Auditory attention switching with listening difficulty: Behavioral and pupillometric measures

Daniel R. McCloy,[a)] Eric Larson, and Adrian K. C. Lee[a),b)]
Institute for Learning and Brain Sciences, 1715 NE Columbia Road, Box 357988, Seattle, Washington 98195-7988, USA

(Received 10 September 2018; revised 24 October 2018; accepted 25 October 2018; published online 12 November 2018)

Pupillometry has emerged as a useful tool for studying listening effort. Past work involving listeners with normal audiological thresholds has shown that switching attention between competing talker streams evokes pupil dilation indicative of listening effort [McCloy, Lau, Larson, Pratt, and Lee (**2017**). J. Acoust. Soc. Am. **141**(4), 2440–2451]. The current experiment examines behavioral and pupillometric data from a two-stream target detection task requiring attention-switching between auditory streams, in two participant groups: audiometrically normal listeners who self-report difficulty localizing sound sources and/or understanding speech in reverberant or acoustically crowded environments, and their age-matched controls who do not report such problems. Three experimental conditions varied the number and type of stream segregation cues available. Participants who reported listening difficulty showed both behavioral and pupillometric signs of increased effort compared to controls, especially in trials where listeners had to switch attention between streams, or trials where only a single stream segregation cue was available.

https://doi.org/10.1121/1.5078618

[VB] Pages: 2764–2771

I. INTRODUCTION

Pupillometry has emerged as a useful tool for studying the effort associated with auditory perception in sub-optimal listening conditions, encompassing both task difficulty (e.g., stimulus degradations) and the listener's application of mental resources to the task (Pichora-Fuller *et al.*, 2016; Winn *et al.*, 2018). However, not all difficult listening situations induce equivalent pupillary responses. Past work has shown that when the task involves simple detection of target speech sounds, stimulus degradations such as reverberation or vocoding incur a behavioral cost in accuracy and/or reaction time, but cause little to no change in the pupillary response; in contrast, requiring the listener to switch attention between auditory streams induces both behavioral cost and changes to the pupillary response (McCloy *et al.*, 2017). This difference has been attributed to the efficacy of listener effort at improving performance: listeners cannot make vocoded stimuli any less degraded by trying harder, but increased listener effort can improve deployment of selective attention.

When stimuli comprise longer spans of speech (so that contextual meaning becomes relevant), several studies have found increased dilation in response to decreased intelligibility. This occurs whether the decline in intelligibility results from increased background noise (Zekveld *et al.*, 2010), using speech as a masker signal (Koelewijn *et al.*, 2012), or spectrally degrading the target sentences (Winn *et al.*, 2015). This change in dilation can be interpreted as a reflection of listener effort stemming from the online construction of sentential meaning from component words, e.g., the effort to reanalyze misperceived words in light of veridical perception of other words in the same sentence (cf. discussion in Winn *et al.*, 2015). Pupillometry has also revealed increased dilation (interpreted as effort) for older listeners and for listeners with impaired function in the auditory periphery (Kuchinsky *et al.*, 2013; Winn *et al.*, 2015; Zekveld *et al.*, 2011). However, many listeners who have clinically normal hearing nevertheless complain of difficulties in acoustically crowded listening conditions (Ruggles *et al.*, 2011), which may stem from supra-threshold coding deficits (Bharadwaj *et al.*, 2014). It is reasonable to hypothesize that their complaints reflect increased listening effort—perhaps resulting from supra-threshold deficits—but to our knowledge, pupillometry has never been used to study listening effort in such populations. This study presents evidence that people who self-report difficulty in challenging listening conditions show both behavioral and pupillometric differences from listeners who do not report such difficulties.

In designing this study we defined a "listening difficulty" group as listeners with clinically normal hearing who self-report difficulty localizing sound sources and/or understanding speech in reverberant or acoustically crowded environments, based on two yes/no screening questions (see Sec. II A for details), and complemented them with age-matched controls who do not report such difficulties. The task was a two-stream target detection task with a pre-trial cue indicating the need to maintain attention to one stream throughout the trial, or to switch attention between streams at a designated mid-trial pause. Listeners were tested in three conditions with differing stream segregation cues: one with two same-voice talkers separated only by simulated

[a)]Also at: Department of Speech and Hearing Sciences, University of Washington, 1417 NE 42nd Street, Box 357988, Seattle, WA 98105-6246, USA.
[b)]Electronic mail: akclee@uw.edu

spatial cues, one with two co-located different-voice talkers, and one where both spatial and voice cues were available to support stream segregation. These cue types (spatial and voice-identity cues, and their combination) were chosen in hopes of determining whether self-reported difficulty with spatial hearing was in fact confined to cases of spatially segregable talkers, or reflected a more general difficulty with selective attention in the presence of competing speech. We analyzed listeners' behavioral and pupillary responses to investigate how those objective measures relate to their self-reported experience of listening difficulty.

II. METHODS

A. Participants

Twelve adults with self-reported listening difficulty (aged 21 to 66 yr) participated in this study, along with 12 control subjects. Listening difficulty was defined by an affirmative answer to either of two screening questions drawn from the Speech, Spatial, and Qualities (SSQ) of Hearing assessment (Gatehouse and Noble, 2004). The questions were:

(1) "Do you have difficulty understanding speech in the presence of background noise or in large rooms that echo?"
(2) "Do you have difficulty determining where a sound came from without having to look?"

Because the pupillary response changes with age (Kumnick, 1954), a control subject was matched to within 2 years of age (at the time of testing) for each of the listening difficulty subjects. Despite the reported listening difficulties, all participants in both groups had normal audiometric thresholds (20 dB hearing level or better at octave frequencies from 250 Hz to 8 kHz). All participants were compensated at an hourly rate, and gave informed consent to participate as overseen by the University of Washington Institutional Review Board.

B. Stimuli

The methods for this study closely follow those in McCloy *et al.* (2017). Stimuli comprised spoken English alphabet letters from the ISOLET v1.3 corpus (Cole *et al.*, 1990) from one female and one male talker. Letters were silence padded, root-mean-square normalized and windowed as described in McCloy *et al.* (2017), except that here the letters were padded to a final duration of 400 ms (instead of 500 ms as in the previous study). Two streams of four letters each were generated for each trial, with a gap between the second and third letters of each stream. The letters "A" and "U" were used only in the pre-trial cues (described below); the target letter was "O" and letters "DEGPV" were non-target items. Target onsets were always separated from each other by at least 1 s (regardless of stream); thus there were at most two O tokens per trial (overall, 1/4 of trials had zero O tokens, 1/2 had one, and 1/4 had two). For trials with only spatial cues, the two streams were the same talker spatialized to left and right sides at ±30° azimuth. For trials with only non-spatial cues, the streams were a male and female voice co-located at either +30° or −30° azimuth. For trials with both cues (the "mixed cue" condition), streams were both spatialized to ±30° and comprised voices from different-gender talkers. Stream spatialization was simulated using binaural room impulse responses (Shinn-Cunningham *et al.*, 2005) truncated to include only the direct impulse response. Unlike McCloy *et al.* (2017), no degradation (vocoding or reverberation) was applied to the stimuli.

C. Procedure

Except where noted, stimulus delivery replicated procedures used in McCloy *et al.* (2017). A diagram of the trial structure is given in Fig. 1. Subjects heard sounds over insert earphones in a darkened soundproof booth, with illumination adjusted to put each subject's baseline pupil dilation in the center of its dynamic range (McCloy *et al.*, 2016). Pupil size was continuously measured with an EyeLink1000 infra-red eye tracker (SR Research, Kanata, ON) at 1000 Hz sampling frequency, with participants' heads stabilized on a chin rest and forehead bar 50 cm from the camera. Participants were instructed to fixate on a white dot centered on a dark screen and maintain this gaze throughout test blocks.

Each trial began with an 800 ms auditory cue (spoken letters "AA" or "AU"); the location and gender of the cue talker conveyed the location and gender of the to-be-attended stream, and the letters spoken indicated whether to maintain attention to that stream throughout the trial (AA cue) or to switch attention to the other talker at the mid-trial gap (AU cue). A 400 ms pause followed the cue, after which the two concurrent 4-letter streams began, with a 600 ms gap between the second and third letters. This mid-trial "switching gap" duration was chosen to ensure adequate time for an auditory attention switch, which has been shown to require 300–400 ms to execute (Larson and Lee, 2013). The task was to respond by button press to the letter O spoken by the target talker while ignoring O tokens spoken by the competing talker (cf. Fig. 1).

There were 8 blocks of 48 trials each, for a total of 384 trials. The gender and location of the cue talker was fixed for each block of trials. The order of blocks and trials was counterbalanced across subjects. The first experimental block was preceded by six short training blocks, exposing listeners gradually to the maintain- and switch-attention conditions and to the spatial, non-spatial, and mixed-cue stimuli. Training blocks were repeated until participants achieved ≥80% of trials correct on mixed-maintain or mixed-switch blocks, ≥70% on mixed blocks containing both maintain and switch trials, and ≥50% on blocks containing a mix of all of the segregation-cue and attentional conditions.

D. Data analysis

Listener responses were labeled as "hits" if a button press occurred between 100 and 800 ms after the onset of an O token in the target stream, and "misses" if there was no response in that window. Responses at any other time during the trial were considered "false alarms," and lack of response in the 100–800 ms window following timing slots lacking

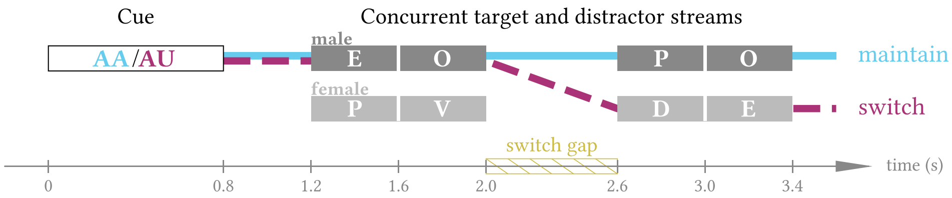

FIG. 1. (Color online) Illustration of "maintain" and "switch" trial types. In the depicted maintain trial (heavy solid line), listeners would hear cue AA in a male voice, attend to the male voice ("EOPO") throughout the trial, and respond twice (once for each O). In the depicted switch trial (heavy dashed line), listeners would hear cue AU in a male voice, attend to the male voice ("EO") for the first half of the trial and the female voice ("DE") for the second half of the trial, and respond once (to the O occurring at 1.6–2.0 s).

target-stream O tokens were considered "correct rejections." Reaction time was recorded for all button presses, but only reaction times for hits were analyzed.

Response accuracy and reaction time were modeled with (generalized) linear mixed-effects regression using the afex package in R (Singmann *et al.*, 2018). A reaction time linear model predicted latency of button press at each timing slot from interactions among trial parameters (maintain/switch attention, spatial/non-spatial/mixed cue) and indicator variables encoding listener group (listening difficulty or control) and timing slot number (four slots per trial, see Fig. 1). A random intercept was also estimated for each participant. Significance of model coefficients was computed via t-tests using the Satterthwaite approximation for degrees of freedom (Satterthwaite, 1946).

The model of response accuracy predicted probability of button press from interactions of attention and segregation-cue trial parameters, indicators for participant group and target/foil presence/absence, and a random intercept for each participant. An indicator for trial slot was not included due to issues with model convergence. This model transformed response probabilities into continuous values suitable for linear modeling using an inverse probit link function, which allows interpretation of coefficient estimates as differences on a d' scale (DeCarlo, 1998; McCloy and Lee, 2015; Sheu *et al.*, 2008). Significance of model coefficients was determined using Wald z-tests.

Pupil diameter recordings were epoched from −0.5 to 5.9 s for each trial, with 0 s defined as the cue onset. Treatment of eye blinks, normalization, and deconvolution of pupil time courses followed the procedures described in McCloy *et al.* (2017). The deconvolved time series can be interpreted as an indicator of mental effort that is time-aligned to the stimulus (i.e., the response latency of the pupillary response has been effectively removed; see McCloy *et al.*, 2016, for a longer discussion of how this measure is interpreted). Statistical comparison of deconvolved pupil dilation time series (i.e., "effort" in Figs. 4 and 5) used a non-parametric cluster-level t-test (Maris and Oostenveld, 2007; McCloy *et al.*, 2016) on the paired differences in deconvolved pupil size between groups (Fig. 4) or attentional conditions (Fig. 5).

E. *Post hoc* analyses

To further test the relationship between pupillary responses and auditory spatial abilities, *post hoc* comparisons were made between summary measures of each subject's pupillary response, and each subject's component scores on the full SSQ of Hearing assessment (Gatehouse and Noble, 2004). Summary pupillometry measures were also compared to each subject's scores on a range of behavioral measures representing binaural health: binaural masking level differences with in-phase noise and in-phase or anti-phase signals (500 Hz signal; N_0S_0 and N_0S_π), frequency modulation detection thresholds (500 Hz; monaural), interaural time- and level-difference detection thresholds, alternating interaural phase detection thresholds (low-frequency temporal fine structure test), and two versions of the Coordinate Response Measure task (spatially separated or co-located talkers). The summary pupillometry measures computed were mean peak amplitude, mean peak latency, and mean area under the curve (AUC). The difference in AUC between attention conditions (switch minus maintain) was also computed.

III. RESULTS

A. Response accuracy

Over all trials, sensitivity (d') ranged across subjects from 1.2 to 4.1 (first quartile 2.2, median 2.7, third quartile 3.2). The model estimated main effects for token type (target, foil, or neither), interactions between token type and each of the trial predictors (maintain/switch attention and spatial/non-spatial/mixed cue) and participant group (listening difficulty/control), and interactions between token type and each pair of predictors.[1] A model that included the four-way interaction of token_type:attention:cue_type:participant_group was also created, but it did not provide a significantly better fit to the data based on nested model comparison using a likelihood ratio test [$\chi^2(6) = 2.95$, $p = 0.82$].

The response accuracy model is summarized in Fig. 2, which shows main effects of each predictor in the left half of each panel, and interactions among predictors in the right half of each panel. The probability of participant response to target items was higher [Fig. 2(A)], and response to foils was lower [Fig. 2(D)] in maintain- versus switch-attention trials. When comparing subject populations, there was a significant difference between the control and listening difficulty groups in spurious responses [where neither target nor foil were present; Fig. 2(H)] but not for responses to target or foil items [Figs. 2(B) and 2(E)]. For the cue type predictor, both spatial and non-spatial cue conditions showed the same pattern relative to the mixed-cue condition: reduced

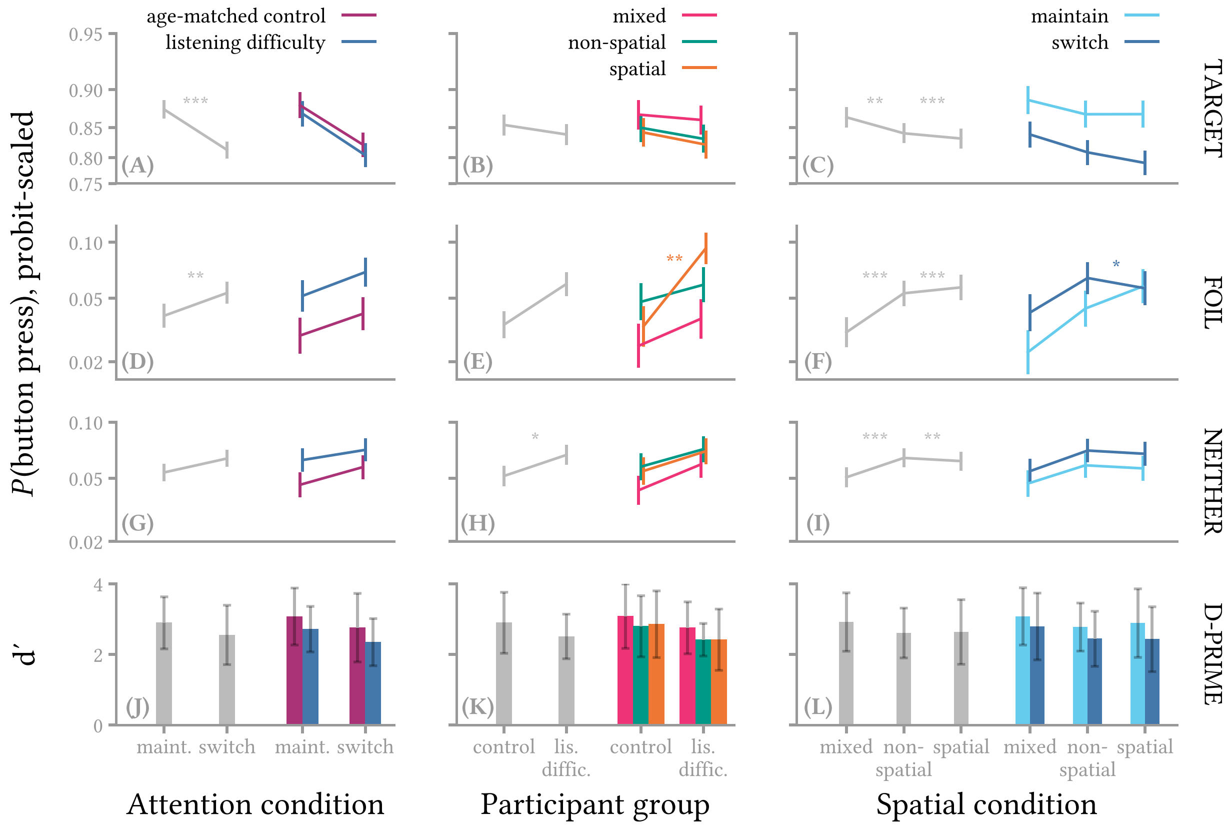

FIG. 2. (Color online) Summary of main effects and interactions in the model of response accuracy. The top three rows reflect probability of button-press response to either targets, foil items, or non-target non-foil items (cf. row labels on right side of figure). Lines in panels (A)–(I) connect estimated marginal means (EMMs) for the various predictors; vertical error bars show 95% confidence intervals for the EMMs. In those panels, the left-hand line (light gray line) shows the main effect indicated on the abscissa; the set of colored lines on the right side of each panel illustrates an interaction between the predictor on the abscissa and the predictor indicated by the legend key at the top of each column. The bottom row combines data from the top three rows into a single measure of detection sensitivity (d'), showing mean ±1 standard deviation of d' values across subjects. See text for discussion of each panel. * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$.

response to targets, and elevated response to both foil items and items that were neither target nor foil [Figs. 2(C), 2(F), and 2(I)]. Mean d' values for each condition, which incorporate target, foil, and spurious responses into a single measure of detection sensitivity, are given in the bottom row of Figs. 2(J)–2(L), with error bars indicating standard deviation across subjects.

In general, difficult experimental conditions seemed to increase all types of errors: missed targets, false alarm responses to foil items, and false alarm responses to non-target non-foil items. The largest effect was seen in the condition in which only spatial cues are available: in that condition, responses to foil items were much higher among participants with listening difficulty than among controls [Fig. 2(E), right side]. There was also an unexpected interaction: there was no difference in responses to foil items between maintain- and switch-attention trials when only spatial cues are present, whereas the non-spatial and mixed-cue conditions did show a difference in response rate to foil items between the maintain- and switch-attention trials [Fig. 2(F), right side].

B. Reaction time

Over all correct responses, median reaction time for each subject ranged from 352 to 574 ms after the onset of the target letter. Model coefficients[1] indicated faster reaction times in timing slots 2–4 compared to slot 1 [Fig. 3(D)], slower reaction times in slot 3 in switch-attention trials relative to maintain-attention trials [Fig. 3(E)], slower reaction times in slots 3 and 4 in non-spatial trials relative to mixed-cue trials [Fig. 3(G)], and slower reaction times in slot 4 of non-spatial switch trials among participants with listening difficulty [Fig. 3(H)].

C. Pupillometry

Average deconvolved pupillary responses for the control and listening difficulty groups are shown in Fig. 4, for each of the segregation cue conditions (columns) and attentional conditions (rows). There appears to be a trend toward larger, later peak pupillary responses among the listening difficulty subjects especially in the switch-attention trials [Figs. 4(A)–4(D)], but none of the comparisons shown in each subplot yielded a statistically reliable difference between the groups (p-values ranged from 0.075 to 0.165).

However, if instead we perform a within-subjects comparison of maintain- versus switch-attention trials, a difference between subject groups emerges. Specifically, for the listening difficulty group, there were significant differences in pupil response between maintain- versus switch-attention

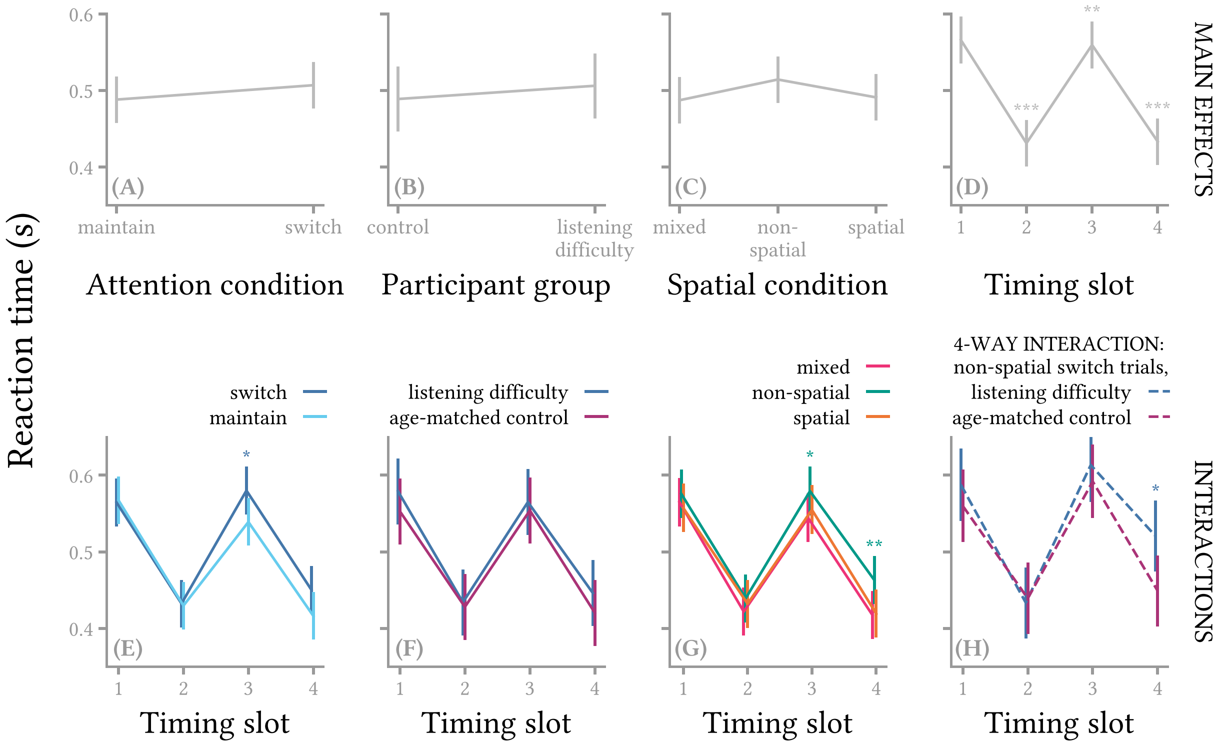

FIG. 3. (Color online) Summary of main effects and significant interactions in the model of reaction time. The top row of panels represent model coefficients for the main effects of attention (A), participant group (B), spatial condition (C), and timing slot (D). Panels (E)–(G) show interactions between timing slot and attention, participant group, and spatial condition (respectively). Panel (H) illustrates the significant four-way interaction, showing how reaction time in slot 4 of non-spatial switch-attention trials is longer for participants with listening difficulty than for age-matched controls. Lines on each plot connect EMMs for the various predictors; vertical error bars show 95% confidence intervals for the EMMs. * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$.

trials that span most or all of the trial time course in all three segregation cue conditions [Figs. 5(A)–5(C)] and also when all three conditions were pooled [Fig. 5(D)], whereas the control group only showed a statistically reliable difference between maintain- and switch-attention trials during the return to baseline at the end of the trial, and only in the two more difficult (spatial / non-spatial) conditions or when conditions were pooled [Figs. 5(E)–5(H)].

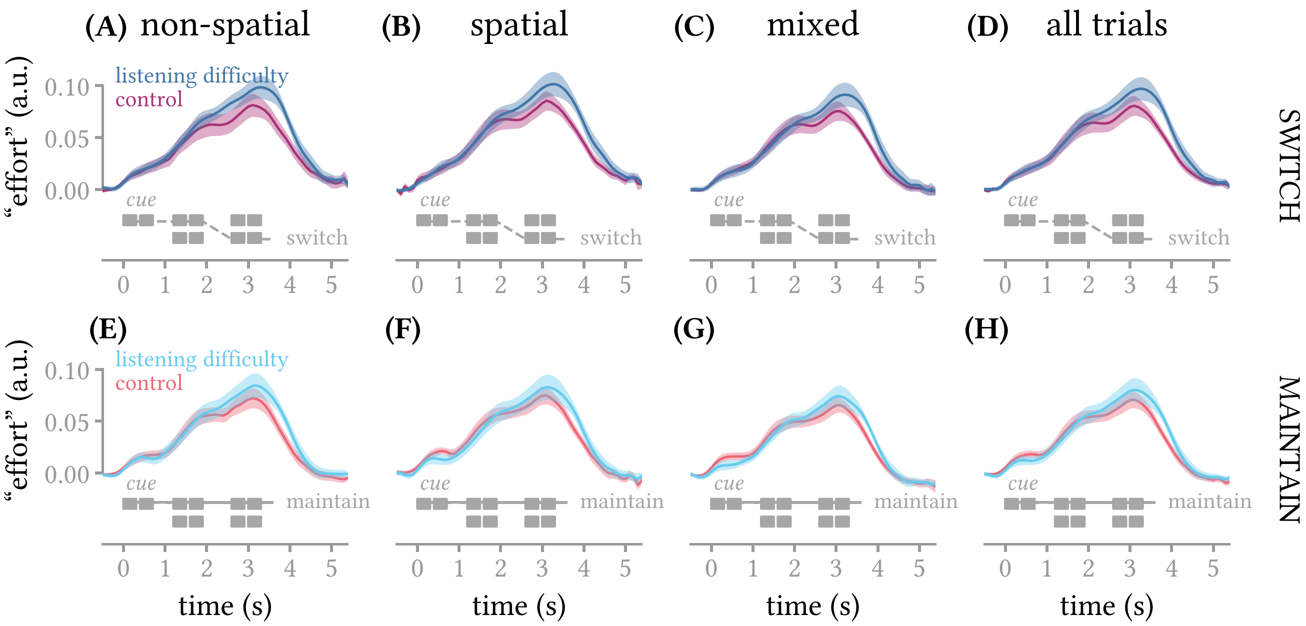

FIG. 4. (Color online) Mean across subjects (±1 standard error of the mean) of deconvolved pupillary response (effort) for listening difficulty versus control groups, in switch- (top row) and maintain-attention trials (bottom row), in the three experimental conditions and pooled across all conditions (columns). In each subplot there were no temporal spans where the difference between groups was found to be significantly non-zero with a false discovery threshold of 0.05.

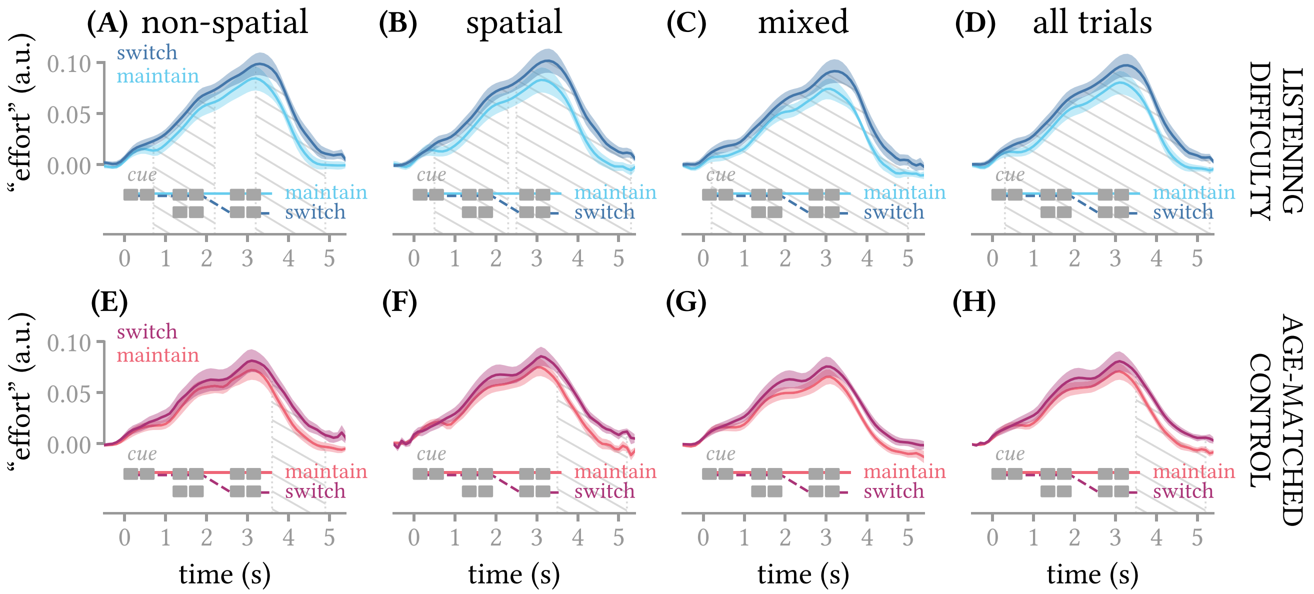

FIG. 5. (Color online) Mean across subjects (±1 standard error of the mean) deconvolved pupillary response (effort) for maintain- versus switch-attention trials in subjects with self-reported listening difficulty (top row) and control subjects (bottom row), in the three cue conditions (left 3 columns) and pooled across all conditions (rightmost column). Hatching indicates temporal spans where the curves in each subplot differ significantly.

D. Summary of results

We compared behavioral performance (response accuracy and reaction time) and pupillometry in an attention switching task in audiometrically normal listeners who self-report listening difficulty, and in age-matched controls. Overall, behavioral differences between participant groups were limited: participants with listening difficulty were more likely to respond to non-target non-foil items in general, more likely to respond to foils in the spatial condition, and tended to have slower response times in the final timing slot of non-spatial switch-attention trials. Other behavioral results (not related to listener group) recapitulated past findings from McCloy *et al.* (2017): switching attention leads to slower response times after the switch and also reduced accuracy; trials with only one stream segregation cue led to reduced accuracy compared to trials with two cues; and trials with only talker gender cues led to slower responses in the latter half of the trial. In terms of pupillary response, there was no statistically significant difference between listener groups when compared directly, but within-subject differences between the maintain- and switch-attention conditions did show a larger difference for participants with listening difficulty than for age-matched controls.

E. *Post hoc* analyses

One obvious question that arises from these results is whether the magnitude of pupillary response—or the difference in response between switch- and maintain-attention trials—reflects a supra-threshold deficit in auditory spatial abilities. To answer this question, we compared summary measures for each subject's pupillary response in the maintain- and switch-attention trials against each subject's SSQ component scores and against a range of behavioral measures representing binaural health (see Sec. II E). None of the summary pupillary measures were significantly correlated with any of the binaural health measures, or with the component scores on the SSQ assessment. Moreover, none of the summary pupillary measures showed a significant difference between participant groups in paired-samples *t*-tests (see Table I).

Nevertheless, the behavioral and pupillometric results of this study suggest that there is a meaningful difference between populations defined solely on the basis of two binary self-report questions regarding spatial hearing and crowded listening environments. The fact that those differences are uncorrelated with measures of auditory spatial abilities may indicate that the pupillary response reflects some other (non-spatial) aspect of auditory system function, or that the relationship between the pupillary response and the various measures of binaural health is sufficiently subtle as to require a much larger sample of listeners to detect.

IV. DISCUSSION

This study compared behavioral and pupillometric measures between groups of audiometrically normal subjects who either do or do not report difficulty localizing sound sources and/or understanding speech in reverberant or

TABLE I. Paired samples *t*-tests of summary pupillometry measures by listener group (listening difficulty subjects versus age-matched controls). AUC = area under the pupil response curve.

Measure	*t*	*p*
AUC: switch	1.19	0.259
AUC: maintain	0.965	0.3551
AUC: switch–maintain	1.172	0.266
peak ampl.: maintain	0.681	0.5102
peak ampl.: switch	1.195	0.2573
peak latency: maintain	0.949	0.363
peak latency: switch	1.076	0.3051

acoustically crowded environments. Specifically, we compared behavioral performance (response accuracy and reaction time) and pupillometry in an attention switching task, with three conditions varying the available stream segregation cues. The goal was to determine whether pupillometry could serve as an objective measure to index self-reported listening difficulty.

Overall, we saw a significant pupillary response difference between switch- and maintain-attention trials in the listening difficulty group but not in the control group. We also saw behavioral differences between groups in certain conditions, though there was no clear "main effect of group" on response accuracy or reaction time. Together, the behavioral and pupillometric findings suggest that the difference between groups was a rather subtle one—not a terribly surprising finding given that all participants were audiometrically normal and individually age-matched.

The apparent failure to reproduce the pupillometric results of McCloy *et al.* (2017) (i.e., the finding that pupil dilation is greater on switch-attention trials, and begins to diverge from maintain-attention trials as soon as the pre-trial cue is heard) in the control population is less worrisome than it might seem: the prior study involved stimulus degradations that may have increased difficulty in the switch-attention trials more strongly than the maintain-attention trials, either by specifically impacting the ability to re-deploy selective attention after a switch, or by raising the baseline effort level such that even listeners with normal hearing were overtaxed in the switch-attention task. In contrast, the present study involved no stimulus degradations, so the more modest difference in pupillary response between switch- and maintain-attention trials in the control population may mean that switching attention is simply not especially effortful for those listeners (which accords with their negative responses to the screening questions about difficulty with spatial hearing and crowded auditory environments).

It is noteworthy that the *post hoc* group comparison *t*-tests did *not* find a significant difference between groups based on the various summary measures of the pupillary response (peak amplitude, peak latency, or AUC; cf. Table I), even when comparing the switch-minus-maintain AUC values (i.e., each subject's gap between the average pupillary response to switch-attention trials versus maintain-attention trials). At first glance this would seem to contradict the difference between groups seen in Fig. 5. However, the statistical test illustrated in Fig. 5 shows that, among subjects with listening difficulty, the average size of the gap is big enough to conclude that it is statistically non-zero throughout most or all of the trial, whereas in the control subjects the average size of the gap is generally not big enough to draw that conclusion. In contrast, the *post hoc* *t*-test of AUC values looks at total gap size for each subject and compares the distribution of gap sizes between the two groups, which is more closely analogous to the statistical comparisons shown in Fig. 4.

Also of interest is the temporal pattern of differences in the pupillary responses, especially regarding the difference between maintain- and switch-attention trials (Fig. 5). Specifically, in the listening difficulty group, we see a replication of previous findings (McCloy *et al.*, 2017) showing that the difference in pupillary response between maintain- and switch-attention trials begins when the pre-trial cue is heard, suggesting that the pupillary response in switch-attention trials reflects preparation or planning for upcoming attention switches. Interestingly, in the spatial and mixed conditions, there is some indication of acausality in the pupillary response: the divergence of the maintain- and switch-attention pupil traces appears to occur during or even slightly before the cue letter that indicates the attentional condition for each trial. It is likely that the acausality is due to the parameters of the deconvolution kernel, which suggests that future studies using this technique may benefit from estimating kernel parameters separately for each subject, rather than relying on published parameter estimates (cf. discussion in McCloy *et al.*, 2016). It is also noteworthy that the pupil traces for the control subjects seem to show faster recovery [i.e., a more pronounced dip in pupillary response between the two halves of the trial; this is easiest to see in Figs. 4(A)–4(D)]. The significance of this observation is not known and merits further investigation.

Taking a wider perspective, many questions remain about supra-threshold auditory deficits, the pupillary response, and the experience of listening difficulty or effort. There is some precedent in the literature for a relationship between pupillary response and self-reported listening effort (e.g., Zekveld *et al.*, 2011), though in most cases self-reported measures are taken in the same session as the behavioral and physiological measures (in some cases immediately after each trial block). Thus it is perhaps a little surprising that two simple questions regarding listening difficulty were sufficient to yield detectable group differences in both behavioral and physiological responses in our data, especially given that the participants in the listening difficulty group were all asked the questions during initial screenings more than 6 weeks prior to their performance of the experimental task. That said, the questions were quite specific to sound localization and listening in multitalker environments, which may have increased their diagnostic utility over more open-ended self-report measures. Screening questions such as these may prove to be valuable tools for future studies.

Still, an important question remains regarding the causes of listening difficulty. If all we had was listeners' self-report, we would not know whether an objective physiological difference underlied listeners' assessments of their own abilities. What the current study shows is that there are indeed behavioral and physiological differences to be found that are associated with differences in self-assessment. However, the current study does not establish that those physiological differences are in fact the common cause driving both differences in behavior and differences in self-assessment: given that pupil dilation reflects a wide range of task-related variables such as memory demands (Taylor, 1981), mathematical complexity (Hess and Polt, 1964), stimulus degradation (Winn *et al.*, 2015), and various linguistic properties of words or sentences (Papesh and Goldinger, 2012; Zekveld *et al.*, 2010), it is entirely possible that a listener's *belief* that they are bad at certain kinds of listening tasks might sufficiently increase their arousal during such tasks as to elevate the magnitude of their evoked pupil response. At the same time, low self-assessment of listening abilities could likewise reduce behavioral performance in psychophysics tasks, analogous to the effects of internalized stereotypes on test performance (e.g., Riciputi and Erdal, 2017; Spencer *et al.*, 1999; Steele and Aronson, 1995). Put another way, we cannot tell definitively whether differences in self-assessment of listening abilities *cause* both the behavioral and pupillometric differences, or whether differences in self-assessment *result from* observed differences in behavioral ability, which in turn may arise from physiological differences. Clearly, further work is needed to refine our understanding of the experience of effortful listening.

ACKNOWLEDGMENTS

Portions of this work were supported by the National Institutes of Health (NIH) Grant No. R01-DC013260 to A.K.C.L., Grant No. T32-DC005361 to the University of Washington, and NIH LRP awards to E.L. and D.R.M. The authors are grateful to Ross Maddox for providing questionnaire responses and binaural health measures on the participants in this study.

[1] See supplementary material at https://doi.org/10.1121/1.5078618 for model summary tables.

Bharadwaj, H. M., Verhulst, S., Shaheen, L., Liberman, M. C., and Shinn-Cunningham, B. G. (**2014**). "Cochlear neuropathy and the coding of supra-threshold sound," Front. Syst. Neurosci. **8**, 26.

Cole, R. A., Muthusamy, Y., and Fanty, M. (**1990**). "The ISOLET spoken letter database," CSETech 205, 1-8, available at https://digitalcommons.ohsu.edu/csetech/205 (Last viewed November 5, 2018).

DeCarlo, L. T. (**1998**). "Signal detection theory and generalized linear models," Psychol. Methods **3**(2), 186–205.

Gatehouse, S., and Noble, W. (**2004**). "The speech, spatial and qualities of hearing scale (SSQ)," Int. J. Audiol. **43**(2), 85–99.

Hess, E. H., and Polt, J. M. (**1964**). "Pupil size in relation to mental activity during simple problem-solving," Science **143**(3611), 1190–1192.

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., and Kramer, S. E. (**2012**). "Processing load induced by informational masking is related to linguistic abilities," Int. J. Otolaryngol. **2012**, 865731.

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

Kumnick, L. S. (**1954**). "Pupillary psychosensory restitution and aging," J. Opt. Soc. Am. **44**(9), 735–741.

Larson, E. D., and Lee, A. K. C. (**2013**). "Influence of preparation time and pitch separation in switching of auditory attention between streams," J. Acoust. Soc. Am. **134**(2), EL165–EL171.

Maris, E., and Oostenveld, R. (**2007**). "Nonparametric statistical testing of EEG- and MEG-data," J. Neurosci. Meth. **164**(1), 177–190.

McCloy, D. R., Larson, E. D., Lau, B., and Lee, A. K. C. (**2016**). "Temporal alignment of pupillary response with stimulus events via deconvolution," J. Acoust. Soc. Am. **139**(3), EL57–EL62.

McCloy, D. R., Lau, B. K., Larson, E., Pratt, K. A. I., and Lee, A. K. C. (**2017**). "Pupillometry shows the effort of auditory attention switching," J. Acoust. Soc. Am. **141**(4), 2440–2451.

McCloy, D. R., and Lee, A. K. C. (**2015**). "Auditory attention strategy depends on target linguistic properties and spatial configuration," J. Acoust. Soc. Am. **138**(1), 97–114.

Papesh, M. H., and Goldinger, S. D. (**2012**). "Pupil-BLAH-metry: Cognitive effort in speech planning reflected by pupil dilation," Atten. Percept. Psychophys. **74**(4), 754–765.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., and Wingfield, A. (**2016**). "Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL)," Ear Hear. **37**, 5S–27S.

Riciputi, S., and Erdal, K. (**2017**). "The effect of stereotype threat on student-athlete math performance," Psychol. Sport Exerc. **32**, 54–57.

Ruggles, D., Bharadwaj, H., and Shinn-Cunningham, B. G. (**2011**). "Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication," Proc. Natl. Acad. Sci. U.S.A. **108**(37), 15516–15521.

Satterthwaite, F. E. (**1946**). "An approximate distribution of estimates of variance components," Biometrics Bull. **2**(6), 110–114.

Sheu, C.-F., Lee, Y.-S., and Shih, P.-Y. (**2008**). "Analyzing recognition performance with sparse data," Behav. Res. Meth. **40**(3), 722–727.

Shinn-Cunningham, B. G., Kopco, N., and Martin, T. J. (**2005**). "Localizing nearby sound sources in a classroom: Binaural room impulse responses," J. Acoust. Soc. Am. **117**(5), 3100–3115.

Singmann, H., Bolker, B., Westfall, J., and Aust, F. (**2018**). "Afex: Analysis of Factorial Experiments," https://CRAN.R-project.org/package=afex (Last viewed November 5, 2018).

Spencer, S. J., Steele, C. M., and Quinn, D. M. (**1999**). "Stereotype threat and women's math performance," J. Exp. Soc. Psychol. **35**(1), 4–28.

Steele, C. M., and Aronson, J. (**1995**). "Stereotype threat and the intellectual test performance of African Americans," J. Pers. Soc. Psychol. **69**(5), 797–811.

Taylor, J. S. (**1981**). "Pupillary response to auditory versus visual mental loading: A pilot study using super 8-mm photography," Percept. Motor Skill. **52**(2), 425–426.

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

Winn, M. B., Wendt, D., Koelewijn, T., and Kuchinsky, S. E. (**2018**). "Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started," Trends Hear. **22**, 1–32.

Zekveld, A. A., Kramer, S. E., and Festen, J. M. (**2010**). "Pupil response as an indication of effortful listening: The influence of sentence intelligibility," Ear Hear. **31**(4), 480–490.

Zekveld, A. A., Kramer, S. E., and 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**(4), 498–510.