Reading is disrupted by intelligible background speech: Evidence from eye-tracking

Martin R. Vasilev[1*]

Simon P. Liversedge[2]

Daniel Rowan[3]

Julie A. Kirkby[1]

Bernhard Angele[1]

[1]Department of Psychology

Bournemouth University, United Kingdom

[2]School of Psychology,

University of Central Lancashire, United Kingdom

[3]Institute of Sound and Vibration Research,

University of Southampton, United Kingdom

*Corresponding author

Phone: +44 7835202606

Email: mvasilev@bournemouth.ac.uk

Word count: 19300

Abstract

It is not well understood whether background speech affects the initial processing of words during reading or only the later processes of sentence integration. Additionally, it is not clear how eye-movements support text comprehension in the face of distraction by background speech and noise. In the present research, participants read single sentences (Experiment 1) and short paragraphs (Experiments 2-3) in four sound conditions: silence, speech-spectrum Gaussian noise, English speech (intelligible to participants), and Mandarin speech (unintelligible to participants). Intelligible speech did not affect the lexical access of words and had a limited effect on the first-pass fixations of words. However, it led to more regressions and more re-reading fixations compared to both unintelligible speech and silence. The results suggested that the distraction is mostly semantic in nature, and there was only limited evidence for a contribution of phonology. Finally, intelligible speech disrupted comprehension only when participants were prevented from re-reading previous words. These findings suggest that the semantic properties of irrelevant speech can disrupt the ongoing reading process, but that this disruption occurs in the post-lexical stages of reading when participants need to integrate words to form the sentence context and to construct a coherent discourse of the text.

Keywords: reading, eye-movements, auditory distraction, background speech, noise

Word count: 199

Public Significance Statement

Listening to irrelevant speech in the background often disrupts reading efficiency. To better understand why this occurs, we recorded participants' eye-movements as they read single sentences and short paragraphs in conditions of speech and noise. We found that irrelevant speech is distracting when participants can process its meaning. The meaning of the speech sound did not affect the lexical identification of words in the text, but it resulted in greater re-reading of previous words. This increase in re-reading behaviour was found to occur because participants attempt to maintain the immediate comprehension of the text in the distracting reading conditions. Once they were no longer able to re-read words, comprehension was negatively affected.

Previous research has indicated that background speech has a direct influence on eye-movements during reading (Cauchard, Cane, & Weger, 2012; Hyönä & Ekholm, 2016; Yan, Meng, Liu, He, & Paterson, 2017). However, it is currently not well understood whether background speech influences the early stages of word processing or if its effect is constrained only to the later processes of sentence integration. Additionally, it is not clear what properties of the speech sound give rise to distraction and how eye-movements support text comprehension when reading in such auditory conditions. In the present research, we investigated how intelligible speech affects sentence and paragraph reading, and whether it influences the lexical processing of words.

There are two main theories that may explain how speech can disrupt reading (see Hughes, 2014 for a discussion of other tasks, such as serial recall). According to the *phonological disruption* view (Salamé & Baddeley, 1982, 1989), background speech automatically gains access to the phonological loop component of working memory capacity (Baddeley, 2003) and thus interferes with the encoding and retrieval of visually presented items (Larsen & Baddeley, 2003). In Baddeley and Hitch's (1974, 1994) model, the loop consists of a phonological store where auditory information can be stored for a period of 1-2 s, and an articulatory rehearsal process, which helps maintain this information. However, the phonological loop is not reserved for processing auditory information. Rather, visually presented items are also converted into a phonological code that is then fed into the store (Baddeley, 2000). Because of this, the irrelevant speech can interfere with the storing and retrieval of visual information, thus causing distraction. In this theory, the phonological loop acts as a filter that lets in speech sounds, but filters out other non-speech sounds such as acoustical noise (Salamé & Baddeley, 1987). Therefore, this view predicts that any speech sound (intelligible or not) would interfere

with reading. Acoustical noise, on the other hand, would not cause interference because it does not gain access to the phonological loop.

Martin, Wogalter, and Forlano (1988) failed to find evidence for the phonological disruption hypothesis in a reading comprehension task. In their experiments, intelligible speech (English) disrupted comprehension significantly more than unintelligible speech (Russian). Additionally, speech consisting of random words was also found to be more disruptive than speech consisting of random non-words. To account for these results, Martin et al. argued that intelligible speech causes *semantic disruption*. They hypothesized that, because reading for comprehension involves extracting the meaning of the text, the semantic content of the irrelevant speech can interfere with this process. Therefore, the semantic disruption view predicts that background speech would disrupt reading only when it is intelligible. A similar theory is the interference-by-process account (Marsh, Hughes, & Jones, 2008, 2009), according to which background speech causes disruption because processing the meaning of the speech relies on the same process that is used by the main task (i.e., extracting the meaning of the text that is being read). Because both views make the same prediction in the present research, we will consider them together as theories of semantic disruption.

Previous behavioral studies on the effect of background sounds on reading have painted a mixed picture (for a review, see Vasilev, Kirkby, & Angele, 2018). For example, while some of them have found that intelligible speech is detrimental to reading and proofreading performance (Jones, Miles, & Page, 1990; Martin et al., 1988; Sörqvist, Halin, & Hygge, 2010), others have failed to find such an effect (Haka et al., 2009; Landström, Söderberg, Kjellberg, & Nordström, 2002; Ljung, Sörqvist, & Hygge, 2009; Venetjoki, Kaarlela-Tuomaala, Keskinen, & Hongisto, 2006). Similarly, studies on the effect of acoustical noise on reading in adults have also resulted

in mixed findings. Some of them have found no evidence that acoustical noise is detrimental to reading comprehension (Gawron, 1984; Jahncke, Hygge, Halin, Green, & Dimberg, 2011; Veitch, 1990), while others have found that it can be detrimental to some people depending on their personality characteristics (Furnham, Gunter, & Peterson, 1994; Ylias & Heaven, 2003). Therefore, the evidence from behavioral studies is inconclusive, but it suggests that at least some sounds may be disruptive to reading.

Eye-tracking Evidence

One limitation of behavioral studies is that they have focused only on the end product of reading (i.e., comprehension). However, recording participants' eye-movements makes it possible to investigate how the reading process unfolds in time and to uncover subtle auditory disruption effects that may not be apparent in comprehension measures. A better understanding of the time course of these effects is also crucial for developing theoretical frameworks that can explain how auditory stimuli interfere with the reading process. While theories of semantic and phonological disruption make very specific predictions about the types of speech sounds that should disrupt reading, these predictions are mostly descriptive in nature and they do not tell us which aspects of the reading process are affected. Therefore, eye-tracking evidence has the potential to advance our theoretical understanding by making it possible to formulate more precise and quantitative predictions in a reading task.

There are only a few studies to date that have investigated the effects of background speech and acoustical noise on eye-movements during reading. In one study, Johansson, Holmqvist, Mossberg, and Lindgren (2012) found that background sounds recorded from a café did not influence fixation durations or fixation probabilities during reading. In contrast, Cauchard, Cane, and Weger (2012) found that participants had longer gaze durations, longer

reading and re-reading times, and made more fixations in the presence of intelligible background speech compared to silence. However, their study was confounded by an additional manipulation in which participants' reading was interrupted on half of the trials for one minute by an unrelated task. This in turn may have influenced their reading behavior.

More recently, Hyönä and Ekholm (2016) reported a series of experiments that investigated how background speech affects reading of syntactically complex sentences. In Experiment 1, they found that listening to intelligible speech (Finnish) did not result in significantly longer fixation durations compared to either speech in an unfamiliar language (Italian) or silence. In this sense, the authors did not find evidence for the phonological disruption hypothesis. In the remaining experiments, Hyönä and Ekholm found that scrambled Finnish speech is more disruptive than both silence and normal, non-scrambled speech. The scrambled Finnish speech was created by randomizing the order of words in the text and reading them aloud with an intonation that resembles coherent speech. The authors also found that scrambled speech created from the to-be-read text was not more distracting than scrambled speech created from an unrelated text. Additionally, scrambled speech from an unrelated text that was semantically, but not syntactically, anomalous was just as distracting as scrambled speech that was both semantically and syntactically anomalous. These results point to two conclusions. First, they suggest that scrambled speech is disruptive not because of similarity in the semantic content between the speech and the text, but because both sources of information are calling on the same semantic processes for analyzing meaning (Hyönä & Ekholm, 2016). Second, they also suggest that the syntactic anomaly of scrambled speech does not *per se* make it more distracting to readers.

Finally, Yan, Meng, Liu, He, and Paterson (2017) investigated distraction effects by background speech in readers of Mandarin Chinese. Participants read single sentences with a target word lexical frequency manipulation in three background sound conditions: intelligible (i.e., Mandarin) speech, meaningless speech (the same speech scrambled in 60 ms segments), and silence. The scrambling method used in this study did not leave the individual words intact as in Hyönä and Ekholm's (2016) experiments, but it preserved the general acoustic variation that is present in meaningful speech. Yan et al. found that intelligible speech resulted in longer reading times, more fixations, and more regressions compared to both meaningless speech and silence. Additionally, the otherwise ubiquitous lexical frequency effect was eliminated in the two speech conditions, but only for the first fixation duration on the target word. This suggests that background speech may have a very early influence on the language processing system by delaying access to the lexical representation of words.

Present Experiments

The few available eye-tracking studies to date have provided the first clues as to how intelligible speech may disrupt reading. With the exception of Hyönä and Ekholm's (2016) Experiment 1, all previous studies seem to suggest that intelligible speech leads to an increase in re-reading fixations. However, it is not immediately clear what properties of background speech give rise to the disruption. For example, it is not known whether the disruption is due only to the semantic properties of speech, or if phonology also plays a role. While the manipulation in Hyönä and Ekholm's (2016) Experiment 1 could make this theoretical distinction, the authors reported no disruption by either intelligible or unintelligible speech. Therefore, their results did not provide support for either the phonological or semantic disruption hypothesis. One possible explanation for this finding is that the foreign (i.e., unintelligible) speech material used in their

study was taken from a language course, while the native speech was an excerpt from a novel. Therefore, the lack of a statistically significant difference may have occurred because the two speech sounds potentially differed in properties such as intonation, content, and rate of speech. The present research made a more stringent test of the semantic and phonological disruption theories by using intelligible and unintelligible speech that are more closely matched on these variables, and by including an acoustical noise condition that contains no phonological information.

Additionally, there is conflicting evidence about which stages of the reading process are influenced by intelligible speech. For example, Hyönä and Ekholm (2016) reported that the effect of scrambled speech was mostly evident in re-reading fixations, whilst Yan et al. (2017) observed the same effect for intelligible speech. These findings suggest that the effect of background speech is mostly evident in second-pass reading measures. However, Cauchard et al. (2012) also reported an effect on gaze durations, and Yan et al. found that intelligible speech eliminated the frequency effect for first fixation durations. The last two findings seem to suggest that the early stages of word processing may also be affected. If the initial processing of words is disrupted, this may occur because the semantic properties of speech interfere with accessing the lexical information of words. This is an important theoretical question that has not been addressed in an alphabetical language before.

Finally, unlike previous behavioural studies, none of the eye-tracking experiments so far have found disruption by background speech in reading comprehension. This result is surprising because it raises the question of what properties of background speech are responsible for the disruption observed in eye-movements. If this disruption is purely phonological in nature, it is likely to occur during the initial stages of word processing when the phonological information of

written words is registered into the phonological store. As a result, such disruption would not be generally expected to affect later comprehension processes and impair participants' comprehension accuracy. However, if the disruption in eye-movements is due to the semantic properties of the speech, it is not clear why comprehension remains unaffected given that semantic processing of the text is important for its comprehension. It is possible that previous eye-tracking studies may have used questions that were too easy to answer or, alternatively, that participants may have been able to compensate for any disruption in comprehension by making more regressions and re-reading fixations. These possibilities have not been examined so far. Therefore, it is not well-understood how eye-movements support the immediate text comprehension when listening to intelligible speech in the background.

We report three experiments that examined the effect of background speech on eye-movements and comprehension processes during reading. In Experiment 1, we investigated how background speech affects the lexical processing of words when reading single sentences. In Experiment 2, we examined its effect on comprehension accuracy and online integration processes when reading short passages. In Experiment 3, we explored the role of re-reading behavior in maintaining immediate text comprehension by preventing participants from re-reading previous words and sentences in the same passages.

Experiment 1

The first goal of Experiment 1 was to investigate whether the phonological or semantic properties of speech (or some combination of the two) is responsible for the disruption observed in eye-movements during reading. We used a paradigm in which participants read single sentences that were presented concurrently with the sounds. Importantly, participants heard the sound stimuli only for the duration that they were actually reading, thus reducing potential

habituation effects (Banbury & Berry, 1997). Additionally, the speech stimuli were carefully matched and consisted of single declarative sentences that were unrelated to each other. This was similar to the reading stimuli, which also consisted of unrelated declarative sentences. Furthermore, only naturally-occurring speech was used (i.e., without any scrambling) and this speech was spoken at a consistent rate throughout the whole experiment. Finally, because participants' comprehension was assessed immediately after reading a sentence, it was possible to test whether background sounds have an immediate effect on reading comprehension. This is an important question because most behavioural studies to date have had a delay between reading the text and the subsequent comprehension assessment (e.g., due to other tasks intervening in between; Martin et al., 1988) and any observed differences may not be due to deficits in immediate text comprehension (see Sörqvist et al., 2010).

The present study used four background sound conditions to differentiate between the phonological and semantic disruption accounts: Gaussian noise filtered to have an amplitude spectrum similar to that of long-term average speech (referred to as 'speech-spectrum noise'), Mandarin speech, English speech, and silence (the control condition). The speech-spectrum noise did not contain any scrambled speech. Rather, it imitated the spectral frequencies that are present in natural speech, but without containing any phonological or semantic information. According to the phonological distraction account (Salamé & Baddeley, 1982, 1987), irrelevant speech should disrupt the ongoing reading process regardless of whether it is intelligible or unintelligible because it automatically gains access the phonological loop of working memory. However, speech-spectrum noise would not cause such disruption because it does not gain access to the phonological loop. Therefore, if the disruption is phonological in nature, we would expect English speech to be more distracting than speech-spectrum noise, but equally as distracting as

Mandarin speech. On the other hand, if the disruption is semantic in nature (Marsh et al., 2008, 2009; Martin et al., 1988), we would expect English speech to be more distracting than Mandarin speech because participants can understand the former language but not the latter.

It should be noted that Mandarin phonology differs from English phonology in certain ways, such as the use of distinct tones, the smaller number of syllables, the lack of polysyllabic words, and the high number of homophones (Duanmu, 2006). Nevertheless, the phonological disruption account (Salamé & Baddeley, 1982, 1987) predicts that interference occurs because the irrelevant speech gains access to the phonological loop and not because of specific properties of the speech itself. In fact, greater phonological similarity between the irrelevant speech and the visual stimuli in the main task does not increase distraction (Jones & Macken, 1995; Larsen, Baddeley, & Andrade, 2000; LeCompte & Shaibe, 1997). Therefore, the actual language of the irrelevant speech is often not thought to be of critical importance, and distraction has been observed with a range of different languages, including Arabic (Baddeley & Salamé, 1986; Salamé & Baddeley, 1987), German (Colle & Welsh, 1976), Russian (Klatte, Lee, & Hellbruck, 2002), and Japanese (Ellermeier & Zimmer, 1997), to name a few.

One possibility is that the disruption by intelligible speech is not either entirely semantic or entirely phonological in nature, but rather some combination of the two. To test for this possibility, we will distinguish between two versions of the phonological disruption account. In the strong version, any distraction effects are attributed to phonology alone. As a result, English speech should to be more distracting than Noise but equally as distracting as Mandarin speech. In the weaker version, phonology is responsible for some, but not all distraction effects. Therefore, the weaker version predicts that Mandarin should be more distracting than speech-spectrum

noise (indicating some contribution of phonology), but less distracting than English speech (indicating that the rest of the disruption effect can be attributed to semantic interference).

The second goal of Experiment 1 was to test whether intelligible speech interferes with the lexical processing of words. Yan et al.'s (2017) findings suggest that intelligible speech may disrupt lexical processing in readers of Mandarin, but, interestingly, this effect was found only in first fixation durations. This suggests that the disruption of lexical access by intelligible speech is limited only to the very first fixation on words. In Experiment 1, we tested whether lexical processing is affected in readers of English by manipulating the lexical frequency of a target word in each sentence. Previous research has shown that lower frequency words are fixated longer than higher frequency words (Inhoff & Rayner, 1986; Rayner, 2009). Therefore, as the frequency effect reflects the difficulty inherent in the lexical access of words, the present study tested whether intelligible speech interferes with lexical access. For example, in any model of word identification where word representations accrue activation constantly (e.g., Morton, 1969; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), we might expect that English speech makes it harder to accumulate activation in order to identify a word compared to the other sound conditions. In this case, we should find a stronger word frequency effect in this condition compared to the other background input conditions because low frequency words require more activation for lexical access than high frequency words. In this sense, we would expect the disruption effect of intelligible English speech to be greater for low frequency words than for high frequency words.

Summary of Predictions

The following predictions were tested in the present experiment:

H1: If the disruption by intelligible speech is entirely phonological in nature, English speech should be more distracting than Silence and Noise, but equally as distracting as Mandarin speech (strong form of phonological interference).

H1.2: If the disruption by intelligible speech is only partially phonological in nature, Mandarin speech should be more distracting than Noise (weaker form of phonological interference).

H2: If the disruption by intelligible speech is entirely semantic in nature, English speech should be more distracting than Silence, Noise, and Mandarin; additionally, prediction H1.2 above should not be supported by the data (strong form of semantic interference).

H2.1: If the disruption by intelligible speech is a combination of semantic and phonological interference, English speech should be more distracting than Silence, Noise, and Mandarin; additionally, prediction H1.2 above should also be supported by the data (combination of phonological and semantic interference).

H3: If intelligible speech interferes with the lexical access of words, there should be greater disruption by English speech for low frequency compared to high frequency words.

Based on the available evidence (e.g., Hyönä & Ekholm, 2016; Yan et al., 2017), we expected to find support for predictions H2 and H3 above.

Method

Participants. Forty university students (70% female) participated for course credit or a payment of £8. Their mean age was 22.4 years (*SD*= 5.2 years; range: 18-40 years). All participants were native speakers of English, reported normal or corrected-to-normal vision, normal hearing, and no prior diagnosis of reading disorders. Participants were naïve as to the

purpose of the experiment. None of them had any knowledge of Mandarin. Ethical approval was obtained from Bournemouth University (protocol No. 11663).

The statistical power of our design was 0.831 for an average effect size of d= 0.47 based on the method described in Westfall (2015). The expected value of d= 0.47 was determined by calculating the effect size for all disruption effects by background speech reported in Hyönä and Ekholm (2016) and then taking their average. As the current power exceeds the recommended value of 0.80 (Cohen, 1988), our experiment was sufficiently powered to detect auditory disruption effects by background speech.

Materials. The reading material consisted of 128 English sentence frames (see Figure 1b for an example). Their average length was 13.2 words. Each sentence frame had a target word position which could contain either a low-frequency or a high-frequency word (picked using the SUBTLEX-UK database; Van Heuven, Mandera, Keuleers, & Brysbaert, 2014). The target word was never one of the first or last three words in the sentence frame. The target words were an equal number of adjectives and nouns. High and low frequency target words were matched on word length, bigram frequency, and neighbourhood size using the N-watch software (Davis, 2005). This information is presented in Table 1. Additionally, cloze-task predictability norms were obtained from 21 students who did not participate in the eye-tracking study. High and low frequency target words did not differ significantly in their predictability given the preceding sentence frame, $t(127) = 0.97$, $p = 0.33$.

Table 1

Descriptive Statistics for the Target Words in Experiment 1

	High-frequency words				Low-frequency words			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Word length (in letters)	5.6	1.1	3	7	5.6	1.1	3	7
Lexical frequency[1]	160	146	46	779	3	2	0.06	10
Bigram token frequency	1282	925	129	5173	1279	994	83	7050
Neighbourhood size	2.8	3.8	0	22	2.8	3.6	0	20
Predictability	0.01	0.04	0	0.29	0.01	0.03	0	0.24

[1]in counts per million.

Auditory stimuli. The sound stimuli consisted of three types of sound: speech-spectrum noise, English speech, and Taiwanese Mandarin speech. The English speech was taken from the BKB corpus (Bench, Kowal, & Bamford, 1979). The corpus consists of short spoken sentences that last for about 1-2 seconds (e.g. “The house had nine rooms.”). Thirty-two sound files were created by concatenating seven speech sentences and removing the silence gaps. Each speech sentence appeared only once in the sound files. In half of the speech sound files, the speaker was female; in the other half, the speaker was male. The speech-spectrum noise was created by filtering Gaussian noise by the average amplitude spectrum of the English BKB sentences in male voice.

Thirty-two Mandarin sound files were created in the same way as the English ones. The speech sentences were taken from Kuo (2006), who translated 240 sentences from the BKB (Bench et al., 1979) and IHR (MacLeod & Summerfield, 1990) corpora. Therefore, the Mandarin speech sentences were intended for the same audience and had the same sentence structure as the English ones. The average speech rate in the experiment was matched between the English

speech (*M*= 3.16 words per second) and the Mandarin speech (*M*= 3.08 words per second) condition, $t(62)= 1.10$, $p= 0.28$.

In Experiment 1, the four sound conditions (Silence, Noise, Mandarin, and English) were presented in blocks of 32 sentences. The sentences within each block appeared in random order. The order of the blocks and the assignment of sound conditions to the sentences were counterbalanced with a full Latin square design. The frequency of the target word was also counterbalanced.

Apparatus. An Eyelink 1000 was used to record participants' eye-movements. Viewing was binocular, but only the right eye was recorded. The sampling frequency was 1000 Hz. Participants rested their head on a chin-and-forehead rest. The sound stimuli were administered through noise-canceling headphones (Bose QuietComfort 25) at 59-61 dB(A). The sound level was measured with a RadioShack digital meter (model 33-2055) over a 2-minute interval. The amplitude resolution of the sounds was 32 bits. The sampling frequency was 22 kHz for the English speech and speech-spectrum noise, and 44 kHz for the Mandarin speech.

The experiment was run using the EyeTrack 0.7.10h software (Stracuzzi, 2004) on a PC with Microsoft Windows XP. The stimuli were presented on a 20-inch Mitsubishi Diamond Pro 2070 monitor with a screen resolution of 1024 x 768. The sentences were displayed in Courier New font and appeared as black text over white background on a single line in the middle of the screen. The number of pixels per letter was 11. Participants sat 60 cm away from the monitor and at this distance each letter subtended approximately .40º of visual angle.

Procedure. Participants were instructed to focus on what they were reading and to ignore any sounds they may hear. Participants wore the headphones throughout the whole experiment.

Three-point calibration of the eye-tracker was performed at the beginning of the experiment and it was then repeated as required. The calibration error was kept at < .30º of visual angle. The experiment started with six practice trials, followed by the experimental trials. The trial presentation is illustrated in Figure 1. The experiment lasted for 30-40 minutes.

Trials began with a drift check, after which a black square appeared with a 50-pixel offset from the left edge of the screen. Once participants fixated the square, the sentence was presented, with the first letter of the first word at the center of the square. The onset of the background sound was simultaneous with the onset of the sentence. Participants used a button on a gamepad controller to terminate the trial once they finished reading the sentence. However, there was a trial timeout that corresponded to the length of the speech sound that was playing. In other words, if a participant did not terminate the trial by pressing a button, the trial ended automatically when the speech sound finished playing. For the English and Mandarin sound conditions, the timeout corresponded to the length of the individual speech files (between 9.2-12.6 s). The same timeouts were randomly assigned to the sentences in the silence and noise conditions. There was a yes/no comprehension question after 34% of trials. For example, in the sentence “The house was immediately recognisable by its green fence and big windows.”, the question was: “Did the house have small windows? Yes/ No”.

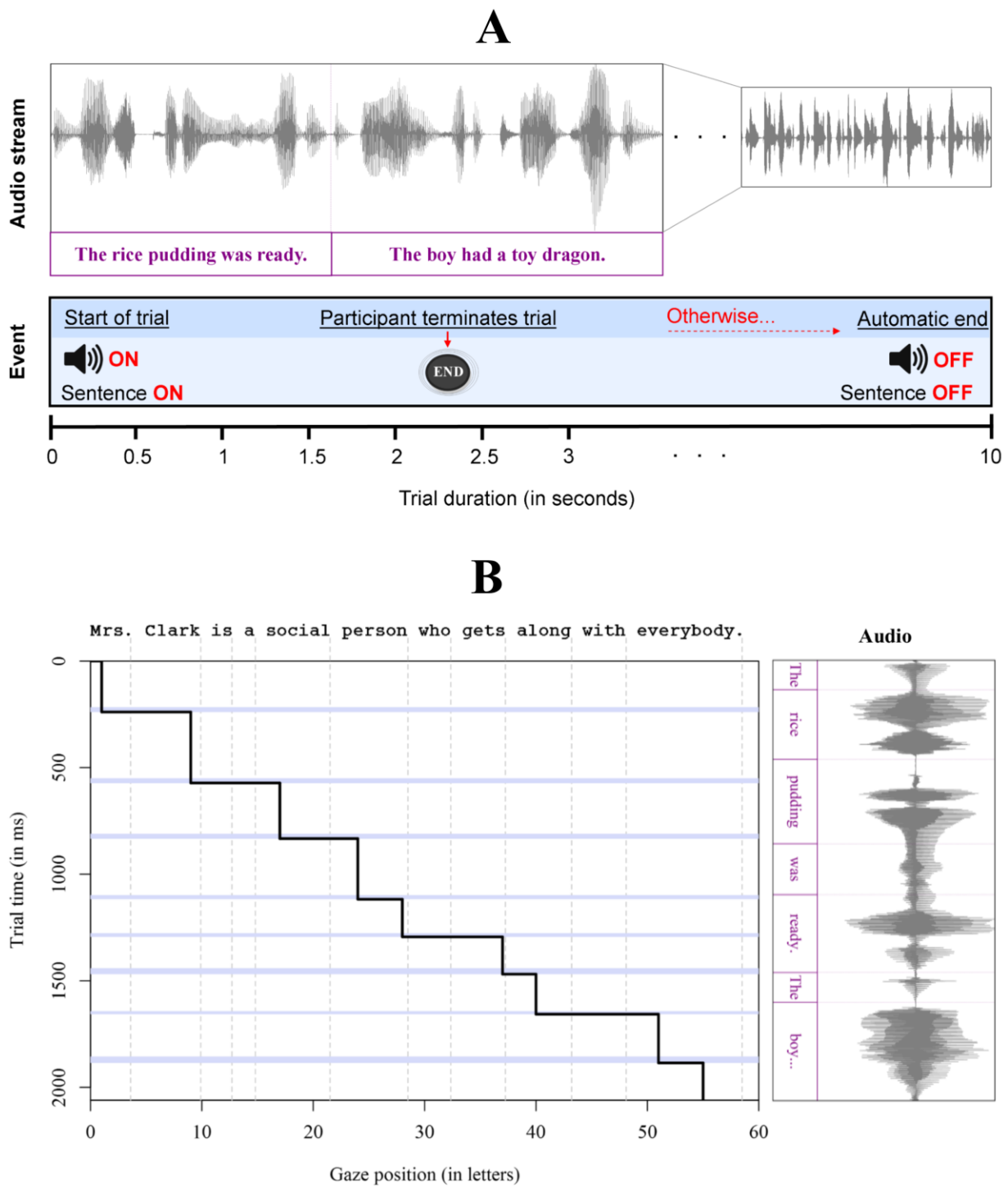

Figure 1. An illustration of the stimuli presentation. Panel **A** shows the events during the trial and the speech sound that was playing. The sentence and the speech sound were simultaneously presented at the start of the trial. Trials were normally terminated by the participant by pressing

the button. If the participant did not press the button, the trial was automatically terminated when the sound stopped playing. Panel **B** shows the timeline (including gaze position and auditory input) of a sample trial that was terminated by the participant. Horizontal blue lines show the saccades and the right-hand side shows the audio that was playing while they were reading. Vertical dotted lines indicate the word boundaries. In the sample sentence, the target word ("social") is high frequency; in the low frequency condition it was replaced by the word "chatty".

Data analysis. Several measures of global reading were analysed in the present study: total sentence reading time (the sum of all fixations on the sentence), fixation duration, probability of regression, saccade length, and number of fixations. In addition to this, the three standard local fixation duration measures were computed for the target word: 1) first fixation duration (FFD; the duration of the first fixation on the word); 2) gaze duration (GD; the sum of all fixations on the word before moving to another word); and 3) total viewing time (TVT; all fixations on the word, including second-pass reading). Finally, comprehension accuracy was also analyzed between the sound conditions. We also report post-hoc analyses in the Supplemental Materials of how the effect of background sound on total sentence reading time changed as the experiment progressed.

The data were analyzed with (Generalized) Linear Mixed Models ((G)LMMs) by using the "lme4" package v.1.1-12 (Bates, Machler, Bolker, & Walker, 2014) in R 3.3.0 (R Core Team, 2016). *P*-values for LMM models were calculated with the lmerTest package v.2.0-33 (Kuznetsova, Brockhoff, & Christensen, 2017). Fixation durations were log-transformed in all analyses. Treatment contrasts were used for the effect of background sound where English speech was the baseline. Low and high frequency target words were coded as 0.5 and -0.5, respectively. Additionally, to test whether phonology may account for some, but not all of the

disruption effects, a separate comparison between Mandarin and Noise was done. The results were adjusted for multiple comparisons with the Holm-Bonferroni procedure (Holm, 1979) to avoid an increase in Type 1 error probability because of this additional comparison. Background sound was entered as a fixed effect in the models; frequency was also a fixed effect in the target word analyses. Random intercepts, as well as random slopes for the sound condition were specified for subjects and items (Baayen, Davidson, & Bates, 2008; this corresponds to the "maximum" model for the main variable used for inferences, see Barr, Levy, Scheepers, & Tily, 2013)[1]. Results were considered statistically significant if the adjusted *p*-values were ≤ 0.05.

Results

The average trial duration was 3.8 s (*SD*= 1.74 s). There were 0.5% of trials where timeout was reached before participants pressed the end button and these were excluded from the data. Furthermore, 5.2% of the fixation duration data were excluded because of blinks. Additionally, trials in which FFD was above 800 ms, GD was above 2000 ms, or TVT was above 3000 ms were removed as outliers from all analyses (0.1% of data). The number of outliers excluded per condition did not differ significantly (χ^2 (2) = 0.4, *p*= 0.82). If fixation duration was an outlier in any of the three measures, the whole trial was removed from the analysis. Fixations shorter than 80 ms that occurred within one letter space of another fixation were combined with that fixation.

Comprehension accuracy. Comprehension accuracy was 94% in the silence condition, 93% in the noise condition, 93% in the Mandarin speech condition, and 91% in the English

[1] The following random slopes for background sound were removed due to convergence failure: random slope for items for saccade length, GD, and TVT; random slope for both participants and items for regression probability and number of first-pass fixations.

speech condition. There were no significant differences in comprehension accuracy across the sound conditions (all *p*s ≥ 0.20). Auditory speech sounds did not appear to affect comprehension accuracy which remained high across all conditions.

Global reading. Descriptive statistics of global reading on the whole sentence are presented in Table 2. The total sentence reading time was significantly longer in English speech compared to Silence (*b*= -0.07, *SE*= 0.03, *t*= -2.52, *p*= 0.03, *d*= -0.23), Noise (*b*= -0.12, *SE*= 0.03, *t*= -4.61, *p*< 0.001, *d*= -0.27) and Mandarin speech (*b*= -0.06, *SE*= 0.02, *t*= -2.61, *p*= 0.02, *d*= -0.14). The remaining analyses indicated that this was due to more second-pass fixations in English speech compared to all other sound conditions (Silence: *b*= -0.24, *SE*= 0.07, *z*= -3.36, *p*= 0.001, *d*= -0.14; Noise: *b*= -0.41, *SE*= 0.08, *z*= -5.37, *p*< 0.001, *d*= -0.18; Mandarin: *b*= -0.22, *SE*= 0.05, *z*= -3.99, *p*< 0.001, *d*= -0.10). As Table 2 shows, there was no difference in the number of first-pass fixations (all *p*s ≥ 0.80). English speech also resulted in a significantly greater regression probability compared to all other sound conditions (Silence: *b*= -0.09, *SE*= 0.02, *z*= -3.52, *p*< 0.001, *d*= -0.03; Noise: *b*= -0.14, *SE*= 0.03, *z*= -5.46, *p*< 0.001, *d*= -0.04; Mandarin: *b*= -0.08, *SE*= 0.02, *z*= -3.31, *p*= 0.002, *d*= -0.05). There were no significant differences in saccade length (all *p*s ≥ 0.35) or word landing position (all *p*s ≥ 0.13) across the sound conditions.

The planned comparison between Mandarin and Noise indicated that participants made significantly more second-pass fixations in Mandarin compared to Noise (*b*= -0.20, *SE*= 0.08, *z*= -2.43, *p*= 0.02, *d*= 0.09). However, as Table 2 shows, this effect was in part driven by the slightly better reading performance under Noise compared to Silence. No other differences between Noise and Mandarin were significant (all *p*s> 0.052). In summary, the results supported most strongly hypothesis H2, which stated that disruption by intelligible speech is only semantic in

nature. Hypothesis H2.1, which stated that the disruption has both a semantic and a phonological component, received only limited support because evidence for partial contribution of phonology (H1.2) was found in only one measure (number of second-pass fixations).

Table 2

Mean of Global Reading Measures per Background Sound Condition in Experiment 1 (SDs in Parentheses)

Sound condition	Total sentence reading time (in ms)	Word landing position (in letters)	Saccade length (in letters)	Regression probability	Number of fixations (per word)		
					1st-pass	2nd-pass	Total
Silence	3040 (1244)	2.81 (2.14)	8.86 (8.11)	.23 (.42)	1.03 (.57)	.48 (.77)	1.51 (.84)
Noise	2960 (1354)	2.86 (2.15)	8.72 (7.69)	.22 (.41)	1.04 (.56)	.44 (.74)	1.48 (.82)
Mandarin	3150 (1426)	2.85 (2.16)	8.91 (8.38)	.23 (.42)	1.04 (.59)	.51 (.82)	1.55 (.92)
English	3370 (1616)	2.86 (2.16)	8.73 (8.15)	.24 (.43)	1.03 (.61)	.62 (.93)	1.65 (1.02)

Target word. Fixation durations on the target word are shown in Figure 2a, and the results of the LMMs are shown in Table 3. There were robust frequency effects on the target word. However, contrary to hypothesis H3, the contrasts between English speech and the remaining sound conditions failed to interact with target word frequency[2]. Consistent with the results from global reading measures, the effect of English speech was not found on first-pass measures, but only on TVT, which includes re-fixations during second-pass reading. This is because English speech resulted in a greater number of re-reading fixations. English speech resulted in longer TVT compared to Silence (*d*= -0.15) and Noise (*d*= -0.12). The difference between English and Mandarin for TVT (*d*= -0.09) did not reach significance on the target word,

[2] In order to test the possibility that the target word analysis did not have sufficient statistical power to detect an interaction effect, frequency norms were obtained for all words in the sentence. The frequencies were then entered into a model that included all the fixations for all words in the sentence. The results (presented in the Supplemental Material) were consistent with the target word analyses and showed no significant interactions with lexical frequency.

but it was significant in the analysis of all words in the sentence (see the Supplemental Materials). No differences between Mandarin and Noise were significant (all *p*s ≥ 0.16). Therefore, the fixation duration analyses supported hypothesis H2, which stated that the disruption by intelligible speech is only semantic in nature.

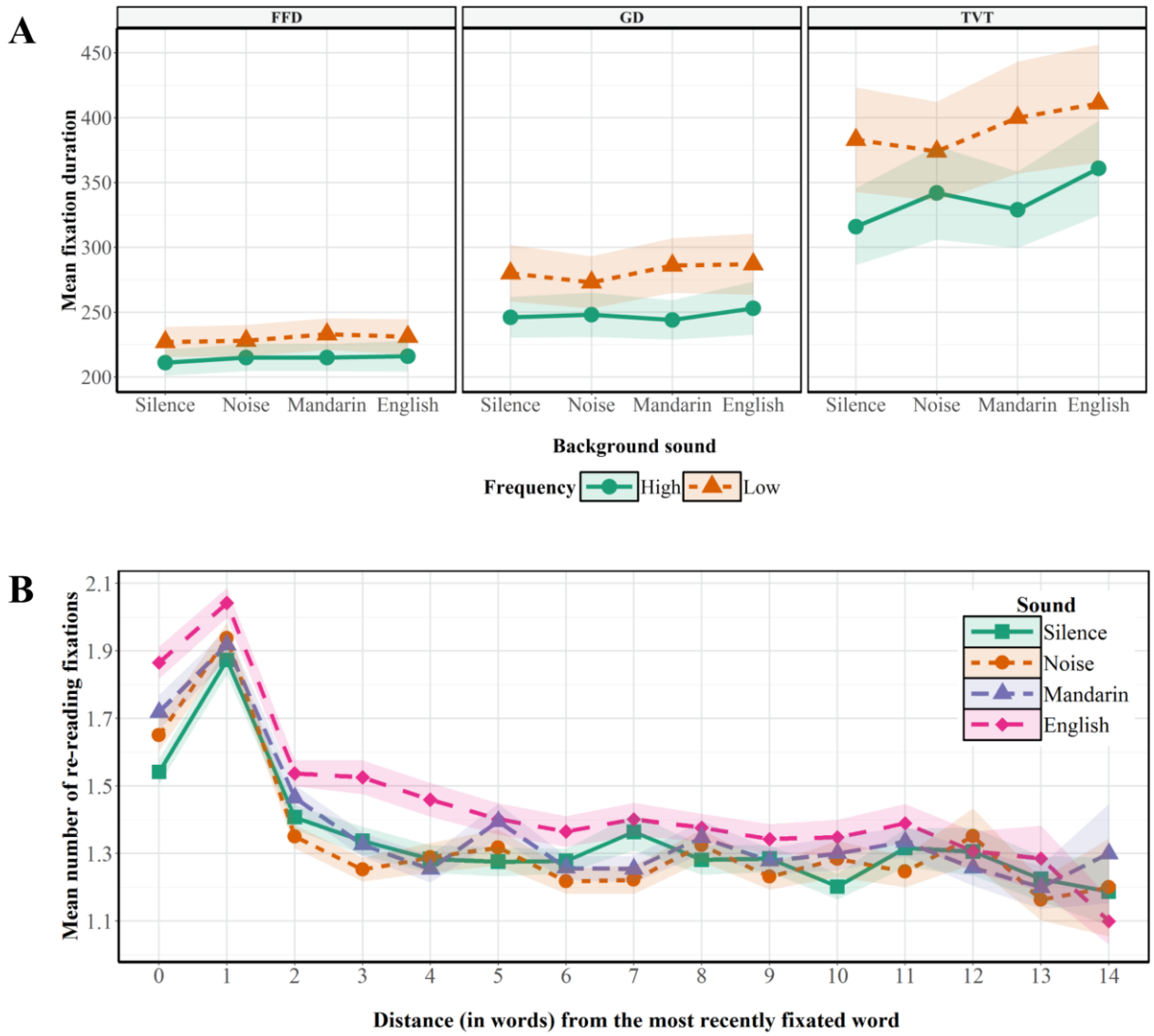

Figure 2. Mean descriptive statistics for Experiment 1. Panel **A**: Fixation durations on the target word for the different background sound conditions, broken down by target word frequency.

Panel **B**: Position of re-reading fixations for the different sound conditions in Experiment 1 as a function of distance from the most recently fixated word. Shading shows the standard error.

Table 3

LMMs for Fixation Durations on the Target Word in Experiment 1

Fixed effect	FFD				GD				TVT			
	b	SE	t	p	b	SE	t	p	b	SE	t	p
Intercept	5.35	.02	265.9	**<.001**	5.49	.03	167.2	**<.001**	5.78	.05	119.02	**<.001**
Freq	.05	.02	2.90	**<.01**	.11	.02	4.80	**<.001**	.12	.03	4.33	**<.001**
Eng vs Slc	-.02	.02	-.93	.36	-.01	.02	-.52	.61	-.08	.03	-2.90	**.01**
Eng vs Noise	<-.01	.01	-.08	.93	-.02	.02	-.72	.47	-.07	.03	-2.64	**.02**
Eng vs Mnd	.02	.02	1.06	.36	.01	.02	.38	.70	-.04	.02	-1.46	.30
Freq: Eng vs Slc	.02	.03	.86	.72	.01	.03	.32	.75	.05	.04	1.28	.40
Freq: Eng vs Noise	<-.01	.03	-.03	.98	-.02	.03	-.66	.51	-.02	.04	-.52	.60
Freq: Eng vs Mnd	.02	.03	.87	.72	.03	.03	.79	.43	.02	.04	.53	.60

Note: Freq: Lexical frequency. Eng: English. Slc: Silence. Mnd: Mandarin. Statistically significant *p*-values are formatted in bold.

Post-hoc analysis. We also conducted some exploratory analyses to investigate where re-reading fixations occurred in the sentence because many of the effects in the present analyses were due, at least in part, to an increase in second-pass fixations. In this analysis, we compared the number and distance of re-reading fixations that were made after the start of a regression until participants made a progressive fixation (i.e., until they fixated a new word in the sentence that they had not already fixated). To determine the location of re-reading fixations, we calculated their distance (in words) in relation to the most recently fixated word in the sentence before the regression (see the Supplemental Materials for an illustration of the method). If English interferes with the integration of recently-read words into the sentence context (i.e., "local" disruption), we would expect re-reading fixations to occur in close proximity to the source of the difficulty, that is, the most recently fixated word in that sentence. In contrast, if this

disruption is due to a failure to maintain the representation of the previous part of the sentence in working memory, we would expect that fixations will be more distant from the most recently fixated word, presumably in order to re-activate the previous sentence context.

The results from the analysis are plotted in Figure 2b. The number of re-reading fixations decreased with increasing distance from the most recently fixated word in the sentence (b= -0.03, SE= 0.005, z= -6.68, p< 0.001). Critically, however, English interacted significantly with distance (b= 0.01, SE= 0.005, z= 2.12, p= 0.03), thus showing that the mean difference between English and silence became smaller with increasing distance. This trend is apparent in Figure 2b where a clear increase in the number of re-reading fixations can be seen only when the distance was five words or less. Therefore, re-reading fixations were mostly constrained to words that were close to the most recently fixated word in the sentence.

Discussion

Experiment 1 investigated disruption effects by intelligible speech on reading single sentences. There were two main questions of the study: (1) is the disruption semantic or phonological in nature (or some combination of the two)? And (2) does intelligible speech affect the lexical processing of words? In terms of the first question, English speech increased the overall sentence reading time compared to silence. This was found to be mostly caused by making more regressions and more second-pass fixations when re-reading words. Experiment 1 provided support for the theoretical prediction that this disruption effect is semantic in nature (Marsh et al., 2008, 2009; Martin et al., 1988). English speech resulted in longer sentence reading times compared to Mandarin speech, and this arose due to readers making more regressions and more re-reading fixations.

Because English speech was consistently more disruptive than Mandarin speech, this provides evidence against the strong form of the phonological disruption view (hypothesis H1), which predicted that any speech sound (intelligible or not) would cause interference because it gains access to the phonological loop of working memory capacity (Salamé & Baddeley, 1982). Nevertheless, there was limited support for the view that phonology may account for some, but not all, of the disruption effects (hypothesis H1.2). This was because Mandarin speech led to more second-pass fixations compared to Noise. However, this effect warrants further replication as it was found in only one measure and it was partially driven by the fact that participants made fewer fixations in Noise compared to Silence. This is especially because a facilitation effect of acoustical noise has generally not been reported in previous studies (e.g. Johansson, 1983; Landström et al., 2002; Martin et al., 1988). Overall, the present results are largely consistent with Hyönä and Ekholm's (2016) Experiment 1, in the sense that the authors did not find any evidence to support the phonological disruption account. Therefore, taken together with Hyönä and Ekholm's (2016) findings, the present results suggest that phonology plays little if any role in auditory distraction by intelligible speech. In this sense, while we acknowledge that there was a hint in the data for a contribution of phonology, the pattern of results is most readily explained by hypothesis H2, which predicted that the disruption effects are entirely semantic in nature.

The results from the global reading measures agree with those of Yan et al. (2017), who also reported longer sentence reading times, more fixations and greater regression probability with intelligible speech in the background. However, Experiment 1 provided greater insight by showing that the increase in fixations was entirely due to more re-reading fixations. Additionally, the present results advance our theoretical understanding of disruption by intelligible speech by showing that these effects are due to the semantic content of the speech. Therefore, one of the

novel contributions of Experiment 1 was to show that semantic disruption is observed in eye-movement measures when comparing naturally-occurring speech sounds: English speech, which could be processed semantically by participants, led to greater disruption in second-pass reading measures compared to Mandarin speech, which could not be processed semantically.

The second aim of Experiment 1 was to investigate whether lexical processing is affected by background speech. Contrary to hypothesis H3, the results indicated that intelligible speech did not make the lexical access of low frequency words more difficult. Indeed, robust frequency effects were observed in all background sound conditions. On the surface, this result may appear to be contrary to Yan et al.'s (2017) finding that intelligible speech eliminated the frequency effect in FFD for Mandarin readers. However, Yan et al. also observed the same effect for meaningless (i.e., scrambled) speech. This in turn argues against disruption to lexical access due to semantic inconsistencies because the two speech conditions did not differ between one another. Therefore, both Yan et al.'s study and Experiment 1 provide converging evidence that the semantic properties of speech do not affect lexical access of words during normal reading.

Experiment 1 also showed that the initial reading of words was not influenced by English speech, as evidenced by the lack of effects in first-pass reading measures. This suggests that the progressive reading of sentences proceeded normally and was not affected by intelligible speech. Because the disruption effects were found in measures of second-pass reading, Experiment 1 suggests that intelligible speech disrupted reading on a more global level, as participants made more re-reading fixations and more regressions compared to unintelligible (Mandarin) speech.

The post-hoc analysis of re-reading fixations provided important insight into the nature of the disruption to processing that intelligible speech caused. Even though this analysis was not pre-planned and should be considered as exploratory, the results suggest that English speech

made it more difficult to integrate recently-read words into the sentence context. This was because the increase in re-reading fixations occurred in close proximity to the initial, first-pass fixations on words, presumably, those words that were the source of processing difficulty (i.e. the origin of the regression). Sentence comprehension is assumed to involve the retrieval of concepts from memory that are used to inform and construct the meaning of the sentence in relation to broader general world knowledge. Also, such knowledge is used to generate expectations and understand new concepts (Griffiths, Steyvers, & Tenenbaum, 2007), as well as to disambiguate sentential ambiguities. However, because auditory English speech and written English sentences both convey semantic meaning, it seems likely that the observed processing difficulty derives from disruption to semantic processes associated with the construction of a representation of sentential meaning.

It seems likely that there are two possible causative accounts for such disruption: it may arise due to competition, or even conflict (i.e., inconsistency) between the two representations of meaning (one deriving from the auditory speech and the other from text reading); alternatively, the processing cost may derive from the cognitive burden associated with processing two, rather than one, sources of sentential meaning. Hyönä and Ekholm (2016) tested the first alternative by presenting scrambled speech that consisted either of the text that participants were reading or of an unrelated text. They found that the two scrambled speech conditions did not differ between one another, which led them to suggest that the observed semantic interference is not due to competing semantic representations between the text and the speech sound. The second interpretation would be consistent with both Hyönä and Ekholm's (2016) results and the interference-by-process account (Marsh et al., 2008, 2009), which predicts that disruption occurs because both the speech and the written text rely on the same process for analysing meaning.

A further interesting finding from Experiment 1 was that none of the background sounds impaired participants' comprehension of the sentences, thus suggesting that whilst the efficiency with which readers were able to construct a representation of sentential meaning was reduced, readers were still able to attain an understanding of the sentence that they were reading. This is consistent with previous eye-tracking studies (Cauchard et al., 2015; Hyönä & Ekholm, 2016; Yan et al., 2017), but not with other behavioral studies (e.g. Martin et al., 1988; Sörqvist et al., 2010). Given that there was evidence for semantic disruption in the eye-movement measures, why have none of the eye-tracking studies so far found effects in comprehension accuracy? Indeed, because extracting the semantic content of the sentence is crucial for comprehension, it might be argued that a semantic disruption effect should also be found in comprehension accuracy measures.

One possible way to explain this apparent inconsistency is that the comprehension questions in previous eye-tracking studies may have been quite easy to answer, whereas those from behavioural studies may have been more taxing. Indeed, almost all eye-tracking studies investigating reading share something in common: comprehension assessment is carried out through the presentation of questions requiring a binary "yes/no" answer, and the average comprehension accuracy is almost always 80% or better. In this sense, it is possible that no difference in comprehension accuracy was found because the questions were not as challenging as those used in behavioural studies. If this is the case, then comprehension accuracy should be disrupted when questions are more difficult and probe a deeper level of text comprehension.

An alternative explanation is that the immediate comprehension of short texts is not disrupted by intelligible speech, regardless of the difficulty of questions. If this is the case, then the disruption observed in the eye-movement measures must be due to a transient difficulty in

processing the meaning of the sentence, which readers can overcome and still achieve approximately the same level of comprehension. In Experiment 2, we manipulated the difficulty of comprehension questions in order to rule out the possibility that the lack of disruption in comprehension accuracy was due to the questions being too easy to answer.

Experiment 2

The first aim of Experiment 2 was to test whether the lack of disruption in comprehension accuracy in Experiment 1 occurred because the questions were not challenging enough. In this study, short paragraphs were used because they offer a more ecologically-valid reading task and allow for greater opportunity to construct comprehension questions that are more demanding of readers. Additionally, paragraphs make it possible to study online integration processes beyond those necessary for single sentences. In Experiment 2, a question difficulty manipulation was added in which participants either answered easy questions that were comparable in their difficulty to those used in Experiment 1 or more difficult questions that required a deeper level of text understanding.

The second aim of the experiment was to test whether intelligible speech disrupts the integration of information across multiple sentences. In other words, is the disruption by intelligible speech limited only to the individual sentences that make up the text, or is there additional disruption due to integrating information across multiple sentences? Interestingly, Cauchard et al. (2012) reported that intelligible speech led to significantly longer sentence look-back times (i.e., greater re-reading of previous sentences), which accounted for 27% of the overall increase in reading time in their experiment. This suggests that the integration of information across sentences may also be affected. However, Hyönä and Ekholm (2016) found a difference in look-back times only in one out of four experiments: more specifically, scrambled

intelligible speech led to longer look-back times compared to the silence condition in their Experiment 3. Therefore, more evidence is required to better understand how intelligible speech may affect the integration of meaning across sentences.

There were two comprehension difficulty conditions in Experiment 2: 1) an easy condition in which the questions could usually be answered by recognising words and phrases from the text; and 2) a difficult condition in which the questions required understanding the meaning of the paragraph to answer. The easy questions were comparable to those used in Experiment 1 and in previous eye-tracking research. The difficult questions required comprehending the main topics of meaning in the paragraph and making inferences based on that meaning. The question difficulty manipulation was modelled after Wotschack and Kliegl's (2013) study in which comprehension of single sentences was assessed with either a multiple-choice question that could typically be answered by visual word recognition alone ("easy" condition) or with a more difficult question in which the answers had less verbatim overlap with the sentence ("difficult" condition). In the present study, the answers to difficult questions were paraphrased in their entirety and thus finding the correct answer required a deeper understanding of the paragraph's meaning. If English speech affects only deeper levels of text comprehension, there should be an interaction between English speech and question difficulty, with greater disruption in comprehension accuracy on the difficult compared to the easy questions.

The same four background sound conditions were used as in Experiment 1. Based on the findings of Experiment 1, we expected to observe more re-reading fixations and more regressions when the text was read in the auditory context of English speech compared to both Mandarin speech and silence. Additionally, we expected that English speech would lead to more regressions to previously-read sentences and to longer sentence look-back times. This was

because we expected that English speech would disrupt the integration of the currently-read sentence into the context of previously-read sentences, thus prompting participants to re-visit previous sentences more often.

Predictions

The same predictions of the phonological disruption (Salamé & Baddeley, 1982, 1987) and semantic disruption theories (Marsh et al., 2008, 2009; Martin et al., 1988) from Experiment 1 were again tested in the present experiment:

H1: If the disruption by intelligible speech is entirely phonological in nature, English speech should be more distracting than Silence and Noise, but equally as distracting as Mandarin speech (strong form of phonological interference).

H1.2: If the disruption by intelligible speech is only partially phonological in nature, Mandarin speech should be more distracting than Noise (weaker form of phonological interference).

H2: If the disruption by intelligible speech is entirely semantic in nature, English speech should be more distracting than Silence, Noise, and Mandarin; additionally, prediction H1.2 above should not be supported by the data (strong form of semantic interference).

H2.1: If the disruption by intelligible speech is a combination of semantic and phonological interference, English speech should be more distracting than Silence, Noise, and Mandarin speech; additionally, prediction H1.2 above should also be supported by the data (combination of phonological and semantic interference).

Consistent with the results from Experiment 1, we expected that hypothesis H2 would be most strongly supported by the data. Additionally, based on the question difficulty manipulation, we predicted that:

> **H3**: English speech should disrupt comprehension accuracy only when participants are answering difficult, but not easy, comprehension questions.

Method

Participants. Forty-eight Bournemouth University students (69 % female) participated for course credit or a payment of £10. Their mean age was 19.8 years (*SD*= 1.7 years; range: 18 - 27 years). None of them had participated in Experiment 1. Ethical approval for Experiments 2 and 3 was obtained from Bournemouth University (protocol No. 14005). The statistical power of Experiment 2 was 0.859 based on the same average effect size used for the power calculation in Experiment 1 (*d*= 0.47). This indicates that Experiment 2 was also sufficiently powered.

Materials and design. The reading materials consisted of 24 paragraphs (see the Supplemental Materials for the whole set of stimuli). Each paragraph was four sentences long and had an average length of 89.7 words (*SD*= 6.2 words; range: 77 to 103 words). The topic of the paragraphs was usually a short description of a person, a place or an event. Real names and specific details were avoided to prevent participants from using their prior knowledge to answer the questions. An example paragraph is provided below:

> Many tourists visiting the land-locked country were not aware of the pristine lake that was situated near its eastern border. Because it was surrounded by a forest and there were no major roads going there, the lake was mostly known only by the locals. However, with its crystal-clear waters and unforgettable scenery, the unspoiled lake was a dream place to

relax. According to one local legend, the lake's water had rejuvenating powers and many people from the region would go there in the summer for a swim.

Each paragraph contained two yes/no questions that could be answered by visual word recognition alone ("easy" condition), and two multiple-choice questions with four answers that required understanding the meaning of the whole paragraph to answer ("difficult" condition). An example of the easy questions is "Did the lake have unforgettable scenery? Yes/No". An example of the difficult questions is:

What can be said about the water in the lake?

1) It was murky and shallow

2) It was believed to alleviate stress and chronic medical conditions

3) It was believed to make you feel younger and more energetic

4) It was thought to be suitable for drinking

The answers to multiple-choice questions were paraphrased to prevent participants from recognizing words or phrases from the paragraph in order to find the correct answer (Wotschack & Kliegl, 2013). In the easy question condition, one question was based on the first two sentences of the paragraph, and the other question was based on the last two sentences. In the difficult question condition, both questions required a more general understanding of the paragraph because their answers were paraphrased. This manipulation was modeled after Wotschack and Kliegl's (2013) study. There was some variability in how the difficult questions were formulated: while some of them were based on more specific topics from the paragraph, others were more general and required participants to indicate which statement from four

alternatives was True/ False given the paragraph. However, the answers to all questions were paraphrased and therefore required a deeper understanding of the paragraph in order to find the correct answer and to eliminate the alternatives.

Ten undergraduate students who did not take part in the eye-tracking experiment participated in a pilot study in which they read the paragraphs, answered the comprehension questions, and rated the difficulty of questions on a scale from 1 (easy) to 5 (difficult). Each of the comprehension questions appeared on a separate screen and participants could not go back to re-read the text to help them answer the questions. The two difficulty conditions were presented in separate blocks that were counterbalanced across participants. Because the easy questions had only two answers and the difficult questions had four answers, participants' comprehension was analysed as accuracy above chance level. This controlled for the difference in chance level performance between the easy (50%) and difficult (25%) questions. Comprehension accuracy was significantly better on the easy (M= 43.7.8 %; SD= 16.6%) compared to the difficult questions (M= 31.2 %; SD= 34.6%), t= -0.06., SE= 0.02, t= -3.72, p< 0.001. This shows that participants understood the paragraphs sufficiently well in both question difficulty conditions. Additionally, questions in the difficult condition (M= 2.70; SD= 1.33) were rated as significantly more difficult than questions in the easy condition (M= 1.61; SD= 1.02), b= 1.06, SE= 0.09, t= 10.98, p < 0.001. Finally, participants spent more time reading the paragraphs in the difficult questions' block (M= 34.8 s; SD= 14.48 s) compared to the easy questions' block (M= 30.9 s; SD= 10.13 s), b= 3.48, SE= 1.43, t= 2.43, p= 0.01.

The speech stimuli were taken from the same two corpora used in Experiment 1. Six English and six Mandarin sound files were created by concatenating 40 unique speech sentences;

each speech file lasted for at least 60 s[3]. Silence gaps were removed to create a continuous stream of speech. Half of the files contained speech that was spoken by a female actor and the remaining half contained speech spoken by a male actor. The English and Mandarin conditions were matched on average rate of speech (English speech: 3.09 words per second; Mandarin speech: 3.08 words per second). The same speech-spectrum noise as in Experiment 1 was used.

The two question difficulty conditions were presented in separate blocks. Within each question difficulty block, the different sound conditions were also blocked. The assignment of paragraphs to conditions and the order of experimental blocks were counterbalanced with a full Latin square design. At the start of each question difficulty block, there were two practice paragraphs (read in silence) that were used to introduce participants to the different type of comprehension questions.

Apparatus. The equipment was the same as in Experiment 1. The paragraphs appeared with a 50-pixel offset on the x axis and 150-pixel offset on the y axis of the screen. The text was double-spaced and aligned to the left. Line breaks occurred at the empty space between words, but with the condition that there should be at least 50 pixels to the right of the last letter on the line. All paragraphs fitted on a single screen. The auditory stimuli were presented at the same sound intensity level as in Experiment 1. Participants pressed buttons on a gamepad controller to terminate the trial and to answer the comprehension questions.

Procedure. Participants were calibrated on a 9-point calibration grid. The calibration accuracy was monitored with a drift check before each trial and participants were recalibrated

[3] Half of the Mandarin speech sounds were looped for the last 2s because the sentences were not long enough to create 60 s of unique speech. The looped speech was reached on only one trial and the seven fixations that occurred during that time were removed from further analysis.

whenever necessary. The average calibration error was kept at ≤ 0.4 º. Each trial started with a black gaze box that appeared at 50 pixels on the x-axis and 150 pixels on the y-axis of the screen. Once participants fixated the box, the paragraph appeared on the screen, with the first letter of the first sentence presented in the middle of where the box was. The onset of the background sound was simultaneous with the appearance of the paragraph on the screen. Each question difficulty block started with the two practice paragraphs. Participants were not informed about the difficulty of the questions prior to the experiment and were simply told that some of them will require a yes/no answer, while others will require a multiple-choice answer. The paragraphs and each of the comprehension questions appeared for a maximum of 60 s on the screen. This duration was determined to be sufficient based on the pilot results. The experiment lasted for about 40-50 minutes.

Data analysis. A few measures of global reading were analyzed: paragraph reading time, number of first- and second-pass fixations, intra-sentence, inter-sentence regression probability, saccade length, and saccade landing position. In Experiment 2, we use the term "intra-sentence" regression to denote the probability of making a regression within the currently-read sentence. This is the traditional measure of regression probability that was reported in Experiment 1 and in most of the existing literature. In contrast, "inter-sentence" regression refers to cases where participants regress to a previously-read sentence. This distinction was introduced to test whether background speech disrupts only the integration of text information within sentences or also integration between sentences. Additionally, sentence re-reading time and sentence look-back time were also analysed. Sentence re-reading time was defined as the sum of all re-reading fixations within the currently-read sentence before the eyes moved on to the next sentence (Liversedge, Paterson, & Pickering, 1998). Sentence look-back time was defined as the sum of

all re-reading fixations in a sentence when participants regress back from a subsequent sentence (Hyönä, Lorch, & Rinck, 2003). Furthermore, the three local measures of word reading were also analyzed: FFD, GD, and TVT. In the analyses of local reading measures, all words in all sentences were included. Finally, comprehension accuracy was analyzed as accuracy above chance level due to the different chance levels in the two question difficulty conditions (50% for the easy questions and 25% for the difficult questions). Two separate models are reported for participants (b_1) and items (b_2) because analysing the data in terms of comprehension accuracy above chance level requires calculating the mean accuracy for each condition and then subtracting the chance level performance from it.

The data were analyzed with (G)LMMs by using the "lme4" package v.1.1-12 (Bates et al., 2014) in the R statistical software v.3.3.1 (R Core Team, 2016). Background sound and question difficulty were entered as fixed effects in the models. Random intercepts, as well as random slopes for background sound and question difficulty were specified for both participants and items (Baayen, Davidson, & Bates, 2008; Barr, Levy, Scheepers, & Tily, 2013). Due to convergence failure, the following random slopes were removed: background sound was removed as a random slope for items for saccade length, number of first fixations, gaze duration, and sentence re-reading time; question difficulty was removed as a random slope for items for inter-regression probability and saccade landing position. Treatment contrasts were used for the background sound condition (with English speech as the baseline). Sum contrasts were used for the question difficulty condition (-1: easy; 1: difficult). Fixation durations were log-transformed in all analyses. Similar to Experiment 1, the results were corrected for multiple comparisons due to the additional comparison between Mandarin and Noise. Results were considered statistically significant if the adjusted *p*-values were ≤ 0.05.

Results

Comprehension accuracy. The results for comprehension accuracy are presented in Figure 3a. There was a main effect of question difficulty (b_1= 0.33, *SE*= 0.03, *t*= 9.89, *p*< 0.001; b_2= 0.33, *SE*= 0.03, *t*= 9.49, *p*<0.001; *d*= -0.41), indicating that comprehension was significantly lower on the difficult compared to the easy questions. However, there was no significant difference in comprehension accuracy between English and Silence, English and Noise, or Mandarin and Noise (all *p*s > 0.12). The difference between English and Mandarin was significant by subjects (b_1= 0.06, *SE*= 0.02, *t*= 2.51, *p*= 0.03), but not by items (b_2= 0.06, *SE*= 0.03, *t*= 2.07, *p*= 0.10). Therefore, there were generally no significant differences in comprehension accuracy between the sound conditions and the hint of an effect in the comparison between English and Mandarin was driven by the slightly better accuracy in Mandarin compared to Silence. Contrary to hypothesis H3, there were also no significant interactions between background sound and question difficulty for any of the comparisons (all *p*s ≥ 0.61). In summary, English speech did not impair comprehension accuracy in Experiment 2. Even though difficult questions resulted in significantly lower accuracy compared to easy questions, the accuracy on any of the sounds did not interact with question difficulty. In this sense, there was no support for the suggestion that English speech disrupts comprehension accuracy only for the difficult, but not for the easy questions.

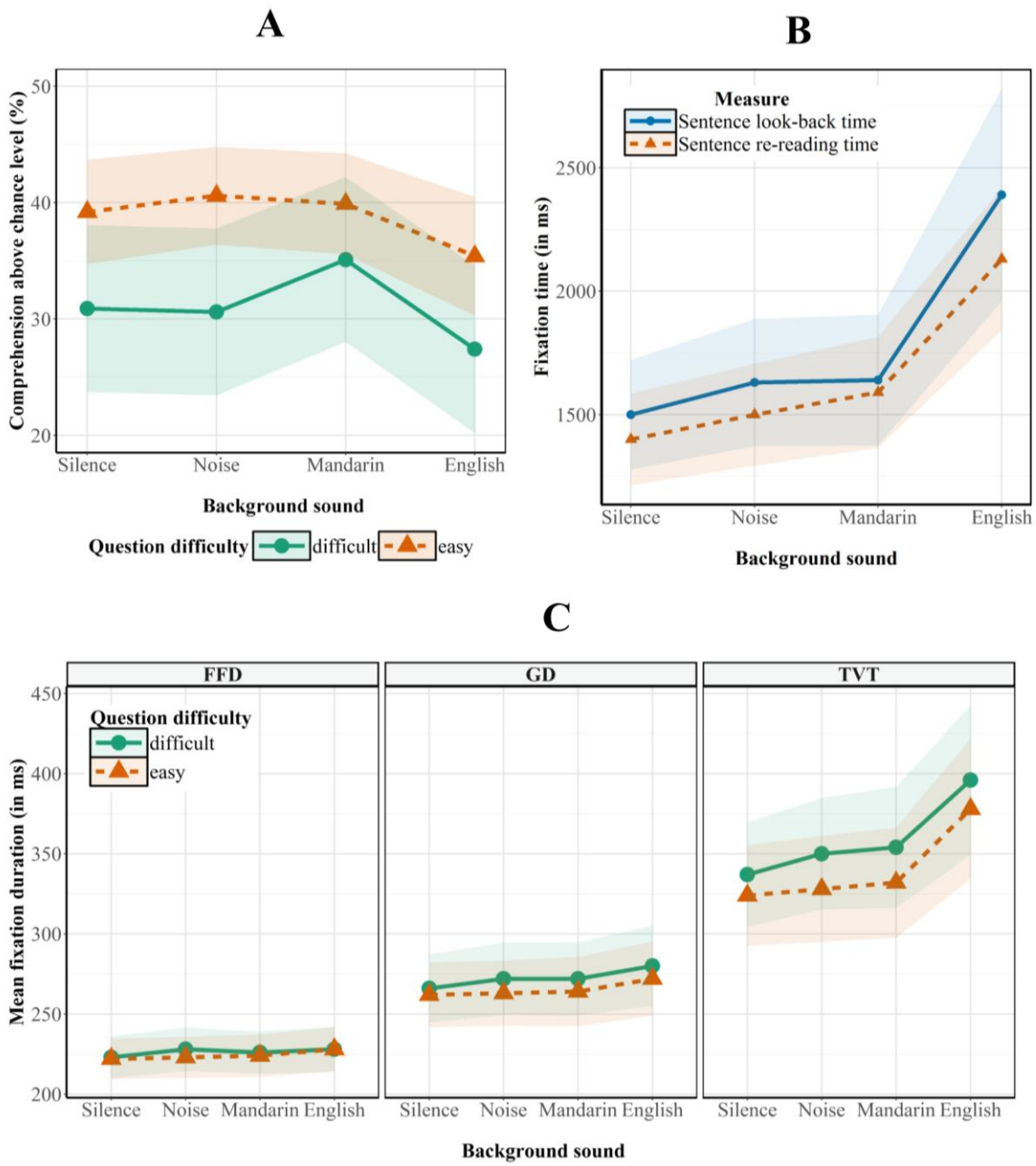

Figure 3. Mean descriptive statistics for Experiment 2. Panel **A**: comprehension accuracy above chance level. Panel **B**: sentence re-reading time and sentence look-back time. Panel **C**: Local word-level reading measures. Shading indicates the standard error.

Although there was no significant difference in comprehension accuracy between English and Silence, it is not immediately obvious why the lack of effect occurred. It is important to

determine whether there is no true difference in comprehension accuracy when the text is read in Silence and under conditions of English speech (i.e., the null hypothesis is true), or alternatively, whether such a difference does exist (i.e., the alternative hypothesis is true), but the present experiment was not sufficiently powered to detect it. We used Bayes factors to discriminate between these two possibilities (see Dienes, 2014, 2016). Bayes factor regression analyses (Rouder & Morey, 2012) were carried out with the "BayesFactor" R package (Morey, Rouder, & Jamil, 2015). This test yields a Bayes factor, which is the posterior odds of the null and the alternative hypothesis, given the data. Bayes factors greater than 1 favor the alternative hypothesis, whereas Bayes factors smaller than 1 favor the null hypothesis. The default prior width of r= $\sqrt{2}/2$ was used from the package. We show in the Supplemental Materials that the choice of prior did not influence the conclusions from these analyses.

The comparison between English speech and Silence in comprehension accuracy showed substantial evidence in support of the null hypothesis of no difference (subjects: BF= 0.18; items: BF= 0.21; see Jeffreys, 1961; Wetzels et al., 2011). Additionally, the analysis favoured the null hypothesis of no interaction between question difficulty and the contrast between English and Silence (subjects: BF= 0.15; items: BF= 0.21). The remaining contrasts between English and Mandarin, English and Noise, and Mandarin and Noise also favoured the null hypothesis of no difference and no interaction with question difficulty (range of BFs: 0.12 - 0.44). Therefore, the Bayes factor analysis suggested that there was no true mean difference in the contrast between English and Mandarin that was significant by subjects in the LMM analysis above. A Bayes factor analysis of the comprehension accuracy data from Experiment 1 also supported the null hypothesis of no difference between English and Silence (BF= 0.17) or English and Mandarin (BF= 0.12), which is also in line with the present results. In summary, the Bayes factor analyses

provided direct evidence that there is no difference in comprehension accuracy between English speech and Silence. They also confirmed the LMM results by showing that the disruption in comprehension by English speech is not modulated by the difficulty of the questions.

Pre-processing of eye-tracking data. Fixation durations were manually pre-processed with the EyeDoctor software (Stracuzzi & Kinsey, 2009) to align the vertical position of fixations (whenever necessary), and to remove blinks from the data (5.81 % of all fixations). Fixations shorter than 80 ms that occurred within one letter of another fixation were combined with that fixation. Any remaining fixations shorter than 80 ms were excluded (1.4 % of the data). Additionally, fixations greater than 1000 ms were excluded as outliers (0.11 % of the data). Furthermore, for the analyses of word time reading measures, FFD longer than 1000 ms, GD longer than 2200 ms, or TVT longer than 3000 ms were discarded as outliers (0.1 % of the data). Although cut-offs of 800 ms for FFD and 2000 ms for GD are typically used in single-line reading studies (e.g. Risse & Kliegl, 2014; Schotter, Lee, Reiderman, & Rayner, 2015), using them resulted in a highly disproportionate number of outliers excluded per sound condition (χ2 (3) = 14.548, p= 0.002). Increasing the cut-offs by 200 ms ensured there were no significant differences in the number of outliers excluded per condition (χ^2 (3) = 4.09, p= 0.27), while still removing the longest fixation durations that may not reflect normal reading[4]. This was also justified by the fact that participants were reading paragraphs which naturally contained longer compound words that are less commonly used in single-line reading studies such as Experiment 1 (see the Supplemental Materials).

[4] A re-analysis of the data with the outlier cut-offs from Experiment 1 did not change the main results or the conclusions from the analysis.

Global reading measures. The descriptive statistics for global reading measures are presented in Tables 4 and 5. The results from the (G)LMMs are presented in Table 6 for all dependent measures, with the exception of saccade landing position, for which the results are reported in the text. English speech resulted in significantly longer paragraph reading time (*d*= -0.47), greater intra-sentence regression probability (*d*= -0.14), and more second-pass fixations (*d*= -0.15) compared to Silence. The difference between English and Noise was significant for paragraph reading time (*d*= -0.37), saccade length (*d*= 0.02), intra-sentence regression probability (*d*= -0.13), and number of second-pass fixations (*d*= -0.14). The contrast between English and Mandarin was significant for paragraph reading time (*d*= -0.36), saccade length (*d*= 0.02), intra-sentence regression probability (*d*= -0.09), number of first-pass fixations (*d*= -0.05) and number of second-pass fixations (*d*= -0.11). There were no differences in saccade landing position for any of the experimental conditions (all *p*s ≥ 0.07).

Table 4

Mean Descriptive Statistics of Global Reading Measures in Experiment 2 (SDs in parenthesis)

Sound condition	Question difficulty	Paragraph reading time (in s)	Intra-sentence regression probability	Inter-sentence regression probability	Number of fixations (per word)		
					1st-pass	2nd-pass	Total
Silence	difficult	25.9 (8.70)	.25 (.43)	.11 (.31)	.81 (.83)	.28 (.61)	1.09 (1.02)
Silence	easy	24.3 (8.00)	.25 (.43)	.08 (.27)	.78 (.79)	.26 (.61)	1.05 (0.99)
Noise	difficult	27.0 (9.79)	.27 (.44)	.11 (.32)	.83 (.85)	.30 (.66)	1.13 (1.08)
Noise	easy	24.6 (9.59)	.24 (.43)	.09 (.28)	.79 (.81)	.26 (.59)	1.04 (1.01)
Mandarin	difficult	26.9 (8.85)	.27 (.45)	.12 (.32)	.81 (.86)	.31 (.68)	1.12 (1.11)
Mandarin	easy	25.0 (9.50)	.27 (.44)	.08 (.27)	.77 (.79)	.28 (.65)	1.05 (1.01)
English	difficult	30.5 (11.54)	.32 (.47)	.15 (.36)	.85 (.99)	.40 (.84)	1.25 (1.33)
English	easy	28.8 (10.54)	.31 (.46)	.12 (.33)	.82 (.88)	.36 (.80)	1.18 (1.23)

The comparison between Mandarin and Noise revealed a significant difference only for intra-sentence regression probability (b= 0.11, SE= 0.05, z= -2.29, p= 0.022, d= -0.03). There were no significant differences for any other measures (all $ps \geq 0.07$). Therefore, similar to Experiment 1, the results supported most strongly hypothesis H2, which stated that disruption effects by intelligible speech are only semantic in nature. There was limited evidence in support of hypothesis H2.1, which stated that the disruption by intelligible speech has both a semantic and a phonological component. However, similar to Experiment 1, this limited support for a contribution of phonology was found in only one measure (intra-sentence regression probability), and even this measure was not the same as the one from Experiment 1 (number of second-pass fixations).

Table 5

Mean Saccade Length and Saccade Landing Position in Experiments 2 and 3 (in Letters; SDs in Parenthesis)

Experiment 2				Experiment 3				
Sound	Question difficulty	Saccade length	Landing position	Sound	Question difficulty	Reading condition	Saccade length	Landing position
Silence	difficult	8.47 (5.63)	2.90 (2.29)	Silence	difficult	normal	9.16 (6.48)	2.71 (2.34)
Silence	easy	8.47 (5.48)	2.88 (2.28)	Silence	difficult	mask	8.82 (6.21)	2.85 (2.42)
Noise	difficult	8.50 (5.74)	2.90 (2.30)	Silence	easy	normal	9.08 (6.94)	2.66 (2.37)
Noise	easy	8.50 (5.37)	2.87 (2.28)	Silence	easy	mask	8.88 (6.32)	2.72 (2.39)
Mandarin	difficult	8.42 (5.70)	2.94 (2.33)	English	difficult	normal	8.88 (6.53)	2.75 (2.38)
Mandarin	easy	8.52 (5.72)	2.83 (2.24)	English	difficult	mask	8.90 (6.58)	2.80 (2.38)
English	difficult	8.30 (5.71)	2.93 (2.31)	English	easy	normal	8.96 (6.60)	2.69 (2.33)
English	easy	8.47 (5.63)	2.85 (2.25)	English	easy	mask	8.78 (6.45)	2.74 (2.38)

Table 6

Results from (G)LMMs on Global Measures of Reading in Experiment 2

Effect	Paragraph reading time				Saccade length				Intra-sentence regression probability			
	b	SE	t	p	b	SE	t	P	b	SE	z	p
Intercept	3.32	.05	63.9	**<.001**	8.52	.21	40.4	**<.001**	-.86	.06	-13.6	**<.001**
Eng vs Slc	-.18	.04	-4.64	**<.001**	.19	.10	1.89	.13	-.32	.06	-5.61	**<.001**
Eng vs Noise	-.14	.02	-5.77	**<.001**	.23	.09	2.69	**.02**	-.32	.06	-5.79	**<.001**
Eng vs Mnd	-.13	.02	-5.43	**<.001**	.20	.07	2.79	**.02**	-.22	.05	-4.58	**<.001**
Diff	.03	.02	1.55	.12	-.06	.05	-1.41	.32	.04	.02	1.66	.20
Diff: Eng vs Slc	-.02	.02	-.90	.74	.05	.05	.96	.67	-.02	.02	-0.78	.43
Diff: Eng vs Noise	.02	.02	1.08	.56	.06	.05	1.23	.44	.04	.02	1.92	.09
Diff: Eng vs Mnd	.02	.02	.89	.75	.01	.05	.16	.87	<-.01	.02	-.16	.87

Effect	Inter-sentence regression probability				Number of 1st-pass fixations				Number of 2nd-pass fixations			
	b	SE	z	p	b	SE	z	P	b	SE	z	p
Intercept	-2.84	.25	-11.5	**<.001**	-.21	.04	-5.65	**<.001**	-1.07	.07	-15.4	**<.001**
Eng vs Slc	-.25	.24	-1.04	.59	-.03	.02	-1.78	.15	-.35	.05	-7.11	**<.001**
Eng vs Noise	-.27	.19	-1.43	.31	-.03	.02	-1.65	.20	-.35	.06	-6.42	**<.001**
Eng vs Mnd	-.14	.18	-.76	.90	-.05	.02	-2.33	**.04**	-.27	.04	-6.16	**<.001**
Diff	.18	.06	3.00	**.01**	.01	.01	1.52	.13	.06	.02	2.39	**.02**
Diff: Eng vs Slc	-.01	.03	-.19	.85	.01	.01	.57	.91	-.01	.02	-0.69	.49
Diff: Eng vs Noise	.11	.03	3.46	**.001**	.01	.01	.69	.91	.04	.02	2.61	**.02**
Diff: Eng vs Mnd	.01	.03	.35	.72	.01	.01	1.51	.26	.01	.02	.63	.53

Note: Eng: English. Slc: Silence. Mnd: Mandarin. Diff: question difficulty. Statistically significant *p*-values are formatted in bold.

The results also showed a significant main effect of question difficulty for two of the dependent measures. Participants made more inter-sentence regressions (d= 0.10) and more second-pass fixations (d= 0.05) when answering difficult compared to easy questions. These results show that the block of paragraphs with difficult questions prompted participants to adopt

a more careful reading strategy, in which they made more re-reading fixations, and regressed more often to previous words and sentences. Additionally, the contrast between English speech and Noise interacted significantly with question difficulty for inter-sentence regression probability and number of second-pass fixations. For both measures, the interaction was due to the fact that the difference between English speech and Noise was smaller in the difficult compared to the easy question condition.

One question of particular interest in Experiment 2 was how intelligible speech affects the integration of information across sentences. To determine this, we compared the disruption in sentence re-reading time and sentence look-back time. If disruption is limited only to the currently-read sentence, there should be a disruption only in sentence re-reading time, but not in look-back time. On the other hand, if intelligible speech affects sentence integration processes, such a disruption should also be observed in look-back time. The descriptive statistics are plotted in Figure 3b. English speech resulted in longer sentence re-reading time compared to Silence (b= -0.38, SE= 0.04, t= -9.14, p< 0.001, d= -0.43), Noise (b= -0.34, SE= 0.05, t= -7.62, p< 0.001, d= -0.36), and Mandarin (b= -0.26, SE= 0.04, t= -7.10, p< 0.001, d= -0.30). However, the difference between Mandarin and Noise was not significant (b= -0.08, SE= 0.05, t= -1.61, p= 0.12, d= 0.06). There were no differences in look-back time between any of the sound conditions (all ps ≥ 0.16). This suggests that the increase in re-reading behaviour was mostly constrained to the currently-read sentence as the difference in look-back time did not reach statistical significance. Nevertheless, it should be noted that English speech resulted in a numerically similar increase in both sentence re-reading time and sentence look-back time (see Figure 3b). An examination of the subject means indicated that there was greater between-subject variability in sentence look-back times, which may have contributed to the lack of a significant difference in that measure.

Therefore, the difference in the results between the two measures may be more quantitative than qualitative in nature.

Word-level reading measures. The descriptive statistics for local fixation duration measures are shown in Figure 3c. English speech resulted in significantly longer TVT compared to all other sound conditions (Silence: *b*= -0.10, *SE*= 0.02, *t*= -5.31, *p*< 0.001, *d*= -0.21; Noise: b= -0.09, SE= 0.02, t= -5.48, d= -0.17; Mandarin: b= -0.08, SE= 0.02, t= -5.21, d= -0.16). English speech also resulted in longer GD compared to both Silence (*b*= -0.03, *SE*= 0.01, *t*= -3.54, *p*= .002, *d*= -0.08) and Mandarin (b= -0.02, SE= 0.01, t= -3.01, p= 0.01, d= -0.05). The only significant difference in FFD was between English speech and Silence (*b*= -0.02, *SE*= 0.01, *t*= -2.33, *p*= 0.05, *d*= -0.05). Therefore, the disruption effects in TVT from Experiment 1 were replicated; additionally, there was also some evidence for disruption in first-pass reading measures (FFD and GD). Consistent with Experiment 1, there were no differences between Mandarin and Noise in word-level reading measures (all *p*s ≥ 0.56). In summary, the analysis of local word-level reading measures supported hypothesis H2, which stated that the disruption effect by intelligible speech is only semantic in nature. Contrary to hypotheses, H1.2 and H2.1, there was no evidence for a contribution of phonology.

Furthermore, there was a significant effect of question difficulty for TVT (*d*= 0.08), which indicated that TVT was longer when participants were answering difficult compared to easy questions. Finally, question difficulty interacted significantly with the comparison between English and Noise for FFD. This was because FFD was longer in English speech compared to Noise, but only when the questions were easy to answer (*d*= 0.05). There were no other significant interactions between question difficulty and background sound (all *p*s ≥ 0.1).

Discussion

Experiment 2 investigated the effect of intelligible background speech on comprehension accuracy and online integration processes during paragraph reading. The eye-movement measures replicated the disruption effects of intelligible speech found in measures of second-pass reading in Experiment 1. In fact, the amount of disruption was greater than what was observed in the single-sentence reading paradigm of Experiment 1. This was because, on average, the size of the effects in Cohen's *d* was 76 % greater in the comparison between English speech and Silence and 84% greater in the comparison between English speech and Mandarin speech. Additionally, unlike Experiment 1, there was some evidence that intelligible speech also disrupted first-pass reading. More specifically, gaze durations were longer in English speech compared to both Mandarin speech and Silence, and first fixation durations were also longer in English speech compared to Silence (but not compared to Mandarin). Participants also made more first-pass fixations in English speech compared to Mandarin (but not compared to Silence).

In this sense, the disruption in paragraph reading was greater than the disruption in sentence reading (Experiment 1) because the magnitude of the effects in second-pass measures was greater and there was at least some evidence that first-pass reading measures were also affected. Because reading connected sentences requires the construction of a discourse model of the text (see Gernsbacher & Foertsch, 2000; O'Brien & Cook, 2015), the greater magnitude of the disruption in paragraph reading may be due to a difficulty in constructing a coherent discourse of the paragraph (Kehler, 2004; Wolf & Gibson, 2005). Additionally, the increase in text context may also explain why an effect in first-pass measures of reading was observed in Experiment 2, but not in Experiment 1.

While the text stimuli were longer in Experiment 2 and participants may have had more opportunity to go back and re-read the text, the probability of making a regression within the

current sentence was comparable in the two experiments (23% in Experiment 1 vs. 25% in Experiment 2 in the silence condition). Additionally, the probability of making a regression to previous sentences (9.5% in the silence condition) was more than twice as low, thus suggesting that such regressions were not as common as regressions within the currently-read sentence. Therefore, the stronger effects in measures of second-pass reading are not likely to be explained by the text stimuli being longer. In Experiment 2, participants also made 22.1% fewer first-pass fixations and 40.2% fewer second-pass fixations compared to Experiment 1. However, at the same time, fixation durations increased by 5.7 % for FFD and by 9.7 % for TVT across all conditions. This suggests that, compared to Experiment 1, participants made fewer but longer fixations in both first-pass and second-pass reading.

Similar to Experiment 1, the results provided strong evidence for the semantic disruption account (Marsh et al., 2008, 2009; Martin et al., 1988). This was because English speech resulted in greater disruption compared to all other sound conditions in measures of both second-pass reading and first-pass reading (gaze durations). Therefore, because English speech resulted in a greater disruption compared to Mandarin speech, there was again no support for the strong form of the phonological disruption account (H1; Salamé & Baddeley, 1982, 1987), which stated that any disruption is due only to the phonology of speech. However, there was limited support for the weaker version of the phonological disruption account (H1.2) because Mandarin speech resulted in greater intra-sentence regression probability compared to Noise. This suggests that there may be limited contribution of phonology to the disruption effects by intelligible speech (which would be consistent with hypothesis H2.1), but this was found in only one measure and the same effect was not observed in Experiment 1 in that same measure. Therefore, the present findings are again most readily accounted by hypothesis H2, which stated that the disruption by

intelligible speech is only semantic in nature. We will revisit the role of phonology in distraction by intelligible speech in the General Discussion, but for now we note that there was only limited evidence in support of a contribution by phonology.

One of the contributions of Experiment 2 was that it investigated how information is integrated across multiple sentences. Generally speaking, there was no evidence to suggest that the integration of information across sentences is disrupted because participants made more regressions to previous sentences when listening to English speech in the background compared to silence. Additionally, the time that they spent re-reading the sentence during such regressions (i.e., look-back time) did not differ between the sound conditions. This is largely consistent with Hyönä and Ekholm's (2016) findings, because the authors also reported no effects in look-back times in three out of their four experiments (the only significant difference was between silence and scrambled speech in Experiment 3). Furthermore, there was no difference between (non-scrambled) intelligible speech and silence in Hyönä and Ekholm's (2016) Experiments 1 and 3, which is also in agreement with the present results. Interestingly, Cauchard et al.'s (2012) finding that intelligible speech led to longer sentence look-back times is contrary to both the present findings and Hyönä and Ekholm's (2016) results. Therefore, further research is required to determine the boundary conditions under which such an effect is observed. We speculate that this discrepancy could potentially be due to differences in the speech stimuli or the text that participants were reading. These are potential mediating factors that have not been thoroughly investigated so far in studies on auditory distraction by intelligible speech.

While the difference was not significant, it is also worth noting that English speech resulted in a numerically greater look-back time compared to Silence and this difference was similar in its numerical magnitude to the disruption effect in sentence re-reading time. An

examination of the participant means indicated that there was considerable between-subject variability. Because of this, future studies might investigate whether individual differences may modulate the effect of intelligible speech on sentence look-back time. For example, the time that participants spend re-reading previous sentences could be related to their ability to suppress the irrelevant background speech (see Sörqvist, Halin, et al., 2010; Sörqvist, Ljungberg, & Ljung, 2010). At any rate, the present study suggests that the increase in re-reading behaviour in response to intelligible speech is mostly constrained to the currently-read sentence and likely does not also extend to previously-read sentences. Therefore, the observed disruption in second-pass reading in the present research is likely not related to a difficulty in integrating text meaning across multiple sentences. Rather, it likely reflects a transient difficulty in integrating the meaning of individual words within the current sentence in order to form the meaning of that sentence.

Although intelligible speech resulted in a considerable disruption of eye-movements, comprehension accuracy remained unaffected in both question difficulty conditions. This suggests that participants could maintain a similar level of text comprehension with English speech in the background, even when the questions probed a deeper level of text understanding. This points to the fact that the disruption observed in eye-movement measures in the English speech condition reflects participants' attempt to successfully attain comprehension in the distracting reading conditions. The results from eye-movement measures provide converging evidence to the same effect. The experimental block with difficult comprehension questions led to a change in eye-movement behaviour, which was characterised by more regressions to previous sentences and longer word re-reading times. However, the disruption effect by English speech did not interact with question difficulty, thus suggesting that the amount of disruption did

not depend on the task demands imposed by the question difficulty manipulation. In this sense, there was no evidence that the disruption effect in eye-movement measures increased in the block with difficult questions. Rather, participants were able to adapt to the different task demands, and the magnitude of the disruption was proportional to these demands.

The effect of question difficulty on eye-movements further suggests that participants can make strategic decisions about the nature of the reading task and adjust their reading behavior accordingly. For example, the increase in number of fixations and the probability of making a regression to previous sentences in the condition with difficult questions could be due to an attempt to engage in more effective discourse processing in order to develop a richer representation of the meaning of the text. This may occur in response to the expectation that participants will be asked more difficult and more detailed comprehension questions. Similar evidence of such "meta" control over eye-movements has also been found in response to the type of text that participants are reading. For example, participants make more regressions and have longer fixation durations when reading scientific texts compared to reading newspaper articles or light fiction (Rayner, Pollatsek, Ashby, & Clifton, 2012).

Finally, because the difficult questions received a difficulty rating of 2.7 on a 5-point scale in the pilot study, it could be argued that the lack of interaction between question difficulty and background sound in comprehension accuracy could be due to the difficult questions still not being challenging enough. However, the fact that the block with difficult questions prompted participants to read the paragraphs more carefully clearly suggests that the difficult questions were more challenging than the easy ones. Additionally, the difficulty rating was subjective in nature and thus may not perfectly correlate with participants' actual performance (i.e., one can

judge the questions to be easy and still answer them incorrectly)[5]. Therefore, even though the difficult questions were still fairly challenging, future studies may wish to utilise even more difficult questions. However, it is worth noting that if the questions are so difficult that accuracy is close to chance-level performance, they will have a poor psychometric sensitivity to detect any potential auditory distraction effects.

Experiment 3

In Experiment 2, participants could re-read the paragraphs as they wished until their allocated time was over. Therefore, the lack of difference in comprehension accuracy between the silence and English speech conditions may have occurred because participants were able to compensate for the experienced distraction by making more regressions and more second-pass fixations. In other words, the increase in re-reading behaviour may occur because participants are actively trying to comprehend the passages at the same level as when they read them in silence. However, because intelligible speech occasionally leads to a transient interference in processing the meaning of the text, participants may need to temporarily interrupt the progressive reading of the text to resolve this interference before moving on to the unexplored text. We will refer to this explanation as the *distraction re-reading* hypothesis.

Previous evidence showing that eye-movements are sensitive to online processing difficulty (see Rayner, 1998, 2009) lends some plausibility to this hypothesis. For example, regressive eye-movements play an important role in resolving temporary sentence ambiguities (e.g., Frazier & Rayner, 1982; Meseguer, Carreiras, & Clifton, 2002). Additionally, the number of fixations that participants make is a sensitive measure of text difficulty and regressive eye-

[5] The point-biserial correlation between accuracy and difficulty rating in the pilot data was $r = -0.48$ overall ($r = -0.36$ on the difficult and $r = -0.34$ on the easy questions). This supports the view that the difficulty rating is only moderately related to participant's performance on the comprehension assessment.

movements increase when there are inconsistencies in the text (Rayner, Chace, Slattery, & Ashby, 2006). Furthermore, regressions seem to allow for additional word processing to occur, which can subsequently influence participants' understanding of the sentence (Booth & Weger, 2013).

There is also more direct evidence showing that regressions support comprehension. For example, Schotter, Tran, and Rayner (2014) used a new manipulation (the so-called *trailing mask* paradigm) to prevent participants from re-reading previous words during a regression. In this paradigm, words are masked by a string of 'x's once participants move to the right of them, thus rendering re-reading useless. Schotter et al. found that preventing participants from re-reading words in a sentence had a negative effect on their comprehension. In the light of these findings, we hypothesized that the increase in regressions and re-reading fixations in the intelligible speech condition is crucial for maintaining the immediate text comprehension in the face of distraction. We expected that comprehension would be compromised if participants could no longer re-read previous words in the text.

In Experiment 3, we used Schotter et al.'s (2014) trailing mask paradigm to prevent participants from re-reading previous words and sentences. The experiment had a 2 x 2 x 2 within-subject design with the following factors: background sound (English speech vs silence), reading condition (normal text vs trailing mask text), comprehension question difficulty (easy vs difficult). To preserve statistical power and because the critical comparison for the present hypothesis is between silence and English speech, the Mandarin and speech-spectrum noise conditions were removed. We expected that English speech will disrupt comprehension compared to the silence condition, but only in the trailing mask condition where no re-reading is possible. Additionally, similar to Experiment 2, we also predicted that the disruption in

comprehension accuracy in the trailing mask condition will be greater for the difficult compared to easy questions. Finally, we also expected to replicate the disruption effects by English speech in measures of second-pass reading from the previous experiments. In summary, the predictions were:

H1: English speech will disrupt comprehension accuracy only when participants cannot re-read previous text in the trailing mask condition (distraction re-reading hypothesis).

H2: In the trailing mask condition, English speech will disrupt comprehension more on the difficult compared to the easy questions.

Method

Participants. Forty-eight Bournemouth University students participated for course credit or a payment of £10 (60.4% female). None of them had participated in the previous experiments. Their mean age was 20.6 years (*SD*= 2.4 years; range: 18-32 years). Two more participant were tested, but their data were excluded due to tracking problems. The study had the same statistical power as Experiment 2 and was therefore sufficiently powered.

Materials and design. The same reading materials from Experiment 2 were used. The English speech was taken from the BKB (Bench et al., 1979) and IHR (MacLeod & Summerfield, 1990) corpora. Twelve 60 s speech files were created by concatenating between 40 to 42 unique speech sentences each and removing the silence gaps between sentences. Half of the sound files contained speech spoken by a male British English speaker and the remaining half contained speech spoken by a female British English speaker.

There were two reading conditions in the experiment: *normal text* (i.e., with no visual changes on the screen) and *trailing mask* text. In the trailing mask condition, each word in the text was permanently masked by a string of ‘x’s once participants made a saccade to the right of

it (see Figure 4b for an illustration). The empty spaces between words were kept in the masked text, which preserved its general outline. This type of masking was identical to the one used by Schotter et al. (2014).

Because Experiment 3 used paragraphs instead of single sentences, it was necessary to extend Schotter et al.'s (2014) trailing mask manipulation for use in a multiple-line reading paradigm[6]. This was needed as the error in tracking the vertical position of the eye can cause incorrect triggering of the display changes in the experiment. Pilot testing indicated that the least obtrusive way to implement this was to add a gaze-contingent check (a small square) at the end of each line that participants had to fixate to indicate they had finished reading the current line. At the start of each trial, only the first line was visible. Once the gaze-contingent check at the end of the first line was triggered, the square disappeared and the next line was automatically revealed[7]. This procedure was then repeated until the whole paragraph had been presented (see Figure 4a for an illustration of the method). To avoid delays due to having to fixate exactly within the square, the line check was triggered immediately after participants' gaze moved to

[6] It should be noted that Olkoniemi, Johander, and Kaakinen (2018) have recently also used the trailing mask paradigm in a paragraph-reading study. However, in their experiment the trailing mask was triggered at the sentence level and not word-by-word as in the present research. Additionally, participants manually triggered the mask by pressing a button.

[7] To ensure that the trailing mask is accurately triggered on the next line, the display changes started when participants made a rightwards (i.e., progressive) saccade to a new word. This was necessary as the return sweep saccade from the end of the previous line to the beginning of the next line can sometimes undershoot the line start, which may be followed by a corrective saccade to the left (Andriessen & de Voogd, 1973; Hofmeister, Heller, & Radach, 1999; Rayner, 1998). Such undershoot fixations are generally not thought to be related to text processing (Abrams & Zuber, 1972) and are much shorter than the average fixation during reading. In Experiment 3, participants landed short of the line start and made a corrective saccade to the left on 41.1% of all line crosses. The average duration of the undershoot fixation was 110 ms (*SD*= 59 ms). The advantage of allowing readers to make a return sweep to the next line was that it kept the reading process more natural. This approach was preferred because a pilot study in which participants had to fixate a gaze box at the start of each new line was found to be too disruptive to the reading process due to the delays in triggering the gaze boxes.

right of the last word on the line (i.e., the square and the space around it simply acted as a catchment area).

The text stimuli were presented in this way in all trials to keep the reading conditions constant across the experiment. Similar to Experiment 2, the background sound and question difficulty conditions were presented in separate blocks. The order of items within each block was randomised. Similar to Schotter et al.'s (2014) experiment, the normal text and trailing mask text trials were intermixed within blocks, but participants received a cue before the start of each trial that told them what type of text they will be reading. In the present study, a black gaze box at the start of each trial indicated that participants will be reading normal text, whereas a blue gaze box indicated that they will be reading the trailing mask text. All blocks and conditions were counter-balanced with a full Latin square design across participants.

Apparatus. The equipment was the same as in Experiment 2, except for the following differences. The experiment was programmed in Matlab 2014a (MathWorks, 2014) with the Psychophysics toolbox (Brainard, 1997; Pelli, 1997) and Eyelink libraries (Cornelissen, Peters, & Palmer, 2002). The text was displayed with the same dimensions and spatial layout as in Experiment 2, but some lines were made shorter to make enough space for the fixation check at the end of each line. The fixation check was a 16 x 16-pixel black square that was situated 3 letter spaces (33 pixels) to the right of the last word on the line. All paragraphs fitted on a single screen. The display changes were completed, on average, within 9.12 ms of the eye moving to the right of each individual word (*SD*= 1.98 ms).

Procedure. Participants were tested individually in a 45-minute session. They were told that the paragraph will be revealed line by line and that they will need to fixate a small square at the end of each line to reveal the next line. Furthermore, participants were informed that the

words in some paragraphs will be masked by ‘x’s after they have read them, but that they should try to read the text as normally as possible. They were also told that the colour of the gaze box will indicate what type of text they will be reading.

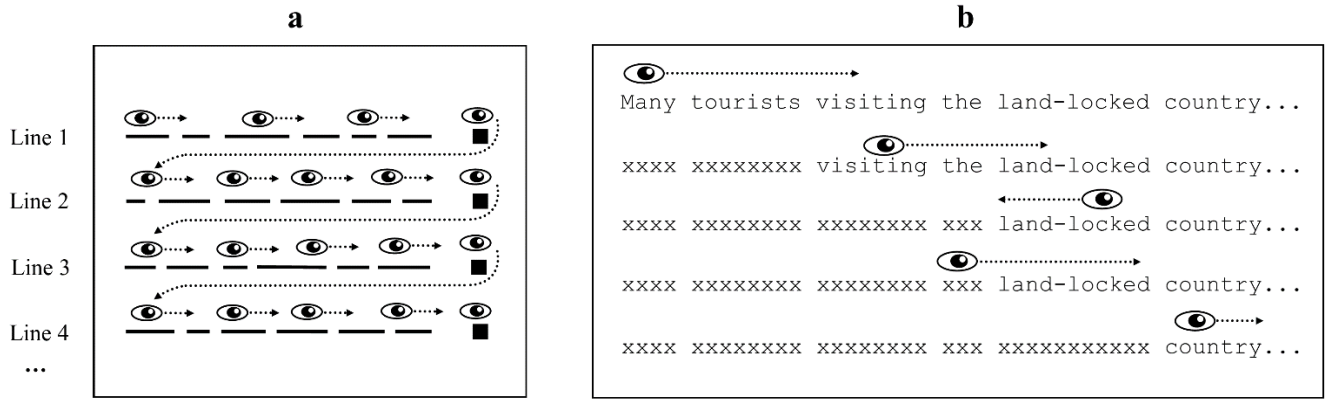

Figure 4. An illustration of the text stimuli presentation in Experiment 3. Panel **a** shows a schematic representation of the line-by-line text presentation (with horizontal lines denoting the text). At the start of each trial, only the first line was visible. Participants then revealed each new line of text by fixating a small black square at the end of each line until the whole paragraph was revealed. Panel **b** shows an example of the trailing mask reading condition. Words were permanently masked by a string of ‘x’s once the eye moved to the right of each word.

Each question difficulty block started with two practice trials. One practice trial was displayed in the normal text condition, while the other one was displayed in the trailing mask condition. In the trailing mask condition, each word was masked after participants made a saccade to the right of it. This was accomplished by placing an invisible boundary (Rayner, 1975) located at the first pixel after the end of each word. Once the boundary was crossed, the word was permanently masked for the remainder of the trial. Participants clicked the left button of the mouse to terminate the trial and to select the correct answer to the comprehension questions. Trials could be terminated only after all lines had been revealed.

Data analysis. The experiment had a 2 (background sound: English speech vs silence) x 2 (reading condition: trailing mask text vs normal text) x 2(comprehension question difficulty: easy vs difficult) design. The same global reading measures from Experiment 2 were analysed: paragraph reading time, number of first- and second-pass fixations, intra-sentence, inter-sentence regression probability, saccade length, and saccade landing position. Additionally, FFD, GD, and TVT were analysed as local word-level measures. Sum contrast coding was used for all variables: background sound (silence: -1; English: 1), reading condition (trailing mask: -1; normal text: 1), comprehension question difficulty (easy: -1; difficult: 1). Participants and items were added as random intercepts in all analyses (Baayen et al., 2008). Background sound, reading condition, and question difficulty were added as random slopes for participants and items in all analyses (Barr et al., 2013). However, due to convergence failure, question difficulty was removed as a random slope for items in the inter-sentence regression probability model. Additionally, question difficulty was removed as a random slope for both participants and items in the landing position model.

Results

The fixation data were pre-processed in the same way as in Experiment 2. Overall, 7.16% of all observations were removed (4.47% due to blinks, 2.4% due to fixations smaller than 80 ms, and 0.29% due to outliers). There were no significant differences in the number of outliers excluded per condition (all ps ≥ 0.11).

Comprehension accuracy. The descriptive statistics for comprehension accuracy are presented in Figure 5. There was a main effect of question difficulty (b_1= -0.06, SE= 0.01, t= -5.62, p< 0.001; b_2= -0.06, SE= 0.01, t= -4.97, p< 0.001; d= -1.67), indicating that comprehension was significantly lower on the difficult compared to the easy questions. Additionally, there was a main effect of background sound (b_1= 0.02, SE= 0.01, t= 2.24, p= 0.03; b_2= 0.02, SE= 0.01, t=

2.08, p= 0.04; d= 0.33), which shows that accuracy was significantly lower in the English speech compared to the Silence condition. Furthermore, the main effect of reading condition was also significant (b_1= 0.03, SE= 0.01, t= 3.56, p= 0.001; b_2= 0.03, SE= 0.01, t= 3.15, p= 0.002; d= 0.49), indicating that comprehension was lower in the trailing mask compared to the normal reading condition.

In line with the distraction re-reading hypothesis (H1), there was a significant interaction between background sound and reading condition (b_1= -0.03, SE= 0.01, t= -3.67, p< 0.001; b_2= -0.03, SE= 0.01, t= -3.15, p= 0.002). This was due to accuracy being lower in English speech compared to Silence, but only in the trailing mask (d= -0.65) and not in the normal reading condition (d= 0.12). However, the three-way interaction with question difficulty was not significant (ps≥ 0.87), which shows that the magnitude of the disruption did not differ as a function of the difficulty of questions. This is contrary to hypothesis H2.

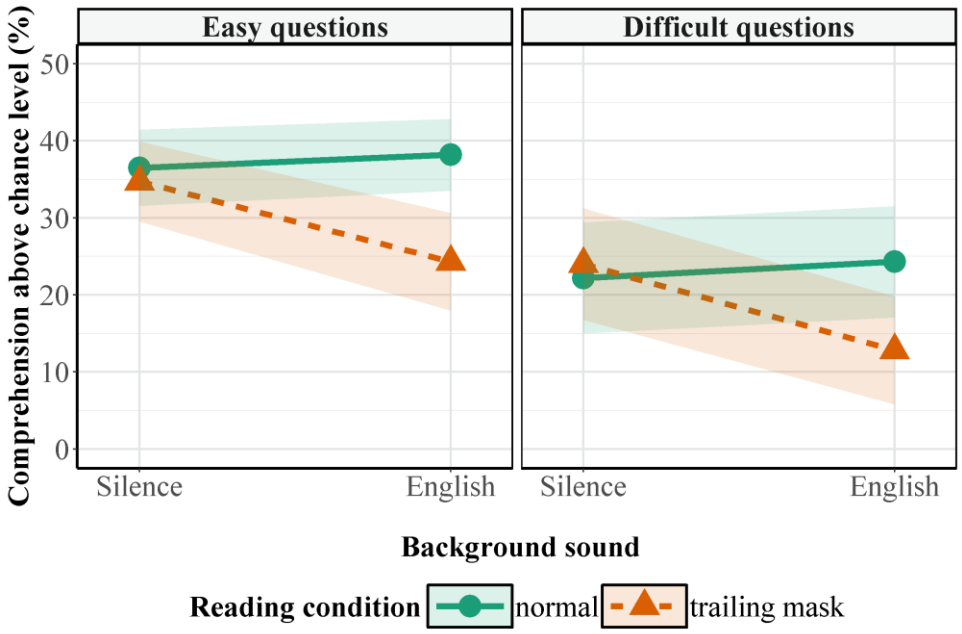

Figure 5. Mean comprehension accuracy above chance level in Experiment 3. Shading indicates the standard error.

Bayes factor regression analysis (Morey et al., 2015; Rouder & Morey, 2012) supported the alternative hypothesis of an interaction between background sound and reading condition (subjects: BF= 11.43; items: BF= 9.40). There was strong evidence in support of the alternative hypothesis that comprehension accuracy is disrupted by English speech in the trailing mask condition (subjects: BF= 27.81; items: BF= 13.70). Conversely, the null hypothesis of no difference in comprehension accuracy between English speech and Silence was supported for the normal reading condition (subjects: BF= 0.13; items: BF= 0.18). Consistent with the LMM analysis, the null hypothesis of no interaction between background sound, reading condition and question difficulty was supported (subjects: BF= 0.11; items: BF= 0.14). Sensitivity analyses using a range of realistic priors indicated that the results were not influenced by the chosen prior distribution (r= $\sqrt{2}/2$; see the Supplementary Materials). In summary, both the Bayes factor and LMM analyses support the distraction re-reading hypothesis, which predicted that comprehension accuracy would be disrupted by English speech only when participants cannot selectively re-read the text. Furthermore, the disruption in comprehension did not differ between the easy and difficult questions.

Global reading measures. The descriptive statistics are presented in Table 5 for saccade length and saccade landing position, and in Table 7 for all other global reading measures. The results from (G)LMMs are presented in Table 8, except for saccade landing position, which is reported in the text. English speech resulted in significantly longer paragraph reading time (*d*= 0.24), greater intra-sentence regression probability (*d*= 0.21), and more second-pass fixations (*d*= 0.04) compared to Silence. Additionally, the trailing mask condition resulted in significantly shorter paragraph reading time (*d*= 0.50), smaller intra-sentence (*d*= 0.19) and inter-sentence (*d*= 0.16) regression probability, fewer first-pass (*d*= 0.11) and second-pass (*d*= 0.18) fixations

compared to the normal reading condition. Furthermore, the trailing mask condition caused saccades to land further away from the beginning of the word than the normal reading condition (*b*= -0.03, *SE*= 0.01, *t*= -3.13, *p*= .003, *d*= -0.03). Likewise, saccades also landed further away from the beginning of the word when participants were answering difficult compared to easy questions (*b*= 0.04, *SE*= 0.01, *t*= 4.83, *p*<0.001, *d*= 0.03).

Table 7

Mean Descriptive Statistics of Global Reading Measures in Experiment 3 (SDs in Parenthesis)

Sound condition	Reading condition	Paragraph reading time (in s)	Fixation duration (in ms)	Intra-sentence regression probability	Inter-sentence regression probability	Number of fixations (per word)		
						1st-pass	2nd-pass	Total
		Easy questions						
Silence	normal	28 (8.3)	219 (94)	.25 (.43)	.08 (.27)	.80 (.8)	.27 (.66)	1.07 (1.02)
Silence	mask	25.3 (6.5)	235 (114)	.20 (.40)	.05 (.21)	.73 (.75)	.18 (.59)	.91 (.94)
English	normal	29.8 (9.9)	224 (101)	.30 (.46)	.08 (.28)	.80 (.82)	.34 (.86)	1.14 (1.19)
English	mask	25.5 (6.9)	239 (117)	.19 (.39)	.05 (.22)	.73 (.83)	.17 (.76)	.90 (1.17)
		Difficult questions						
Silence	normal	28.4 (7.7)	222 (98)	.26 (.44)	.10 (.30)	.81 (.82)	.29 (.81)	1.10 (1.14)
Silence	mask	25.1 (6)	239 (118)	.20 (.40)	.05 (.22)	.72 (.78)	.18 (.69)	.90 (1.06)
English	normal	31.5 (9.2)	226 (103)	.30 (.46)	.09 (.29)	.82 (.83)	.35 (.78)	1.17 (1.16)
English	mask	25.8 (8.1)	239 (118)	.20 (.40)	.04 (.21)	.71 (.76)	.18 (.61)	.89 (.96)

Table 8

Results from (G)LMMs for Global Reading Measures in Experiment 3

Effect	Dependent measure											
	Paragraph reading time				Saccade length				Intra-sentence regression probability			
	b	SE	t	p	b	SE	t	p	b	SE	z	p
Intercept	10.2	.03	309.7	**<.001**	9.09	.19	47.92	**<.001**	-1.37	.08	-15.9	**<.001**
Sound	.02	.006	3.40	**.002**	-.03	.04	-.71	.47	.05	.02	2.53	**.01**
Diff	.01	.008	1.39	0.17	<.01	.05	.03	.97	.03	.02	1.52	.12
RC	.07	.01	6.10	**<.001**	.05	.07	.78	.43	.23	.03	8.47	**<.001**
Sound: Diff	.006	.004	1.44	0.15	<.01	.02	.28	.77	.01	.008	1.52	.12
Sound: RC	.02	.004	4.16	**<.001**	-.05	.02	-2.42	**.01**	.06	.008	7.5	**<.001**
Diff: RC	.01	.004	2.47	**.01**	-.01	.02	-.47	.63	.001	.008	.14	.88
Sound: Diff: RC	.004	.004	.92	.36	-.04	.02	-1.79	.07	-.01	.008	-1.57	0.11
	Inter-sentence regression probability				Number of 1st-pass fixations				Number of 2nd-pass fixations			
	b	SE	z	p	b	SE	z	p	b	SE	z	p
Intercept	-2.88	.07	-38.5	**<.001**	-.28	.02	-12.0	**<.001**	-1.65	.08	-19.03	**<.001**
Sound	-.04	.04	-1.14	.25	-.001	.005	-.15	.88	.05	.02	2.55	**.01**
Diff	.05	.04	1.29	.19	.001	.006	.25	.79	.03	.02	1.58	0.11
RC	.22	.04	4.82	**<.001**	.05	.009	6.03	**<.001**	.29	.03	9.75	**<.001**
Sound: Diff	-.05	.01	-3.45	**.001**	<.001	.004	-.09	.92	.005	.007	.67	.49
Sound: RC	-.01	.01	-.67	.49	.002	.004	.44	.65	.06	.007	8.72	**<.001**
Diff: RC	.04	.01	2.87	**.004**	.01	.004	2.81	**.005**	.01	.007	1.53	.12
Sound: Diff: RC	-.01	.01	-.91	.35	.005	.004	1.30	.19	-.007	.007	-1.12	.26

Note: Sound: background sound. Diff: question difficulty. RC: reading condition. Statistically significant *p*-values are formatted in bold.

The interaction between background sound and question difficulty reached significance for inter-sentence regression probability. This was due to participants making fewer inter-

sentence regressions in the English compared to Silence condition, but only when answering difficult comprehension questions (*d*= -0.02). Additionally, background sound interacted significantly with reading condition for paragraph reading time, saccade length, intra-sentence regression probability, and number of second-pass fixations. This was due to participants taking longer to read the paragraphs (*d*= 0.27), making more intra-sentence regressions (*d*= 0.09), more second-pass fixations (*d*= 0.08), and having shorter saccade length (*d*= -0.03) in the English speech compared to the Silence condition, but only when reading was normal and not when the text had a trailing mask. Therefore, these interactions replicate the results from Experiment 2 by showing that English speech disrupts these measures under normal reading conditions.

Furthermore, question difficulty interacted significantly with reading condition for paragraph reading time, inter-sentence regression probability, and number of first-pass fixations. This was due to longer paragraph reading times (*d*= 0.12), greater inter-sentence regression probability (*d*= 0.05), and more first-pass fixations (*d*= 0.02) when participants were answering difficult, as opposed to easy questions, but only in the normal reading condition. This again replicates the question difficulty effects from Experiment 2 by showing that answering difficult comprehension questions leads to a change in reading behaviour that is characterized by more fixations and more regressions to previous sentences.

Word-level reading measures. The descriptive statistics for word-level reading measures are displayed in Figure 6 and the LMM results are shown in Table 9. Consistent with Experiment 2, English speech resulted in significantly longer fixation durations for all three measures compared to Silence (FFD: *d*= 0.04; GD: *d*= 0.04; TVT: *d*= 0.06). Additionally, the trailing mask resulted in significantly longer FFD (*d*= -0.05) and GD (*d*= -0.08) compared to the normal reading condition. This indicates that reading the text with a trailing mask prolonged the

duration of first-pass fixations on words. Conversely, the trailing mask condition resulted in significantly *shorter* TVT (d= 0.08) compared to the normal reading condition. The opposite effect was due to participants making fewer second-pass fixations in the trailing mask condition (which count towards TVT), presumably because the masked text did not provide any useful information and participants developed the strategy of avoiding it.

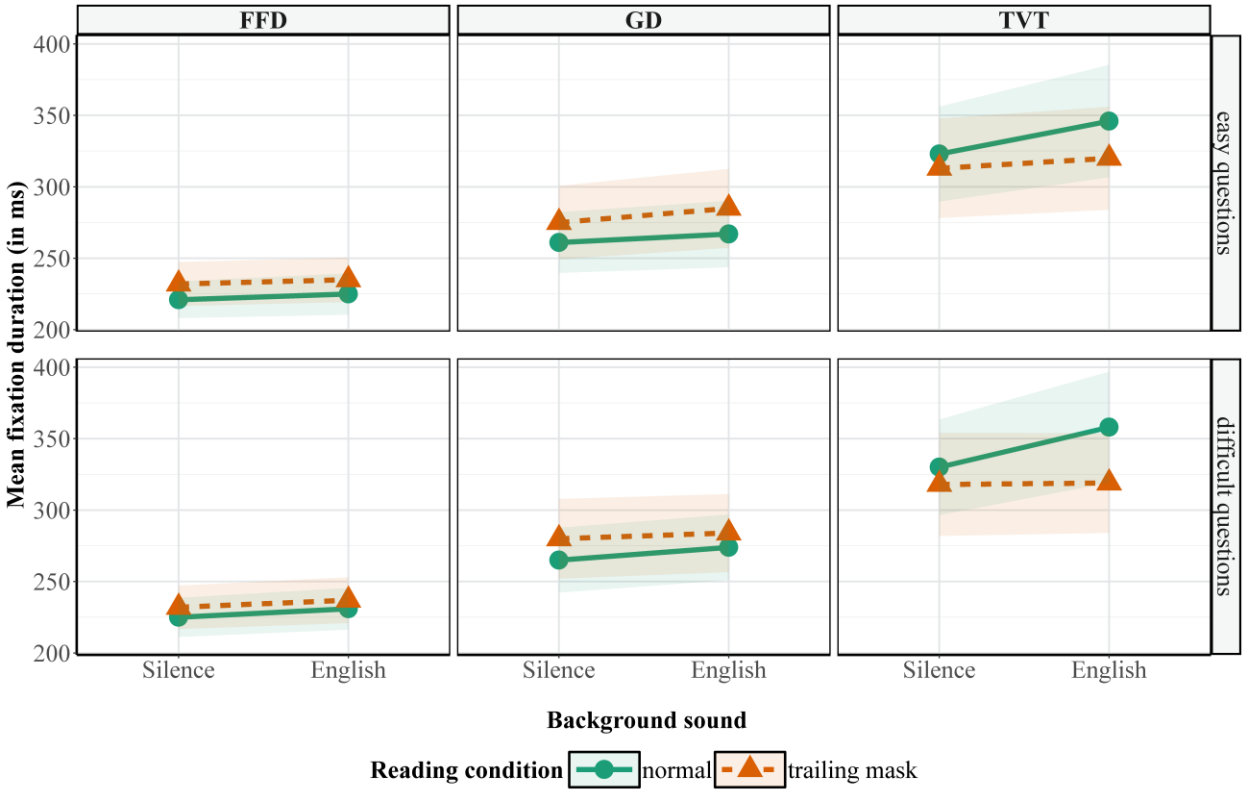

Figure 6. Mean descriptive statistics for local word-level reading measures in Experiment 3. Shading indicates the standard error.

Background sound interacted significantly with reading condition for TVT, but not for FFD or GD. This was due to TVT being longer in the English speech condition compared to the silence condition (d= 0.10), but only when reading was normal. This replicates Experiments 1-2 where TVT was also disrupted by English speech under normal reading conditions. The lack of

interaction between background sound and reading condition for FFD and GD is not surprising because these measures include fixations that were made during the first-pass reading of words. Because viewing conditions during first-pass reading did not differ between the trailing mask and the normal reading condition, background sound manifested itself as a main effect rather than as an interaction.

Table 9

Results from LMMs for Local Word-Level Reading Measures in Experiment 3

Effect	FFD				GD				TVT			
	b	SE	t	p	b	SE	t	p	b	SE	t	p
Intercept	5.35	0.01	369.7	**<.001**	5.47	.02	324.2	**<.001**	5.61	.02	270.7	**<.001**
Sound	.006	.003	2.26	**.03**	.009	.003	2.68	**.01**	.02	.004	3.62	**.001**
Diff	.005	.003	1.97	.054	.006	.003	1.76	.08	.009	.005	1.97	.055
RC	- .01	.003	-3.85	**<.001**	- .01	.004	-3.37	**.002**	.03	.007	4.36	**<.001**
Sound: Diff	.001	.002	.67	.49	<-.001	.002	-.006	.99	.001	.002	.34	.73
Sound: RC	.001	.002	.75	.45	.001	.002	.49	.61	.01	.002	5.22	**<.001**
Diff: RC	.003	.002	2.26	**.02**	.003	.002	1.39	.16	.005	.002	2.55	**.01**
Sound: Diff: RC	.001	.002	.47	.63	.004	.002	1.97	**.049**	.003	.002	1.55	.12

Note: Sound: background sound. Diff: question difficulty. RC: reading condition. Statistically significant *p*-values are formatted in bold.

Additionally, there was a significant two-way interaction between question difficulty and reading condition for both FFD and TVT. This also replicates the results from Experiment 2 by showing that FFD (d= 0.05) and TVT (d= 0.04) were longer when participants were answering difficult compared to easy questions, but only in the normal reading condition (which was equivalent to the reading mode in Experiment 2). Finally, there was a significant three-way

interaction between background sound, question difficulty and reading condition for GD. This was due to GD being longer in English speech compared to Silence for all conditions, except in the trailing mask condition when the comprehension questions were difficult.

Discussion

Experiment 3 tested the distraction re-reading hypothesis, which predicted that intelligible speech will have a negative effect on the immediate comprehension of short paragraphs only when participants cannot go back to selectively re-read the text. The results supported this hypothesis because comprehension accuracy was significantly disrupted when re-reading of previous words was prevented in the trailing mask condition, but no such disruption occurred in the normal reading condition. At the same time, English speech resulted in a significant disruption of second-pass measures during normal reading, thus replicating the results from Experiments 1-2. Therefore, the present results suggest that the increase in re-reading behaviour when listening to intelligible speech is related to maintaining an accurate immediate comprehension of the paragraphs. As there was no significant interaction with question difficulty, it appears that the disruption in comprehension accuracy occurs regardless of whether participants are answering easy or difficult questions. This is consistent with the results from Experiment 2. Finally, Experiment 3 also replicated the question difficulty effect on eye-movement measures from Experiment 2, which showed that participants made more fixations, more regressions to previous sentences and had longer TVT when answering difficult compared to easy questions.

While the main findings from Experiment 2 were replicated, there may be a few apparent inconsistencies regarding the measures in which the effects were found. Before considering

them, it is important to note that a direct replication of the intelligible speech and question difficulty effects from Experiment 2 can be shown in this experiment by a significant two-way interaction between each of the two factors and reading condition. This is because only the conditions with normal text presentation (and not the trailing mask one) corresponded to the reading conditions from Experiment 2. On the other hand, a main effect of background sound or question difficulty shows that the respective effect was observed in both the normal and the trailing mask condition. This is still consistent with the findings from Experiment 2, but it would suggest that the effect is not limited only to normal reading.

The effect of intelligible speech in Experiment 3 was observed in the same dependent measures as Experiment 2, apart from saccade length, which did not differ between the English and silence condition in Experiment 2. Nevertheless, the difference in Experiment 2 was still in the expected direction and English speech also differed significantly from both Mandarin speech and Noise in that experiment. Additionally, while there was no interaction between background sound and reading condition for FFD and GD in the present experiment, the main effect of background sound was significant for both variables. This is still consistent with the results from Experiment 2 because it suggests that first-pass fixation durations generally increased in the English speech condition regardless of whether the text was normal or had a trailing mask. This is not surprising because the trailing mask manipulation had no effect on the first-pass fixations of words. Therefore, first-pass fixation durations generally increased in the presence of intelligible speech regardless of the reading condition. Finally, the only inconsistent finding with respect to question difficulty was that this effect was found in number of first-pass fixations instead of number of second-pass fixations. However, while not significant, the mean difference in the number of second-pass fixations was still in the expected direction.

In summary, Experiment 3 found evidence that regressions and re-reading fixations allow readers to maintain the immediate comprehension of short paragraphs when listening to intelligible speech in the background. This suggests that readers use regressive eye-movements to resolve temporary comprehension difficulties that arise from semantic interference due to the irrelevant speech sound (Marsh et al., 2008, 2009). While the present results demonstrate the link between regressive saccades and immediate text comprehension when reading under distracting conditions, they do not exclude the possibility that comprehension may still be negatively affected even if selective re-reading of the text is possible. Clearly, there is nothing that prevents readers from making regressions to previous words and sentences in everyday life situations. Additionally, re-reading has also been possible in previous studies that have shown disruption in comprehension accuracy by intelligible speech (e.g., Baker & Madell, 1965; Martin et al., 1988; Sörqvist, Halin, et al., 2010). This is not necessarily inconsistent with the present results because they only show that readers can maintain the immediate comprehension of short paragraphs that are fairly easy to understand for skilled readers. For example, it is possible that the strategy of selectively re-reading the previous text may not be enough to compensate for semantic disruption when readers are processing longer and more complex texts (e.g., university-level textbooks). This is a possibility that needs to be explored by future research.

General Discussion

In the first two experiments, there was clear evidence that intelligible speech disrupts eye-movements during reading. This result is consistent with previous evidence showing that intelligible speech (both coherent and scrambled) results in attentional distraction that is detectible at the level of eye fixations (Cauchard et al., 2012; Hyönä & Ekholm, 2016; Yan et al., 2017). In Experiment 1, the lexical processing of words was not influenced by intelligible

speech, but participants had greater difficulty integrating words into the sentence context. Experiment 2 extended these results by showing that the disruptive effect of intelligible speech appears to be limited mostly to the currently-read sentence. At the same time, participants' immediate comprehension was not affected, even when the comprehension questions were more difficult to answer. Finally, Experiment 3 showed that comprehension accuracy was disrupted by intelligible speech only when participants could not re-read previous words and sentences.

The present research showed that disruption effects by intelligible speech were consistently observed in measures of second-pass reading (total viewing time, intra-sentence regression probability, and number of second-pass fixations) in all three experiments. However, Experiments 2 and 3 also revealed effects in first-pass reading measures (first fixation duration and gaze duration). This disruption of first-pass measures in paragraph reading may be due to the greater text context and the need for discourse processing that is not required when reading single unconnected sentences. Therefore, the present results raise the possibility that intelligible speech may become more distracting when the text context increases. This could be because readers find it more difficult to maintain sustained attention on their task for longer periods of time, which would be necessary when reading connected text.

The present research found strong support for semantic disruption by background speech in eye-movements during reading. This is consistent with the semantic disruption account by Martin et al. (1988) and the interference-by-process account by Marsh et al. (2008, 2009). Additionally, the present results are also in line with Hyönä and Ekholm's (2016) experiments, which also pointed towards distraction due to semantic interference from processing the meaning of the speech sound. In contrast, the present research found no support for the strong form of the phonological disruption account (Salamé & Baddeley, 1982, 1987) that any speech sound should

be equally distracting because it gains access to the phonological loop. Nevertheless, two effects suggested a possible contribution of phonology. In Experiment 1, unintelligible speech (Mandarin) resulted in more second-pass fixations compared to noise, and in Experiment 2 unintelligible speech resulted in more regressions within the currently-read sentence compared to noise.

It is worth considering these two findings in more detail to examine what role phonology may play in distraction by intelligible speech. First, the effect from Experiment 1 was partially driven by the fact that participants made fewer second-pass fixations in noise compared to silence. This was confirmed by the lack of significant difference between Mandarin and silence (p= 0.72), which suggests that the effect reached significance because the means in the Mandarin and Noise condition were going in the opposite direction in relation to the silence baseline. Additionally, this effect was not replicated in Experiment 2, which further raises questions about its generalizability across different types of reading materials.

Furthermore, even though there was a significant difference in intra-sentence regression probability between Mandarin and Noise in Experiment 2, the lack of increase in number of second-pass fixations suggests that participants did not actually spend more time re-reading words in the sentence (this was also confirmed by a lack of difference in sentence re-reading time between Mandarin and noise in Experiment 2). In other words, participants in Experiment 2 were more likely to regress back within the current sentence in Mandarin speech compared to noise, but they did not actually spend more time processing words again. To some extent, this may argue against an explanation of disrupted word processing or sentence integration by Mandarin speech because participants would have likely made more re-reading fixations to recover from the disruption (as was the case when they listened to English speech). However, the

increase in regression probability without an associated increase in re-reading fixations could potentially suggest that the unintelligible Mandarin speech may have elicited some type of attention orienting response (e.g., Sokolov, 2001). This could be either due to its perceptual novelty or to some unexpected prosodic features that were present in the speech. At present, this remains a speculation that needs to be tested by future research.

Because the unintelligible Mandarin speech in the present studies contained distinct tones that are not present in English speech (Duanmu, 2006), it could be argued that the two effects above may be due to differences in pitch. The present research cannot exclude this possibility and further work is required to rule out this alternative explanation. Nevertheless, it should be noted that this explanation is at odds with the common finding that native speakers of atonal languages such as English often have difficulties in distinguishing between Mandarin tones (e.g., Kiriloff, 1969; Morett & Chang, 2015; see also Wang, Spence, Jongman, & Sereno, 1999). In summary, the two significant differences between Mandarin and Noise present only limited support for a partial contribution of phonology in distraction by intelligible speech in eye-movements. This conclusion is largely in agreement with Hyönä and Ekholm's (2016) Experiment 1, where no evidence for phonological disruption was found.

Even though the present research found that intelligible speech consistently disrupts second-pass reading, the magnitude of the effects was small. This suggests that intelligible speech results only in a mild reduction of reading efficiency. This is consistent with a recent meta-analysis of auditory distraction effects in reading comprehension, where a very similar range of effect sizes was observed (Vasilev et al., 2018). We speculate that the magnitude of effects may be larger in certain participant population. For example, children may show larger effects due to their poorer control of attention and their ability to filter out task-irrelevant stimuli

(Doyle, 1973; Gomes, Molholm, Christodoulou, Ritter, & Cowan, 2000; Plude, Enns, & Brodeur, 1994). This is a question that needs to be tested in future studies.

While the present findings are consistent at the basic level with the semantic disruption accounts of Martin et al. (1988) and Marsh et al. (2008, 2009), these theories do not make specific predictions about how intelligible speech affects eye-movements during reading. Therefore, the present experiments provide a more detailed account of how the semantic properties of background speech affect the decisions of when and where to move the eyes next. One of the key findings was that the semantic properties of background speech did not disrupt the low-level lexical identification of individual words in the sentence. This finding points to the fact that intelligible speech affects only the post-lexical stages of language processing. While there was evidence for a general slowing down of language processing that was shown by the longer first-pass reading measures (Experiments 2-3), progressive reading behaviour remained relatively unaffected. This was evidenced by the lack of disruption in oculomotor measures, such as saccade landing position. While there was some evidence for a disruption in saccade length, the magnitude of the effects was very small. This suggests that participants likely did not experience great difficulty in progressing through the text and reading new words. Instead, the semantic properties of the irrelevant speech likely created a temporary difficulty in constructing the semantic meaning of the sentence and forming a coherent text discourse. This in turn may have prompted participants to make more regressions in order to resolve the difficulty before they continue reading new words.

The present results also provide insights into how the disruption by intelligible speech could be simulated in computational models of eye-movement control during reading. For example, a recent version of the E-Z Reader model (Reichle, Warren, & McConnell, 2009) has

attempted to simulate effects of higher-level language processing on eye-movements. Reichle et al. (2009) introduced a new post-lexical integration stage that reflects the processing associated with integrating the currently fixated word into higher-level language representations, such as the syntactic structure of the sentence. In this framework, the present results could be modelled by implementing a post-lexical parameter that checks for interference in integrating the meaning of the last few words in the context of the text that has been read so far. The detection of such interference by intelligible speech would then be associated with greater probability of making a regression to previous words in order to overcome this transient processing difficulty and continue with the progressive reading of the text.

While there was robust disruption by intelligible speech in eye-movement measures, comprehension accuracy in the first two experiments remained unaffected. This suggests that intelligible speech does not degrade the meaning of the text that has been read, at least in the short term and when reading single sentences or short paragraphs. Even though a number of behavioural experiments have reported a disruption in comprehension accuracy (e.g., Baker & Madell, 1965; Halin, 2016; Martin et al., 1988; Sörqvist, Halin, et al., 2010), the pressent research is not necessarily inconsistent with such studies because it only shows that the immediate comprehension of short sentences and paragraphs is not affected by intelligible speech when participants can re-read previous words and sentences. This difference in the results is not likely to be explained by the greater difficulty of comprehension questions in previous studies because the average accuracy was 34.1% above chance level in the studies cited above (range: 21.2- 43.3%). The average accuracy above chance level on the difficult questions in the present research was 31% in Experiment 2 and 23% in the normal reading condition of

Experiment 3. Therefore, the difficult questions were slightly more challenging than the questions used in previous studies.

There are a few possible reasons why a disruption in comprehension may have been observed in previous research. For example, intelligible speech may only disrupt the transfer of text meaning to long-term memory. In fact, many behavioural experiments have had a delay between the reading task and the comprehension assessment, often even with other tasks in between (e.g., Boman, 2004; Knez & Hygge, 2002; Martin et al., 1988). Additionally, the present research used text stimuli that were relatively short and easy to understand. Therefore, it may be the case that intelligible speech disrupts the comprehension of longer and more complex texts that require making inferences between different paragraphs or larger topics of meaning. Finally, the speech stimuli were also relatively simple and they may not have been very engaging to our participants. Therefore, it may be more difficult to maintain comprehension of the text when the intelligible speech is more engaging. This could be because engaging speech makes it harder to selectively attend to the text and filter out the irrelevant speech sound. There is some evidence to suggest that the content of the speech may influence the amount of distraction. For example, hearing only one side of a telephone conversation is more distracting than hearing both sides of the conversation, presumably because the former type of speech is less predictable than the latter (Emberson, Lupyan, Goldstein, & Spivey, 2010; Marsh et al., 2018). In a similar fashion, engaging speech may be more likely to attract attention away from the main task and thus lead to a greater disruption in comprehension. These are all avenues that need to be explored by future research.

Behavioural studies have also shown that intelligible speech can disrupt performance on other tasks, such as free recall, that require the use of semantic processing (Marsh et al., 2008,

2009; Marsh, Perham, Sörqvist, & Jones, 2014; Marsh, Sörqvist, Hodgetts, Beaman, & Jones, 2015; see Marsh & Jones, 2010 for a review). One task that is more similar to reading and also requires the retrieval of concepts from semantic memory is verbal fluency (e.g., retrieving examples of the semantic category "animals"). Consistent with the interference-by-process account, Jones, Marsh, and Hughes (2012) showed that verbal, but not phonemic, fluency is disrupted by intelligible speech. The former task relies on semantic processing, while the latter does not. Interestingly, the present research suggests that, unlike verbal fluency, reading is not disrupted at the stage of retrieving word concepts from semantic memory. Rather, this disruption occurs later when participants need to combine the meaning of individual words to comprehend the sentence and to build a coherent discourse of the text.

The lack of disruption in retrieving word concepts provides support for the interference-by-process account (Marsh et al., 2008, 2009), which stipulates that the nature of the main task determines when intelligible speech is distracting. In the context of verbal fluency, the task is to retrieve word concepts from semantic memory according to a certain rule. In contrast, reading imposes different task demands because retrieving the concepts of individual words is not enough for comprehension- readers also need to combine these concepts to form the meaning of the sentence. The interference-by-process account can also explain why the amount of disruption in eye-movement measures was greater in a paragraph-reading task compared to a sentence-reading one. When reading paragraphs, there is a greater emphasis on semantic processing and comprehension because the text is more complex. Additionally, participants also need to combine the meaning of all sentences in order to form a coherent discourse of the text.

Finally, the present findings also have practical implications for educational and work settings where irrelevant speech is often present. For example, intelligible speech is a common

problem in open-plan offices and other shared work areas that are characterized by poor acoustical privacy (Haapakangas, Hongisto, Eerola, & Kuusisto, 2017; Haapakangas, Hongisto, Hyönä, Kokko, & Keränen, 2014; Schlittmeier & Liebl, 2015). As a result, irrelevant speech from nearby workers or phone conversations can have a negative impact on reading and other office tasks that rely on processing the meaning of written text (e.g., proofreading or copying written information). The present results suggest that intelligible speech will result in slower reading due to the need for greater re-reading of previous words. This has implications for job performance as workers will generally need more time to complete reading tasks if intelligible speech is present in the background. Additionally, comprehension deficits may also occur if workers do not have enough time to engage in effective re-reading of previous text in order to compensate for the experienced distraction. More research in applied settings is required to test directly the magnitude of disruption in reading performance among workers in open-plan offices.

In summary, the present findings suggest that intelligible speech does not affect the lexical retrieval of words. Rather, the disruption occurs later when readers need to integrate the meaning of new words into the sentence context. Additionally, the amount of disruption in eye-movement measures depended on the demands of the reading task, with short paragraphs leading to greater disruption compared to single sentences. The present research also showed that intelligible speech can disrupt the ongoing reading process even when comprehension remains unaffected. This highlights the utility of eye-tracking to detect subtle auditory disruption effects that may not be captured by measures of comprehension accuracy. Finally, the increase in re-reading behaviour appears to be important for maintaining the immediate comprehension of the text because comprehension was compromised when participants read the paragraphs in a format that prevented them from selectively re-reading previous words and sentences. This suggests that

regressions play a key role in maintaining comprehension of the text and allow readers to recover from transient attentional distraction.

Acknowledgements

Martin Vasilev was supported by a PhD studentship from Bournemouth University. Simon Liversedge acknowledges support from ESRC Grant ES/R003386/1. The data files, R scripts, and reading stimuli used in the present research (Vasilev, Liversedge, Rowan, Kirkby, & Angele, 2019) are openly available at: https://osf.io/jvsm8/

References

Abrams, S. G., & Zuber, B. L. (1972). Some temporal characteristics of information processing during reading. *Reading Research Quarterly*, *8*(1), 40–51. Retrieved from http://www.jstor.org/stable/746979

Andriessen, J. J., & de Voogd, A. H. (1973). Analysis of eye movement patterns in silent reading. *IPO Annual Progress Report*, *8*, 29–34. Retrieved from http://alexandria.tue.nl/tijdschrift/IPO 8.pdf

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baddeley, A. D. (2000). The phonological loop and the irrelevant speech effect: some comments on Neath (2000). *Psychonomic Bulletin & Review*, *7*(3), 544–549. https://doi.org/10.3758/BF03214369

Baddeley, A. D. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189–208. https://doi.org/10.1016/S0021-9924(03)00019-4

Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of Learning and Motivation* (Vol. 8, pp. 47–89). https://doi.org/10.1016/S0079-7421(08)60452-1

Baddeley, A. D., & Hitch, G. J. (1994). Developments in the concept of working memory. *Neuropsychology*, *8*(4), 485–493. https://doi.org/10.1037/0894-4105.8.4.485

Baddeley, A. D., & Salamé, P. (1986). The unattended speech effect: Perception or memory?

Journal of Experimental Psychology: Learning, Memory, and Cognition, *12*(4), 525–529. https://doi.org/10.1037/0278-7393.12.4.525

Baker, R. W., & Madell, T. O. (1965). A continued investigation of susceptibility to distraction in academically underachieving and achieving male college students. *Journal of Educational Psychology*, *56*(5), 254–258. https://doi.org/10.1037/h0022467

Banbury, S., & Berry, D. C. (1997). Habituation and dishabituation to speech and office noise. *Journal of Experimental Psychology: Applied*, *3*(3), 181–195. https://doi.org/10.1037//1076-898X.3.3.181

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. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D. M., Machler, M., Bolker, B. M., & Walker, S. C. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bench, J., Kowal, Å., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British Journal of Audiology*, *13*(3), 108–112. https://doi.org/10.3109/03005367909078884

Boman, E. (2004). The effects of noise and gender on children's episodic and semantic memory. *Scandinavian Journal of Psychology*, *45*(5), 407–416. https://doi.org/10.1111/j.1467-9450.2004.00422.x

Booth, R. W., & Weger, U. W. (2013). The function of regressions in reading: Backward eye movements allow rereading. *Memory & Cognition*, *41*(1), 82–97.

https://doi.org/10.3758/s13421-012-0244-y

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Cauchard, F., Cane, J. E., & Weger, U. W. (2012). Influence of background speech and music in interrupted reading: An eye-tracking study. *Applied Cognitive Psychology*, *26*(3), 381–390. https://doi.org/10.1002/acp.1837

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Colle, H. A., & Welsh, A. (1976). Acoustic masking in primary memory. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 17–31. https://doi.org/10.1016/S0022-5371(76)90003-7

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256. https://doi.org/10.1037/0033-295X.108.1.204

Cornelissen, F. W., Peters, E. M., & Palmer, J. (2002). The Eyelink Toolbox: Eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 613–617. https://doi.org/10.3758/BF03195489

Davis, C. J. (2005). N-watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, *37*(1), 65–70. https://doi.org/10.3758/BF03206399

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in*

Psychology, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Doyle, A.-B. (1973). Listening to distraction: A developmental study of selective attention. *Journal of Experimental Child Psychology*, *15*(1), 100–115. https://doi.org/http://dx.doi.org/10.1016/0022-0965(73)90134-3

Duanmu, S. (2006). Chinese (Mandarin): Phonology. In *Encyclopedia of Language & Linguistics* (pp. 351–355). Elsevier Science.

Ellermeier, W., & Zimmer, K. (1997). Individual differences in susceptibility to the "irrelevant speech effect." *The Journal of the Acoustical Society of America*, *102*(4), 2191–2199. https://doi.org/10.1121/1.419596

Emberson, L. L., Lupyan, G., Goldstein, M. H., & Spivey, M. J. (2010). Overheard cell-phone conversations: When less speech is more distracting. *Psychological Science*, *21*(10), 1383–1388. https://doi.org/10.1177/0956797610382126

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

Furnham, A., Gunter, B., & Peterson, E. (1994). Television distraction and the performance of introverts and extroverts. *Applied Cognitive Psychology*, *8*, 705–711. https://doi.org/10.1002/acp.2350080708

Gawron, V. J. (1984). Noise: Effect and aftereffect. *Ergonomics*, *27*(1), 5–18.

https://doi.org/10.1080/00140138408963460

Gernsbacher, M. A., & Foertsch, J. A. (2000). Three models of discourse comprehension. In S. Garrod & M. J. Pickering (Eds.), *Language processing* (pp. 283–299). East Sussex, UK: Psychology Press.

Gomes, H., Molholm, S., Christodoulou, C., Ritter, W., & Cowan, N. (2000). The development of auditory attention in children. *Frontiers in Bioscience*, *5*, 108–120. https://doi.org/10.1093/toxsci/kfs057

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Haapakangas, A., Hongisto, V., Eerola, M., & Kuusisto, T. (2017). Distraction distance and perceived disturbance by noise—An analysis of 21 open-plan offices. *The Journal of the Acoustical Society of America*, *141*(1), 127–136. https://doi.org/10.1121/1.4973690

Haapakangas, A., Hongisto, V., Hyönä, J., Kokko, J., & Keränen, J. (2014). Effects of unattended speech on performance and subjective distraction: The role of acoustic design in open-plan offices. *Applied Acoustics*, *86*, 1–16. https://doi.org/10.1016/j.apacoust.2014.04.018

Haka, M., Haapakangas, A., Keränen, J., Hakala, J., Keskinen, E., & Hongisto, V. (2009). Performance effects and subjective disturbance of speech in acoustically different office types- A laboratory experiment. *Indoor Air*, *19*(6), 454–467. https://doi.org/10.1111/j.1600-0668.2009.00608.x

Halin, N. (2016). Distracted while reading? Changing to a hard-to-read font shields against the effects of environmental noise and speech on text memory. *Frontiers in Psychology*, *7*, 1–6.

https://doi.org/10.3389/fpsyg.2016.01196

Hofmeister, J., Heller, D., & Radach, R. (1999). The return sweep in reading. In *Current Oculomotor Research* (pp. 349–357). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4757-3054-8_49

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(6), 65–70. https://doi.org/10.2307/4615733

Hughes, R. W. (2014). Auditory distraction: A duplex-mechanism account. *PsyCh Journal*, *3*(1), 30–41. https://doi.org/10.1002/pchj.44

Hyönä, J., & Ekholm, M. (2016). Background speech effects on sentence processing during reading: An eye movement study. *PloS One*, *11*(3), e0152133. https://doi.org/10.1371/journal.pone.0152133

Hyönä, J., Lorch, R. F. J., & Rinck, M. (2003). Eye movement measures to study global text processing. In J. Hyönä, R. Radach, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (Elsevier, pp. 313–334). Amsterdam. https://doi.org/10.1016/B978-044451020-4/50018-9

Inhoff, A. W., & Rayner, K. (1986). Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, *40*(6), 431–439. https://doi.org/10.3758/BF03208203

Jahncke, H., Hygge, S., Halin, N., Green, A. M., & Dimberg, K. (2011). Open-plan office noise: Cognitive performance and restoration. *Journal of Environmental Psychology*, *31*(4), 373–382. https://doi.org/10.1016/j.jenvp.2011.07.002

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford, UK: Oxford University Press.

Johansson, C. R. (1983). Effects of low intensity, continuous and intermittent noise on mental performance and writing pressure of children with different intelligence and personality characteristics. *Ergonomics*, *26*(3), 275–288. https://doi.org/10.1080/00140138308963341

Johansson, R., Holmqvist, K., Mossberg, F., & Lindgren, M. (2012). Eye movements and reading comprehension while listening to preferred and non-preferred study music. *Psychology of Music*, *40*, 339–356. https://doi.org/10.1177/0305735610387777

Jones, D. M., & Macken, W. J. (1995). Phonological similarity in the irrelevant speech effect: Within- or between-stream similarity? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 103–115. https://doi.org/10.1037/0278-7393.21.1.103

Jones, D. M., Marsh, J. E., & Hughes, R. W. (2012). Retrieval from memory: Vulnerable or inviolable? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(4), 905–922. https://doi.org/10.1037/a0026781

Jones, D. M., Miles, C., & Page, J. (1990). Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory? *Applied Cognitive Psychology*, *4*(2), 89–108. https://doi.org/10.1002/acp.2350040203

Kehler, A. (2004). Discourse coherence. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 241–265). Oxford, UK: Blackwell Publishing.

Kiriloff, C. (1969). On the auditory perception of tones in Mandarin. *Phonetica*, *20*(2–4), 63–67. https://doi.org/10.1159/000259274

Klatte, M., Lee, N., & Hellbruck, J. (2002). Effects of irrelevant speech and articulatory

suppression on serial recall of head and read stimuli. *Psychologische Beiträge*, *44*, 166–186.

Knez, I., & Hygge, S. (2002). Irrelevant speech and indoor lighting: Effects on cognitive performance and self-reported affect. *Applied Cognitive Psychology*, *16*(6), 709–718. https://doi.org/10.1002/acp.829

Kuo, Y.-C. (2006). *Cochlear implants in a tone language: Mandarin Chinese (upublished doctoral thesis)*. University College London, United Kingdom.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Landström, U., Söderberg, L., Kjellberg, A., & Nordström, B. (2002). Annoyance and performance effects of nearby speech. *Acta Acustica United with Acustica*, *88*(4), 549–553.

Larsen, J. D., & Baddeley, A. D. (2003). Disruption of verbal STM by irrelevant speech, articulatory suppression, and manual tapping: Do they have a common source? *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *56*(8), 1249–1268. https://doi.org/10.1080/02724980244000765

Larsen, J. D., Baddeley, A. D., & Andrade, J. (2000). Phonological similarity and the irrelevant speech effect: Implication for models of short-term verbal memory. *Memory*, *8*(3), 145–157. https://doi.org/10.1080/096582100387579

LeCompte, D. C., & Shaibe, D. M. (1997). On the irrelevance of phonological similarity to the irrelevant speech effect. *The Quarterly Journal of Experimental Psychology*, *50A*(1), 100–119.

Liversedge, S. P., Paterson, K. B., & Pickering, M. J. (1998). Eye movements and measures of reading time. In G. Underwood (Ed.), *Eye guidance in reading and scene perception* (pp. 55–75). Oxford: Elsevier. https://doi.org/10.1016/B978-008043361-5/50004-3

Ljung, R., Sörqvist, P., & Hygge, S. (2009). Effects of road traffic noise and irrelevant speech on children's reading and mathematical performance. *Noise & Health*, *11*(45), 194–198. https://doi.org/10.4103/1463-1741.56212

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology*, *24*, 29–43. https://doi.org/10.3109/03005369009077840

Marsh, J. E., Hughes, R. W., & Jones, D. M. (2008). Auditory distraction in semantic memory: A process-based approach. *Journal of Memory and Language*, *58*(3), 682–700. https://doi.org/10.1016/j.jml.2007.05.002

Marsh, J. E., Hughes, R. W., & Jones, D. M. (2009). Interference by process, not content, determines semantic auditory distraction. *Cognition*, *110*(1), 23–38. https://doi.org/10.1016/j.cognition.2008.08.003

Marsh, J. E., & Jones, D. M. (2010). Cross-modal distraction by background speech: What role for meaning? *Noise & Health*, *12*(49), 210–216. https://doi.org/10.4103/1463-1741.70499

Marsh, J. E., Ljung, R., Jahncke, H., MacCutcheon, D., Pausch, F., Ball, L. J., & Vachon, F. (2018). Why are background telephone conversations distracting? *Journal of Experimental Psychology: Applied*, *24*(2), 222–235. https://doi.org/10.1037/xap0000170

Marsh, J. E., Perham, N., Sörqvist, P., & Jones, D. M. (2014). Boundaries of semantic

distraction: Dominance and lexicality act at retrieval. *Memory & Cognition*, *42*, 1285–1301. https://doi.org/10.3758/s13421-014-0438-6

Marsh, J. E., Sörqvist, P., Hodgetts, H. M., Beaman, C. P., & Jones, D. M. (2015). Distraction control processes in free recall: Benefits and costs to performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 118–133. https://doi.org/10.1037/a0037779

Martin, R. C., Wogalter, M. S., & Forlano, J. G. (1988). Reading comprehension in the presence of unattended speech and music. *Journal of Memory and Language*, *27*(4), 382–398. https://doi.org/10.1016/0749-596X(88)90063-0

MathWorks. (2014). Matlab R2014a [Computer software]. Natick, Massachusetts, USA.

Meseguer, E., Carreiras, M., & Clifton, C. (2002). Overt reanalysis strategies and eye movements during the reading of mild garden path sentences. *Memory & Cognition*, *30*(4), 551–561. https://doi.org/10.3758/BF03194956

Morett, L. M., & Chang, L. Y. (2015). Emphasising sound and meaning: pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, *30*(3), 347–353. https://doi.org/10.1080/23273798.2014.923105

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). BayesFactor: Computation of Bayes Factors for Common Designs [R package version 0.9.12-2]. Retrieved from https://cran.r-project.org/package=BayesFactor

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178. https://doi.org/10.1037/h0027366

O'Brien, E. J., & Cook, A. E. (2015). Models of discourse comprehension. In A. Pollatsek & R. Treiman (Eds.), *The Oxford handbook of reading* (pp. 217–231). New York, USA: Oxford University Press.

Olkoniemi, H., Johander, E., & Kaakinen, J. K. (2018). The role of look-backs in the processing of written sarcasm. *Memory and Cognition*. https://doi.org/10.3758/s13421-018-0852-2

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897X00366

Plude, D. J., Enns, J. T., & Brodeur, D. (1994). The development of selective attention: A life-span overview. *Acta Psychologica*, *86*(2–3), 227–272. https://doi.org/10.1016/0001-6918(94)90004-3

Rayner, K. (1975). The perceptual span and peripheral cues in reading. *Cognitive Psychology*, *81*(7), 65–81. https://doi.org/10.1016/0010-0285(75)90005-5

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506. https://doi.org/10.1080/17470210902816461

Rayner, K., Chace, K. H., Slattery, T. J., & Ashby, J. (2006). Eye movements as reflections of comprehension processes in reading. *Scientific Studies of Reading*, *10*(3), 241–255. https://doi.org/10.1207/s1532799xssr1003

Rayner, K., Pollatsek, A., Ashby, J., & Clifton, C. J. (2012). *Psychology of reading* (2nd ed.). New York, USA: Psychology Press.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1–21. https://doi.org/10.3758/PBR.16.1.1

Risse, S., & Kliegl, R. (2014). Dissociating preview validity and preview difficulty in parafoveal processing of word n +1 during reading. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 653–668. https://doi.org/10.1037/a0034997

Rouder, J. N., & Morey, R. D. (2012). Default bayes factors for model selection in regression. *Multivariate Behavioral Research*, *47*(6), 877–903. https://doi.org/10.1080/00273171.2012.734737

Salamé, P., & Baddeley, A. D. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 150–164. https://doi.org/10.1016/S0022-5371(82)90521-7

Salamé, P., & Baddeley, A. D. (1987). Noise, unattended speech and short-term memory. *Ergonomics*, *30*(8), 1185–1194. https://doi.org/10.1080/00140138708966007

Salamé, P., & Baddeley, A. D. (1989). Effects of background music on phonological short-term memory. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *41*(1–A), 107–122. https://doi.org/10.1080/14640748908402355

Schlittmeier, S. J., & Liebl, A. (2015). The effects of intelligible irrelevant background speech in offices- Cognitive disturbance, annoyance, and solutions. *Facilities*, *33*(1/2), 61–75. https://doi.org/10.1108/F-05-2013-0036

Schotter, E. R., Lee, M., Reiderman, M., & Rayner, K. (2015). The effect of contextual constraint on parafoveal processing in reading. *Journal of Memory and Language*, *83*, 118–139. https://doi.org/10.1016/j.jml.2015.04.005

Schotter, E. R., Tran, R., & Rayner, K. (2014). Don't believe what you read (only once): Comprehension is supported by regressions during reading. *Psychological Science*, *25*(6), 1218–1226. https://doi.org/10.1177/0956797614531148

Sokolov, E. N. (2001). Orienting response. In *International Encyclopedia of the Social & Behavioral Sciences* (pp. 10978–10981). Elsevier Science Ltd. https://doi.org/10.1016/B0-08-043076-7/03536-1

Sörqvist, P., Halin, N., & Hygge, S. (2010). Individual differences in susceptibility to the effect of speech on reading comprehension. *Applied Cognitive Psychology*, *24*(1), 67–76. https://doi.org/10.1002/acp.1543

Sörqvist, P., Ljungberg, J. K., & Ljung, R. (2010). A sub-process view of working memory capacity: Evidence from effects of speech on prose memory. Memory, 18(3), 310–326. http://doi.org/10.1080/09658211003601530

Stracuzzi, D. J. (2004). EyeTrack (Version 0.7.10h) [Computer software]. Retrieved from http://blogs.umass.edu/eyelab

Stracuzzi, D. J., & Kinsey, J. D. (2009). EyeDoctor (Version 0.6.5) [Computer Software]. Retrieved from http://blogs.umass.edu/eyelab

Team, R. C. (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org/

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Vasilev, M. R., Liversedge, S. P., Rowan, D., Kirkby, J. A., & Angele, B. (2019, April 27). Reading is disrupted by intelligible background speech: Evidence from eye-tracking (data and materials). https://doi.org/10.17605/OSF.IO/JVSM8

Vasilev, M. R., Kirkby, J. A., & Angele, B. (2018). Auditory distraction during reading: A Bayesian meta-analysis of a continuing controversy. *Perspectives on Psychological Science*, 174569161774739. https://doi.org/10.1177/1745691617747398

Veitch, J. A. (1990). Office noise and illumination effects on reading comprehension. *Journal of Environmental Psychology*, *10*(3), 209–217. https://doi.org/10.1016/S0272-4944(05)80096-9

Venetjoki, N., Kaarlela-Tuomaala, A., Keskinen, E., & Hongisto, V. (2006). The effect of speech and speech intelligibility on task performance. *Ergonomics*, *49*(11), 1068–1091. https://doi.org/10.1080/00140130600679142

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*(6), 3649–3658. https://doi.org/10.1121/1.428217

Westfall, J. (2015). PANGEA: Power analysis for general anova designs. *Unpublished Manuscript*. Retrieved from http://jakewestfall.org/publications/pangea.pdf

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011).

Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Wolf, F., & Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, *31*(2), 249–287. https://doi.org/10.1162/0891201054223977

Wotschack, C., & Kliegl, R. (2013). Reading strategy modulates parafoveal-on-foveal effects in sentence reading. *Quarterly Journal of Experimental Psychology*, *66*(3), 548–562. https://doi.org/10.1080/17470218.2011.625094

Yan, G., Meng, Z., Liu, N., He, L., & Paterson, K. B. (2017). Effects of irrelevant background speech on eye movements during reading. *The Quarterly Journal of Experimental Psychology*, 1–20. https://doi.org/10.1080/17470218.2017.1339718

Ylias, G., & Heaven, P. C. L. (2003). The influence of distraction on reading comprehension: A big five analysis. *Personality and Individual Differences*, *34*(6), 1069–1079. https://doi.org/10.1016/S0191-8869(02)00096-X