty within a single session, the current study aims to examine training-related changes in behavior and in the pupillary response from Baseline to Follow-up testing sessions.

An Arrington Research PC-60 ViewPoint EyeTracker sampled pupil size as a proportion of screen size at 60 Hz (www.ArringtonResearch.com). Only the major axis of the ellipse fit to the pupil was used in analyses, as this metric does not change with eye movements (see Kuchinsky et al., 2013 for a confirmatory analysis). Additionally, fixations made per trial were classified with modified MATLAB scripts from the GazeAlyze toolbox (Berger, Winkels, Lischke, & Höppner, 2012) to verify that the average number of fixations did not change from Baseline to Follow-Up for either group [Training, $t(13) = 1.23$, $p = .24$, $d = .34$; Control $t(14) = 1.20$, $p = .25$, $d = .32$].

Analyses

Behavioral—Behavioral analyses were conducted on recognition scores (correct vs. incorrect) and RTs from the pupillometry task. RTs represented the span of time from word stimulus onset to when the participant touched the screen to indicate the word that they had heard. Trials for which the RT was more than 2.5 SD slower than the individual's mean were excluded from all analyses. Of 192 trials, this criterion resulted in an average of 2.97% excluded trials [$SD = .69\%$] per participant.

A logistic multilevel model using the R (R Development Core Team, 2011) package lme4 (Bates, Maechler, & Bolker, 2011) was performed to examine the impact of training (Group x Session) on performance for each trial. Random effects terms included a random slope of Session for participants and trials and a random slope of Session x SNR for SNR analyses. A similar analysis was conducted by specifying a Gaussian distribution for the log-transformed reaction time (log RT) data. Sub-models were created for the Training and Control groups separately in order to clarify interaction effects. When significant effects of Group were obtained, covariates were included to control for individual differences that varied by groups. In particular, the groups differed significantly in pure tone average across .5, 1.0, and 2.0 kHz frequencies (PTA; Figure 1) and in the variances for Picture Vocabulary standard score (Table 1).

Pupillometry—Preprocessing methods and analysis techniques from Kuchinsky et al. (2013) were used in the current study and were based on methods similar to those reported elsewhere in the literature (Piquado et al., 2010; Zekveld et al., 2010, 2011).

Preprocessing: The eyetracker generates data for the major axis length of the pupil at each time sample occurring every 16.67ms (60 Hz sampling rate) as well as a code indicating the quality of the data collected. Samples on which the pupil was not accurately located or for which the pupil aspect ratio was smaller than one standard deviation below the mean for the entire trial (i.e., indicating a blink; Piquado et al., 2010) were removed. Using data from the previous and subsequent samples that met data quality criteria, these points were linearly interpolated across. Trials requiring more than 50% of the data to be interpolated were omitted from analyses (Siegle et al., 2003), resulting in the removal of 11% of trials on average [$SD = 15\%$] for both sessions. Of the analyzed trials, only 23% of samples were interpolated on average [$SD = 5\%$] and this percentage did not differ based on the group or session being tested [Session $\times$ Group, $F(1,27) = 1.96$, $p = .17$, $\eta^2 = .07$; Session, $F(1,27) = .74$, $p = .40$, $\eta^2 = .03$; Group, $F(1,27) = .03$, $p = .86$, $\eta^2 = .001$]. Data were then smoothed with a 5-sample moving average.

To control for the significant correlation between the mean and standard deviation of pupil size across trials [$M r = .61$, $SD r = .25$ across individuals], each time point was divided by the mean of the entire trial time series (i.e., from trial onset to RT). This allowed for consistent scaling of pupil size values across both trials and participants. Baseline correction was then performed to examine the extent to which pupil size in response to word identification increased relative to a period in which only background noise was played (1000 ms prior to the word stimulus onset). The preprocessed time series data were averaged within each participant and session. For each trial, a total of 113 time samples (1883.71 ms) following the word stimulus onset were entered into pupillometry analyses. This length of time ensured that average pupil size values were obtained for every condition and participant, but was well below the average RT, thus limiting the impact of RT on variation in the pupil response tail (Kuchinsky et al., 2013).

Growth curve analysis: Growth Curve Analysis (GCA) is a statistical technique that is more appropriate for time series data than traditionally used ANOVAs on time-binned data (Kalénine, Mirman, Middleton, & Buxbaum, 2012; Magnuson, Dixon, Tanenhaus, & Aslin, 2007; Mirman, Dixon, & Magnuson, 2008). We have previously shown that GCA is particularly sensitive to variation in the pupil response with task demands compared to traditional analyses (Kuchinsky et al., 2013). By fitting orthogonal polynomial terms to time series data, the impact of experimental conditions on variation in the shape of the pupil response curve could be independently assessed.

The polynomial terms that describe the shape of the response function are detailed in previous research (Kalénine, Mirman, Middleton, & Buxbaum, 2012; Mirman et al., 2008). The Intercept term describes the overall mean of the pupil response. The Linear term describes the slope of the pupil response with more positive values indicating larger pupil size toward the tail compared to the beginning of the response. The Quadratic term describes the shape of the primary curve inflection point, with more positive values denoting a more linear, flatter distribution and more negative values representing a more peaked response. Finally, the Cubic term captures the extent to which there is a secondary inflection point in the data. In previous work, this secondary inflection has been shown to result in the primary inflection point being shifted in time and asymmetric, with more positive values indicating

an earlier and more transient (steeply rising and falling) peak response (e.g., Kalénine, Mirman, & Buxbaum, 2012; Kalénine, Mirman, Middleton, & Buxbaum, 2012; Kuchinsky et al., 2013; Mirman et al., 2008). The specific impact of the Cubic effect in the current data set is depicted in figures in the Results section.

GCA was implemented using R (R Development Core Team, 2011) and the nlme package (Pinheiro, Bates, DebRoy, Sarkar, & R Development Core Team, 2011). Fixed effects (polynomial and experimental regressors) and random effects (error terms) were modeled to predict the corrected pupil response data. Random effects were allowed to vary for the polynomials of theoretical interest (random slopes for Intercept through Cubic terms) across participants within each SNR condition and session.

We then analyzed how experimental conditions (Session: Baseline vs. Follow-Up; and Group: Training vs. Control) modulated the fixed effect polynomial regressors by determining a best-fitting model of the pupillary data. The most comprehensive full model included the highest order polynomial of interest (e.g., Session x Group x Cubic term) as well as all lower order polynomials that make up that effect (e.g., Session x Group, Session x Cubic, Group x Cubic, Session, Group, and Cubic). Through backwards elimination model testing, a significant effect was obtained if a factor's removal from the full model resulted in a significant decrease in the -2 times the log likelihood ratio (-2LL), as given by the log likelihood ratio (L Ratio). This was done iteratively beginning with the highest and then the lower order interactions. A lower order interaction and main effect was always included, regardless of its significance, if it was subsumed by a higher order interaction that significantly contributed to model fit. L Ratio values presented in the text were always calculated by comparing a model that contained a stated effect of interest compared to a model that was identical except it did not contain that effect. To examine the extent to which these patterns of results were dependent on the SNR condition, model testing was also performed for a model containing all SNR x Session x Group x Polynomial interactions. Though the results present SNR as a continuous predictor indicative of the +/- 2 dB manipulation relative to a baseline SNR, the same pattern of results was obtained when treating SNR as a contrast coded, three-level categorical variable (Easier, Training, Harder).

To clarify significant interactions, sub-models were tested in which the effect of Session was evaluated separately for the Training and Control groups. Individual difference covariates that were observed to vary between groups (PTA and Picture Vocabulary) were mean centered and included in any models for which group effects were obtained.

As in our previous study, correct trials were the focus of analyses because of the potentially unknown sources of error for a given incorrect trial, including lapses in attention and misperceptions of the word. Additionally, in the current study we aimed to examine the extent to which pupillometry provides a unique training outcome metric, over and above word recognition and RT. Thus, only correct trials were entered into GCA analyses and an average log RT for each participant for each session was included as a covariate. Across all participants, out of a possible 11136 trials, a total of 9689 (87%) met the quality criteria and of those, 4629 trials (48%) were correct trials that were included in these analyses. Each

session contributed half of the data (49 and 51%, for the first and second session, respectively).

Results

Behavioral

Word identification—Across all SNRs, word recognition significantly improved for the Training group [Session, $\chi^2 = 8.78$, $p = .003$, $\beta = .29$; Baseline, $M = 46.65\%$, $SD = 1.48$; Follow-Up, $M = 48.88\%$, $SD = 1.37$] and not the Control group [Session, $\chi^2 = .35$, $p = .55$, $\beta = .05$; Baseline, $M = 46.63\%$, $SD = 2.08$; Follow-Up, $M = 47.08\%$, $SD = 2.29$]. The marginal two-way interaction that described this relationship [Session x Group, $\chi^2 = 3.12$, $p = .08$, $\beta = .24$] was invariant to SNR [Session x Group x SNR, $\chi^2 = .31$, $p = .58$, $\beta = .08$], individual differences in hearing loss [Session x Group x PTA, $\chi^2 = 1.49$, $p = .68$, $\beta = -.02$], and vocabulary [Session x Group x Picture Vocabulary, $\chi^2 = .63$, $p = .89$, $\beta = .01$]. A significant main effect of PTA indicated that individuals with better hearing identified more words, irrespective of Group [$\chi^2 = 4.57$, $p = .03$, $\beta = -.01$]. No such effect was observed for Picture Vocabulary [$\chi^2 = .14$, $p = .71$, $\beta = .002$].

Reaction time—RT provided a unique outcome measure in addition to changes in word identification (Figure 2b) when examining only correctly identified word trials and accounting for variation in SNR conditions. All participants, regardless of group, responded more quickly with increasingly easy SNR conditions [SNR, $\chi^2 = 24.99$, $p < .0001$, $\beta = .01$] and at Follow-Up compared to Baseline testing [Session, $\chi^2 = 7.70$, $p = .006$, $\beta = -.02$]. However, this difference was even larger for the Training group [Follow-Up, $M = 2.37$ s, $SD = .37$; Baseline, $M = 2.55$ s, $SD = .33$] than the Control group [Follow-Up, $M = 2.33$ s, $SD = .23$, vs. Baseline, $M = 2.45$ s, $SD = .37$; Session x Group, $\chi^2 = 7.70$, $p = .006$, $\beta = -.02$]. This RT training effect did not differ by SNR [Session x Group x SNR, $\chi^2 = .04$, $p = .83$, $\beta = .003$].

Pupillometry

We first asked to what extent there were significant training-related changes in pupil dynamics irrespective of SNR presentation (Group x Session x Polynomial; Table 2A). Model testing presented in Table 2 lead to the selection of a model that maximally fit the pupil response data (Figure 3). This best-fitting model included a significant Group x Session x Cubic interaction [parameter estimate $\beta = 370.35$, $t = 2.95$, $p = .003$]. Separate sub-model tests for the Training and Control group data revealed that only the Training group had a faster peaking pupil response that fell off more rapidly at Follow-Up compared to Baseline testing [Session x Cubic $\beta = 281.96$, $t = 2.31$, $p = .02$]. This pattern was not observed for the Control group [Session x Cubic $\beta = -78.78$, $t = -.74$, $p = .46$]. Figure 4 demonstrates the impact of the Cubic term by comparing the pupil response that is predicted by a model that contains the Cubic term to one that does not.

There was also a significant Group x Session x Intercept interaction [$\beta = 111.46$, $t = 2.85$, $p = .005$]. Sub-model tests demonstrated that only the Training group had an overall larger pupil response at Follow-Up compared to Baseline [Session x Intercept $\beta = 79.75$, $t = 2.40$, p

= .02] whereas the Control group did not show this pattern [Session x Intercept $\beta = -19.50$, $t = -.69$, $p = .49$].

We then tested to what extent training-related changes in the pupil response were dependent on familiarity with the Training SNR and thus limited to that condition (SNR x Group x Session x Polynomial; Table 2B). Model testing revealed that no effects were modulated by the inclusion of an SNR covariate, and thus the training effects were consistent across conditions.

While training did not impact the peakedness of the pupil response (no Group x Session x Quadratic interaction), there was a significant Group x Quadratic effect [Training vs. Control, $\beta = -465.95$, $t = 90.07$, $p < .0001$]. As shown in Figure 3, this effect characterized the overall flatter pupil response for the Control compared to the Training group, irrespective of Session. We demonstrate below the extent to which this effect was driven by individual differences factors and was independent of the training effects of interest.

Controlling for session differences: RT—We included a mean-centered log RT covariate in the best-fitting model to ensure significant effects of Session on the pupil response were not impacted by Session differences in RT. Indeed, including a Session x Group x Cubic x log RT term did not improve model fit [L Ratio = .33, $p = .56$], nor did an interaction with the intercept term [L Ratio = .11, $p = .74$].

Controlling for group differences: PTA and Picture Vocabulary—As described in the Methods, two individual difference metrics varied between the Training and Control groups: the means for PTA and the variances for Picture Vocabulary standardized scores. These measures were not statistically correlated with one another [$r(27) = -.21$, $p = .28$]. We examined the extent to which significant Group effects on pupil dynamics were mitigated by including these factors as covariates, specifically for 1) the earlier shifted, more transient pupil peak (Session x Group x Cubic) and overall larger pupil response (Session x Group x Intercept) following training, and 2) the more peaked pupil response for the Training than the Control group (Group x Quadratic).

1. Neither Picture Vocabulary nor PTA significantly modulated the training-related Cubic or Intercept effects. Including a Session x Group x Cubic x Individual Difference term did not improve model fit [Picture Vocabulary, L Ratio = .24, $p = .63$; PTA, L Ratio = .74, $p = .39$], nor did including a Session x Group x Intercept x Individual Difference term [Picture Vocabulary, L Ratio = .26, $p = .61$; PTA, L Ratio = 1.25, $p = .26$]
2. Controlling for individual differences mitigated the Group x Quadratic interaction. For both factors, including a Quadratic x Individual Difference covariate improved model fit [Picture Vocabulary, L Ratio = 22.62, $p < .0001$; PTA, L Ratio = 15.71, $p < .0001$]. Thus, overall, participants with worse hearing and lower vocabulary knowledge exhibited a flatter (more positive Quadratic) pupil response [Picture Vocabulary, $\beta = -16.65$, $p < .0001$; PTA, $\beta = 19.44$, $p = .0004$]. When these significant interactions were added to the best-fitting model, the strength of the

original Group x Quadratic parameter [$\beta = -465.95, t = -90.07, p < .0001$] was substantially reduced [$\beta = -251.39, t = -2.69, p = .007$].

Individual differences in pupil dynamics—In this study, we observed changes in average pupil dilation (Intercept) and the timing of the peak pupillary response (Cubic) while examining correct trials and controlling for RT. By extracting random effects terms from the best-fitting model, we were also able to examine the extent to which a given individual's results vary from the average pattern (stronger or weaker effects than the group fitted model; e.g., Mirman et al., 2008) and whether the strength of these effects covary and relate to other outcome measures.

Individuals who exhibited the greatest change in the speed of the peak pupil response (Cubic) also showed the largest increase in average pupil size (Intercept) [Training, $r(12) = .59, p = .03$; Control, $r(13) = .51, p = .05$]. Thus, the same individuals contributed to the training-related changes on the Intercept and Cubic, rather than different subgroups driving one effect or the other. Additionally, Training individuals who showed the largest change in average pupil size (Intercept) also showed the greatest reduction in RT [Training, $r(12) = -.83, p < .001$; Control, $r(13) = -.14, p = .63$]. There were no other relationships among individual differences in the Cubic and Intercept effects and RT or word recognition performance for either group [all $|r| < .38, p > .19$].

Discussion

We observed that speech-perception training for older adults with hearing loss resulted in improved word recognition in noise and changes in the pupillary response, an indicator of cognitive demand. Importantly, the pupil response peaked more rapidly and average pupil size was larger following training even when words were correctly identified and at SNRs that were untrained. The results of this study suggest that using independent, objective measures of behavioral performance and of physiological demand has the potential to provide a more comprehensive assessment of variation in cognitive effort with hearing loss and speech-perception training. Additionally, the results provide further evidence that benefits of this speech-perception training protocol generalize to untrained listening conditions (here, SNR; new talkers, Burk & Humes, 2008; Humes et al., 2008), and thus is not solely due to familiarity with the acoustic signal (e.g., Nygaard & Pisoni, 1998).

Across multiple training sessions, participants identified spectrally shaped (“aided”) speech in background noise and received auditory and visual feedback. Because all stimuli were spectrally shaped according to each participant's hearing profile, audibility was ensured despite variation in detection thresholds. We extended previous studies using this speech-perception training protocol (Burk et al., 2006; Burk & Humes, 2007, 2008; Humes et al., 2009) in part by including a Control group with demographic and cognitive profiles that were largely matched to a group of older adult trainees with hearing loss and by testing SNRs that were not trained. This design provided a control for repeated exposure to test materials.

As predicted, only Training group participants identified more word stimuli at Follow-Up compared to Baseline testing. Using an orthographic Visual World Paradigm, participants selected a target word in noise among four alternatives that were matched to the target on a number of dimensions. While this design resulted in consistent training-related improvements across individuals, performance averaged near 50%. It may have been very difficult for participants to narrow down the alternatives beyond two options, particularly as older adults may be especially susceptible to competing stimuli (e.g., Ben-David et al., 2011; Helfer & Straub, 2014; Reville & Spieler, 2012). However, training benefits could be observed even within the restricted range of performance improvements. In fact, even when only analyzing correctly identified word trials, trainees responded more quickly at Follow-Up than at Baseline compared to controls. Thus while repeated testing (as examined via the Control group) alone did not result in improved percent recognition scores, it could have contributed to the changes in RTs as all participants tended to speed up across sessions (e.g., familiarization with using a touchscreen). Additionally, the overall slow (approximately 2.5 s) RT may in part be due to a limitation of the current study: touchscreen responses required gross, target-oriented arm movements that have been previously observed to be slower and more variable for older adults (e.g., Yan, Thomas & Stelmach, 1998). Future studies that utilize other response methods (e.g., button press) or collect measurements of motor control (e.g., the Purdue Pegboard test; Tiffin & Asher, 1948) may be able to account for this factor.

Because of the noted changes in subjective reports of listening effort that accompany hearing loss (Ahlfstrom et al., 2014; Cox & Alexander, 1992; Humes et al., 2001; Humes, 1999; Sweetow & Sabes, 2006), we also extended previous work by Humes and colleagues by including an additional training outcome measure of effort. Specifically, we monitored pupil dilation, which has been used extensively as an objective indicator of moment-to-moment changes in cognitive demands. The peak pupil response has been observed to be smaller and/or rise more quickly with decreasing task difficulty (Beatty & Lucero-Wagoner, 2000; Kahneman et al., 1969; Kuchinsky et al., 2013; Landgraf, van der Meer, & Krueger, 2010; Nuthmann & van der Meer, 2005; Zekveld et al., 2010). However, studies do not necessarily report changes in both the peak size and timing of the pupil response, suggesting that these measures may characterize different aspects of task processing difficulty. Indeed, computational models of the pupillary response include separate parameters that determine peak pupil amplitude and timing (Hoeks & Levelt, 1993; Wierda, Rijn, Taatgen, & Martens, 2012). In the current experiment (also see Kuchinsky et al., 2013), GCA revealed independent changes in the peakedness (Quadratic), average size (Intercept), and the delay and transience of the response (Cubic). Specifically, speech-perception training impacted the time course and overall size, but not the peak magnitude of pupil dilation. Pupil dilation was on average larger and the pupil response peaked more quickly following training across all SNRs, even when controlling for recognition performance, RT, and individual differences.

The timing of pupillary dilation may provide an especially informative outcome measure because it has been associated with a particular neural mechanism, allowing us to better understand not just *what* changes with training but *how*. Pupil dilation has been correlated with the firing of neurons within the locus coeruleus (LC; Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010; Rajkowski, Kubiak, & Aston-Jones, 1993), which is a region of the brainstem

that regulates states of attention in order to support task performance (Aston-Jones & Cohen, 2005). As there are no known direct pathways between the LC and iris muscles, this functional link is thought reflect the parallel downstream influence of a common source, suggesting that pupil dilation may be used as a “reporter variable” to noninvasively index LC functioning in humans (Aston-Jones & Cohen, 2005; Gilzenrat et al., 2010; Rajkowski et al., 1993).

In particular, the peaked, task-evoked pupil response is thought to index a phasic pattern of LC firing, while overall average pupil dilation is indicative of a tonic pattern. These distinct patterns of LC neuron activity are critical to adapting attentional states to optimize task performance, such that transitions from tonic to phasic firing are associated with a switch from explorative to focused attention (Aston-Jones & Cohen, 2005). Specifically, phasic activity has been shown to be sensitive to task difficulty in a target detection task (Rajkowski, Majczynski, Clayton, & Aston-Jones, 2004). The LC phasic response was more delayed with more difficult stimulus discrimination, even more so than behavioral RT. Mirroring these results, in Kuchinsky et al. (2013), the pupil response was more delayed with decreasing SNR, did not increase with response competition difficulty, and was independent of variation in RT. These parallels suggest that the training-related speeding of the pupil response in the current study is indicative of faster discrimination of speech in noise.

The tonic pattern of firing of the LC has been linked to variation in average levels of pupil dilation across an epoch and tied to non-focused attention (Aston-Jones & Cohen, 2005). Specifically, average dilation (here, given by the Intercept) has been shown to decrease with decreasing attentiveness and arousal (Gilzenrat et al., 2010; Jepma & Nieuwenhuis 2011; Kahneman & Peavler, 1969). This suggests that training also resulted in increased arousal and potentially increased attentional focus for the training group. We additionally observed that individuals who exhibited a larger training-related change in average pupil size (Intercept) also exhibited larger changes in the speed of the pupil response (Cubic) as well as their behavioral response (RT), which aligns with the notion that engaging in focused attention facilitates stimulus processing.

In this study, participants not only had to discriminate speech from noise, but also select the aurally presented target from a set of visually displayed alternatives. Indeed, Kuchinsky et al. (2013) observed that during correct word recognition in particularly challenging listening conditions, greater response competition was associated with increased peakedness of the pupil response. Thus, the effort engaged to support correct decision-making may be reflected in the peak amplitude of the pupil response (i.e., Quadratic). More generally, in experiments that simultaneously collected pupillometry and functional imaging data, increasing task-evoked pupil dilation has been associated with greater activity in regions within attention-related cortex, including anterior cingulate (Critchley, Tang, Glaser, Butterworth, & Dolan, 2005), frontal (Siegle et al., 2003), and parietal cortices (Landgraf et al., 2010). These regions are generally engaged with increasing task difficulty, when monitoring and adapting performance (Kerns et al., 2004) and when redirecting attention to task-relevant information (Corbetta & Shulman, 2002).

To summarize the hypothesized significance of the obtained pupillary metrics, average pupil size across a trial may index general arousal levels (Intercept), peak pupil latency (Cubic) may index the discriminability of target stimuli, and the height of the peak of the pupil response (Quadratic) may reflect the ability to allocate decision-making related effort to select a target from amongst competing responses. These cognitive processes are not independent, for example, as the difficulty of discriminating speech from noise can impact the difficulty of selecting that target from amongst competing alternatives. Indeed, Kuchinsky et al. (2013) observed an interaction between SNR and response selection difficulty on the peak pupil response amplitude (Quadratic). Thus, continued human and animal model research will be critical to validating the link between pupil dynamics and the interactive cognitive and neural processes give rise to task-related effort.

The extent to which cognitive effort is experienced during task performance depends not only on task difficulty but also on individuals' cognitive abilities (Kahneman, 1973; Navon & Gopher, 1979). The maximum pupil size tends to be associated with moderate levels of task difficulty, with dilation asymptoting or even decreasing as the task becomes too challenging (Cabestrero et al., 2009; Granholm, Asarnow, Sarkin, & Dykes, 1996; Peavler et al., 1974; Zekveld & Kramer 2014). In the current study, we controlled for variation in difficulty and were thereby able to independently analyze individual difference effects. Selecting an SNR for each participant ensured that the pupillometry task was moderately difficult across all individuals (Figure 2): well above chance (25%) but just under 50% word stimuli identification accuracy on average. When further controlling for variation in performance by only analyzing correctly identified word stimuli, we found that individual differences impacted the peakedness of the pupil response (Quadratic), such that participants with better hearing and vocabulary knowledge displayed larger peaks. Importantly, this was observed 1) across individuals, 2) across testing sessions, and 3) independently of the training effect on other pupillary metrics.

Our finding that poorer hearing and smaller vocabularies led to a less peaked pupil response is consistent with Zekveld and colleagues' (2011) finding in middle-aged adults that individuals with normal hearing (versus a hearing-impaired group) and with better word vocabulary exhibited larger peak pupil dilation when listening to speech in noise. Zekveld et al. and others suggest that, when faced with a challenging task, individuals who have sufficient perceptual and cognitive abilities are able to exert additional effort to perform successfully, and thus exhibit robust neural and pupillary responses (Neubauer & Fink, 2009; van der Meer et al., 2010; Zekveld et al., 2011). Because speech and noise were spectrally shaped in the current study, the overall poorer word recognition and a flatter pupillary response observed with higher PTAs were likely driven by degraded peripheral encoding or central limitations in cortical processing, rather than reduced audibility. Indeed, neuroimaging studies have shown that adults with even mild to moderate hearing loss exhibit reduced gray matter volume in primary auditory cortex (Eckert, Cute, Vaden, Kuchinsky, & Dubno, 2012; Husain et al., 2011; Peelle, Troiani, Grossman, & Wingfield, 2011). Thus, poorer hearing and smaller vocabularies would limit decision-making or response selection processing, thereby resulting in reduced peakedness or smaller peak pupil dilation. Critically, these individual differences did not diminish training-related changes in word recognition or the timing of the pupil response.

Limitations of the current study stem from our use of a relatively passive Control group. While Controls did not differ from the trainees on many demographic and cognitive factors, there were experimental differences that should be systematically examined or controlled in future work. Specifically, the Control group did not receive training in the laboratory and thus had less contact with experimenters, less practice using the touchscreen display, and were not blinded to their study group. One consequence is that we cannot rule out the interpretation that changes in RT for the Training group resulted from differences in touchscreen practice. Future studies should consider employing an active control group, for example who perform another form of training in the laboratory and receive similar experience with response methods.

While Control participants were contacted weekly to maintain contact with experimenters, another consequence of this design is that they may have experienced a different degree of study engagement. Future studies that collect measures of motivation and engagement may help explain some of the variance that our models did not capture. In particular, while including individual differences in vocabulary and hearing loss reduced differences in the pupil response peaks between groups, individual differences in motivation may further reduce this variance as peak pupil dilation has previously been shown to increase with greater motivational incentive (Chiew & Braver, 2013).

In conclusion, by collecting behavioral and physiological measures, we better characterized the benefits that older adults with hearing loss receive from speech-perception training. Trainees not only recognized more words in noise across varying SNRs, but changes in the timing of the pupil response suggest that they also more rapidly discriminated speech from background noise. Simple modifications to tests already in use (e.g., videonystagmography) may allow audiologists to evaluate the feasibility of including pupillometry as a training and hearing aid outcome measure. Understanding changes in effortful speech processing may thus benefit clinicians in selecting optimal treatment strategies and benefit patients in gaining a better understanding of the outcomes they may expect from intervention.

Acknowledgments

We thank Gary Aston-Jones for valuable input on an earlier version of this manuscript. The contributions of William Bologna, Scott Davis, Sarah Hall, Jenny West and the study participants to data collection are also greatly appreciated. Work was supported in part by a Hearing Health Foundation Centurion Clinical Research Award, the National Institute on Deafness and other Communication Disorders (P50 DC00422), Indiana University, and was conducted in a facility constructed with support from Research Facilities Improvement Program (C06 RR14516) from the National Center for Research Resources, NIH. This project was also supported by the South Carolina Clinical and Translational (SCTR) Institute, with an academic home at the Medical University of South Carolina, NIH/NCRR (UL1 RR029882).

References

- Ahlstrom JB, Horwitz AR, Dubno JR. Spatial separation benefit for unaided and aided listening. *Ear & Hearing*. 2014; 35(1):72–85.10.1097/AUD.0b013e3182a02274 [PubMed: 24121648]
- American National Standards Institute. American National Standard Specifications for Audiometers. New York: Author; 2004.
- Anderson Gosselin P, Gagné JP. Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research*. 2011; 54(3):944–58.10.1044/1092-4388(2010/10-0069)

- Aston-Jones G, Cohen JD. An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience*. 2005; 28:403–50.10.1146/annurev.neuro.28.061604.135709
- Bates, D.; Maechler, M.; Bolker, B. lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-39. 2011. <http://CRAN.R-project.org/package=lme4>
- Beatty, J.; Lucero-Wagoner, B. The pupillary system. In: Cacioppo, JT.; Tassinary, LG.; Berntson, GG., editors. *Handbook of psychophysiology*. 2. New York: Cambridge University Press; 2000. p. 142-162.
- Ben-David BM, Chambers CG, Daneman M, Pichora-Fuller MK, Reingold EM, Schneider BA. Effects of aging and noise on real-time spoken word recognition: evidence from eye movements. *Journal of Speech, Language, and Hearing Research*. 2011; 54:243–262.10.1044/1092-4388(2010/09-0233)
- Berger C, Winkels M, Lischke A, Höppner J. GazeAlyze: a MATLAB toolbox for the analysis of eye movement data. *Behavioral Research Methods*. 2012; 44(2):404–19.10.3758/s13428-011-0149-x
- Bologna WJ, Chatterjee M, Dubno JR. Perceived levels of listening effort for a tonal task with contralateral competing signals. *Journal of the Acoustical Society of America*. 2013; 134(4):352–358.10.1121/1.482080
- Brainard DH. The Psychophysics Toolbox. *Spatial Vision*. 1997; 10:433–436. [PubMed: 9176952]
- Burk MH, Humes LE. Effects of training on speech recognition performance in noise using lexically hard words. *Journal of Speech, Language, and Hearing Research*. 2007; 50(1):25–40.10.1044/1092-4388(2007/003)
- Burk MH, Humes LE. Effects of long-term training on aided speech-recognition performance in noise in older adults. *Journal of Speech, Language, and Hearing Research*. 2008; 51(3):759–71.10.1044/1092-4388(2008/054)
- Burk MH, Humes LE, Amos NE, Strauser LE. Effect of training on word-recognition performance in noise for young normal-hearing and older hearing-impaired listeners. *Ear & Hearing*. 2006; 27(3):263–78.10.1097/01.aud.0000215980.21158.a2 [PubMed: 16672795]
- Cabestrero R, Crespo A, Quirós P. Pupillary dilation as an index of task demands. *Perceptual and Motor Skills*. 2009; 109(3):664–678.10.2466/pms.109.3.664-678 [PubMed: 20178265]
- Chiew KS, Braver TS. Temporal dynamics of motivation-cognitive control interactions revealed by high-resolution pupillometry. *Frontiers in Psychology*. 2013; 4(15):1–15.10.3389/fpsyg.2013.00015 [PubMed: 23382719]
- Corbetta M, Shulman GL. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*. 2002; 3(3):201–15.10.1038/nrn755
- Cox RM, Alexander GC. Maturation of hearing aid benefit: objective and subjective measurements. *Ear & Hearing*. 1992; 13(3):131–41.10.1097/00003446-199206000-00001 [PubMed: 1397752]
- Critchley HD, Tang J, Glaser D, Butterworth B, Dolan RJ. Anterior cingulate activity during error and autonomic response. *NeuroImage*. 2005; 27(4):885–95.10.1016/j.neuroimage.2005.05.047 [PubMed: 15996878]
- Crosson, B.; DeBoe, J.; Leber, W. *Psychological Assessment Resources*. 1990. Visual search and attention test: Professional manual.
- Downs DW. Effects of hearing and use on speech discrimination and listening effort. *The Journal of Speech and Hearing Disorders*. 1982; 47(2):189–93. [PubMed: 7176597]
- Dreschler WA, Verschuure H, Ludvigsen C, Westermann S. ICRA noises: Artificial noise signals with speech-like spectral and temporal properties for hearing instrument assessment. *International Journal of Audiology*. 2001; 40(3):148–157.10.3109/00206090109073110
- Eckert MA, Cute SL, Vaden KI, Kuchinsky SE, Dubno JR. Auditory cortex signs of age-related hearing loss. *Journal of the Association for Research in Otolaryngology*. 2012; 713:703–713.10.1007/s10162-012-0332-5 [PubMed: 22618352]
- Engelhardt PE, Ferreira F, Patsenko EG. Pupillometry reveals processing load during spoken language comprehension. *Quarterly Journal of Experimental Psychology*. 2010; 63(4):639–45.10.1080/17470210903469864

- Fraser S, Gagné JP, Alepins M, Dubois P. Evaluating the effort expended to understand speech in noise using a dual-task paradigm: the effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research*. 2010; 53(1):18–33.10.1044/1092-4388(2009/08-0140)
- Gatehouse S, Gordon J. Response times to speech stimuli as measures of benefit from amplification. *British Journal of Audiology*. 1990; 24(1):63–68.10.3109/030053690009077843 [PubMed: 2317602]
- Gilzenrat MS, Nieuwenhuis S, Jepma M, Cohen JD. Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective & Behavioral Neuroscience*. 2010; 10(2):252–69.10.3758/CABN.10.2.252
- Granholm E, Asarnow RF, Sarkin AJ, Dykes KL. Pupillary responses index cognitive resource limitations. *Psychophysiology*. 1996; 33(4):457–61.10.1111/j.1469-8986.1996.tb01071.x [PubMed: 8753946]
- Hällgren M, Larsby B, Lyxell B, Arlinger S. Speech understanding in quiet and noise, with and without hearing aids. *International Journal of Audiology*. 2005; 44(10):574–583.10.1080/14992020500190011 [PubMed: 16315448]
- Helfer KS, Staub A. Competing speech perception in older and younger adults: Behavioral and eye-movement evidence. *Ear & Hearing*. 2014; 35(2):161–170.10.1097/aud.0b013e3182a830cf [PubMed: 24351611]
- Héту R, Riverin L, Lalande N, Getty L, St-Cyr C. Qualitative analysis of the handicap associated with occupational hearing loss. *British Journal of Audiology*. 1988; 22(4):251–264.10.3109/03005368809076462 [PubMed: 3242715]
- Hicks CB, Tharpe AM. Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research*. 2002; 45(3):573–584.10.1044/1092-4388(2002/046)
- Hoeks B, Levelt WJM. Pupillary dilation as a measure of attention: a quantitative system analysis. *Behavior Research Methods, Instruments, & Computers*. 1993; 25(1):16–26.10.3758/BF03204445
- Hornsby BWY. The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear & Hearing*. 2013; 34(5):523–534.10.1097/AUD.0b013e31828003d8 [PubMed: 23426091]
- Humes LE. Dimensions of hearing aid outcome. *Journal of the American Academy of Audiology*. 1999; 10(1):26–39. [PubMed: 9921723]
- Humes LE, Burk MH, Strauser LE, Kinney DL. Development and efficacy of a frequent-word auditory training protocol for older adults with impaired hearing. *Ear & Hearing*. 2009; 30(5):613–27.10.1097/AUD.0b013e3181b00d90 [PubMed: 19633564]
- Humes LE, Garner CB, Wilson DL, Barlow NN. Hearing-aid outcome measured following one month of hearing aid use by the elderly. *Journal of Speech, Language, and Hearing Research*. 2001; 44(3):469–486.10.1044/1092-4388(2001/037)
- Husain FT, Medina RE, Davis CW, Szymko-Bennett Y, Simonyan K, Pajor NM, Horwitz B. Neuroanatomical changes due to hearing loss and chronic tinnitus: a combined VBM and DTI study. *Brain Research*. 2011; 1369:74–88.10.1016/j.brainres.2010.10.095 [PubMed: 21047501]
- Jepma M, Nieuwenhuis S. Pupil diameter predicts changes in the exploration–exploitation trade-off: evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience*. 2011; 23(7):1587–1596.10.1162/jocn.2010.21548 [PubMed: 20666595]
- Kahneman, D. *Attention and Effort*. Englewood Cliffs, NJ: Prentice-Hall Inc; 1973.
- Kahneman D, Beatty J. Pupil diameter and load on memory. *Science*. 1966; 154(3756):1583–1585.10.1126/science.154.3756.1583 [PubMed: 5924930]
- Kahneman D, Peavler WS. Incentive effects and pupillary changes in association learning. *Journal of Experimental Psychology*. 1969; 79(2 pt 1):312–318.10.1037/h0026912 [PubMed: 5785645]
- Kahneman D, Tursky B, Shapiro D, Crider A. Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experimental Psychology*. 1969; 79(1):164–167.10.1037/h0026952 [PubMed: 5785627]
- Kalénine S, Mirman D, Buxbaum LJ. A combination of thematic and similarity-based semantic processes confers resistance to deficit following left hemisphere stroke. *Frontiers in Human Neuroscience*. 2012; 6(106):1–12.10.3389/fnhum.2012.00106 [PubMed: 22279433]

- Kalénine S, Mirman D, Middleton EL, Buxbaum LJ. Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2012; 38(5):1274–1295.10.1037/a0027626
- Kerns JG, Cohen JD, MacDonald AW, Cho RY, Stenger VA, Carter CS. Anterior cingulate conflict monitoring and adjustments in control. *Science (New York, NY)*. 2004; 303(5660):1023–6.10.1126/science.1089910
- Kleiner M, Brainard D, Pelli D. What's new in Psychtoolbox-3? *Perception*. 2007;36. ECVF Abstract Supplement.
- Koelewijn T, Zekveld A, Festen JM, Kramer SE. Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear & Hearing*. 2011; (1990):1–10.10.1097/AUD.0b013e3182310019
- Kramer SE, Kapteyn TS, Festen JM, Kuik DJ. Assessing aspects of auditory handicap by means of pupil dilatation. *Audiology*. 1997; 36:155–164.10.3109/00206099709071969 [PubMed: 9193733]
- Kuchinsky SE, Ahlstrom JB, Vaden KI, Cute SL, Humes LE, Dubno JR, Eckert MA. Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*. 2013; 50(1):23–34.10.1111/j.1469-8986.2012.01477.x [PubMed: 23157603]
- Laeng B, Ørbo M, Holmlund T, Miozzo M. Pupillary Stroop effects. *Cognitive Processing*. 2011; 12(1):13–21.10.1007/s10339-010-0370-z [PubMed: 20865297]
- Landgraf S, van der Meer E, Krueger F. Cognitive resource allocation for neural activity underlying mathematical cognition: A multi-method study. *ZDM*. 2010; 42(6):579–590.10.1007/s11858-010-0264-7
- Lin FR, Thorpe R, Gordon-Salant S, Ferrucci L. Hearing loss prevalence and risk factors among older adults in the United States. *The Journals of Gerontology Series A, Biological Sciences and Medical Sciences*. 2011; 66(5):582–90.10.1093/gerona/66.5.582
- Lin FR, Yaffe K, Xia J, Xue QL, Harris TB, Purchase-Helzner E, Simonsick EM. Hearing loss and cognitive decline in older adults. *JAMA Internal Medicine*. 2013; 173(4):293–9.10.1001/jamainternmed.2013.1868 [PubMed: 23337978]
- Magnuson JS, Dixon J, Tanenhaus M, Aslin R. The dynamics of lexical competition during spoken word recognition. *Cognitive Science*. 2007; 31(1):133–156.10.1080/03640210709336987 [PubMed: 21635290]
- McCoy SL, Tun PA, Cox LC, Colangelo M, Stewart RA, Wingfield A. Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*. 2005; 58(1):22–33.10.1080/02724980443000151
- McQueen JM, Viebahn MC. Tracking recognition of spoken words by tracking looks to printed words. *Quarterly Journal of Experimental Psychology*. 2007; 60(5):661–71.10.1080/17470210601183890
- Metselaar M, Maat B, Krijnen P, Verschuure H, Dreschler W, Feenstra L. Comparison of speech intelligibility in quiet and in noise after hearing aid fitting according to a purely prescriptive and a comparative fitting procedure. *European Archives of Otorhinolaryngology*. 2008; 265(9):1113–1120.10.1007/s00405-008-0596-x [PubMed: 18246361]
- Miller JD, Watson CS, Kewley-Port D, Sillings R, Mills WB, Burleson DF. SPATS: Speech Perception Assessment and Training System. *The Journal of the Acoustical Society of America*. 2007; 122(5):3063.10.1121/1.2942927 [PubMed: 20228894]
- Mirman D, Dixon JA, Magnuson JS. Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*. 2008; 59(4):475–494.10.1016/j.jml.2007.11.006 [PubMed: 19060958]
- Nachtegaal J, Kuik DJ, Anema JR, Goverts ST, Festen JM, Kramer SE. Hearing status, need for recovery after work, and psychosocial work characteristics: Results from an internet-based national survey on hearing. *International Journal of Audiology*. 2009; 48(10):684–91.10.1080/14992020902962421 [PubMed: 19863354]
- Navon D, Gopher D. On the economy of the human-processing system. *Psychological Review*. 1979; 86(3):214–255.10.1037//0033-295X.86.3.214

- Neubauer AC, Fink A. Intelligence and neural efficiency. *Neuroscience and Biobehavioral Reviews*. 2009; 33(7):1004–23.10.1016/j.neubiorev.2009.04.001 [PubMed: 19580915]
- Nuthmann A, van der Meer E. Time's arrow and pupillary response. *Psychophysiology*. 2005; 42(3): 306–17.10.1111/j.1469-8986.2005.00291.x [PubMed: 15943685]
- Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. *Perception & Psychophysics*. 1998; 60(3):355–376.10.3758/BF03206860 [PubMed: 9599989]
- Peavler WS. Pupil size, information overload, and performance differences. *Psychophysiology*. 1974; 11(5):559–566.10.1111/j.1469-8986.1974.tb01114.x [PubMed: 4415394]
- Peelle JE, Troiani V, Grossman M, Wingfield A. Hearing loss in older adults affects neural systems supporting speech comprehension. *Journal of Neuroscience*. 2011; 31(35):12638–12643.10.1523/JNEUROSCI.2559-11.2011 [PubMed: 21880924]
- Pelli DG. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*. 1997; 10:437–442.10.1163/156856897x00366 [PubMed: 9176953]
- Pichora-Fuller MK. Perceptual effort and apparent cognitive decline: Implications for audiologic rehabilitation. *Seminars in Hearing*. 2006; 27(4):284–293.10.1055/s-2006-954855
- Pichora-Fuller MK, Schneider BA, Daneman M. How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*. 1995; 97(1):593–608.10.1121/1.412282 [PubMed: 7860836]
- Picou EM, Ricketts TA, Hornsby BWY. How hearing aids, background noise, and visual cues influence objective listening effort. *Ear & Hearing*. 2013;1–13.10.1097/AUD.0b013e31827f0431
- Pinheiro, J.; Bates, B.; DebRoy, S.; Sarkar, D. R Development Core Team. nlme: Linear and nonlinear mixed effects models. R package version 3.1–101. 2011. <http://cran.r-project.org/web/packages/nlme>
- Piquado T, Isaacowitz D, Wingfield A. Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*. 2010; 47(3):560–9.10.1111/j.1469-8986.2009.00947.x [PubMed: 20070575]
- R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011. Retrieved from <http://www.r-project.org>
- Rajkowski J, Kubiak P, Aston-Jones G. Correlations between locus coeruleus (LC) neural activity, pupil diameter and behavior in monkey support a role of LC in attention. *Society for Neuroscience Abstracts*. 1993; 19:974.
- Rajkowski J, Majczynski H, Clayton E, Aston-Jones G. Activation of monkey locus coeruleus neurons varies with difficulty and performance in a target detection task. *Journal of Neurophysiology*. 2004; 92(1):361–71.10.1152/jn.00673.2003 [PubMed: 15028743]
- Revill KP, Spieler DH. The effect of lexical frequency on spoken word recognition in young and older listeners. *Psychology & Aging*. 2011; 27(1):80–87.10.1037/a0024113 [PubMed: 21707175]
- Salthouse TA, Toth J, Daniels K, Parks C, Pak R, Wolbrette M, Hocking KJ. Effects of aging on efficiency of task switching in a variant of the Trail Making Test. *Neuropsychology*. 2000; 14(1): 102–111.10.1037//0894-4105.14.1.102 [PubMed: 10674802]
- Salverda AP, Tanenhaus MK. Tracking the time course of orthographic information in spoken-word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 2010; 36(5): 1108–17.10.1037/a0019901
- Saunders GH, Chisolm TH, Abrams HB. Measuring hearing aid outcomes--not as easy as it seems. *Journal of Rehabilitation Research & Development*. 2005; 42(4 Suppl 2):157–168.10.1682/JRRD.2005.01.0001 [PubMed: 16470471]
- Siegle GJ, Steinhauser SR, Stenger VA, Konecky R, Carter CS. Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. *NeuroImage*. 2003; 20(1):114–124.10.1016/S1053-8119(03)00298-2 [PubMed: 14527574]
- Sweetow RW, Sabes JH. The need for and development of an adaptive Listening and Communication Enhancement (LACE) Program. *Journal of the American Academy of Audiology*. 2006; 17(8): 538–58.10.3766/jaaa.17.8.2 [PubMed: 16999250]
- Tiffin J, Asher EJ. The Purdue pegboard: Norms and studies of reliability and validity. *Journal of Applied Psychology*. 1948; 32(3):234–247.10.1037/h0061266 [PubMed: 18867059]

- Tun PA, McCoy S, Wingfield A. Aging, hearing acuity, and the attentional costs of effortful listening. *Psychology and Aging*. 2009; 24(3):761–6.10.1037/a0014802 [PubMed: 19739934]
- Van der Meer E, Beyer R, Horn J, Foth M, Bornemann B, Ries J, Wartenburger I. Resource allocation and fluid intelligence: Insights from pupillometry. *Psychophysiology*. 2010; 47(1):158–69.10.1111/j.1469-8986.2009.00884.x [PubMed: 19761522]
- Viljanen A, Kaprio J, Pyrkko I, Sorri M, Pajala S, Kauppinen M, Rantanen T. Hearing as a predictor of falls and postural balance in older female twins. *Journal of Gerontology, A Biological Science and Medical Science*. 2009; 64A(2):312–317.10.1093/gerona/gln015
- Wechsler, D. Wechsler Memory Scale (WMS-III). San Antonio, TX: The Psychological Corporation; 1997.
- Wechsler, D. Wechsler Abbreviated Scale of Intelligence. San Antonio, TX: The Psychological Corporation; 1999.
- Wierda SM, Van Rijn H, Taatgen NA, Martens S. Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proceedings of the National Academy of Sciences*. 2012; 109(22):8456–8460. www.pnas.org/cgi/doi/10.1073/pnas.1201858109.
- Wild CJ, Yusuf A, Wilson DE, Peelle JE, Davis MH, Johnsrude IS. Effortful listening: The processing of degraded speech depends critically on attention. *The Journal of Neuroscience*. 2012; 32(40):14010–21.10.1523/JNEUROSCI.1528-12.2012 [PubMed: 23035108]
- Woodcock, RW.; McGrew, KS.; Mather, N. Woodcock-Johnson III Tests of Cognitive Abilities. Itasca, IL: Riverside Publishing; 2001a.
- Woodcock, RW.; McGrew, KS.; Mather, N. Woodcock-Johnson III Tests of Achievement. Itasca, IL: Riverside Publishing; 2001b.
- Yan JH, Thomas JR, Stelmach GE. Aging and rapid aiming arm movement control. *Experimental Aging Research*. 1998; 24(2):155–168.10.1080/036107398244292
- Zekveld AA, Kramer SE, Festen JM. Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear & Hearing*. 2010; 31(4):480–90.10.1097/AUD.0b013e3181d4f251 [PubMed: 20588118]
- Zekveld AA, Kramer SE, Festen JM. Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear & Hearing*. 2011; 32(4):498–510.10.1097/AUD.0b013e31820512bb [PubMed: 21233711]
- Zekveld AA, Kramer SE. Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*. 2014; 51(3):277–284.10.1111/psyp.12151 [PubMed: 24506437]

Figure 1.

Audiograms depict participants' mild-to-moderate, sloping high-frequency sensorineural hearing loss. Individual audiograms are represented by solid (Training) and dashed (Control) lines with the means at each frequency for each group represented by filled (Training) and open (Control) circles.

Figure 2.

A) Only Training participants significantly identified more word stimuli in noise from Baseline to Follow-Up testing. B) Reaction time (RT) provided an additional outcome metric of training beyond word identification. Though all participants generally responded more quickly at Follow-Up testing, trainees demonstrated a larger reduction in RT than Controls. Error bars represent the standard error of the mean.

Figure 3.

Raw data and fitted pupil size values from the best-fitting GCA model (where 0 represents baseline pupil size prior to word stimulus presentation) are plotted as a function of time from target word stimulus onset. Only trials for which participants correctly identified words are included. The top panels plot the Control and Training groups' fitted values as blue and red lines with error bars representing the standard error of the mean. The standardized pupil response values (i.e., output of preprocessing) are shown in black to demonstrate the models' fit of the raw data. These four fitted lines are then plotted in the same panel for direct comparison across Groups and Sessions. Only the Training group displayed a significantly earlier peaking and more transient pupil response (Cubic term) and larger overall dilation average (Intercept) at Follow-Up compared to Baseline.

Figure 4.

To interpret the significant Group x Session x Cubic interaction on the pupil response, two models are compared: the predicted pupil size (fitted effects) from the best-fitting model, presented in Figure 3 (black lines); and the predicted pupil size from same model with the Cubic terms removed (gray lines). This comparison demonstrates that the Cubic effect captures the faster peaking pupil response exhibited only by the Training group.

Table 1

Cognitive measures for Training and Control group participants. Standardized or percentile scores were used whenever possible (all except for Connections).

Construct	Measure	Control Group		Training Group		Group Difference		
		M	SD	M	SD	t(27)	p	d
Intelligence	IQ ¹	.73	.26	.80	.24	-.73	.47	-.28
Visual Attention	Visual Search and Attention ²	.58	.30	.47	.30	.96	.35	.37
Processing Speed	Connections - Complex ³	10.08	3.29	10.27	3.28	-.15	.88	-.06
Episodic Memory	Total Memory ⁴	98.93	13.47	100.71	14.87	-.34	.74	-.13
Working Memory	Numbers Reversed ⁵	108.73	16.63	109.71	16.98	-.16	.88	-.06
Language-Related Processing	Sound Blending ⁵	105.53	11.35	109.50	9.83	-1.00	.33	-.38
	Incomplete Words ⁵	115.27	12.97	108.79	10.96	1.45	.16	.56
	Letter-Word Identification ⁶	108.20	8.17	106.57	7.31	.56	.58	.22
	Word Attack ⁶	101.27	8.57	101.71	11.23	-.12	.90	-.05
	Picture Vocabulary ⁶	93.00	7.03	97.57	11.62	-1.29	.21	-.50 [†]

[†] Levene's Test indicated unequal variances between the groups ($F = 6.24, p = .02$), though the mean difference was not significant, even with adjusted degrees of freedom, $t(21) = -1.27, p = .22, d = -.55$.

¹ Wechsler (1999)

² Crosson, DeBoe, & Leber (1990)

³ Salthouse et al. (2000)

⁴ Wechsler (1997)

⁵ Woodcock, McGrew, & Mather (2001a)

⁶ Woodcock, McGrew, & Mather (2001b)

Table 2

Model tests examine the impact of training on each component of the pupil response. The first row of each section presents the full model's fit of the data ($-2 * \log \text{likelihood}$; $-2LL$). Subsequent rows present the extent to which model fit is significantly worse (L Ratio) after the removal of a given interaction term. Tests are grouped into (A) effects of training that are consistent across SNR and (B) effects of training that vary across SNR. Significant effects ($p < .05$) are bolded.

Interaction	Model Testing		
	$-2LL$	L Ratio	p
(A) Group x Session x Polynomial	-112679		
Intercept	-112679	5.16	.02
Linear	-112679	.14	.70
Quadratic [†]	-112679	.37	.54
Cubic	-112684	4.92	.03
(B) SNR x Group x Session x Polynomial	-112674		
Intercept	-112675	1.10	.29
Linear	-112674	.21	.65
Quadratic	-112674	.03	.85
Cubic	-112674	.02	.90

[†] Following the removal of the non-significant Group x Session x Quadratic interaction, a Group x Quadratic effect was observed to significantly improve model fit [L Ratio = 30.41, $p < .0001$].