JSLHR

Research Article

Effects of Second Language Proficiency and Linguistic Uncertainty on Recognition of Speech in Native and Nonnative Competing Speech

Alexander L. Francis,[a] Laura J. Tigchelaar,[b] Rongrong Zhang,[c] and Adriana Zekveld[d,e]

Purpose: The purpose of this study was to investigate the effects of 2nd language proficiency and linguistic uncertainty on performance and listening effort in mixed language contexts.
Method: Thirteen native speakers of Dutch with varying degrees of fluency in English listened to and repeated sentences produced in both Dutch and English and presented in the presence of single-talker competing speech in both Dutch and English. Target and masker language combinations were presented in both blocked and mixed (unpredictable) conditions. In the blocked condition, in each block of trials the target–masker language combination remained constant, and the listeners were informed of both prior to beginning the block. In the mixed condition, target and masker language varied randomly from trial to trial. All listeners participated in all conditions. Performance was assessed in terms of speech reception thresholds, whereas listening effort was quantified in terms of pupil dilation.
Results: Performance (speech reception thresholds) and listening effort (pupil dilation) were both affected by 2nd language proficiency (English test score) and target and masker language: Performance was better in blocked as compared to mixed conditions, with Dutch as compared to English targets, and with English as compared to Dutch maskers. English proficiency was correlated with listening performance. Listeners also exhibited greater peak pupil dilation in mixed as compared to blocked conditions for trials with Dutch maskers, whereas pupil dilation during preparation for speaking was higher for English targets as compared to Dutch ones in almost all conditions.
Conclusions: Both listener's proficiency in a 2nd language and uncertainty about the target language on a given trial play a significant role in how bilingual listeners attend to speech in the presence of competing speech in different languages, but precise effects also depend on which language is serving as target and which as masker.

Interest in research on listening effort has increased considerably in recent years (McGarrigle et al., 2014; Pichora-Fuller et al., 2016). One of the dominant perspectives in this developing field is that understanding speech in difficult contexts depends on the application of limited resources such as selective attention or working memory (Pichora-Fuller et al., 2016; see also discussion by Strauss & Francis, 2017). As demand on these resources increases, performance decreases and subjective effort increases concomitantly, accompanied by increases in psychophysiological responses associated with cognitive effort, such as pupil dilation (Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014). Although the relationship between resource demand and effort is well established (Kahneman, 1973; Pichora-Fuller et al., 2016), less is known about what characteristics of a given effortful listening task actually give rise to the sensation of effort. Understanding the source(s) of listening effort in normally hearing listeners will ultimately contribute to the development of methods to alleviate effort for nonnative listeners and individuals with hearing or cognitive impairments that may greatly increase the effort of listening even under relatively benign circumstances (Krause, Kennedy, & Nelson, 2014; Schmidtke, 2016).

One of the most common methods for increasing listening effort in the laboratory is to ask listeners to repeat

[a]Department of Speech, Language & Hearing Sciences, Purdue University, West Lafayette, IN
[b]Universiteit Utrecht, the Netherlands
[c]Department of Statistics, Purdue University, West Lafayette, IN
[d]VU University Medical Center, Amsterdam, the Netherlands
[e]Linnaeus Centre, Linköping University, Sweden
Correspondence to Alexander L. Francis: francisa@purdue.edu

Editor-in-Chief: Frederick (Erick) Gallun
Editor: Daniel Fogerty
Received July 3, 2017
Revision received December 29, 2017
Accepted March 26, 2018
https://doi.org/10.1044/2018_JSLHR-H-17-0254

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

target speech heard in the presence of some sort of background noise, particularly masking speech. In these conditions, effort arises from the listeners' need to accommodate the complex interaction between properties of the target and masking speech. More specifically, a preponderance of recent research has begun to suggest that understanding speech in competing speech involves high-level linguistic processing and the application of resource-limited cognitive systems, such as working memory and attention (Arlinger, Lunner, Lyxell, & Pichora-Fuller, 2009; Pichora-Fuller et al., 2016; Rönnberg et al., 2013; though cf. Füllgrabe & Rosen, 2016). It is the engagement of these systems that, in turn, results in the well-documented sensation of effort (cf. Ohlenforst et al., 2017; Westbrook & Braver, 2016) and increases in physiological markers of effort such as pupil dilation (Koelewijn, Zekveld, Festen, & Kramer, 2012) that arise in adverse listening conditions (see also discussion by Strauss & Francis, 2017). However, although it is clear that something in the complex combination of target, masker, and listener characteristics contribute to listening effort, more research is necessary to identify how effort arises.

Understanding speech in the presence of competing speech is generally considered as a particular case of informational masking (Kidd & Colburn, 2017; Kidd, Mason, Richards, Gallun & Durlach, 2008). Properties of informational masking signals distract or detract in some way from processing the target speech in a manner that goes beyond simple acoustic interference (energetic masking). Work by Shinn-Cunningham and colleagues (e.g., Shinn-Cunningham, 2008; Shinn-Cunningham & Best, 2008) characterizes informational masking in terms of failure of distinct cognitive processing mechanisms. Specifically, informational masking may arise either (or both) when there is a failure in attentional mechanisms involved in stream segregation and/or object selection. It may also arise when listeners must resort to postperceptual mechanisms of inference and deduction to compensate for difficulties related to the ability to accurately process information from the target language (Shinn-Cunningham & Best, 2008). In other words, informational masking arises in contexts in which there is demand on either external attention (i.e., stream segregation, object selection) or (also) on internal attention (see terminological discussion by Strauss & Francis, 2017 and also similar concepts shown in Edwards, 2016, Figure 2; Wingfield & Tun, 2007, Figure 1). Although it is obvious that higher-level information such as linguistic knowledge must play a role in guiding internal attention, it is not yet clear whether or how linguistic knowledge might affect the direction of external attention or the degree to which these influences might interact.

Here we present the results of an experiment in which listeners repeated target sentences heard in the presence of a masking sentence. Targets were produced in two languages, native and nonnative, and masking sentences were also spoken in either the native or nonnative language, with all combinations of target and masker language heard by all participants. Our interest was primarily in the degree to which second language proficiency, as determined by a short standardized test of English grammar and vocabulary administered prior to beginning the listening task, might affect listeners' ability to segregate target from masking streams (i.e., external attention) under conditions of greater or lesser uncertainty about target and masker language.

Target–Masker Linguistic Similarity and Language Experience

Some of the first studies to demonstrate an effect of masker language did so by comparing speech recognition when the listener's native language was masked either by speech in the same language or in an unfamiliar language. Results showed clearly superior performance when stimuli were presented in unfamiliar language maskers over native language ones (Garcia Lecumberri & Cooke, 2006; Van Engen & Bradlow, 2007). This result was even observed when the unintelligible (unfamiliar language) masking speech was produced by the same talker in phrases with comparable spectrotemporal properties to those used for the intelligible (native language) masking speech (Freyman, Balakrishnan, & Helfer, 2001). Subsequent research suggested that part of the difficulty arose from the linguistic similarity of the target and masking languages, such that listeners experienced greater interference from maskers that were linguistically more similar to the language of the target stimuli (Calandruccio, Brouwer, Van Engen, Dhar, & Bradlow, 2013). Moreover, studies with bilingual listeners showed that such effects of masker intelligibility were likely gradient and depended at least in part on the listener's proficiency in the second language. For example, Calandruccio and Zhou (2014) studied so-called balanced bilingual listeners living in the United States who were equally proficient in both English and Greek. When tested on the perception of English targets, these listeners showed a comparable degree of interference from both English and Greek when either was used as a masker. In other words, both the dominant (English) and nondominant (Greek) native language can serve as equally distracting maskers for balanced bilinguals. In contrast, Van Engen (2010) showed that late learners of a second language (English) continued to experience greater interference from their native language (Mandarin Chinese) than from English, although they did show significant interference from English when listening to English targets in either English or Mandarin masking speech. In other words, listeners appear to experience more interference from maskers produced in languages they know well than in languages they have learned more recently (and presumably know less well). Finally, a recent study by Dai, McQueen, Hagoort, and Kösem (2017) suggests that at least part of the increase in distraction with increasing familiarity can be explained by learned intelligibility. In this study, they presented target speech in the context of noise-vocoded speech that was initially difficult to understand. They then trained listeners to better understand vocoded speech and showed that masking interference from that speech increased after training. Taken together, the results of these various studies suggest that the more familiar a

listener is with the language and/or acoustic properties of the masking speech affects, the more interference it causes.

In addition, however, proficiency in the target language also affects performance with respect to understanding the target language. In another study corroborating and extending their original target–masker similarity hypothesis, Brouwer, Van Engen, Calandruccio, and Bradlow (2012) played English targets to both Dutch and English listeners in both Dutch and English background speech (two-talker babble). They found that both groups did better when listening to English targets in a Dutch background than they did listening to English targets in an English background, supporting the hypothesis that performance should be better when target and masker are presented in different languages. However, the English listeners received a greater benefit from the target–masker language mismatch than did the Dutch listeners. One reason for this asymmetric interference may be that, for the English listeners, the Dutch masking speech was unfamiliar and unintelligible, but for the Dutch listeners, it was familiar and quite intelligible. In addition, Brouwer et al. (2012) played Dutch listeners Dutch targets in both Dutch and English background speech. In this case, Dutch listeners performed less well in the Dutch-in-Dutch condition than they did in the Dutch-in-English condition, again consistent with the target–masker mismatch hypothesis. However, Dutch listeners' benefit from target–masker mismatch was greater for the Dutch targets than it was for the English ones. In other words, Dutch listeners received more benefit from switching the masker from Dutch to English when listening to Dutch targets (their native language) than they did from switching the masker from English to Dutch when listening to English targets (their less proficient language).

Similarly, Kilman, Zekveld, Hällgren, and Rönnberg (2014) found that two groups of Swedish listeners (one with lower proficiency in English and one with higher proficiency) both performed better with Swedish (native) targets than with English (nonnative) ones and both performed better with English (nonnative) maskers than with Swedish (native) ones. However, the magnitude of the benefit from switching from Swedish to English maskers was much greater for the low English proficiency group, suggesting that they received even less interference from the English masker than did the high proficiency group. Moreover, with nonnative (English) targets, both the low and high proficiency groups showed no difference between Swedish and English masker effects, though the high proficiency group performed better overall than did the low one. This lack of a release from masking when the masker is in a nonnative as opposed to native language is not consistent with the findings of Brouwer et al. (2012) or Van Engen (2010) but accords well with those of Garcia Lecumberri and Cooke (2006). One possible explanation for this is that the speakers in the studies by Brouwer et al. (2012) and Van Engen (2010) were all highly proficient in their second language whereas those in the Garcia Lecumberri and Cooke (2006) study and the Kilman et al. (2014) study may have had more varied proficiency (see discussion by Kilman et al., 2014).

In summary, across many different studies there appears to be a significant reduction in informational masking in cases in which the target and masker are more distinct linguistically (see also Calandruccio et al., 2013) and also for cases in which the target speech is in the listeners' native language rather than a still familiar but nevertheless nonnative language, and these benefits interact (Brouwer et al., 2012). However, greater proficiency in a second language also appears to reduce listeners' ability to ignore information in a masking stream in that language. This suggests that the mechanism(s) by which listeners select target speech properties to attend to can also affect attention to properties of the masking speech.

Uncertainty and Expectation

All of these studies discussed thus far were carried out under conditions in which listeners knew which language the targets and maskers were in on any given trial, minimizing uncertainty about what properties of speech related to the stream to be attended to and what properties belonged to speech to be ignored. In the study of informational masking of nonspeech stimuli, uncertainty, especially about masker properties, has repeatedly been shown to affect performance (Kidd et al., 2008; Watson, 2005). In particular, speech-in-speech masking increases with increasing uncertainty about the semantic content of the masker (Brungart & Simpson, 2004; Freyman, Helfer, & Balakrishnan, 2007), suggesting again that listeners use linguistic–semantic information to facilitate stream segregation. However, no study of informational masking has yet examined the effect of uncertainty about the language of the target or masking speech. This may be an important question because, in bilingual speech communities, "code switching" (the use of multiple languages within a single utterance) is quite common. If code switching introduces linguistic uncertainty that prompts listeners to attend differently to bilingual speech than to monolingual, this may have implications for our understanding of what is and what is not effortful in typical speech communication contexts.

Research on bilingual speakers' perception of speech involving code switching suggests that listeners' expectations about whether they will be hearing speech in only one of their languages versus in more than one can affect processing demand. For example, Olson (2017) found that, when Spanish/English bilingual listeners heard utterances that contained both Spanish and English prior to a specific target word that could be in either English or Spanish, they were equally fast to process the target word regardless of language. However, if the precursor sentence was only in one language, then they were slower to respond if the precursor was in English (their second language) and the target was in Spanish (their first language) than vice versa. These results suggest that listeners changed their allocation of cognitive resources depending on their "language mode"

(Grosjean, 2001), that is, their expectations about whether they were likely to hear one language or more than one. In particular, the additional time required to switch from the second language to the first language in monolingual contexts is typically interpreted as reflecting a greater cognitive cost to disengage from the second language as compared to the first language (Olson, 2017). Because this cost disappears when listeners expect to hear both languages (in bilingual trials), we might predict that, in the present experiment, the expectation of hearing speech in two languages could improve the ability to disengage from and therefore perhaps ignore one of the two languages in multiple-language trials.

To the best of our knowledge, no study has examined the effect of uncertainty or expectation about the target or masker language on perception of speech in competing speech. However, taken together, the results of the previously cited studies suggest that listeners are likely to operate in a constantly bilingual mode in the mixed condition, applying the same degree of listening effort on all trials as they prepare to process or ignore either language on every trial. In this case, we might expect little or no effect of the blocked versus mixed manipulation on trials containing both languages because in both the blocked and mixed conditions listeners will expect two languages in every trial and so will be operating in a bilingual mode on each trial regardless of condition. However, we might expect single-language trials to be different in the mixed condition as compared to the blocked condition because in the mixed condition listeners are prepared for the possibility of hearing two languages of these trials but in the blocked condition listeners are focused on only one language. Whether the mixed-condition, single-language trials should be expected to be more or less effortful than those in the blocked condition depends on whether bilingual listening is more effortful than monolingual listening.

Pupil Dilation as a Measure of Effort

Here we aimed to more precisely identify factors that influence task demands in perception of speech in cross-linguistic competing speech. We examined two dependent measures: performance (speech reception thresholds [SRTs]) and pupil dilation, a psychophysiological measure linked to listening effort (i.e., demand on cognitive capacity during listening). Our goal of supplementing the behavioral SRT measure with a physiological measure was to permit the possibility of distinguishing between conditions exhibiting poor performance that results from listeners "giving up" on the task (cf. Eckert, Teubner-Rhodes, & Vaden, 2016; Hornsby, Naylor, & Bess, 2016; Strauss & Francis, 2017), which would result in a pattern of poor performance but low physiological evidence of effort and conditions that result in listeners simply not being able to perform as well despite trying harder (high effort with poor performance). Pupil dilation has long been used as a tool for investigating cognitive demand (Beatty, 1982; Kahneman, 1973) and has been shown to effectively index a variety of linguistic processing demands (Piquado, Isaacowitz, & Wingfield, 2010) and listening effort (Zekveld et al., 2014). It has also been linked to more specific cognitive mechanisms such as the application of selective attention (Wierda, van Rijn, Taatgen, & Martens, 2012) and working memory demand (Goldinger & Papesh, 2012; Koelewijn, Zekveld, Festen, Rönnberg, & Kramer, 2012), as well as more general psychological phenomena such as arousal and task demand (Kahneman, 1973) and engagement in the task (Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010; Kahneman, Peavler, & Onuska, 1968). Although research on how specific task- and listener-related properties affect pupil dilation is still ongoing, a variety of studies have demonstrated that pupil dilation indexes listening effort from a variety of causes, including masking noise (Zekveld, Kramer, & Festen, 2010) and competing speech (Koelewijn, Zekveld, Festen, & Kramer, 2012), as well as semantic processing (Winn, 2016), divided attention (Koelewijn, Shinn-Cunningham, Zekveld, & Kramer, 2014), and processing speech in multiple languages (Hyönä, Tommola, & Alaja, 1995). Thus, although the precise relationship between pupil dilation and listening effort is not entirely well defined, we expect pupil diameter to increase as a function of the cognitive demands of the listening task but make no specific claim as to the cognitive mechanism(s) being engaged to a greater degree.

Furthermore, in addition to examining peak pupil dilation, as is typical, we also attempt to take the overall shape of the pupil dilation curve into account (cf. Winn, 2016), albeit in a relatively constrained manner. Specifically, based on observation of the shape of the pupil dilation curve, we identify two regions of interest—the peak that is typically employed in analyses of pupil dilation (e.g., Zekveld et al., 2014) and also the period immediately preceding the onset of speech during which listeners are presumably formulating their response in the target language. Here we consider the trough or low point of the curve closest to the listener's response as reflecting the effort involved in formulating a response under the assumption that responses that are easier to formulate will permit a greater decrease in pupil dilation during this period than will responses that are more cognitively demanding to formulate.

Summary

Previous research suggests that bilingual listeners' ability to recognize speech in competing speech is likely to be affected by the linguistic similarity of the two languages and their relative fluency in each language. Here we also investigate the possibility that language mode, prompted by the level of uncertainty about which language the target sentence will be produced in, may also affect processing demands (listening effort). In particular, we are interested in whether there is a processing cost (i.e., greater effort applied) to listening to speech in a bilingual mode. In order to make this comparison, we compare conditions in which listeners must be operating in a bilingual mode but yet derive no advantage from doing so (i.e., when the target and

masker are in the same language) with those in which listening bilingually is advantageous (target and masker language differ) and with those in which listeners are operating in a more strictly monolingual mode (certain that the target and masker will be in the same language).

Method

Participants

Seventeen native speakers of Dutch (11 women, six men, aged 20–41) who also considered themselves moderately proficient in English participated in this experiment and were paid for their time under a protocol approved by the ethics committee of the VU University Medical Center, Amsterdam. Participants were sought from the medical center and surrounding community with the goal of obtaining a sample population with moderate to high proficiency in English as a second language (comparable to that of the studies of Brouwer et al., 2012, or Van Engen, 2010). None reported any history of speech, language, hearing, cognitive, or neural disorders, and all exhibited pure-tone thresholds in both ears ≤ 20 dB HL at the octave frequencies between 500 and 8000 Hz. Pupil dilation data from three participants (two men, one woman) were unreliable, and thus, these were excluded from further analyses. The 14 remaining participants' English competence varied, with scores on the English Comprehension Test from the Swedish National Agency for Education (Kilman et al., 2014) ranging from 3 to 11 out of 12 (M = 9.36, median = 10.0, SD = 2.08). This is a test of general written English comprehension employing short answer questions to assess overall proficiency in English of nonnative speakers. Because the one person who scored a 3 appeared to be an outlier (all other scores were above 7), data from this person were also excluded from further analyses, resulting in a group with a mean English test score of 9.85 (median = 10.0, SD = 1.03). Education levels also varied: Four had completed an intermediate vocational education, four had a higher-level vocational education, and five had completed a university education. To assess Dutch competence, participants read a passage from the Transcriptions of Listening Test 4L/5S from the Dutch Office of Intercultural Evaluation and then answered written comprehension questions about it (see Zekveld, Kramer, Kessens, Vlaming, & Houtgast, 2009, for description of materials). Scores ranged from zero to six errors out of 13 possible. These were converted to number of correct items ranging from 7 to 13 for subsequent analyses (M = 10.85, median = 11.0, SD = 1.72). Age and test scores are shown in Table 1. A comparison of correlations between each of these subject-specific variables showed that English score was significantly correlated with Dutch score, Pearson's r = .60, $t(11)$ = 2.46, p = .031, 95% CI [0.07, 0.86], but not with age, r = −.135, $t(11)$ = −0.45, p = .65, 95% CI [−0.63, 0.45], or education level, r = −.173, $t(11)$ = 0.58, p = .571, 95% CI [−0.66, –0.42].

Table 1. Relevant demographic properties of participants.

Measurement	Age (years)	English test score (correct out of 12)	Dutch test score (correct out of 13)
Mean	31.1	9.8	10.8
SD	6.78	1.03	1.72
Range	20–41	7.5–11	7–13

Stimuli

Stimuli consisted of recordings of 222 meaningful but emotionally neutral sentences in Dutch (Versfeld, Daalder, Festen, & Houtgast, 2000) and English (HINT sentences; Nilsson, Soli, & Sullivan, 1994) spoken by two adult male native speakers of Dutch and English, respectively. We also recorded new versions of both sets of sentences produced by male adults who had roughly similar timbre as judged impressionistically by the experimenters. Mean f0 and f0 ranges were also relatively similar across the four talkers (original and new English 108 vs. 109 Hz; original and new Dutch 109 vs. 139 Hz based on a random selection of four sentences from each talker). The original stimuli were always used as targets, whereas the maskers were always drawn from the newly recorded sets. Stimuli were all recorded and presented at 44.1 kHz, but the original English stimuli appear to have been previously up-sampled, possibly from an original sampling rate of 20 kHz as they contained no energy above about 10.5 kHz. All stimuli were root-mean-square amplitude normalized before presentation.

Procedure

Participants were recruited using Dutch language materials, and the entire study was conducted in Dutch by a native speaker of Dutch who was also very fluent in English. Participants completed hearing and language tests prior to starting the adaptive SRT testing. During SRT testing, the size and location of the participant's left pupil were recorded at a sampling rate of 60 Hz using an SMI iView X RED remote eye-tracking system mounted below a computer screen situated approximately 45 cm directly in front of the participant. Auditory stimuli were presented via headphones (Sony MDR-V900). The eye tracker was calibrated for each individual at the start of the test session. Throughout the experiment, participants were not restrained but were seated in a fixed chair and were instructed to hold their head still against a back rest, to maintain focus on a fixation cross shown on a gray screen, and to blink as little as possible during each listening block. The experimenter continuously monitored a real-life recording of the data and the eyes, allowing immediate corrective action in case of a drop in the data quality due to, for example, excessive blinking or movement. Illumination in the room was set for each participant individually, such that resting pupil dilation was approximately at the midpoint of its dynamic range between bright light (~100 lx)

and near-complete darkness (completely dark except for the residual illumination provided by the turned-off screen and indicator lights on the computers).

Participants completed seven blocks of trials, two without masking (Dutch and English unmasked conditions), four with a consistent target and masking language (Dutch target sentences embedded in a Dutch masking talker, Dutch targets in an English masker, English in English, and English in Dutch), and one mixed block in which the target and masking languages varied randomly from trial to trial. Blocks were separated by a short rest period to allow subjects to relax briefly. There were 24 trials in each of the unmasked and consistent-masker blocks and 60 in the mixed block (to allow for 15 trials of each combination of target and masker language). Each of the trials with masking speech started with 3 s of masking speech only, allowing the participant to recognize the speech to be ignored prior to the onset of target speech and ensuring that the pupillary response to the target speech was not unduly influenced by any orienting-like response to the onset of stimulation. The masking speech then continued 4 s beyond the end of the target sentence. The duration of the target sentence ranged between 1.2 and 2.7 s, such that the entire trial duration ranged from 8.2 to 9.7 s. Participants were instructed to repeat the target sentence or as much of it as they were able to hear as soon as the sound (masker) stopped. Therefore, responses began at least 4 s after the end of the target speech.

Signal level in the unmasked conditions or signal-to-masker ratio (SMR; measured in dB) in the masked conditions was varied adaptively according to a method of adjustment (Levitt, 1971) to achieve an eventual performance of approximately 50% correct (with a trial scored as correct only if participants repeated all words in the sentence correctly in order) using a one-up/one-down adjustment scheme in order to ensure that all participants were performing at roughly equivalent levels of difficulty. Each combination of target and masker language constituted one "run" from which SRT could then be calculated (the four runs consisting of each combination of target and masker language were interleaved in the mixed condition). Overall signal level for the combined masking and target speech was fixed at 65 dB SPL. The starting SMR was −15 dB for all masked speech conditions, and the starting level was 35 dB SPL for the unmasked conditions. At the start of each block, the first sentence was repeatedly presented with increasing SMR or level (step size, 4 dB) until the participants were able to correctly repeat the entire sentence. After this individual starting level was obtained, subsequent sentences were presented in the adaptive procedure (step size, 2 dB SMR and 2 dB SPL for the masked and unmasked sentences, respectively).

Order of conditions was randomized across participants, except that the mixed block was always completed last. This was to ensure that listeners would have had experience with all possible combinations of target and masker language in the experimental context before starting the condition of higher uncertainty. The aim was to avoid confounding effects of uncertainty about target–masker language combination with inexperience with a particular combination. Because the total number of trials in the mixed condition was limited to prevent undue influence of fatigue, every mixed trial (15 of each combination of target and masker language) was intended to be included in the final analysis whereas the first nine trials in each of the blocked conditions could be discarded. Thus, had the mixed condition been presented first, on the first trial listeners would have had little experience with the task as well as being uncertain of what target and masker language would be presented. By choosing a fixed order, we did introduce the possibility of a different confound, namely that listeners might be more fatigued (and thus, the task might require greater effort) or might be more experienced with the task (and thus, the task might require less effort) by the time they came to the mixed condition.

Data Processing

Task performance was quantified as the level (unmasked conditions) or SMR (masked) obtained over the last 15 trials of each block (blocked conditions) or adaptive run (mixed condition). Pupil size was computed on each of these 15 trials in the interval starting 1 s prior to sentence onset (i.e., during the presentation of the masking speech for the masked speech trials) and ending at masker offset for the shortest sentence presented in the set (i.e., at least 5.2 s after target speech offset). Pupil diameters more than 3 *SD*s below the mean diameter of each trial were coded as a blink. Trials in which more than 15% of the samples were coded as blinks were excluded from data analysis. Eye blinks were replaced by linear interpolation starting four samples before and ending eight samples after a blink. The data were passed through a 5-point moving average smoothing filter and were then averaged over trials for each of the conditions. The mean pupil diameters in the first second of the interval were defined as baseline diameter. See Zekveld, Kramer, and Festen (2011) for a detailed description of the analysis. For all measures, blocked and mixed data were analyzed separately. Note that, due to the design of the experiment, there were no unmasked trials in the mixed condition.

Statistical Analyses

Data were analyzed using linear mixed effects modeling (lmer) implemented in nlm4 (v. 1.1-14; Bates, Maechler, Bolker, & Walker, 2015) within the R (v. 3.3.3) programming environment (R Development Core Team, 2017). Significance values were computed using the lmerTest package (v. 2.0-36; Kuznetsova, Brockhoff, & Christensen, 2017). In each case, we began with an initial maximal or nearly maximal model (Barr, Levy, Scheepers, & Tily, 2013) and then selectively removed interactions that did not contribute significantly to the model ($p > .05$) in order of highest to lowest number of factors in the interaction and greatest to smallest p value (rerunning the fitting after each removal), for example, starting with the four-way

interaction and progressing through the three-way and two-way interactions and finally to the individual factors until we arrived at an optimal model including only factors and interactions that contributed significantly to the model as well as those lower-order terms subsumed within higher-level interactions. Thus, when we report, for example, that the interaction between X and Y was a significant component of the model, it may be assumed that both X and Y remained in the model whether or not they were individually significant.

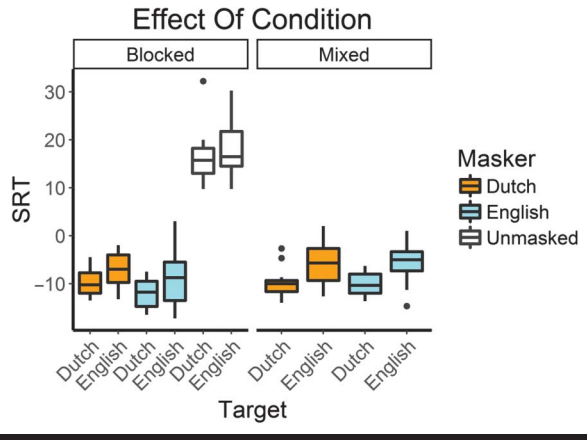
Results

Overall, results were broadly consistent with the expectation that listening effort, as indicated by pupil dilation and also SRTs, would be greater in the mixed condition than in the blocked condition and with English as compared to Dutch targets. English test score also affected responses, with individuals with higher scores generally performing better and/or exhibiting evidence of lower effort, but patterns of interaction were complex and must be examined separately.

SRTs

SRTs (SMR averages over the last 15 trials) for the blocked (leftmost four bars) and mixed conditions (rightmost four bars) are shown in Figure 1. Unmasked scores (thresholds for 50% correct recognition of speech in quiet, in dB SPL) are shown in the middle, unfilled bars. Note that smaller SRT and level values indicate better performance (lower SMR or absolute signal level required to achieve 50% correct). In quiet (middle two bars), participants were marginally better at recognizing Dutch stimuli (M = 16.21, SD = 5.69 dB SPL) as compared to English (M = 17.85, SD = 5.26 dB SPL) by paired t test, $t(12) = 2.12$, $p = .056$, 95% CI [−0.043, 2.927].

Figure 1. Speech reception thresholds (SRTs) for the eight masked conditions and signal-to-masker ratio for the two unmasked conditions (in dB SPL). Dark lines indicate medians; boxes indicate the interquartile range. Whiskers show maximum and minimum values, and circles indicate statistical outliers. Conditions with Dutch maskers are indicated in orange; English maskers are in blue.



In order to examine differences between the eight masked SRT conditions using linear mixed effects modeling, we started with a maximal model that included the categorical factors of condition (mixed vs. blocked [reference]), target (English, Dutch [reference]), and masker (English, Dutch [reference]), and the continuous (uncentered) variable of English test score. The random effects model included condition, target, masker, and subject. English test score was not included in the random model because it is continuous, and the model did not converge when it was included. Dutch test score was not included as either a fixed or random effect because it covaried with English test score. See Appendix A for the formal model description. The backward selection process (see Method above) resulted in an optimal model that included significant contributions (Table 2) from the main effects of condition, target, masker, and English test score and a significant interaction between condition and masker and between condition and English test score.

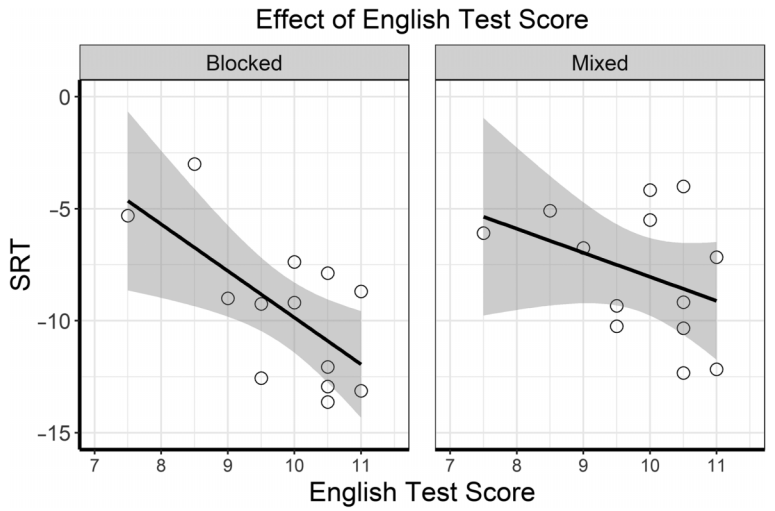
The main effects of condition, target, and masker are observable in Figure 1. The pattern shown in Figure 1 and the positive β_{Mixed} value suggest that listeners needed higher SMRs in the mixed (−7.87 dB, SD = 4.35) as compared to the blocked (−9.54 dB, SD = 4.26) condition. Similarly, trials with English targets (−6.98 dB, SD = 4.88) had higher SRTs (poorer performance) compared to those with Dutch targets (−10.43 dB, SD = 2.95), and trials with Dutch maskers (−8.14 dB, SD = 4.01) had higher SRTs (poorer performance) compared to those with English maskers (−9.26 dB, SD = 4.67). The interaction between condition and masker is not obvious in the figure but may be identified in the overall greater difference in SRTs for Dutch versus English maskers in the blocked condition (−8.47 dB, SD = 3.57 vs. −10.61 dB, SD = 4.69, respectively) as compared to the mixed condition (Dutch: −7.82 dB, SD = 4.45 vs. English: −7.92 dB, SD = 4.34).

Figure 2 shows the significant effect of the continuous variable English test score on SRT, separately by condition. As English test score increases (improves), SRT decreases (improves) in both conditions, suggesting that participants with greater English proficiency were better able to cope with less favorable SMR values. This effect

Table 2. Estimates of fixed effects based on the optimal model for speech reception threshold.

Fixed effects	S	*SE*	*df*	*t*	*p*
(Intercept)	7.20	4.67	21.57	1.54	.138
$Condition_{MIXED}$	−10.35	4.76	32.63	−2.17	.037
$Target_{ENGLISH}$	3.45	0.82	13.00	4.23	.001
$Masker_{ENGLISH}$	−2.13	0.74	30.47	−2.89	.007
English test score	−1.77	0.47	21.44	−3.75	.001
Condition × Masker	2.03	0.86	65.00	2.37	.021
Condition × English test score	1.12	0.48	32.37	2.33	.026

Figure 2. Correlation between English test score (note nonzero left axis) and mean speech reception threshold (SRT) score (dB SPL) for each participant (circles). Dark line indicates linear fit; gray ribbon indicates 95% confidence interval.



appears to be stronger in the blocked as compared to the mixed condition.

Pupil Dilation

Group Patterns

The mean, time-normalized pupil traces were fitted using linear mixed effects modeling including the fixed factors of target, masker, condition, and English test score, as well as first- through fourth-order time variables (i.e., time, time^2, time^3, time^4) and all interactions to allow for the possibility that the shape of the pupil dilation curve might vary across conditions or individuals. The random effects model included the factors of time, target, masker, and condition (see Appendix B for complete model description). Group means by condition are shown in the left two panels in Figure 3 (displayed by block), and the corresponding fitted curves from the model are shown in the right two panels. Curve fitting was used only to simplify the identification of peak and "trough" values (see below) for each participant in each condition, reducing the influence of trivial maxima or minima within individual curves.

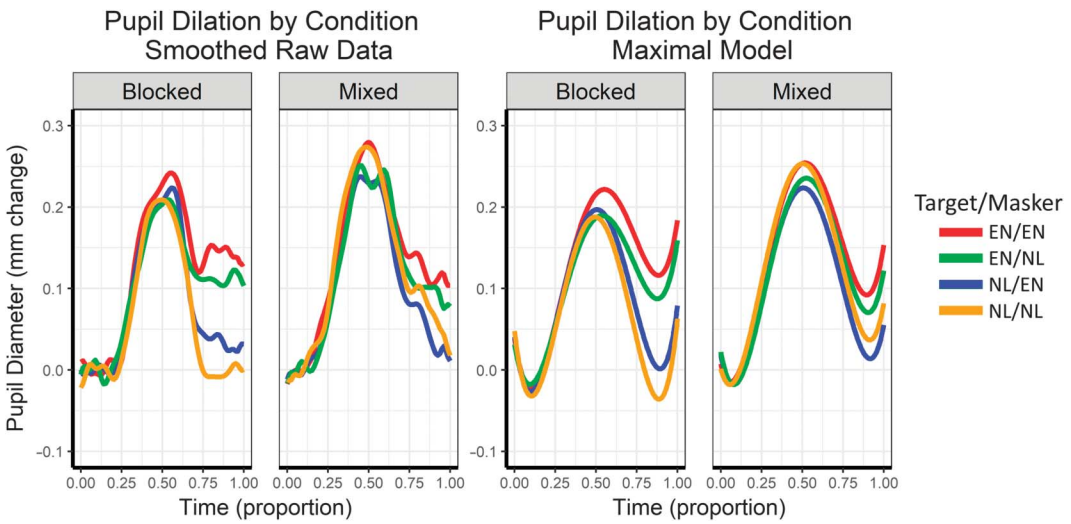
The first observation that may be made is that there appear to be two relatively distinct regions in which the pupil dilation curves differ noticeably from one another: the peak near the middle of the graph, presumably reflecting maximal demand on cognitive resources for target speech recognition, and the trough near the offset where pupil dilation is presumably being affected by the cognitive demand for processing of the recognized target speech as participants prepare to respond (Winn, 2016). In order to investigate the interactions between target, masker, English test score, and pupil dilation at these two locations, individual peak and trough values were extracted from the fitted curves for each individual subject in each combination of target and masker in each condition (blocked and masked).

Peak pupil dilation. Looking at the pattern of peak pupil dilation values from the individual fitted curves alone (Figure 4), it may be observed that the mixed condition exhibits higher peak pupil dilation overall, suggesting greater demand on cognitive processing, given that peak pupil dilation is typically associated with cognitive demand (Zekveld et al., 2014). However, because the mixed condition was always presented last, this could also reflect the influence of time in experiment (e.g., experience, fatigue, or some combination thereof). Looking at more specific differences between combinations of target and masker, in the blocked conditions the English-in-English curve appears to show a markedly greater peak pupil dilation whereas in the mixed condition both Dutch-in-Dutch and English-in-English curves appear to exhibit similarly higher peak pupil dilation values.

In order to test these observations statistically, a linear mixed effects model of fitted curve peak values was computed. Fixed effects of condition (blocked [reference] vs. mixed), target (English vs. Dutch [reference]), masker (English vs. Dutch [reference]), and English test score and all interactions of these factors were included, along with a random effect model consisting of condition, target, masker, and subject. Because all high-level interactions contributed significantly to the model, results from the full model were used. The full model specification and results of the model are shown in Appendix C.

Results of the pairwise (Tukey's honestly significant difference) comparisons showed that, in the blocked condition, peak pupil dilation in the English-in-English combination (0.22 mm, $SD = 0.074$) was significantly greater than in the English-in-Dutch combination (0.19 mm, $SD = 0.045$), $\beta_{Diff} = 0.03$, $SE = 0.005$, $p < .001$, but was not

Figure 3. Comparison of smoothed raw pupil dilation data (left two panels) and optimal fitted curves (right two panels) for masked trials in blocked (left panel in each pair) and mixed (right panels) conditions. EN = English; NL = Dutch.



different from the Dutch-in-Dutch (0.19 mm, SD = 0.057), β_{Diff} = 0.03, SE = 0.013, p = .222, or the Dutch-in-English (0.20 mm, SD = 0.068), β_{Diff} = 0.03, SE = 0.012, p = .458, combinations.

Similarly, in the mixed condition, peak pupil dilation for the English-in-English combination (0.26 mm, SD = 0.074) is significantly greater than the English-in-Dutch combination (0.24 mm, SD = 0.045), β_{Diff} = 0.019, SE = 0.005, p = .023, but not the Dutch-in-English combination (0.23 mm, SD = 0.078), β_{Diff} = 0.031, SE = 0.012, p = .249, or the Dutch-in-Dutch combination (0.25 mm, SD = 0.079), β_{Diff} = 0.001, SE = 0.013, p > .999. Peak pupil dilation for the mixed Dutch-in-Dutch combination differs significantly from the mixed Dutch-in-English combination, β_{Diff} = 0.029, SE = 0.005, p = .001, but not from the mixed English-in-Dutch combination, β_{Diff} = 0.017, SE = 0.012, p = .820.

Comparing combinations of target and masker across blocking condition, there is a significant increase in peak pupil dilation from blocked to mixed for the Dutch-in-Dutch combination (0.19 [0.057] to 0.25 [0.079] mm), β_{Diff} = 0.065, SE = 0.011, p = .001, and the English-in-Dutch combination (0.19 [0.045] to 0.24 [0.065] mm), β_{Diff} = 0.046, SE = 0.011, p = .0022, but not for the English-in-English combination (0.22 [0.074] vs. 0.26 [0.076] mm), β_{Diff} = 0.033, SE = 0.011, p = .158, or Dutch-in-English combination (0.20 [0.068] vs. 0.23 [0.078] mm), β_{Diff} = 0.027, SE = 0.011, p = .323.

Finally, although peak pupil dilation might be expected to be related to performance (i.e., SRT), computing the Pearson's product moment correlation between raw peak pupil dilation (in mm) and SRT (in dB) shows no significant correlations between SRT and fitted curve peak for all combinations of target and masker language in both blocked and mixed conditions (all ps > .10).

Pupil dilation trough. Looking at specific patterns of the lowest point of the trough near the offset (Figure 5) in both the blocked and mixed conditions, the two English target curves show a larger value consistent with the idea that these conditions require greater commitment of effort to prepare for upcoming speech production in that language (or perhaps preparing to respond in Dutch requires less effort). Again, because all high-level interactions contributed significantly to the model, results from the full model were used. The full model specification and results of the model are shown in Appendix D. Pairwise comparisons were calculated on individual trough values as for the peak measurements.

These results showed a significant (p < .001) increase in pupil dilation when changing the target from Dutch to

Figure 4. Peak pupil dilation from fitted curves for each participant for the eight masked conditions. Dark lines indicate medians; boxes indicate the interquartile range. Whiskers show maximum and minimum values, and circles indicate statistical outliers. Conditions with Dutch maskers are indicated in orange; English maskers are in blue.

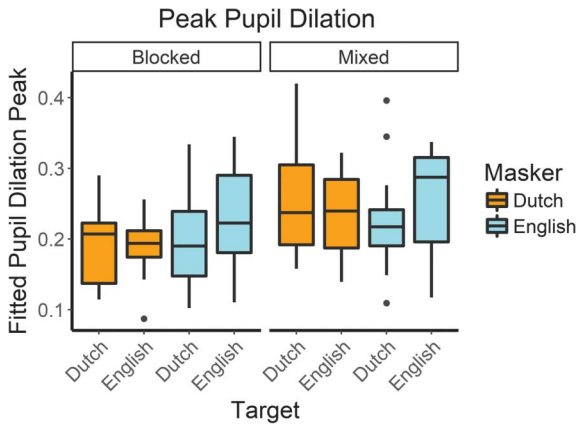
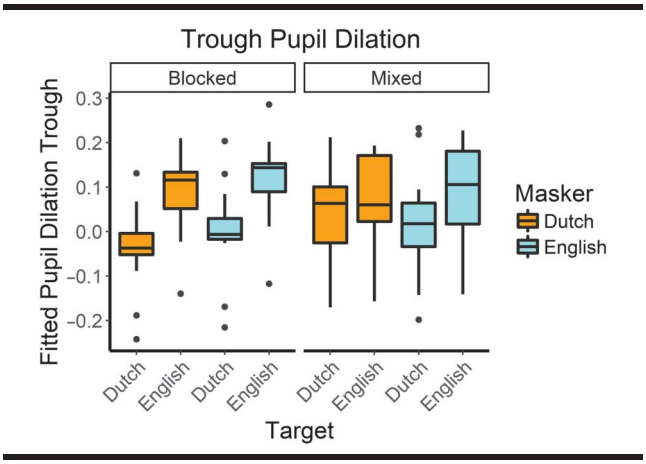


Figure 5. Pupil dilation at the trough from fitted curves for each participant for the eight masked conditions. Dark lines indicate medians; boxes indicate the interquartile range. Whiskers show maximum and minimum values, and circles indicate statistical outliers. Conditions with Dutch maskers are indicated in orange; English maskers are in blue.



English in the blocked condition and in one of the mixed conditions. Specifically, the three significant cases are blocked Dutch-in-Dutch ($M = -0.037$ mm, $SD = 0.100$) versus blocked English-in-Dutch ($M = -0.001$ mm, $SD = 0.109$), $\beta_{Diff} = 0.123$, $SE = 0.012$, $p < .001$, blocked Dutch-in-English versus blocked English-in-English ($M = 0.115$ mm, $SD = 0.101$), $\beta_{Diff} = 0.116$, $SE = 0.012$, $p < .001$, and mixed Dutch-in-English ($M = 0.013$ mm, $SD = 0.125$) versus mixed English-in-English ($M = 0.091$ mm, $SD = 0.119$), $\beta_{Diff} = 0.078$, $SE = 0.012$, $p < .001$. The only nonsignificant such comparison is the mixed condition Dutch-in-Dutch ($M = 0.036$ mm, $SD = 0.112$) versus English-in-Dutch ($M = 0.069$ mm, $SD = 0.111$), $\beta_{Diff} = 0.033$, $SE = 0.012$, $p = .184$. There is also a significant increase in trough value from blocked to mixed condition for the Dutch-in-Dutch combination, $\beta_{Diff} = 0.073$, $SE = 0.012$, $p = .001$. However, no other changes from blocked to mixed are significant ($p > .50$ for all three).

Finally, as for the peak pupil dilation scores, there were no significant correlations between trough pupil dilation values and SRT for any of the combinations of target and masker language in either the blocked and mixed conditions ($p > .08$ for all).

Effects of English Language Proficiency

To examine the interaction between target, masker, block, and English language proficiency in the pupil dilation data, we plotted the peak and trough values of the fitted curve for each combination of target and masker in both blocked and mixed conditions for each individual subject against English test score as shown in Figure 6 and compared them statistically using a Pearson's product moment correlation. The only significant correlations between peak pupil dilation and English test score were in the English-in-English combination (blocked condition), $r = .61$, $p = .028$, and in the Dutch-in-Dutch combination (mixed condition), $r = .58$, $p = .036$. In terms of the bottom of the trough (right two panels in Figure 6), the trend toward higher overall values as English proficiency increases is less obvious, and none of the correlations were significant ($p > .4$ for all).

Discussion

Overall, the present results suggest that both second language proficiency and linguistic uncertainty affect the performance and effort of bilingual listeners when processing speech in competing speech in a manner that interacts to some degree with the target and masker language.

Target and Masker Language

Listeners performed better on trials with Dutch as compared to English targets and with English as compared to Dutch maskers, although these factors did not interact with one another. This suggests that listeners were simply showing the expected effect of native versus nonnative language experience. Although many of the participants in this study were highly fluent in English, it was nevertheless their second language, and as predicted, this made it both harder to recognize and easier to ignore than their native language, Dutch.

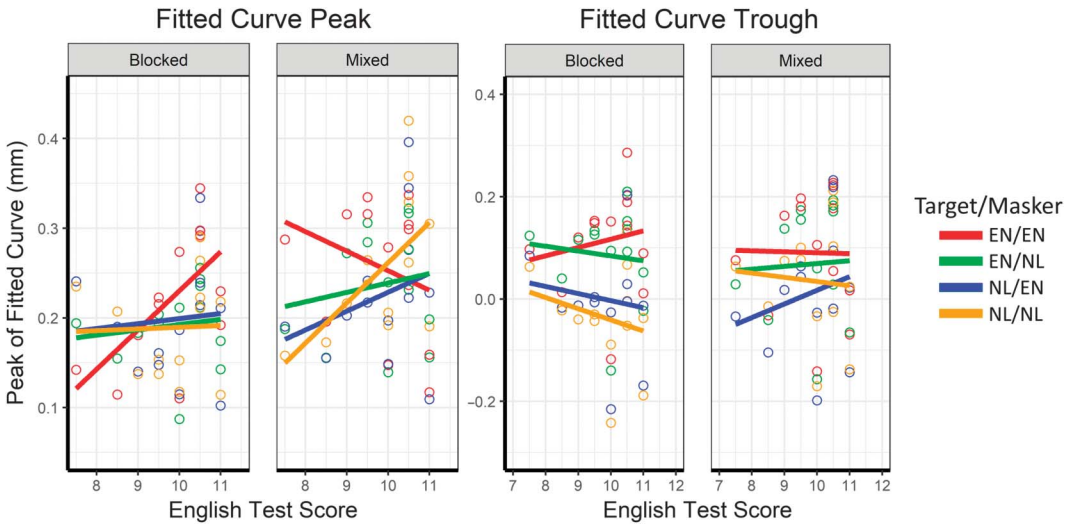
Considering the pupil dilation analyses provides more detailed insight. In particular, the English-in-English combination provoked greater peak pupil dilation as compared to the English-in-Dutch combination in both the blocked and mixed conditions, suggesting either that the Dutch masker was more effortful than the English or that when target and masker languages are matched effort is greatest, consistent with the hypotheses of Calandruccio and colleagues (e.g., Calandruccio et al., 2013). Both of these interpretations are further supported by the finding that, at least in the mixed condition, the Dutch-in-Dutch combination induced greater pupil dilation (hence presumably listening effort) than the Dutch-in-English combination.

English Proficiency

English proficiency showed significant effects in both the SRT and pupil dilation measures: Participants with higher English test scores tended to perform better, though there was no evidence of the expected interactions between target and masker and English test score. The lack of such interactions suggests either that the overall correlation between English test score and SRT simply reflects a broad tendency for people with better English test scores to perform better across the board (i.e., in all conditions) or perhaps that too few participants were included to permit expected interactions to reach significance.

Considering pupil dilation as a measure of listening effort provides some support for the hypothesis that English test score reflects something more general about these

Figure 6. Comparison of English test scores with peak pupil dilation data (left two panels) and pupil dilation at the trough (right two panels) for masked trials in blocked (left panel in each pair) and mixed (right panels) conditions for all subjects. Lines indicate linear fit. EN = English; NL = Dutch.



listeners: Pupil dilation did correlate with English test score in the English-in-English combination in the blocked condition, meaning that listeners with higher English test scores exhibited greater effort in this combination in this condition. This could indicate that they are more susceptible to interference from English language maskers, and this stronger interference outweighs any benefit of English proficiency for English target recognition. On the other hand, in the mixed condition, English test scores were significantly correlated with pupil dilation in the Dutch-in-Dutch combination (only). This might be taken as suggesting that English proficiency somehow also increases listeners' susceptibility to interference from Dutch maskers or reduces the ability to recognize Dutch targets. More likely, however, is that these results reflect the influence of some more basic variable, perhaps one that underlies the (presumably not directly causally related) correlation between Dutch and English proficiency scores. That is, the pattern of test scores in the two languages may simply reflect the influence of a single factor, perhaps a more general cognitive capacity such as IQ that also affects susceptibility to distraction. Alternatively, this correlation could be coincidental. Nevertheless, because it exists it could explain the pattern of overall better performance across conditions for individuals with higher English test scores. That is, because individuals who were more proficient in English were also coincidentally better in Dutch as well, these individuals exhibited better performance across the board. Further research with a larger number of participants with a greater range of linguistic capabilities and more extensive and sophisticated baseline testing would be necessary to disentangle these factors.

It is also possible, however, that the lower proficiency listeners in the present experiment were simply not trying as hard as the higher-proficiency listeners were, or were less engaged in the task overall, or simply had fewer cognitive resources to devote to the task, resulting in both a lower degree of physiological effort and poorer behavioral performance. Previous studies have linked smaller pupil dilations to lower proficiency (Ahern & Beatty, 1981; Koelewijn, Zekveld, Festen, Rönnberg, & Kramer, 2012; Kuchinsky et al., 2013; Wendt, Dau, & Hjortkjær, 2016; Zekveld & Kramer, 2014; Zekveld et al., 2011), consistent with the so-called cognitive resource hypothesis (cf. Zekveld et al., 2011, and references therein), and there is also the possibility that lower-proficiency listeners were simply "giving up" either because they found the task too difficult or because they felt less motivated (Eckert et al., 2016; Hornsby et al., 2016). However, there is insufficient evidence in the present data set to decide between either of these hypotheses, given that the only conditions in which lower-proficiency listeners tended to exhibit lower overall pupil dilation than those with higher proficiency were the blocked English-in-English and the mixed Dutch-in-Dutch cases.

It seems plausible, however, that participants in this study may have been aware that their English proficiency was a factor of interest. Although participants were not explicitly told how they performed on the English test that occurred immediately prior to the listening test until after they completed the entire experiment (and even then, only if they wished to hear it), they may have consciously or unconsciously adjusted some characteristics associated with their performance (e.g., arousal, motivation) based on their (self-assessed) confidence in their performance on the English pretest. Future research is necessary to pin down such effects, but at a minimum, research incorporating physiological measures of effort should take into account psychosocial variables such as participants' self-assessment of pretest performance or other factors that might affect not just performance itself but more individual, context-dependent variables such as motivation (see also discussion

by Hornsby et al., 2016; Pichora-Fuller et al., 2016; Strauss & Francis, 2017). One way to mitigate such effects would be to administer any proficiency tests only after the experimental task has been completed.

Uncertainty and Bilingual Mode Processing

Changing from blocked to mixed has a significant effect on both performance (SRTs) and pupil dilation, and this change interacts with effects of masker and English test score. In particular, the main effect of masker on SRT is driven almost exclusively by differences in the blocked condition. Essentially, it appears that the relative benefit of listening in the presence of an English masker instead of a Dutch masker is eliminated in the mixed condition, which may also be related to the effect of English proficiency on performance. Notably, the effect of English test score on SRT is stronger in the blocked condition (where there is a benefit to having an English masker as opposed to a Dutch masker) than in the mixed condition (where this benefit disappears). This suggests that nonnative language proficiency may contribute to the benefit that derives from being able to more easily ignore a nonnative as opposed to a native masker, but this benefit is attenuated in the mixed condition, even though the masker began before the target. This supports the interpretation that the mixed condition, by introducing linguistic uncertainty, may encourage listeners to operate in a more bilingual mode, reducing the benefit that might otherwise derive from being able to focus only on one language and ignore the other.

On the other hand, pupil dilation increases from the blocked condition to the mixed condition, presumably reflecting greater listening effort, for trials with both Dutch and English targets but only in the presence of a Dutch masker. If listeners are moving to a more bilingual model of listening in the mixed condition, it is possible that this could simultaneously reduce the performance benefit from listening to speech in the presence of a nonnative masker while also increasing the effort of listening to speech in the presence of a native masker.

Finally, the pattern of pupil dilation toward the end of the measurement period is consistent with the interpretation that pupil dilation during this time reflects effort related to preparing to speak in the target language, though perhaps more so when listeners are aware from the start that they will be responding in that language. In the blocked condition, pupil dilation (effort) is higher when the target language is English (i.e., Dutch-in-Dutch < English-in-Dutch, and Dutch-in-English < English-in-English). In the mixed condition, this pattern only obtains for the combination with an English masker (Dutch-in-English < English-in-English), suggesting perhaps that operating in a bilingual mode continues to provide some benefits to ignoring the effects of a nonnative masker when preparing to speak in one's native language.

Note, however, that it is still eminently plausible that time in experiment interacts in some complex way with target, masker, or both, producing the pattern of effects that we interpret here as being due to language mode and/or linguistic uncertainty. Such effects cannot be isolated in the present results because the design confounds time in experiment with condition. Further research specifically counterbalancing block type and timing would be necessary to definitively isolate these factors.

Conclusions

In summary, results of the analysis of performance (SRTs) suggest that both listener's proficiency in a second language and uncertainty about the target language (or language mode) may play a significant role in how listeners attend to speech in the presence of competing speech in different languages. Incorporating pupil dilation into the analysis as a measure related to listening effort indicates that the role of second language proficiency in the present case (at least) may be more complex than the performance data alone would suggest. In particular, results suggest that there may be differences in how second language proficiency interacts with the effects of target and masker language depending on whether we consider the effort involved in recognizing the target speech to begin with (here assessed in terms of peak pupil dilation) or examine effort related to preparing to speak in the target language (here assessed in terms of trough pupil dilation).

Overall, despite the inclusion of participants who differed in English proficiency and who may have differed in terms of their overall motivation to perform, group data as a whole suggest that listeners changed the way they attend to the auditory scene in the mixed block condition. We interpret this as a consequence of uncertainty from trial to trial about target and masker language, prompting a switch to a more bilingual mode of processing, though the present design does not allow us to completely rule out a simpler account related to time in experiment. Moreover, it also seems possible that the administration of language tests closely preceding speech perception testing may have contributed to the differences observed here. As is always the case with results showing a lack of effect, it is also possible that there were simply not enough participants to achieve statistical significance in specific conditions or analyses.

Thus, future research on this topic should include more participants with a wider range of proficiency in English and should include more comprehensive assessments of their linguistic and cognitive capabilities (cf. Kilman et al., 2014), although such testing should be performed after administering the primary speech perception task in order to avoid confounding speech perception performance with individual affective/motivational responses to perceived performance on the assessment task(s). Second, the broader pattern of findings presented here suggests the need for additional research to investigate specific mechanisms by which second language proficiency might affect not just processing of target speech but also the direction of selective attention (affecting inhibition of auditory streams consisting of speech in a particular language) and preparation

of response speech. With bi- and multilingualism as the norm in much of the world (and increasingly common even in the United States), there is a clear need to better understand how listening effort is affected by nonnative language proficiency and other cognitive and linguistic factors associated with multilingual communication. Finally, the present results also suggest that more research is necessary to determine how motivation and other psychosocial factors may affect performance on listening tests and, ultimately, in real-world contexts.

Acknowledgments

This research was partly funded by a Fellowship for Study in a Second Discipline from the Executive Vice President for Research, Purdue University, granted to Alexander L. Francis. We thank Hans van Beek for programming assistance and Matthew Winn for comments on an earlier draft of this manuscript.

References

Ahern, S., & Beatty, J. (1981). Physiological evidence that demand for processing capacity varies with intelligence. In *Intelligence and learning* (pp. 121–128). Boston, MA: Springer.

Arlinger, S., Lunner, T., Lyxell, B., & Pichora-Fuller, K. (2009). The emergence of cognitive hearing science. *Scandinavian Journal of Psychology, 50,* 371–384.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, B., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91,* 276–292.

Best, V., Ozmeral, E. J., & Shinn-Cunningham, B. G. (2007). Visually-guided attention enhances target identification in a complex auditory scene. *The Journal of the Association for Research in Otolaryngology, 8,* 294–304.

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America, 131*(2), 1449–1464.

Brungart, D. S., & Simpson, B. D. (2004). Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty. *The Journal of the Acoustical Society of America, 115*(1), 301–310.

Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S., & Bradlow, A. R. (2013). Masking release due to linguistic and phonetic dissimilarity between the target and masker speech. *American Journal of Audiology, 22*(1), 157–164.

Calandruccio, L., & Zhou, H. (2014). Increase in speech recognition due to linguistic mismatch between target and masker speech: Monolingual and simultaneous bilingual performance. *Journal of Speech, Language, and Hearing Research, 57*(3), 1089–1097.

Dai, B., McQueen, J. M., Hagoort, P., & Kösem, A. (2017). Pure linguistic interference during comprehension of competing speech signals. *The Journal of the Acoustical Society of America, 141*(3), EL249–EL254.

Eckert, M. A., Teubner-Rhodes, S., & Vaden, K. I., Jr. (2016). Is listening in noise worth it? The neurobiology of speech recognition in challenging listening conditions. *Ear and Hearing, 37,* 101S–110S.

Edwards, B. (2016). A model of auditory-cognitive processing and relevance to clinical applicability. *Ear and Hearing, 37,* 85S–91S.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *The Journal of the Acoustical Society of America, 109,* 2112–2122.

Freyman, R. L., Helfer, K. S., & Balakrishnan, U. (2007). Variability and uncertainty in masking by competing speech. *The Journal of the Acoustical Society of America, 121,* 1040.

Füllgrabe, C., & Rosen, S. (2016). On the (un)importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds. *Frontiers in Psychology, 7,* 1268.

Garcia Lecumberri, M. L., & Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America, 119*(4), 2445–2454.

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 252–269.

Goldinger, S. D., & Papesh, M. H. (2012). Pupil dilation reflects the creation and retrieval of memories. *Current Directions in Psychological Science, 21*(2), 90–95.

Grosjean, F. (2001). The bilingual's language modes. In J. Nicol (Ed.), *One mind, two languages: Bilingual language processing* (pp. 1–22). Malden, MA: Blackwell.

Hornsby, B. W., Naylor, G., & Bess, F. H. (2016). A taxonomy of fatigue concepts and their relation to hearing loss. *Ear and Hearing, 37,* 136S–144S.

Hyönä, J., Tommola, J., & Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology, A, 48*(3), 598–612.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Kahneman, D., Peavler, W. S., & Onuska, L. (1968). Effects of verbalization and incentive on the pupil response to mental activity. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 22*(3), 186–196.

Kidd, G., Jr., & Colburn, H. S. (2017). Informational masking in speech recognition. In J. C. Middlebrooks, J. Simon, A. Popper, & R. Fay (Eds.), *The auditory system at the cocktail party, Springer handbook of auditory research* (Vol. 60, pp. 75–109). New York, NY: Springer.

Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational masking. In W. A. Yost, A. N. Popper, & R. R. Fay (Eds.), *Auditory perception of sound sources* (pp. 143–190). New York, NY: Springer Science + Business Media.

Kilman, L., Zekveld, A., Hällgren, M., & Rönnberg, J. (2014). The influence of non-native language proficiency on speech perception performance. *Frontiers in Psychology, 5,* 651.

Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2014). The pupil response is sensitive to divided attention during speech processing. *Hearing Research, 312,* 114–120.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing, 33*(2), 291–300.

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., & Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *International Journal of Otolaryngology, 2012*. https://doi.org/10.1155/2012/865731

Krause, M. O., Kennedy, M. R., & Nelson, P. B. (2014). Masking release, processing speed and listening effort in adults with traumatic brain injury. *Brain Injury, 28*(11), 1473–1484.

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

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26.

Levitt, H. C. C. H. (1971). Transformed up–down methods in psychoacoustics. *The Journal of the Acoustical Society of America, 49*(2B), 467–477.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *International Journal of Audiology, 53,* 433–440.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America, 95*(2), 1085–1099.

Ohlenforst, B., Zekveld, A. A., Jansma, E. P., Wang, Y., Naylor, G., Lorens, A., … Kramer, S. E. (2017). Effects of hearing impairment and hearing aid amplification on listening effort—A systematic review. *Ear and Hearing, 38,* 261–281.

Olson, D. J. (2017). Bilingual language switching costs in auditory comprehension. *Language, Cognition and Neuroscience, 32*(4), 494–513,

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, W. Y., Humes, L. E., … Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing, 37,* 5S–27S.

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology, 47*(3), 560–569.

R Development Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rönnberg, J., Lunner, T., Zekveld, A. A., Sörqvist, P., Danielsson, H., Lyxell, B., … Rudner, M. (2013). The ease of language understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience, 7,* 31.

Schmidtke, J. (2016). The bilingual disadvantage in speech understanding in noise is likely a frequency effect related to reduced language exposure. *Frontiers in Psychology, 7*(678), 1–15.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12*(5), 182–186.

Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends in Amplification, 12,* 283–299.

Strauss, D. J., & Francis, A. L. (2017). Toward a taxonomic model of attention in effortful listening. *Cognitive, Affective and Behavioral Neuroscience, 17*(4), 809–825.

Van Engen, K. J. (2010). Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. *Speech Communication, 52*(11), 943–953.

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America, 121*(1), 519–526.

Versfeld, N. J., Daalder, L., Festen, J. M., & Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *The Journal of the Acoustical Society of America, 107*(3), 1671–1684.

Watson, C. S. (2005). Some comments on informational masking. *Acta Acustica United With Acustica, 91*(3), 502–512.

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in psychology, 7,* 345.

Westbrook, A., & Braver, T. S. (2016). Dopamine does double duty in motivating cognitive effort. *Neuron, 89*(4), 695–710.

Wierda, S. M., van Rijn, H., Taatgen, N. A., & Martens, S. (2012). Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proceedings of the National Academy of Sciences, 109*(22), 8456–8460.

Wingfield, A., & Tun, P. A. (2007). Cognitive supports and cognitive constraints on comprehension of spoken language. *Journal of the American Academy of Audiology, 18*(7), 548–558.

Winn, M. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends in Hearing, 20,* 1–17.

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage, 101,* 76–86.

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology, 51*(3), 277–284.

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

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

Zekveld, A. A., Kramer, S. E., Kessens, J. M., Vlaming, M. S., & Houtgast, T. (2009). User evaluation of a communication system that automatically generates captions to improve telephone communication. *Trends in Amplification, 13*(1), 44–68.

Appendix A

Speech Reception Threshold Model and Results

The model for the speech reception threshold analysis was written as follows (in standard R notation):

```
model = lmer(SRT ~ Condition*Target*Masker*EnglishTestScore +
(Condition + Target + Masker | Subject)
```

Results of the full model of speech reception threshold scores.

Fixed effects	Estimate	*SE*	*df*	*t*	Pr(>|*t*|)
(Intercept)	9.6439	6.1995	49.16	1.556	.12622
$Cond_{MIXED}$	–12.1005	8.7025	64.66	–1.390	.16916
$Target_{ENGLISH}$	7.9955	10.6933	37.58	0.748	.45929
$Masker_{ENGLISH}$	–9.1348	9.3129	54.27	–0.981	.33100
EnglishTestScore	–1.9697	0.6265	49.16	–3.144	.00283
$Cond_{MIXED}$:$Target_{ENGLISH}$	–2.0843	11.6724	65.00	–0.179	.85883
$Cond_{MIXED}$:$Masker_{ENGLISH}$	11.1005	11.6724	65.00	0.951	.34513
$Target_{ENGLISH}$:$Masker_{ENGLISH}$	7.6303	11.6724	65.00	0.654	.51561
$Cond_{MIXED}$:EnglishTestScore	1.2192	0.8794	64.66	1.386	.17040
$Target_{ENGLISH}$:EnglishTestScore	–0.5523	1.0806	37.58	–0.511	.61229
$Masker_{ENGLISH}$:EnglishTestScore	0.6992	0.9411	54.27	0.743	.46069
$Cond_{MIXED}$:$Target_{ENGLISH}$:$Masker_{ENGLISH}$	–2.7859	16.5073	65.00	–0.169	.86651
$Cond_{MIXED}$:$Target_{ENGLISH}$:EnglishTestScore	0.3634	1.1796	65.00	0.308	.75902
$Cond_{MIXED}$:$Masker_{ENGLISH}$:EnglishTestScore	–0.9275	1.1796	65.00	–0.786	.43453
$Target_{ENGLISH}$:$Masker_{ENGLISH}$:EnglishTestScore	–0.7515	1.1796	65.00	–0.637	.52629
$Cond_{MIXED}$:$Target_{ENGLISH}$:$Masker_{ENGLISH}$:EnglishTestScore	0.2960	1.6681	65.00	0.177	.85973

Note. *SE* = standard error; *df* = degrees of freedom; Pr(>|*t*|) = two-tailed *p* value.

Appendix B

Pupil Dilation Model (Curve Fitting)

The model for fitting curves to the raw pupil dilation traces was written as follows (in standard R notation):

```
model = lmer(pupil ~ time + I(time^2) + I(time^3) + I(time^4)
    + time*Target*Masker*EnglishTestScore*Condition +
    I(time^2)*Target*Masker*EnglishTestScore*Condition +
    I(time^3)*Target*Masker*EnglishTestScore*Condition +
    I(time^4)*Target*Masker*EnglishTestScore*Condition +
    (time + Target + Masker + Condition | Subject)
```

Appendix C

Pupil Dilation Peak Model and Results

The model for analyzing peak pupil dilation data was written as follows (in standard R notation):

```
model = lmer(Peak ~ Condition*Target*Masker*EnglishTestScore +
       (Condition + Target + Masker | Subject)
```

Results of the full model of peak pupil dilation values.

Fixed effects	Estimate	*SE*	*df*	*t*	Pr(>|*t*|)
(Intercept)	0.1707	0.1547	13.10	1.104	.28965
$Cond_{MIXED}$	–0.3571	0.1135	13.08	–3.146	.00768
$Target_{ENGLISH}$	–0.0370	0.1197	13.15	–0.308	.76279
$Masker_{ENGLISH}$	–0.0271	0.0468	13.39	–0.580	.57179
EnglishTestScore	0.0019	0.0156	13.10	0.121	.90591
$Cond_{MIXED}$:$Target_{ENGLISH}$	0.3574	0.0090	52.27	39.840	< 2e–16
$Cond_{MIXED}$:$Masker_{ENGLISH}$	0.2319	0.0090	52.27	25.854	< 2e–16
$Target_{ENGLISH}$:$Masker_{ENGLISH}$	–0.3128	0.0090	52.27	–34.865	< 2e–16
$Cond_{MIXED}$:EnglishTestScore	0.0429	0.0115	13.08	3.742	.00244
$Target_{ENGLISH}$:EnglishTestScore	0.0040	0.0121	13.15	0.330	.74644
$Masker_{ENGLISH}$:EnglishTestScore	0.0037	0.0047	13.39	0.778	.45010
$Cond_{MIXED}$:$Target_{ENGLISH}$:$Masker_{ENGLISH}$	0.4445	0.0127	52.27	35.037	< 2e–16
$Cond_{MIXED}$:$Target_{ENGLISH}$:EnglishTestScore	–0.0383	0.0009	52.27	–42.255	< 2e–16
$Cond_{MIXED}$:$Masker_{ENGLISH}$:EnglishTestScore	–0.0274	0.0009	52.27	–30.292	< 2e–16
$Target_{ENGLISH}$:$Masker_{ENGLISH}$:EnglishTestScore	0.0341	0.0009	52.27	37.587	< 2e–16
$Cond_{MIXED}$:$Target_{ENGLISH}$:$Masker_{ENGLISH}$:EnglishTestScore	–0.0426	0.0013	52.27	–33.209	< 2e–16

Note. *SE* = standard error; *df* = degrees of freedom; Pr(>|*t*|) = two-tailed *p* value.

Appendix D

Pupil Dilation Trough Model and Results

The model for analyzing the trough pupil dilation data was written as follows (in standard R notation):

```
model = lmer(Trough ~ Condition*Target*Masker*EnglishTestScore
  + (Condition + Target + Masker | Subject)
```

Results of the full model of trough pupil dilation values.

Fixed effects	Estimate	*SE*	*df*	*t*	Pr(>|t|)
(Intercept)	0.1776	0.2601	13.01	0.683	.5068
$Cond_{MIXED}$	–0.0613	0.1182	13.04	–0.518	.6128
$Target_{ENGLISH}$	0.0015	0.1190	13.06	0.013	.9901
$Masker_{ENGLISH}$	–0.0417	0.0454	13.26	–0.919	.3744
EnglishTestScore	–0.0218	0.0263	13.01	–0.830	.4213
$Cond_{MIXED}$:$Target_{ENGLISH}$	–0.1036	0.0074	52.21	–14.009	< 2e–16
$Cond_{MIXED}$:$Masker_{ENGLISH}$	–0.3244	0.0074	52.21	–43.871	< 2e–16
$Target_{ENGLISH}$:$Masker_{ENGLISH}$	–0.1829	0.0074	52.21	–24.735	< 2e–16
$Cond_{MIXED}$:EnglishTestScore	0.0136	0.0120	13.04	1.141	.2743
$Target_{ENGLISH}$:EnglishTestScore	0.0124	0.0120	13.06	1.028	.3228
$Masker_{ENGLISH}$:EnglishTestScore	0.0079	0.0046	13.26	1.729	.1070
$Cond_{MIXED}$:$Target_{ENGLISH}$:$Masker_{ENGLISH}$	0.6440	0.0105	52.21	61.575	< 2e–16
$Cond_{MIXED}$:$Target_{ENGLISH}$:EnglishTestScore	0.0014	0.0007	52.21	1.819	.0747
$Cond_{MIXED}$:$Masker_{ENGLISH}$:EnglishTestScore	0.0270	0.0007	52.21	36.085	< 2e–16
$Target_{ENGLISH}$:$Masker_{ENGLISH}$:EnglishTestScore	0.0178	0.0007	52.21	23.825	< 2e–16
$Cond_{MIXED}$:$Target_{ENGLISH}$:$Masker_{ENGLISH}$:EnglishTestScore	–0.0600	0.0010	52.21	–56.850	< 2e–16

Note. *SE* = standard error; *df* = degrees of freedom; Pr(>|*t*|) = two-tailed *p* value.