# Through the Pupils' Lens: Multilingual Effort in First and Second Language Listening

Dana Bsharat-Maalouf,[1] Jens Schmidtke,[2] Tamar Degani,[1] and Hanin Karawani[1]



**Objectives:** The present study aimed to examine the involvement of listening effort among multilinguals in their first (L1) and second (L2) languages in quiet and noisy listening conditions and investigate how the presence of a constraining context within sentences influences listening effort.

**Design:** A group of 46 young adult Arabic (L1)–Hebrew (L2) multilinguals participated in a listening task. This task aimed to assess participants' perceptual performance and the effort they exert (as measured through pupillometry) while listening to single words and sentences presented in their L1 and L2, in quiet and noisy environments (signal to noise ratio = 0 dB).

**Results:** Listening in quiet was easier than in noise, supported by both perceptual and pupillometry results. Perceptually, multilinguals performed similarly and reached ceiling levels in both languages in quiet. However, under noisy conditions, perceptual accuracy was significantly lower in L2, especially when processing sentences. Critically, pupil dilation was larger and more prolonged when listening to L2 than L1 stimuli. This difference was observed even in the quiet condition. Contextual support resulted in better perceptual performance of high-predictability sentences compared with low-predictability sentences, but only in L1 under noisy conditions. In L2, pupillometry showed increased effort when listening to high-predictability sentences compared with low-predictability sentences, but this increased effort did not lead to better understanding. In fact, in noise, speech perception was lower in high-predictability L2 sentences compared with low-predictability ones.

**Conclusions:** The findings underscore the importance of examining listening effort in multilingual speech processing and suggest that increased effort may be present in multilingual's L2 within clinical and educational settings.

**Key words:** First language, Listening effort, Multilingualism, Pupillometry, Second language, Speech perception.

(Ear & Hearing 2024;XX;00–00)

## INTRODUCTION

With the rise of multilingualism around the world (Grosjean 2008, 2010; Modiano 2023), many individuals use their second language (L2) in the workplace and in educational settings. These naturalistic environments often include adverse listening conditions, such as noise, which may hinder individuals' ability to perceive speech effectively (Mattys et al. 2012). Despite numerous studies that have documented poorer perceptual performance of multilinguals under adverse listening conditions (Garcia Lecumberri et al. 2010; Scharenborg & van Os 2019; Cowan et al. 2022), the underlying mechanisms of this phenomenon remain relatively poorly understood. Building upon the gaps identified within existing research, the present study aimed to investigate the perceptual difficulties experienced by multilinguals in adverse listening conditions, with a specific focus on the exertion of listening effort. In particular, the present study examined whether listening effort within multilinguals differs between L1 (first language) and L2 in quiet and noisy conditions, and how the presence of constraining sentential context modulates these effects.

Perceptual studies with multilingual individuals show that although these listeners can effectively perceive speech in their languages under quiet listening conditions, their performance significantly declines in their L2 when faced with adverse listening conditions (Mayo et al. 1997; Von Hapsburg et al. 2004; Rogers et al. 2006; Rosenhouse et al. 2006; Weiss & Dempsey 2008; Garcia Lecumberri et al. 2010; Shi & Sanchez 2010; Tabri et al. 2015; Desjardins et al. 2019; Skoe & Karayanidi 2019; Bsharat-Maalouf & Karawani 2022a, b). The perceptual disadvantage of multilinguals under adverse conditions is modulated by a variety of factors, including those pertaining to listeners' proficiency (Shi 2012, 2015; Rimikis et al. 2013; Kilman et al. 2014; Schmidtke 2016; Scharenborg et al. 2018), and acquisition history (Mayo et al. 1997; Meador et al. 2000; Weiss & Dempsey 2008; Shi 2010, 2012; Shi & Sanchez 2010; Regalado et al. 2019). Of relevance to the present study, the amount of contextual information present in the speech material has also been shown to modulate the perceptual performance of multilinguals under adverse listening conditions (Mayo et al. 1997; Van Wijngaarden et al. 2002; Bradlow & Alexander 2007; Warzybok et al. 2015; Krizman et al. 2017; Skoe & Karayanidi 2019; Bsharat-Maalouf & Karawani 2022b). In particular, multilinguals' perceptual difficulties become more pronounced as the complexity of the speech stimuli increases. For instance, Krizman et al. (2017) investigated performance in tones, single words, and sentences, all presented in noise. They found that bilinguals tested in their L2 performed poorer than monolinguals when perceiving sentences, performed similarly when perceiving single words, and performed better when listening to tones. Thus, whereas monolinguals were able to rely on contextual cues during the perceptual process, bilinguals' ability to capitalize on such cues appears to be reduced in noise. Other studies comparing performance on sentences with varying levels of predictability (high and low-predictability sentences) showed that whereas L1 listeners benefited from contextual cues when listening to sentences presented under adverse listening conditions, this benefit was not as evident in L2 listeners (Mayo et al. 1997; Bradlow & Alexander 2007; Shi 2010; Schmidtke 2016; Kousaie et al. 2019; Skoe & Karayanidi 2019; Bsharat-Maalouf & Karawani 2022b). For instance, Bsharat-Maalouf and Karawani (2022b) examined perceptual performance of Arabic–Hebrew multilinguals as they listened to words, as well as to high and low-predictability sentences, all

[1]Department of Communication Sciences and Disorders, University of Haifa, Haifa, Israel; and [2]Haifa Center for German and European Studies, University of Haifa, Haifa, Israel.

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





<zdoi; 10.1097/AUD.0000000000001602>

presented in listeners' L1 and L2 under quiet and noisy conditions. The results showed that a benefit from contextual cues in noise was prominent in L1 sentences but was absent in the L2. It is interesting that this same study did not detect differences between L1 and L2 processing in quiet, nor in perceptual processing of single words presented in noisy conditions. However, the authors suggested that the absence of a perceptual difference in performance does not preclude the possibility that the underlying cognitive processes involved are different across the L1 and the L2. In particular, it is possible that in quiet conditions, multilinguals needed to exert more listening effort in their L2 to achieve similar accuracy as they did in their L1.

Listening effort refers to the deliberate allocation of mental resources to compensate for challenges when carrying out a listening task (Pichora-Fuller et al. 2016). Whereas most models of speech perception work under the assumption of an optimal speech signal, the Ease of Language Understanding (ELU) model explains language understanding under suboptimal listening conditions as they occur, for instance, when noise is present. According to this model, listening effort is linked to the ease with which perceived signals are matched with semantic long-term memory stored representations, and the necessity of engaging explicit cognitive processes, such as working memory, during language understanding (Rönnberg et al. 2008, 2013, 2019, 2021). In favorable listening conditions, the signal is easily matched to stored representations in long-term memory, enabling rapid, automatic, and implicit speech processing. However, under adverse listening conditions, such as in noise, the signal is distorted, making the process of matching the signal to stored representations more challenging. As a result, listening becomes more demanding, requiring the engagement of an explicit processing loop that relies on working-memory resources to aid understanding. This increase in explicit processing is associated with heightened listening effort (Rönnberg et al. 2008, 2013, 2019).

There are several reasons to predict increased listening effort when multilingual individuals process speech in their L2. First, heightened effort in L2 can be attributed to the quality of their stored representations. Multilinguals often experience reduced proficiency, usage, and exposure in their L2 compared with L1 (Rosenhouse et al. 2006; Desjardins et al. 2019; Abbas et al. 2021; Bsharat-Maalouf & Karawani 2022b). According to exemplar models of speech understanding, stored representations in memory are shaped by long-term experience, such that fewer encounters with a word result in less detailed representations in long-term memory (Goldinger 1996; Schmidtke 2016) and weaker links among their subcomponents (Gollan et al. 2005, 2008, 2011, 2015; Sebastián-Gallés et al. 2005). According to the ELU, the process of matching the perceived signal with stored representations is influenced by external factors such as the presence of background noise, as discussed earlier, and by internal factors related to the listener, like the precision of the stored representations (Rönnberg et al. 2013). Thus, in cases where representations are thought to be less precise—such as in multilingual listeners' L2—the ease of matching the incoming signal with stored representations is likely to be lower, requiring greater listening effort.

Another reason to expect increased listening effort in multilingual L2 has to do with the nature of language activation and competition. Specifically, during multilingual language processing, activation of words from both languages occurs, leading to candidate word activation from one language during the processing of stimuli in the other language (Marian & Spivey 2003; Weber & Cutler 2004; Schwartz & Kroll 2006; Shook & Marian 2012, 2013; Blumenfeld & Marian 2013; Chen et al. 2017; Bobb et al. 2020). This language coactivation exhibits an asymmetry, with candidate words from L2 being less activated compared with those from L1 (Broersma & Cutler 2008). According to the ELU model, stored representations consist of multiple attributes, and there is a minimum threshold of these attributes that must be accessed for a particular stored representation to be successfully selected (Rönnberg et al. 2013, 2019). Thus, when the activation of attributes falls below this threshold, some neighboring representations may be retrieved, requiring the engagement of explicit processes to complete lexical access (Rönnberg et al. 2013, 2019). In the context of multilingual L2 listening, increased competition arising from L1 candidate words can intensify the challenge of reaching the necessary threshold for successful representation access, leading to heightened listening effort (Borghini & Hazan 2018).

Lastly, the challenge faced by multilinguals in utilizing contextual cues in their L2, as discussed earlier, can be considered another reason to expect increased listening effort in this language. Listeners typically resort to contextual cues trying to maximize available information when a necessity arises (Skoe & Karayanidi 2019; Corps & Rabagliati 2020). In the ELU model, the use of contextual support may be viewed as a top-down process used to infer missing information when a mismatch occurs. Consequently, the challenge of relying on contextual cues in L2 can lead to less efficient resolution of mismatches, ultimately leading to increased effort. Collectively, considering the less precise stored representations, the dynamics of language coactivation, and the reduced utility of contextual cues, it is reasonable to hypothesize that the demand and use of cognitive resources increase when listening to L2, thereby contributing to heightened listening effort in L2 compared with L1.

Examining multilingual listening effort is important because it can uncover challenges in speech processing that are not apparent in perceptual performance (Picou et al. 2013; Desjardins & Doherty 2014; Picou & Ricketts 2014; Winn et al. 2015; Xia et al. 2015; Brown et al. 2020; McLaughlin & Van Engen 2020; Pielage et al. 2021; Winn & Teece 2021; Baese-Berk et al. 2023). Thus, given the aforementioned factors, which strongly indicate that speech processing in L2 may be more cognitively demanding compared with L1, relying exclusively on perceptual performance may be insufficient to uncover the challenges in speech processing faced by multilinguals. Understanding and addressing these challenges are not only of theoretical importance but also carries practical significance. This is because sustained listening effort has been linked to heightened levels of mental fatigue and stress (Hornsby et al. 2016; Pichora-Fuller 2016; Alhanbali et al. 2017), and reduced multitasking abilities (Wu et al. 2016; Gagne et al. 2017; Kaplan Neeman et al. 2022).

Given the dual importance of examining listening effort from both theoretical and practical perspectives, in recent years there has been a noticeable increase in interest surrounding listening effort among multilingual listeners (Kilman et al. 2015; Borghini & Hazan 2018, 2020; Francis et al. 2018; Lam et al. 2018; Desjardins et al. 2019; Peng & Wang 2019; Visentin et al. 2019; Oosthuizen et al. 2020; Brännström et al. 2021).

Bsharat-Maalouf et al. (2023) conducted a recent review of this literature, identifying its current limitations, and proposing avenues that require further exploration. One notable limitation emphasized in that review pertains to the research design of previous studies, which primarily focused on comparing effort across different listeners. Specifically, to date, the common approach has been to use between-participant comparisons, often contrasting the listening effort of monolinguals with that of bilinguals in their L2. For example, Peng and Wang (2019) showed that when engaged in English perceptual tasks presented in adverse listening conditions, bilinguals who acquired English as their L2 reported significantly higher levels of perceived listening effort when compared with English monolinguals. Lam et al. (2018) further corroborated this trend by demonstrating that listening to English words led to prolonged reaction times, signifying increased effort, in bilinguals for whom English served as their L2, in contrast with monolinguals who spoke English as their L1. As pointed out in Bsharat-Maalouf et al., the comparison between monolinguals and bilinguals could be problematic because it often includes individuals with varying characteristics beyond just their language background. Also, recent studies by De Houwer (2023) and Rothman et al. (2023) have extensively addressed the challenges associated with using monolinguals as a control group when studying multilingual language processing. To alleviate this issue, the present study adopted a within-participant design, testing both perceptual performance and listening effort among Arabic–Hebrew multilinguals while they listen to various speech stimuli (including words, and high and low-predictability sentences) presented in both their L1 and L2 in quiet and noisy conditions.

A second limitation highlighted in the Bsharat-Maalouf et al. (2023) review pertains to the various tools used to test listening effort. In particular, the review shows that many types of measure have been used to examine multilingual listening effort, including subjective ratings and behavioral measures such as dual task paradigms and reaction times (Kilman et al. 2015; Lam et al. 2018; Desjardins et al. 2019; Peng & Wang 2019; Visentin et al. 2019; Oosthuizen et al. 2020; Brännström et al. 2021), as well as the objective tool of pupillometry, measuring changes in pupil responses (Schmidtke 2014; Borghini & Hazan 2018, 2020; Brännström et al. 2021). As the measurement tool used to assess listening effort has the potential to substantially influence the observed findings (Wendt et al. 2016; Alhanbali et al. 2019; Visentin et al. 2022), attention should be given to the selected tool. Guided by previous studies highlighting the sensitivity and reliability of the pupillometry as a measure of listening effort (Giuliani et al. 2021; Neagu et al. 2023) and in line with the Bsharat-Maalouf et al. review, which underscores the consistency of results obtained from this measure in the context of multilingual listening effort, the present study used pupillometry as the prominent index of listening effort (for comprehensive reviews on this tool see Van Engen & McLaughlin 2018; Winn et al. 2018; Zekveld et al. 2018).

Using pupillometry, Schmidtke (2014) showed that during a spoken-word recognition in English task, Spanish (L1)–English (L2) bilinguals exhibited delayed pupil responses, indicative of increased effort, in their L2 compared with English monolingual individuals. Likewise, two studies by Borghini and Hazan (2018, 2020) provided consistent findings among Italian (L1)–English (L2) bilinguals, with bilingual listeners showing greater pupillary dilation, signifying heightened listening effort, in their L2 in comparison to their English monolingual counterparts. However, one confounding variable that was explicitly raised by Borghini and Hazan in relation to the between-participant comparisons was the disparity in cognitive abilities among participants. This disparity could introduce a source of bias, potentially contributing to the observed differences in listening effort between the two groups. Indeed, pupillometry is sensitive to interindividual differences including age, hearing status, motivation, level of fatigue, and cognitive abilities (Zekveld et al. 2011, 2018; Winn et al. 2018). These confounding factors may thus limit the generalizability of the previous pupillometry studies where a between-participant design was used.

To the best of our knowledge, thus far, only the study of Francis et al. (2018) has used a within-participant comparison and utilized pupillometry to examine multilingual listening effort. In that study, listening effort was examined within a group of Dutch (L1)–English (L2) bilinguals while listening to noisy sentences. They found an increase in pupil dilation when changing the target sentences from Dutch to English, indicating heightened listening effort in multilingual L2 compared with L1. The extent to which these effects extend to simpler speech stimuli is unclear given that only sentences were tested in the Francis et al. study. Thus, the present study extended the literature by incorporating a simpler set of speech stimuli consisting of single words. This allowed us to isolate potential differences in listening effort without the influence of contextual cues typically present in sentences, a factor that may differ in L1 compared with L2 (Scharenborg & van Os 2019). In addition, our study extended the work of Francis et al. by examining multilingual listening effort in both quiet and noisy environments, providing an important control for listeners' performance in more ideal conditions across their two languages. It thus allowed examination of whether multilingual listening effort in L2 is increased, even in quiet conditions, where no adverse environmental factors are present. In summary, the first aim of the present study was to investigate differences in listening effort within multilinguals as they listen to single words in both their L1 and L2 under quiet and noisy conditions.

Beyond this core aim, given the challenges multilinguals encounter in benefiting from contextual cues in their L2 (Skoe & Karayanidi 2019; Bsharat-Maalouf & Karawani 2022b), our study also aimed to shed light on the impact of contextual cues within sentences on the listening effort experienced by multilinguals in both of their languages. Whereas this issue has been studied in monolinguals (Desjardins & Doherty 2014; Johnson et al. 2015; Winn 2016; Holmes et al. 2018; Lau et al. 2019; Hunter & Humes 2022; Rovetti et al. 2022), the research conducted by Borghini and Hazan (2020) is the only study to examine how the availability of contextual cues during sentence comprehension in noise affects multilingual listening effort. To manipulate the availability of contextual cues they used plausible and anomalous sentences and examined such effects on monolingual English speakers and Italian native speakers who had learned English as an L2. Surprisingly, their findings showed that a coherent semantic context within sentences did not reduce listening effort for either monolinguals or bilinguals. This was evident by a lack of difference in pupillary dilation when processing plausible sentences compared with anomalous ones. This finding contradicted previous studies conducted with monolinguals, which demonstrated that higher stimulus predictability typically led to a reduction in listening effort (Johnson et al.

2015; Winn 2016; Holmes et al. 2018; Rovetti et al. 2022). The authors suggested that their choice of using anomalous versus plausible sentences rather than high versus low-predictability sentences could have contributed to this unexpected result. In addition, they acknowledged the possibility that presenting plausible and anomalous sentences in separate blocks could have affected participants' anticipation of coherent or incoherent sentences, potentially influencing performance. The authors further suggested that the different signal to noise ratios (SNRs) at which plausible and anomalous sentences were presented, used to ensure comparable levels of intelligibility, may have affected the observed pattern. Because the SNR used for the plausible sentences was overall more challenging compared with the SNR used for anomalous sentences, increased effort was required in listening to plausible sentences, potentially overshadowing any impact of semantic context on effort.

In the present study, we addressed these acknowledged limitations. Specifically, in addition to assessing single words, we included both high and low-predictability sentences presented randomly within the same blocks to minimize listener's expectations. Unlike Borghini and Hazan (2020), we examined listening effort in noisy conditions under a fixed level of noise, ensuring consistent degradation across experimental conditions. However, at the same time, we acknowledged the potential differences in perceptual accuracy when assessing multilinguals in L1 and L2 sentences in noise under the same SNR (Bsharat-Maalouf & Karawani 2022b). Such differences could potentially confound the assessment of listening effort, as poorer perceptual accuracy may result in increased pupillometric measures (Zekveld et al. 2010; Zekveld & Kramer 2014). However, by including a quiet condition assessment in our study, we aimed to examine if the effects observed regarding contextual cues remain consistent even when speech intelligibility remains unaffected and comparable across languages.

In summary, the present study aimed to answer two key research questions. First, does listening effort within multilinguals differ in L1 and L2 in quiet and noisy conditions? Second, how does the presence of a constraining context within sentences influence multilingual listening effort, and do these effects manifest differently in L1 and L2? To examine these questions the present study tested a group of Arabic–Hebrew multilinguals in both Arabic (L1) and Hebrew (L2), presenting words, low-predictability sentences, and high-predictability sentences in both quiet and noise conditions. Perceptual performance was assessed alongside pupillometry to measure the level of listening effort.

## MATERIALS AND METHODS

### Participants

Forty-six young adult Arabic–Hebrew–English multilinguals (40 females, mean age = 23.09 [3.87], years of formal education = 14.86 [1.83]) participated in this study. Data from 2 additional participants were excluded, one due to self-reported hearing impairment and the second due to noncompletion of the experimental task.

All participants grew up in exclusively Arabic-speaking homes and received education in schools where Arabic was the primary instructional language. Around the age of 8, they started learning Hebrew through formal instruction and had some exposure to Hebrew as it is the majority language in the country. At the time of data collection, participants were taking university-level classes taught in Hebrew. Participants learned English as their third language around the age of 9 and had exposure to the language through media resources such as music, television, and watching movies.

Participants demonstrated greater proficiency in Arabic compared with Hebrew as established through self-report data collected via the Multilingual Language Background Questionnaire (Abbas et al. 2024) and objective proficiency tests including semantic fluency (Gollan et al. 2002; Kavé 2005), and a picture naming test (Multilingual Naming Test [MINT] Sprint, Garcia & Gollan 2022). See details later (Background tasks) and Table 1 for participant characteristics.

None of the participants had any prior knowledge or exposure to any language other than Arabic, Hebrew, and English, none reported cognitive or neural disorders, cataracts, or hearing loss, and none had taken any drugs or medications before the experiment. They also had normal or corrected-to-normal vision and reported no history of language or learning disabilities. In addition, exclusion criteria included caffeine consumption less than 3 hours before the testing session. These criteria were essential to exclude any potential confounding variables on perceptual performance or pupil dilation. Participants were recruited through advertisements on social media and around campus. They provided informed consent in accordance with the university's ethics committee and either received course credit or monetary compensation for their participation.

### Overview of Experimental Session

The experiment described here constitutes a single session of a larger study conducted at the lab. Before participation, participants were required to complete a screening form to ensure they met the inclusion criteria. Within the experimental session, participants completed a listening task, during which their pupil size was recorded (as detailed in Experimental task procedure), and two proficiency tests (the semantic fluency task and the MINT Sprint task, see Background tasks), both in Arabic and in Hebrew. Specifically, following the completion of the listening task in each language, participants performed the corresponding proficiency tests in the same language. The order of the two proficiency tests (the semantic fluency task

**TABLE 1. Multilingual participant characteristics (N = 46)**

| | L1 (Arabic) | L2 (Hebrew) |
|---|---|---|
| Age began to learn the language (yrs) | 0 (birth) | 7.59 (0.88) |
| Self-rated proficiency (0–10 scale) | 9.76 (0.40) | 8.13 (1.02) |
| Current exposure (%) | 61.68 (12.76) | 31.63 (10.9) |
| Current use (%) | 51.70 (18.52) | 32.22 (14.33) |
| Semantic fluency (number of items) | 22.52 (4.66) | 16.85 (4.72) |
| Mint Sprint Test (range 0–80) | 67.67 (6.25) | 37.61 (11.9) |

*Proficiency ratings were averaged across productive and receptive language skills (speaking, reading, speech comprehension, and writing). Exposure percentage was averaged across various contexts (work, university, friends, family, and free time), and use percentage was averaged across different activities (speaking, reading, writing, social media, music listening, and TV watching). The percentage for exposure and use does not sum up to 100%, as the remaining proportion accounts for usage and exposure to the English language. Semantic fluency (Gollan et al. 2002; Kavé 2005) scores were obtained by summing items produced for two categories, 1 min per category. Mint Sprint Test (Garcia & Gollan 2022) scores were derived by summing the total words produced in two rounds. In all measures, there were significant differences ($p < 0.01$) between L1 and L2 languages. SDs appear in parentheses.*

and the MINT Sprint task) within each language was counterbalanced. The order of language administration was counterbalanced across participants. In a subsequent session of the experiment, participants filled out a questionnaire to evaluate language and background characteristics. The listening task in each language took approximately 1 hr to complete, while the remaining tasks (proficiency tests and questionnaire) taken together took approximately 20 min. During the experimental session, all instructions were presented in the participants' native language (Arabic) to ensure their comprehension of the listening task and to maintain language consistency across participants. In addition, all communication throughout the session was naturally conducted in Arabic by a native Arabic experimenter.

## Stimuli and Tests

### Main Experimental Task •

#### *Experimental task stimuli.*

#### *Single words*

Arabic and Hebrew single-word nouns (120 in each language) were used in the present study. The stimuli in the different languages were not translation equivalents. No cognates or false cognates were included to avoid any confusion based on phonological similarity across languages (Degani et al. 2018).

The words in both languages were matched in length (number of pronounced phonemes), frequency (counts per million extracted from *arTenTen* and *heTenTen* corpora via Sketch Engine [Kilgarriff et al. 2014]), and normed concreteness (all *ps* ≥ 0.67, see Table 2). Concreteness norms were based on the ratings of five native Arabic speakers who rated the concreteness of Arabic and Hebrew words using a five-point scale (ranging from 1 = not concrete at all to 5 = very concrete).

To further confirm that native Arabic speakers would be familiar with the Hebrew stimuli, 10 additional Arabic students were presented with the list of Hebrew words (120 words in total) and asked to provide translations of each word in their native language, Arabic. The results indicated that each word received accurate translations from at least 8 out of the 10 participants, with an overall 98.9% proper translation rate.

#### *Sentences*

For each single word, two sentences with six words each were created, with the target noun always presented in sentence-final position. One sentence was designated high predictability, and one low predictability as detailed later. The final list of stimuli is provided in Supplemental Digital Content 1, http://links.lww.com/EANDH/B523. Sentences were matched in length (number of syllables) across Arabic and Hebrew (Table 2) and were created to be plausible, and of simple grammatical structure, as verified by 4 native speakers of Arabic and 4 native speakers of Hebrew, who did not take part in the main experiment.

Sentence predictability was established through a norming study, following general procedures outlined in Mor and Prior (2022). The predictability of Arabic sentences was assessed by 36 native Arabic speakers, while the predictability of Hebrew sentences was assessed by 18 native Arabic speakers and 18 native Hebrew speakers. In the norming study, each sentence was presented with the final word replaced by a blank and participants were instructed to complete the sentence with the first word that came to mind. Sentences were taken out and rewritten if fewer than 60% of the participants produced the target word in the high-predictability sentences, or when more than 10% of the participants produced the target word in the low-predictability sentences. For these revised sentences, we then followed the same procedure again until the criteria were met. To avoid priming the target word, each participant in each norming phase was exclusively presented with either high-predictability sentences or low-predictability sentences. All Hebrew sentences were completed by a plausible noun by Arabic participants, indicating the comprehensibility of Hebrew sentences for L2 speakers of the language. The probability of target words collected from Arabic speakers and from Hebrew speakers met the criteria we set, confirming the desired predictability levels. Overall, the predictability of sentences was carefully matched across Arabic and Hebrew, see Table 2.

#### *Recording*

Stimuli were recorded by native female speakers of each respective language. The recordings were made using JBL Tune 500BT headphones equipped with a microphone in a sound-attenuated booth with a 44.1 kHz sample rate and 32-bit resolution. The Arabic and the Hebrew speakers were asked to produce the stimuli at a natural rate with neutral intonation. The Arabic stimuli were recorded in the Southern Levantine Arabic dialect, which is the dialect predominantly spoken by multilingual Arab individuals in the country (Brustad & Zuniga 2019). To ensure consistency in recorded stimuli, Praat software (Boersma & Weenink 2009) was used to adjust the amplitude of each recording, resulting in samples with the same average root-mean-square amplitude. The intelligibility of the recordings was assessed by 2 native speakers for each language and found to be clear and accurate. The durations of single-word recordings, as well as the durations of sentence recordings, were matched between Arabic and Hebrew (all *ps* ≥ 0.25, see Table 2). In addition, within each language, the durations of high and low-predictability sentence recordings were matched (*ps* ≥ 0.27).

**TABLE 2. Single words and sentence characteristics in L1 (Arabic) and L2 (Hebrew)**

| | Single Word | | | | High-Predictability Sentences | | | Low-Predictability Sentences | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Length | Freq. | Concret. | Dur. | Length | Predic. | Dur. | Length | Predic. | Dur. |
| L1 | 5.03 (0.63) | 58.95 (84.84) | 4.84 (0.43) | 0.80 (0.09) | 16.02 (1.46) | 90.11 (10.14) | 3.05 (0.17) | 15.67 (1.56) | 1.11 (2.84) | 3.05 (0.18) |
| L2 | 5.12 (0.64) | 58.42 (77.38) | 4.79 (0.5) | 0.79 (0.09) | 16.46 (1.31) | 89.70 (11.14) | 3.05 (0.23) | 16.1 (1.54) | 1.94 (4.12) | 3.01 (0.28) |
| *p* | 0.67 | 0.82 | 0.78 | 0.54 | 0.127 | 0.34 | 0.91 | 0.19 | 0.149 | 0.25 |

*Word length represents number of phonemes; frequency (Freq.) are counts per million (extracted from Sketch Engine Kilgarriff et al. 2014); concreteness (Concret.) rated on a scale of 1 (low)–5 (high). Sentence length represents the number of syllables per sentence; predictability (Predic.) represents the percentage established in the predictability norming study; duration (Dur.) represents the recorded stimuli's duration, in seconds. No significant differences were found between L1 and L2 stimuli (p > 0.05) across all measures. SDs appear in parentheses.*

***Noise manipulation***

In the listening task, stimuli were presented in quiet and noise. In the noise condition, the recorded stimuli were mixed with speech-shaped noise via Praat software (Boersma & Weenink 2009), at a SNR of 0 dB. To generate the speech-shaped noise, white noise was filtered to match the long-term average spectrum of the stimuli given in each language. The 0 dB SNR was chosen based on findings in the literature suggesting that this level would be challenging but feasible (not too easy nor too difficult) for a range of participant profiles (Garcia Lecumberri et al. 2010), including multilinguals (Bsharat-Maalouf & Karawani 2022b).

Within each language, three lists of the stimuli were created (see Supplemental Digital Content 2, http://links.lww.com/EANDH/B524, for matching details). These lists were then rotated across versions, such that every word was presented twice to each participant, albeit under different listening conditions and context levels. As a result, each participant was presented with all word stimuli twice, and across participants, each word was presented in all possible conditions (quiet/noise by single/low/high).

***Experimental task procedure.*** Before starting the experimental task in each language, participants had a short practice block to familiarize themselves with the listening task. The language of stimuli within the practice block was tailored to the language of the upcoming listening task.

Participants performed the listening task with eight blocks in each language, with language order counterbalanced across participants. Within each language, half of the blocks (four) were quiet, and half included noise (four), in randomized order. Each block consisted of 30 trials, incorporating an even mix of single words, high-predictability sentences, and low-predictability sentences, presented randomly to prevent participants from anticipating the type of stimulus presented. In total, each participant was presented with 240 experimental trials in each language: 80 single words (40 in quiet and 40 in noise), 80 high-predictability sentences (40 in quiet and 40 in noise), and 80 low-predictability sentences (40 in quiet and 40 in noise). Trial order and block order were fully randomized for each participant.

During the listening task participants sat in a dimly lit sound-attenuated booth, in front of a computer screen positioned 65 cm away. A chinrest was used to reduce movement and facilitate reliable pupil size measurement (Winn et al. 2018). Stimuli were binaurally presented to participants via JBL Tune 500BT headphones at a stable intensity. During the listening task, changes in pupil size were recorded using the Eyelink Portable Duo (SR Research, Kanata, Ontario, Canada), monocularly from the pupil of the right eye at a sampling rate of 1000 Hz. To avoid any confounding effect on the pupil dilation, room luminance was stable for all participants and the computer screen maintained a constant gray background color (RGB values: 225, 225, 225).

Participants were first presented with written instructions in their native language (Arabic) about the listening task. Then, a nine-point calibration procedure was initiated and validated. Calibration and validation were followed by the practice block of four trials (including single words and sentences, half presented in quiet and half in noise), and then by the eight experimental blocks, as explained earlier. Before each trial, a drift correction point was displayed to ensure consistent pupil tracking throughout the task. Each trial started with a black fixation cross followed by 1 sec of either silence or speech-shaped noise, depending on the block condition. This allowed for establishing baseline pupil diameter (Winn et al. 2018). The speech stimulus was then played, while the fixation cross remained black. Following stimulus offset, the fixation cross continued to be displayed for an additional 3 sec, accompanied by either silence or noise, based on the block condition. This interval allowed sufficient time for the pupil to reach its maximum dilation (Winn et al. 2018). During the display of the black cross, participants were instructed to maintain their gaze and focus on the cross. Then, the fixation cross was replaced by a question mark (for a maximum time of 5 sec), which signaled participants to repeat the stimulus (single word or sentence) out loud as accurately as possible. Participants were permitted to rest their eyes and shift their gaze when the question mark was displayed. The trial ended with a blank screen displayed for 1.5 sec (Fig. 1). After confirming that the participant was ready to continue, the next trial was initiated by the experimenter. Participants were given a short break after each block (30 trials) but breaks between trials were also permitted in case a participant asked for it. No feedback was given during the experimental blocks.

During data collection, the experimenter was able to monitor the pupil recording visually and intervene if necessary. When needed, the experimenter reminded participants to focus their gaze on the center of the screen, not to blink during the fixation cross, or to adjust their position to enable the eye tracker to detect their pupil.

**Background tasks.**

***Semantic fluency task***

Participants were asked to produce as many words as possible within 1 min in a given language (Gollan et al. 2002; Kavé 2005). Two fixed semantic categories per language were used: occupations and furniture for Arabic, and fruits and sports for Hebrew. The categories for each language were chosen based on a previous norming study that ensured comparability across the pairs of categories. The order of administering the two categories within each language was randomized. During the task, each category was presented on a computer screen, followed by an hourglass indicating the time limit (60 sec). The number of correct words produced for the two categories within each language was summed to obtain a single semantic fluency score for each language (Table 1).

***MINT sprint***

Participants were asked to name a set of 80 pictures displayed on a computer screen (Garcia & Gollan 2022). The pictures were presented in an 8 by 10 grid and were ordered by difficulty, with the easier items appearing on the top rows and the more difficult items at the bottom. Participants were given a time limit of 3 min to name the pictures, as quickly as possible, starting from the top left corner of the screen and progressing through each row. After completing the first pass, participants were given a second pass, with no time limit, during which they could attempt to name any pictures they had skipped in the first round. The same set of 80 pictures was used across the two languages. The total number of words produced in both rounds (ranging from 0 to 80) within each language was summed to give a single total score for that language (Table 1).

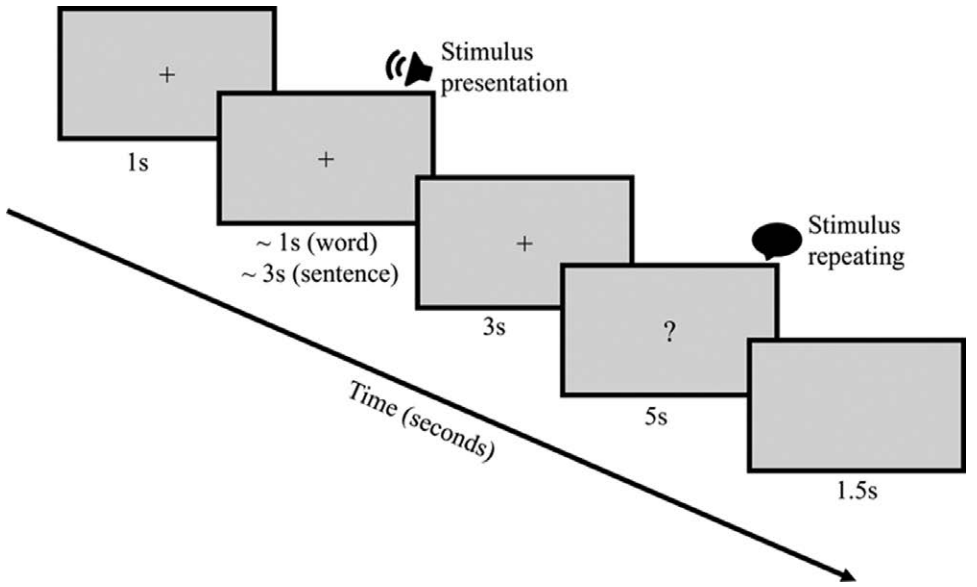


Fig. 1. Example trial sequence in the listening task. After establishing the pupil baseline, participants heard a stimulus, which they were asked to repeat aloud when prompted with a question mark.



***Multilingual Language Background Questionnaire (adapted from The Language Experience and Proficiency Questionnaire)***

Participants completed a computerized questionnaire (available at https://osf.io/preprints/psyarxiv/jfk8b, Marian et al. 2007; Abbas et al. 2024), which collected demographic information, such as age, gender, years of education and parental language, and education background, as well as self-ratings of language use, exposure, and proficiency in all acquired languages (Table 1).

***Data processing.*** Participants' verbal responses were audio-recorded using a mini universal serial bus recorder to allow offline coding. In the case of single words, a score of 1 point was assigned to an exact repetition of the word. For high and low-predictability sentences, a score of 1 point was assigned to each word repeated correctly, and overall sentence accuracy was computed as the number of words correctly repeated out of the total of six target words in the sentence.

Pupil recording was continuous during the experiment. However, the analysis focused on the time window before participants were prompted to repeat the stimulus they heard. This step was taken to avoid potential confounding effects of motor planning and head movements on pupil diameter (Richer & Beatty 1985).

Pupil size was recorded by the system in arbitrary units (au), which were subsequently converted to millimeters (mm) of diameter. The conversion formula that we determined ($pupil_{(mm)} = 0.0022 \times pupil_{(au)} + 1.8702$) was established by recording artificial pupils (black circles printed on white paper) of different known sizes (Wilschut & Mathôt 2022).

A multistep procedure was used to preprocess the pupil data and address missing data, which resulted from participants looking away from the screen or when participants closed their eyes momentarily (i.e., blinking) (Mathôt & Vilotijević 2022). Pupil diameters more than three SDs below the mean diameter of each trial were coded as a blink using the Eyelink Data Viewer software (SR Research Ltd., version 4.3.1). Because blinks are accompanied by partial occlusion of the pupil, which results in unreliable measurements (Siegle et al. 2008; Zekveld et al. 2018), we excluded the 100 msec preceding and following a blink event. Following this procedure, we excluded trials for which more than 25% of observations were missing, resulting in the exclusion of 0.2% of the data. Missing values in the remaining trials were replaced through linear interpolation, that is, the points on either side of a blink were connected by a straight line (Mathôt & Vilotijević 2022). Next, we applied a four-point moving average smoothing filter over the de-blinked data to reduce high-frequency noise (Schmidtke 2018).

Following preprocessing, the average of the last 200 msec of the prestimulus period was used to establish a baseline for each trial (similar to Van Steenbergen & Band 2013; Shechter & Share 2021). This baseline was then subtracted from all subsequent measurements in the trial (baseline corrected values = observed pupil size − baseline) to be able to infer the degree of pupil dilation in response to the stimulus (also called task-evoked pupil response, Mathôt & Vilotijević 2022). Thus, in each trial, we determined two critical pupil outcome measures: peak amplitude (relative to baseline) and peak latency. Peak amplitude represents the maximum positive dilation in a trial, measured from speech onset until 3 sec after stimulus offset, offering insights into the maximum cognitive load experienced (Zekveld et al. 2011; Koelewijn et al. 2014, 2015, 2018). Peak latency denotes the time taken for the peak dilation amplitude to manifest, representing when cognitive resources were deployed (Hyönä et al. 1995). For completeness, the analysis of mean pupil dilation (relative to baseline)—representing the average pupil dilation throughout the entire trial, from speech onset to 3 sec after stimulus offset—is reported in Supplemental Digital Content 3, http://links.lww.com/EANDH/B525. This analysis generally showed patterns consistent with the peak amplitude analysis.

***Data analysis and model structure.*** We conducted separate analyses for single words and sentences. The difference in measurement scales posed potential statistical challenges as perceptual accuracy in single words yielded either 0 or 1, while sentence accuracy could range from 0 to 6. In addition,

the cognitive demands placed on participants by asking them to repeat the stimulus differed between the two types of stimuli. In single-word trials, participants were tasked with the relatively simpler processes of repeating a single word, whereas in sentence trials, they faced the more complex challenge of repeating the entire sentence, requiring sentence-level processing, including reliance on contextual cues. Moreover, differences in duration between trials containing single words and those containing sentences make direct comparisons of changes in pupil dilation difficult to conduct. Therefore, analysis of single-word pupil data aimed to explore listening effort differences in multilinguals' L1 and L2. Conversely, the analysis of sentence-related pupil data focused on understanding how sentential context modulates these effects.

In the single-word pupil data analysis, we considered only trials with correct responses. Thus, any trials where participants did not provide any response or repeated the word incorrectly were excluded from the analysis. At the sentence level, we included pupil responses from trials where participants successfully repeated at least three out of six words from the sentence. We implemented these criteria to strike a balance between ensuring participants paid attention (Zekveld et al. 2010, 2014; Wendt et al. 2018) and retaining a high number of trials. After applying these exclusion criteria, an average of 89% of trials per participant were retained in the analysis (SD = 6.3, range = 58–97%). We set a minimum threshold of 50% valid trials for each condition per participant (i.e., 20 trials) as an inclusion criterion, which was met by all participants.

Perceptual performance and pupil data were analyzed using linear mixed-effect models, which offer the advantage of simultaneously accounting for variance related to participants and items (Brown 2021). Single-word accuracy was analyzed using logistic mixed-effects model due to the binary nature of the responses (0 = incorrect, 1 = correct). For sentence accuracy, which ranged from 0 (= no words repeated) to 6 (all words repeated), we utilized a negative binomial mixed-effects model to account for the count nature of the responses (Hilbe 2011). Pupil data were recorded continuously, so we used linear mixed-effects models with an assumed Gaussian error distribution for analysis.

For single-word models, the fixed effects included listening condition (quiet versus noise, with quiet set as the reference) and language (Arabic versus Hebrew, with Arabic set as the reference), along with their interaction. In addition, to control for fatigue effects (Wang et al. 2018; Jain & Nataraja 2019) trial order was included as a covariate. The random structure of the models included by-participant and by-item intercepts, as well as by-item slope for condition and by-participant slopes for condition and language.

For the sentence models, the same fundamental model structure was maintained, with the addition of the fixed effect of context (high versus low sentences, with low set as the reference), and its interactions with condition and language, as well as by-participant and by-item random slopes for context. Furthermore, in the pupil data sentence model, we included perceptual accuracy as a covariate due to the adoption of fixed noise level (SNR = 0 dB), which has the potential to introduce variations in perceptual accuracy (Bsharat-Maalouf & Karawani 2022b) and consequently confound pupil responses (Zekveld et al. 2010; Zekveld & Kramer 2014). Thus, by including perceptual accuracy as a covariate, we aimed to test listening effort above and beyond differences in perceptual accuracy.

To address convergence issues with the models including the maximal random structure, we used the buildmer function from the buildmer package (version 2.8; Voeten 2019) in R (version 4.2.2; R Core Team 2021). This function uses the (g)lmer function from the lme4 package (version 1.1-32; Bates et al. 2015) to select the random structure using backward stepwise elimination starting from the most complex model and systematically simplifying the random slopes until the model reaches convergence. Once the maximally converging model has been identified, the function calculates $p$ values for all fixed effects based on Satterthwaite degrees of freedom using the lmerTest package (version 3.1-3; Kuznetsova et al. 2017), or the Wald degrees of freedom for binomial distribution. We used the "include" subcommand to maintain all critical fixed effects in the model and allow evaluation of their contribution. To test interactions and examine pairwise comparisons, the selected model was refitted using (g)lmer and followed by the testInteractions function from the phia package (version 0.2-1; De Rosario-Martinez et al. 2015) with Bonferroni adjustments for multiple comparisons.

Model summaries (obtained from the summary function) for perceptual and pupillometry data are presented in Tables 3 and 4. Note that, because fixed effects were dummy-coded, the effects presented in these tables reflect simple effects rather than main effects. The main effects of each fixed variable were obtained from the chi-square test (for perceptual data) and the anova function (for pupil data) and are presented in the text. Significance was evaluated with an alpha level of 0.05. Figure 2 and Supplemental Digital Content 4, http://links.lww.com/EANDH/B526, present descriptive statistics for perceptual performance and pupillometry. Supplemental Digital Content 5, http://links.lww.com/EANDH/B527, provides a summary table of the results.

## RESULTS

### Single Word—Perception and Pupillometry

**Perception •** Significant main effects were observed for listening condition [$\chi^2$ (1) = 250.870, $p < 0.001$] and language [$\chi^2$ (1) = 21.336, $p < 0.001$] in the perception of single words, demonstrating better performance in quiet compared with noise and for L1 compared with L2 single words (Fig. 2). The interaction between listening condition and language did not reach statistical significance [$\chi^2$ (1) = 2.474, $p = 0.115$].

**Pupillometry (Peak Amplitude) •** Significant main effects were observed for listening condition [$F$(1,48.051) = 38.736, $p < 0.001$] and language [$F$(1,49.372) = 16.384, $p < 0.001$] on peak amplitude. A significant interaction between listening condition and language [$F$(1,48.57) = 4.816, $p = 0.033$] qualified these effects. Pairwise comparisons with Bonferroni corrections showed that L2 pupil amplitudes were significantly larger compared with L1 in both quiet (value = −0.013, $\chi^2$ = 5.640, $p = 0.035$) and noise (value = −0.028, $\chi^2$ = 18.340, $p <$ 0.001), but that the effect was more pronounced in noise, see Figures 2 and 3.

**Pupillometry (Peak Latency) •** A significant main effect was observed for listening condition on peak latency [$F$(1,45.2) = 5.262, $p = 0.026$], indicating delayed peaks in noise compared with quiet. Neither the main effect of language [$F$(1,6447.1) = 1.000, $p = 0.317$] nor the listening condition by language interaction [$F$(1,6450.3) = 0.471, $p = 0.492$] reached statistical

**TABLE 3. Model summary for single words perception and pupillometry**

| | Perception | | | Peak Amplitude | | | | Peak Latency | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *b* (SE) | *z* | *p* | *b* (SE) | *t* | *p* | | *b* (SE) | *t* | *p* |
| Fixed effects | | | | | | | | | | |
| Intercept | 6.868 (0.443) | 15.476 | **<0.001** | 0.197 (0.007) | 25.289 | **<0.001** | | 1641.77 (48.620) | 33.767 | **<0.001** |
| Condition$_{(noise)}$ | −3.835 (0.398) | −9.619 | **<0.001** | 0.016 (0.004) | 3.809 | **<0.001** | | 72.50 (46.840) | 1.548 | 0.125 |
| Language$_{(Hebrew)}$ | −0.339 (0.580) | −0.584 | 0.558 | 0.013 (0.005) | 2.375 | **0.021** | | −44.05 (35.210) | −1.251 | 0.211 |
| Condition$_{(noise)}$: language$_{(Hebrew)}$ | −0.851 (0.541) | −1.573 | 0.116 | 0.014 (0.006) | 2.195 | **0.033** | | 35.86 (52.220) | 0.687 | 0.492 |
| Control variable | | | | | | | | | | |
| Trial order | 0.200 (0.058) | 3.451 | **0.001** | −0.012 (0.001) | −9.353 | **<0.001** | | −24.28 (14.550) | −1.669 | 0.095 |
| | Var. (SD) | Corr. | | Var. (SD) | Corr. | | | Var. (SD) | Corr. | |
| Random effects | | | | | | | | | | |
| Item$_{(intercept)}$ | 2.575 (1.605) | | | 0.001 (0.005) | | | | 0.002 (0.044) | | |
| Item: condition$_{(noise)}$ | — | | | 0.001 (0.013) | −0.03 | | | — | | |
| Participant$_{(intercept)}$ | 0.198 (0.445) | | | 0.002 (0.051) | | | | 79,700 (282.300) | | |
| Participant: language$_{(Hebrew)}$ | — | | | 0.001 (0.033) | −0.05 | | | — | | |
| Participant: condition$_{(noise)}$ | — | | | 0.001 (0.021) | 0.08 | 0.29 | | 40,050 (200.100) | −0.12 | |
| Participant: language$_{(Hebrew)}$: condition$_{(noise)}$ | — | | | 0.001 (0.034) | 0.27 | −0.38 | −0.39 | — | | |

*Fixed effects reflect simple effects relative to the reference level when other variables are at their reference level without correction for multiple comparisons. For main effects, see $\chi^2$, F, and p values in the text. For the mean pupil model refer to Supplemental Digital Content 3, http://links.lww.com/EANDH/B525. Bold values indicate effects that are statistically significant, with p values less than 0.05.*

**TABLE 4. Model summary for sentences perception and pupillometry**

| | Perception | | | Peak Amplitude | | | Peak Latency | | |
|---|---|---|---|---|---|---|---|---|---|
| | *b* (SE) | *z* | *p* (z) | *b* (SE) | *t* | *p* (t) | *b* (SE) | *t* | *p* (t) |
| Fixed effects | | | | | | | | | |
| Intercept | 1.761 (0.011) | 150.234 | **<0.001** | 0.402 (0.013) | 29.936 | **<0.001** | 2690.78 (88.71) | 30.334 | **<0.001** |
| Condition$_{(noise)}$ | −0.152 (0.014) | −10.710 | **<0.001** | 0.027 (0.005) | 5.244 | **<0.001** | 466.27 (66.97) | 6.963 | **<0.001** |
| Language$_{(Hebrew)}$ | −0.018 (0.014) | −1.337 | 0.181 | 0.042 (0.009) | 4.640 | **<0.001** | 565.50 (87.89) | 6.434 | **<0.001** |
| Context$_{(high)}$ | 0.017 (0.013) | 1.281 | 0.200 | 0.005 (0.003) | 1.562 | 0.118 | 95.92 (50.65) | 1.894 | 0.058 |
| Condition$_{(noise)}$: language$_{(Hebrew)}$ | −0.303 (0.021) | −14.452 | **<0.001** | −0.015 (0.004) | −3.070 | **0.002** | −492.82 (76.10) | −6.476 | **<0.001** |
| Condition$_{(noise)}$: context$_{(high)}$ | 0.059 (0.019) | 3.030 | **0.002** | −0.003 (0.004) | −0.784 | 0.433 | −159.89 (72.31) | −2.211 | **0.027** |
| Language$_{(Hebrew)}$: context$_{(high)}$ | 0.001 (0.019) | 0.074 | 0.941 | 0.006 (0.004) | 1.474 | 0.141 | −11.55 (71.43) | −0.162 | 0.871 |
| Condition$_{(noise)}$: language$_{(Hebrew)}$: context$_{(high)}$ | −0.172 (0.029) | −5.816 | **<0.001** | −0.001 (0.006) | −0.149 | 0.881 | 237.62 (106.33) | 2.235 | **0.025** |
| Control variable | | | | | | | | | |
| Trial order | 0.016 (0.003) | 4.355 | **<0.001** | −0.017 (0.001) | −16.709 | **<0.001** | −57.86 (18.63) | −3.105 | **0.001** |
| Perception* | — | — | — | −0.002 (0.001) | −2.433 | **0.014** | −86.70 (15.43) | −5.617 | **<0.001** |
| | Var. (SD) | Corr. | | Var. (SD) | Corr. | | Var. (SD) | Corr. | |
| Random effects | | | | | | | | | |
| Item$_{(intercept)}$ | 0.001 (0.027) | | | 0.001 (0.009) | | | 18,131 (134.6) | | |
| Participant$_{(intercept)}$ | 0.001 (0.041) | | | 0.008 (0.089) | | | 291,011 (539.5) | | |
| Participant: language$_{(Hebrew)}$ | — | | | 0.003 (0.056) | 0.09 | | 222,367 (471.6) | −0.16 | |
| Participant: condition$_{(noise)}$ | — | | | 0.001 (0.027) | 0.14 | 0.05 | 77,385 (278.2) | −0.25 | 0.10 |

*Fixed effects reflect simple effects relative to the reference level when other variables are at their reference level without correction for multiple comparisons. For main effects, see $\chi^2$, F, and p values in the text. For the mean pupil model refer to Supplemental Digital Content 3, http://links.lww.com/EANDH/B525. Bold values indicate effects that are statistically significant, with p values less than 0.05.*

**Perceptual accuracy was included as a covariate in the pupillometry models.*

significance (see Supplemental Digital Content 5, http://links.lww.com/EANDH/B527).

**Summary of Single-Word Results** • Listening in quiet conditions was easier compared with noisy conditions, supported by perceptual and pupillometry findings. While perceptual measures showed no differences between L1 and L2 single words, differences between languages emerged in peak amplitude, indicating increased effort in L2 compared with L1 in both quiet and noise. Moreover, differences in peak amplitudes between L1 and L2 became more pronounced when listening to words presented in noise compared with words presented in quiet.

### Sentences—Perception and Pupillometry

**Perception** • Significant main effects were observed for listening condition [$\chi^2$ (1) = 1549.209, $p < 0.001$] and language [$\chi^2$ (1) = 486.596, $p < 0.001$] in the perception of sentences. Furthermore, a main effect of context was observed [$\chi^2$ (1) = 3.767, $p = 0.05$]. These effects were qualified by significant two-way interactions

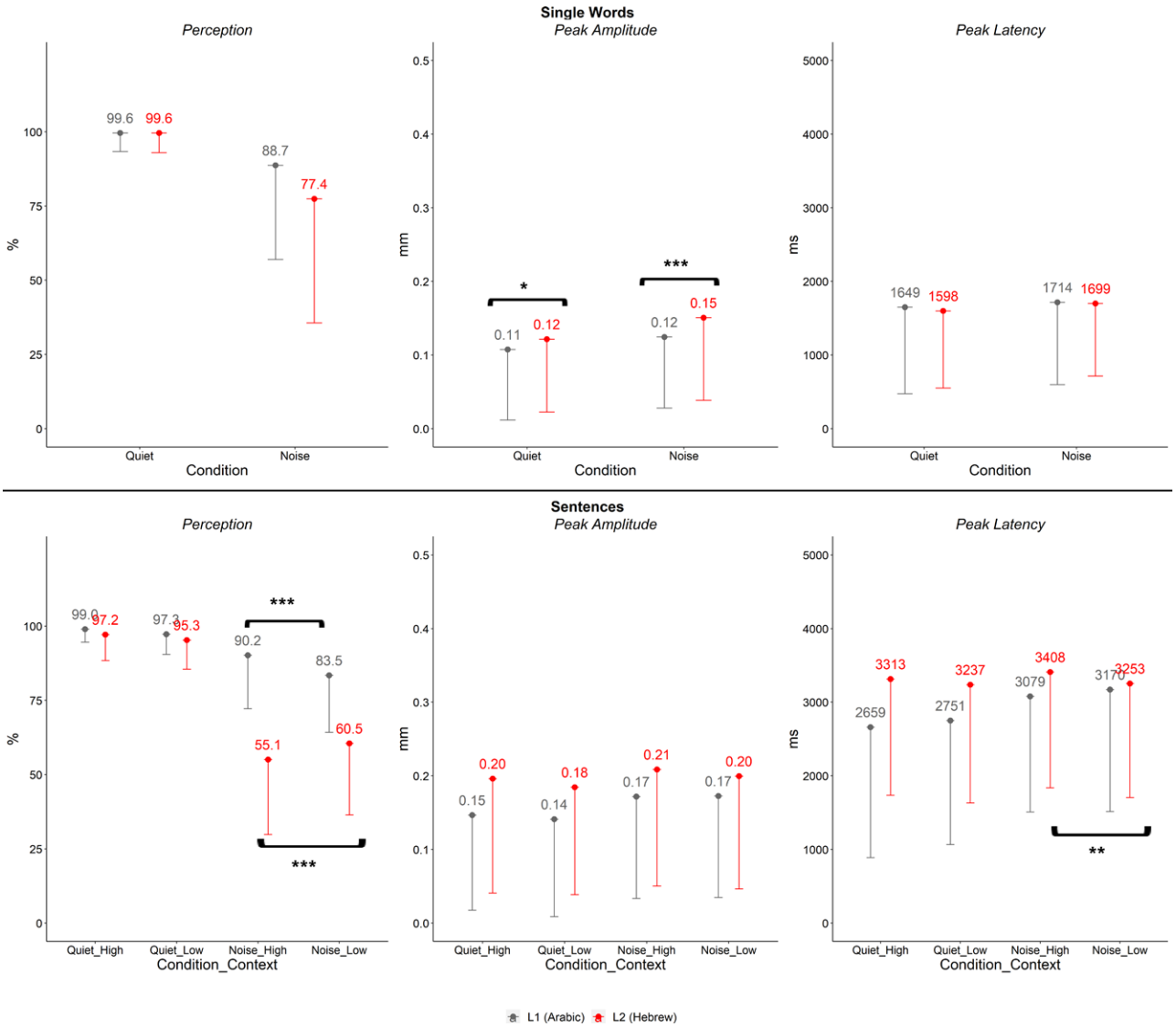


Fig. 2. Descriptive statistics for perceptual accuracy, peak amplitude, and peak latency for single words (upper panels) and sentences (lower panels) in both quiet and noise conditions. Error bars represent SD. Asterisks denote significant higher-order interactions; for simple main effects or interactions, refer to Supplemental Digital Content 5, http://links.lww.com/EANDH/B527. Significant differences are indicated by ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

between listening condition and language [$\chi^2$ (1) = 677.096, $p < 0.001$), context and language [$\chi^2$ (1) = 22.852, $p < 0.001$) as well as a three-way interaction involving language, listening condition, and context [$\chi^2$ (1) = 33.633, $p < 0.001$], see Supplemental Digital Content 5, http://links.lww.com/EANDH/B527. Pairwise comparisons with Bonferroni corrections revealed no significant differences between high and low-predictability sentences in the quiet condition in L1 (value = 0.982, $\chi^2 = 1.642$, $p = 0.799$) or in L2 (value = 0.981, $\chi^2 = 1.904$, $p = 0.670$). In noise, however, significant differences between sentence types emerged, revealing distinct patterns in L1 compared with L2. As shown in Figure 2, in noise, multilinguals in L1 had significantly better accuracy in high-predictability sentences compared with low-predictability sentences (value = 0.925, $\chi^2 = 28.865$, $p < 0.001$), but in L2, accuracy in high-predictability sentences was significantly worse than in low-predictability sentences (value = 1.098, $\chi^2 = 28.177$, $p < 0.001$).

**Pupillometry (Peak Amplitude) •** Significant main effects were observed for listening condition [$F(1,54.9) = 15.422$, $p < 0.001$], language [$F(1,47.9) = 18.888$, $p < 0.001$] and context [$F(1,12965.2) = 13.999$, $p < 0.001$] on peak amplitude, see Supplemental Digital Content 5, http://links.lww.com/EANDH/B527. These effects were qualified by a significant two-way interaction between listening condition and language [$F(1,13075.9) = 17.517$, $p < 0.001$]. Pairwise comparisons with Bonferroni corrections showed that in quiet, significantly larger peak amplitudes (value = −0.045, $\chi^2 = 26.960$, $p < 0.001$) were observed in L2 compared with L1, but these language-based differences became smaller in the noise condition (value = −0.029, $\chi^2 = 11.042$, $p = 0.001$). The context-by-language interaction [$F(1,12972.8) = 3.351$, $p = 0.06$] and the three-way interaction between language, listening condition, and context did not reach significance [$F(1,12954.8) = 0.022$, $p = .881$].

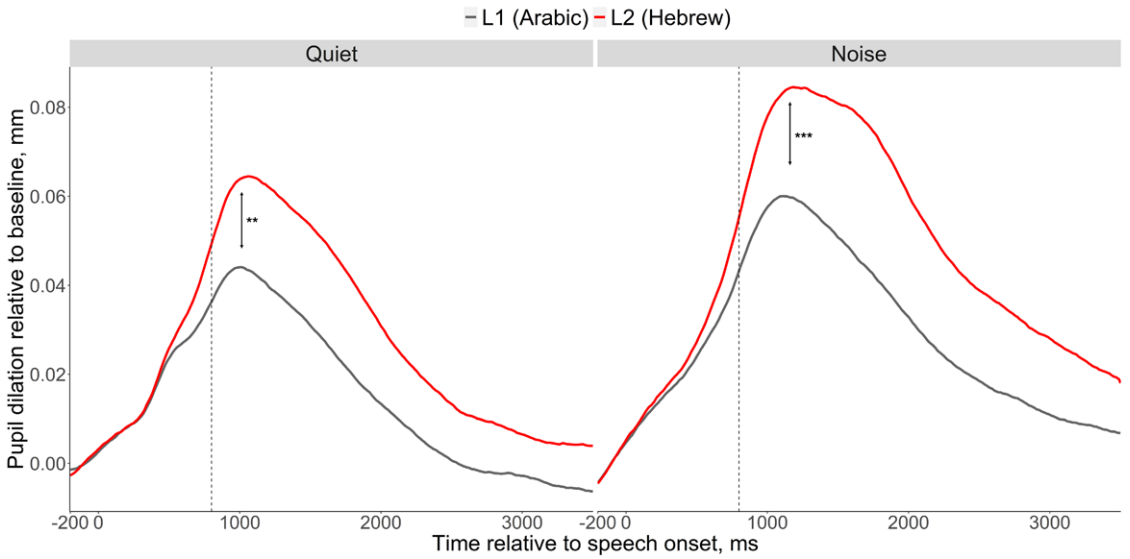


Fig. 3. Mean pupil response over time (in msec) in quiet and noise for L1 (Arabic) and L2 (Hebrew) single words. On the *x* axis −200 represents the baseline period and 0 denotes word onset. The vertical dashed line represents word offset. Asterisks denote significant higher-order interactions; for simple main effects or interactions, refer to Supplemental Digital Content 5, http://links.lww.com/EANDH/B527. ***$p < 0.001$; **< 0.001.



**Pupillometry (Peak Latency)** • Significant main effects were observed for listening condition [$F(1,63.7) = 13.925$, $p < 0.001$], language [$F(1,52.9) = 23.088$, $p < 0.001$] and context [$F(1,12971.9) = 6.858$, $p = 0.008$] on peak latency. These effects were qualified by significant two-way interaction between listening condition and language [$F(1,13066.5) = 42.106$, $p < 0.001$], context and language [$F(1,12981.5) = 4.072$, $p = 0.04$] as well as a three-way interaction involving language, listening condition, and context [$F(1,12964.6) = 4.993$, $p = 0.025$] see Supplemental Digital Content 5, http://links.lww.com/EANDH/B527. Pairwise comparisons with Bonferroni corrections revealed that in L1, peak latency did not differ between high and low-predictability sentences in quiet (value = −95.915, $\chi^2 = 3.586$, $p = 0.232$) or in noisy conditions (value = 63.979, $\chi^2 = 1.521$, $p = 0.869$). In contrast, in L2, peak latency was affected by predictability in noisy conditions, suggesting increased effort in noisy high-predictability sentences compared with low-predictability sentences. Specifically, while in quiet, the differences in peak latencies between high-predictability sentences and low-predictability sentences were not significant (value = −84.362, $\chi^2 = 2.795$, $p = 0.378$), in noise, the peak latency of high-predictability sentences was significantly more delayed than that of low-predictability sentences (value = −162.088, $\chi^2 = 7.560$, $p = 0.023$), see Figures 2 and 4.

**Summary of Sentence Results** • In line with the findings of single words, listening to sentences in quiet was easier compared with the noise condition. Larger pupil dilation and delayed peak responses were observed during the processing of L2 sentences (even after accounting for differences in perceptual accuracy). Context had no influence on accuracy in quiet for either L1 or L2, possibly due to a ceiling effect. However, in noisy conditions, increased contextual support facilitated accuracy in the L1 but resulted in decreased accuracy in the L2. Further, contextual modulation did not influence pupil measures in L1. In L2, however, more effort was exerted on high-predictability sentences compared with low-predictability sentences.

## DISCUSSION

The present study aimed to examine the listening effort exerted by multilingual individuals in quiet and noisy listening conditions in their L1 and L2. Using single words, our objective was to examine differences in listening effort without the influence of contextual cues available to the listeners in sentences. Conversely, by utilizing high and low-predictability sentences, we aimed to examine how the presence of a constraining context influences multilinguals' listening effort, and whether these effects manifest differently in L1 and L2. To answer these questions, the present study tested a group of Arabic (L1)–Hebrew (L2) young adult multilinguals on their perceptual performance, as well as on the listening effort they exerted in each of the two languages, using pupillometry.

Our findings, summarized in Supplemental Digital Content 5, http://links.lww.com/EANDH/B527, and further elaborated upon in the following sections, showed that, as expected, listening in quiet was easier than listening in noise. Specifically, perceptual accuracy was significantly better in quiet than in noise, with smaller and earlier pupil responses, indicative of reduced listening effort. Multilinguals showed differences in perceptual performance between their L1 and L2, particularly notable in noisy conditions, and under heightened stimulus demands. Specifically, while perceptual accuracy reached ceiling levels in L1 and L2 in quiet conditions and was comparable in L1 and L2 in single words presented in noise, differences emerged when encountering sentences in noise, with lower accuracy for L2 sentences compared with L1 sentences. Pupil measures indicated increased listening effort for L2 stimuli compared with L1 stimuli. This difference was evident even in the quiet condition, where perceptual accuracy was similar and at the ceiling for both languages.

In addition, in noisy conditions, increased contextual support enhanced accuracy in the L1 but resulted in decreased accuracy in the L2. Contrary to our original hypothesis, contextual modulation did not influence pupil measures in L1, while

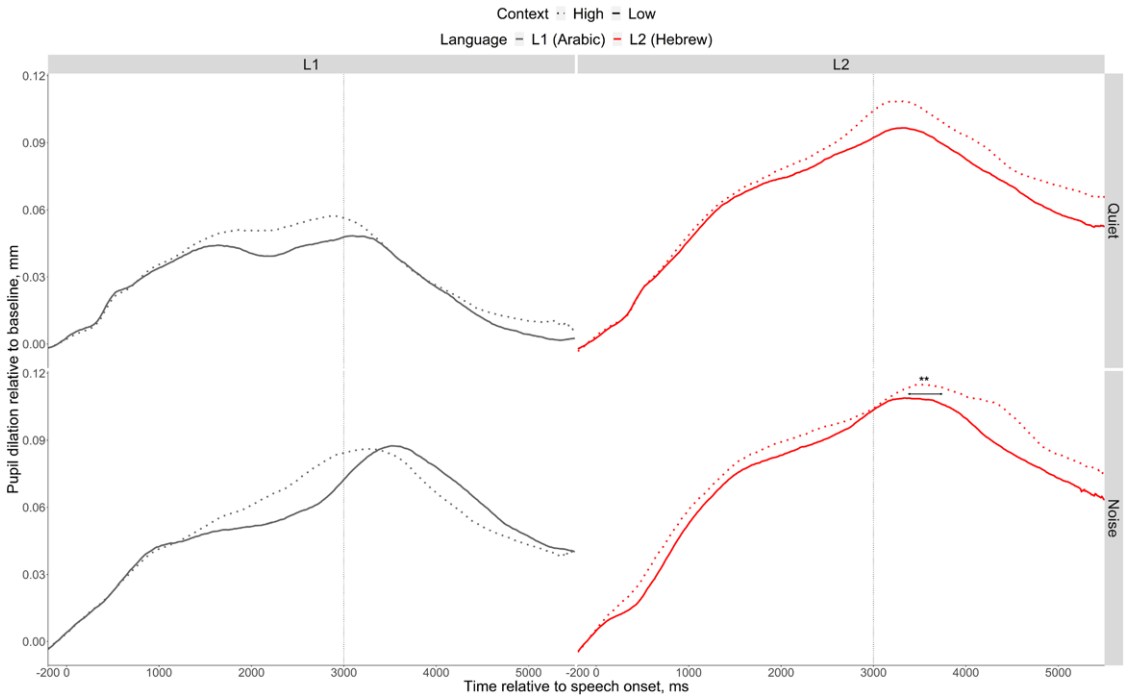


Fig. 4. Mean pupil response over time (in msec) for L1 (Arabic) and L2 (Hebrew) high and low-predictability sentences in quiet and noisy conditions. On the *x* axis −200 represents the baseline period, and 0 denotes sentence onset. The vertical dashed line represents sentence offset. Asterisks denote significant higher-order interactions; for simple main effects or interactions, refer to Supplemental Digital Content 5, http://links.lww.com/EANDH/B527. $^{**}p < 0.01$.

in L2, more effort was exerted in listening to high-predictability sentences compared with low-predictability sentences.

### Effects of Noise and Language on Speech Perception and Effort

Our findings showed a notable decline in speech perception when individuals are exposed to background noise, accompanied by an increased effort during the listening process. These findings are consistent with prior research (Shimizu et al. 2002; Klatte et al. 2010; Koelewijn et al. 2012; Mattys et al. 2012; Zekveld & Kramer 2014; Borghini & Hazan 2018; Bsharat-Maalouf & Karawani 2022b; Neagu et al. 2023) and align with the ELU model (Rönnberg et al. 2008, 2013, 2019). In addition, multilinguals exhibited differences in processing L1 and L2 sentences, with lower accuracy in L2 compared with L1, particularly under adverse listening conditions. This pattern is consistent with the results of prior work (Rogers et al. 2006; Rosenhouse et al. 2006; Weiss & Dempsey 2008; Shi & Sanchez 2010; Tabri et al. 2015; Desjardins et al. 2019; Bsharat-Maalouf & Karawani 2022a, b). Of note, the interaction between language and condition did not reach significance in the single-word perceptual analysis, suggesting a similar detrimental effect of noise on L1 and L2. While a direct comparison between single words and sentences was not feasible in the present study, these findings imply that when dealing with simpler stimuli, such as single words in the present study, the differences between L1 and L2 perception may be less pronounced or even absent. However, when the complexity of the speech stimulus increases (as seen in sentences), multilinguals in their L2 may encounter more perceptual challenges within noisy conditions (see also Krizman et al. 2017; Bsharat-Maalouf & Karawani 2022b).

The present pupillometry results further suggest increased effort in L2, evident in larger peak amplitudes as well as longer peak latencies compared with L1. This aligns with existing research comparing bilinguals to monolinguals (Schmidtke 2014; Borghini & Hazan 2018, 2020; Lam et al. 2018; Desjardins et al. 2019; Peng & Wang 2019; Visentin et al. 2019; Brännström et al. 2021), demonstrating increased listening effort in bilinguals' L2. However, one contribution of our study lies in the utilization of the within-participant design, wherein multilingual individuals served as their own control, reducing confounding variables often seen in between-participant comparisons (Borghini & Hazan 2018; Bsharat-Maalouf et al. 2023). This approach is particularly important for pupillometry, given its sensitivity to interindividual differences (Zekveld et al. 2011, 2018; Winn et al. 2018).

Nonetheless, the within-participant design necessitated the use of different stimuli across languages, which could have contributed to the observed differences between L1 and L2. To mitigate this concern, in addition to carefully matching Arabic and Hebrew stimuli, we supplemented our findings with data from a control group of native Hebrew young adults, who completed the listening task in Hebrew (their L1). As shown in Supplemental Digital Content 6, http://links.lww.com/EANDH/B528, comparative analyses showed significant differences between the multilingual participants (tested in Hebrew, their L2) and the control group (tested in Hebrew, their L1), with lower perceptual accuracy (in noise) and higher listening effort for multilinguals. These findings show that performance was not driven by language-specific characteristics, but instead follow L1 versus L2 and resemble previous studies utilizing a between-participant comparison examining the same language across participants (Schmidtke 2014; Borghini & Hazan 2018, 2020; Lam et al. 2018; Desjardins et al. 2019; Peng & Wang

2019; Visentin et al. 2019; Brännström et al. 2021). Moreover, findings from Bsharat-Maalouf and Karawani's (2022b) study on multilinguals with backgrounds akin to our sample suggested that differences in performance between Arabic and Hebrew among Arabic (L1)–Hebrew (L2) individuals are unlikely attributable to stimuli variations. Thus, given the findings of the control group of native Hebrew speakers and after considering the findings of Bsharat-Maalouf and Karawani, it is unlikely that the L1-L2 differences observed here can be attributed to variations in the language of the stimuli (Arabic versus Hebrew).

Our study extended the research conducted by Francis et al. (2018), the only previous study to utilize pupillometry and within-participant comparisons when testing multilinguals' listening effort. While Francis et al. focused solely on sentences in noisy conditions, our study exemplified increased effort in L2 with single words when the involvement of contextual support is unlikely. Moreover, by including both quiet and noisy conditions, our findings showed greater L2 effort even when multilinguals listened to stimuli in the absence of external noise. Thus, even in the absence of adverse listening conditions, L2 listening appeared to necessitate a higher cognitive effort compared with L1. As suggested in the introduction, the increased listening effort in L2 may be related to factors that hold the potential to impact the matching process in the ELU model, either independently or interactively. One such factor could be the quality of stored representations. As presented in Table 1, multilinguals exhibited lower language experience in L2 compared with L1 across various metrics, including the age of language acquisition, proficiency, exposure time, and usage patterns, all of which could contribute to less precise representations in L2 compared with L1. This disparity in language experience may hamper the ease of matching the incoming signal with stored representations (Rönnberg et al. 2008, 2013, 2019), increasing listening effort in L2. In addition, the heightened competition due to language co-activation, particularly evident in L2 (Blumenfeld & Marian 2013; e.g., Marian & Spivey 2003; Weber & Cutler 2004; Shook & Marian 2012, 2013; Chen et al. 2017) can introduce difficulties in accessing a specific stored representation, requiring the engagement of explicit processes to complete lexical access (see also, Zhang & Samuel 2018). Thus, the findings from our single-word data suggest that either the quality of stored representations or the heightened competition due to language co-activation, or possibly both, could be contributing factors to the increased effort observed in L2. To examine the role of the third proposed factor, namely multilinguals' ability to rely on top-down contextual support, we turn to examine our findings from the sentence data.

### Effect of Context on Speech Perception and Effort

Our perceptual findings indicated that the effect of contextual cues was not consistent across listening conditions. In particular, in quiet, there were no differences in accuracy between high and low-predictability sentences in either the L1 or the L2. This could possibly be due to a ceiling effect, as performance in quiet was rather high. However, under noisy conditions, more contextual cues within the sentences appeared to enhance accuracy in L1 but to diminish accuracy in L2. The difficulty faced by multilingual individuals in utilizing contextual cues in their L2 under adverse listening conditions aligns with previous studies (Mayo et al. 1997; Bradlow & Alexander 2007; Shi 2010; Schmidtke 2016; Kousaie et al. 2019; Skoe & Karayanidi 2019; Bsharat-Maalouf & Karawani 2022b). Notably, earlier studies outlined this difficulty by either revealing no differences across context levels (Bsharat-Maalouf & Karawani 2022b) or highlighting a persistence of the predictability context benefit in the L2 of multilinguals, albeit to a lesser extent than observed in monolinguals (Schmidtke 2016; Skoe & Karayanidi 2019). Here, in contrast, we observed interference caused by the presence of context, resulting in lower accuracy in high-predictability sentences compared with low-predictability sentences in the L2. This discrepancy might be explained by nuanced differences in experimental conditions and task demands. For example, unlike Bsharat-Maalouf and Karawani (2022b), we used speech-shaped noise instead of babble noise, potentially influencing masking effects and perceptual performance. In addition, diverging from previous research that primarily emphasized the repetition of the final word in a sentence (Schmidtke 2016; Skoe & Karayanidi 2019), our study required participants to repeat the entire sentence. This deliberate choice aimed to prevent participants from solely concentrating on the last word, which would have compromised the utility of the pupil data collected. This higher load on memory may have overshadowed the anticipated advantages of contextual cues in the L2. These methodological nuances necessitate further investigation to understand the factors influencing multilingual perceptual performance under diverse conditions.

The present study goes beyond testing the effect of context on perceptual performance to examine the involvement of listening effort when processing stimuli with high contextual support. The presence of contextual cues within sentences may offer two alternative possibilities regarding listening effort. The first possibility would be that listening to high-predictability sentences would demand less effort, as the presence of context facilitates easier access to sentence meaning when compared with low-predictability sentences (Winn 2016). According to this hypothesis, the presence of contextual cues should limit the number of candidate words, thus reducing the lexical search space (Rovetti et al. 2022), and the likelihood of mismatches (Rönnberg et al. 2019). Conversely, a second possibility is that listening to high-predictability sentences may be associated with increased listening effort compared with low-predictability sentences, as the reliance on contextual strategies necessitates greater reliance on higher-level processing in the ELU model.

Our pupil findings suggest that sentential context affected multilinguals' listening effort in the L2, but not in their L1. In particular, in L1, no significant differences were observed in pupil measures between high and low-predictability sentences, deviating from the expectations of the two possibilities mentioned earlier. Thus, the hypothesis by which listeners may capitalize on contextual cues to reduce the lexical search space (Rovetti et al. 2022) or the likelihood of mismatches (Rönnberg et al. 2019) may not hold true for multilingual individuals processing their native language, and may be limited to monolingual speakers only. Considering the presence of language co-activation, multilingual individuals might experience heightened competition even in their L1, relative to monolingual speakers. Thus, whereas in monolinguals, the presence of contextual cues may succeed in limiting the number of candidate words, reducing the lexical search space and mismatches, in the case of multilinguals, even within their L1, the efficiency of contextual cues

may be reduced (for no context effects in multilinguals' L1 during visual processing see Norman & Degani 2024). This idea aligns with the suggestion by Van Assche et al. (2016), who, in their review, highlighted studies indicating that the semantic constraint of a sentence does not necessarily restrict multilingual language co-activation.

In addition, it is plausible that both hypothesized processes regarding context and listening effort are occurring simultaneously, thereby offsetting each other and resulting in no change in effort in L1. On the one hand, contextual cues in L1 might help reduce the lexical search space, simplifying the retrieval of sentence meaning. On the other hand, the reliance on these contextual strategies may necessitate greater higher-level processing. This increased cognitive demand could counteract the benefits of reduced lexical search, leading to no significant differences in listening effort between high and low-predictability sentences in multilinguals' L1.

It is interesting that the pupil data in the present study did indicate contextual influences on listening effort in the L2. In particular, increased effort was observed in high-predictability sentences compared with low-predictability sentences, supporting the idea that relying on context in high-predictability sentences engages explicit cognitive processes, requiring additional effort. At the same time, the lower effort observed in low-predictability sentences may indicate participants' disengagement from the task upon recognizing the lower predictability of those sentences. The mental states of listeners, including motivation and engagement, have been suggested to significantly influence listening effort (Zekveld & Kramer 2014; Koelewijn et al. 2015; Ohlenforst et al. 2017, 2018; Ayasse et al. 2021; Micula et al. 2021, 2022; Relaño-Iborra et al. 2022), as highlighted by the Model of Listening Engagement (Herrmann & Johnsrude 2020) and the Framework for Understanding Effortful Listening (Pichora-Fuller et al. 2016). These models emphasize that when individuals perceive listening goals as unattainable, their engagement and motivation may decrease, influencing the allocation of cognitive resources (even when such resources are available). In the present study, we attempted to maintain consistent motivation and engagement in high and low-predictability sentences by randomizing their presentation within each block, following the limitation highlighted in Borghini and Hazan (2020). Nonetheless, it should be noted that high and low-predictability sentences used in this study differed in the first part of the sentence (the words leading up to the target word, see Supplemental Digital Content 1, http://links.lww.com/EANDH/B523). Consequently, it is plausible that the reduced content in the first part of low-predictability sentences, compared with high-predictability ones, might lead participants to engage less when listening to the low-predictability sentences, thereby reducing their effort. To delve deeper into this possibility, future studies should go beyond merely randomizing the presentation order of high and low-predictability sentences. They should explore participants' motivation levels while listening to these sentences and examine how this factor modulates listening effort. Furthermore, as the present study, along with Borghini and Hazan, marks the initial attempts to explore how contextual cues affect listening effort among multilinguals, further research in this area is essential to validate and expand upon our findings. Future directions should include comparative analyses of how contextual cues influence multilingual listening effort in their L1 compared with monolinguals. In addition, integrating additional measures of effort would be advantageous, allowing for a comprehensive examination of potential differences in L1 that may not have been observed using the pupillometry measure alone. This is crucial, given that different measures of effort may capture distinct dimensions of the construct (Alhanbali et al. 2019).

### Dissociations Between Perception and Listening Effort

Our findings underscore a notable dissociation between the concepts of speech perception and listening effort, affirming the study by Winn and Teece (2021). Specifically, in quiet conditions where perceptual performance was at ceiling, we observed increased listening effort when processing L2 relative to L1 speech stimuli. In noisy environments, this heightened listening effort for L2 stimuli persisted even when considering trials with accurate perception, as evident in the single-word analysis, and when accounting for perceptual differences, as demonstrated in the sentence analysis. Furthermore, when sentential contextual cues were considered, our results emphasize that listening effort does not always align with an individual's perceptual performance. Specifically, in L1, the perceptual outcomes revealed advantages in high-predictability sentences compared with low-predictability sentences in noise, but measures of listening effort across both sentence types did not differ significantly. Further, in L2, despite the heightened effort observed when listening to high-predictability sentences compared with low-predictability sentences, this increased effort did not translate into enhanced perceptual performance. Together, these findings underscore the significance of exploring listening effort beyond speech perception, as these two concepts may provide valuable insights into the challenges faced by multilinguals in speech processing, which may not always be aligned with each other.

### Implications, Limitations, and Future Directions

The multilingual participants tested in the present study were enrolled in university courses conducted in their L2 at the time of testing. While these participants may appear to uphold a good level of perceptual performance in such an environment, the current results suggest that this comes at a greater cognitive cost than that required when processing L1. This finding bears significant implications for both educational and clinical contexts, as the increased effort needed in L2 may have negative consequences, such as increased mental fatigue and diminished ability for multitasking. This holds particular relevance in today's increasingly globalized world, where a growing number of individuals live, work, and socialize in environments where their L2 is predominant. In addition, the growing presence of multilingual individuals in clinical settings (Douglas 2011; Bunta et al. 2016; Hisagi et al. 2024) highlights the importance of effectively understanding and addressing the increased effort involved in L2 listening. Audiologists, speech-language pathologists, and educational institutions can implement various approaches to alleviate this increased effort. These may include providing assessment materials and instructions in the multilingual's L1 whenever possible, offering breaks to prevent mental fatigue, and integrating visual aids in the multilingual's L1. Moreover, mitigating adverse conditions, such as reducing environmental noise or encouraging multilinguals to avoid sources of noise whenever feasible, can substantially reduce listening effort for L2 listeners. By implementing these

approaches, better support for the diverse linguistic needs of multilingual populations can be provided, leading to more accurate and equitable outcomes in clinical and educational contexts.

Furthermore, the significance of accounting for listening effort becomes particularly evident when reflecting on the design of our study. In the present study, we introduced speech-shaped noise as the adverse listening condition, and we chose sentences as our intricate speech stimuli for examination. These choices were made to target perceptual performance that is challenging but manageable for multilinguals. However, it is crucial to acknowledge that real-world scenarios involve more complex forms of adverse conditions, like babble noise, accented speech, and reverberation, in conjunction with more complex speech stimuli, such as narratives, lectures, and conversations. Thus, it is reasonable to assume that in real-life situations, multilingual individuals in their L2 may need to exert substantially greater effort than what our present study suggests, underscoring the need for future research examining a broader spectrum of real-life communicative contexts involving greater acoustic and linguistic complexities.

Some limitations of the present study should be considered. While the current findings underscore heightened listening effort in L2 compared with L1, the potential influence of nuances within the linguistic background and experience of multilingual individuals in their L2 cannot be dismissed. These sources of variability may modulate listening effort, suggesting a compelling avenue for further exploration. Also, native Hebrew speakers—who speak fewer languages than our participants—demonstrated higher accuracy in their L1 compared with our multilingual participants tested in their L1 (see perceptual accuracy in Supplemental Digital Content 6, http://links.lww.com/EANDH/B528, compared with Fig. 2). This finding aligns with Bsharat-Maalouf and Karawani (2022b), which suggests that as multilinguals speak more languages, their perceptual performance in L1 may decline due to reduced exposure to each language or greater competition between known languages. However, this relationship in terms of listening effort remains unexplored. Therefore, further investigation into this aspect offers a promising direction for future research.

## CONCLUSIONS

Using pupillometry, this study revealed increased listening effort in multilingual individuals when processing words and sentences in their L2 compared with L1, in both quiet and noisy conditions. Notably, contextual cues within sentences, particularly in multilingual L2, had an additional impact, with high-predictability sentences resulting in increased effort compared with low-predictability sentences. However, despite this increased effort, perceptual performance did not show improvement, as indicated by lower accuracy in high-predictability sentences compared with low-predictability sentences in the noisy condition. These findings highlight the critical role of assessing listening effort in uncovering challenges in multilingual speech processing.

## ACKNOWLEDGMENTS

The authors thank the study participants, and Lamees Natour for her help with data collection and coding.

This study was supported by the Israel Science Foundation (ISF; grant number 2031/22 awarded to H.K.). A student fellowship was awarded from the University of Haifa to D.B.-M. The funding organizations had no role in the design and conduct of the study; in the collection, analysis, and interpretation of the data; or in the decision to submit the article for publication; or in the preparation, review, or approval of the article.

D.B.-M., T.D., and H.K. designed the study. D.B.-M. performed experiments and collected the data. D.B.-M. and J.S. analyzed the data. All authors interpreted the results. D.B.-M. wrote the original article. J.S., T.D., and H.K. revised the paper. All authors discussed the final version of the manuscript and commented on it at all stages. T.D. and H.K. provided resources, funding, editing, and supervision.

Parts of the study were presented at the 46th Association Research of Otolaryngology (ARO) midwinter meeting in February 2023, the 48th Boston University Conference on Language Development (BUCLD) in November 2023, and at the Psychonomic Society's 64th Annual Meeting in November 2023.

Ethical IRB approval was obtained from the ethical committee of the University of Haifa (No. 243/18).

All data required to reach the stated conclusions are presented in the paper. Additional data related to this paper may be found in Supplementary Digital Content, and in the OSF link: https://osf.io/zkvu3/?view_only=17eb1bf44cf54dbcbd3d568c1697f693.

The authors have no conflicts of interest to disclose.

Address for correspondence: Hanin Karawani, Department of Communication Sciences and Disorders, University of Haifa, Eshkol Tower, Office 815, 199 Aba Khoushy Avenue, Haifa 3498838, Israel. E-mail: hkarawani@staff.haifa.ac.il

Received January 16, 2024; accepted September 14, 2024

## REFERENCES

Abbas, N., Degani, T., Elias, M., Prior, A., Silawi, R. (2024). *Multilingual language background questionnaire in Hebrew*. https://doi.org/10.31234/osf.io/jfk8b.

Abbas, N., Degani, T., Prior, A. (2021). Equal opportunity interference: Both L1 and L2 influence L3 morpho-syntactic processing. *Front Psychol*, *12*, 673535.

Alhanbali, S., Dawes, P., Lloyd, S., Munro, K. J. (2017). Self-reported listening-related effort and fatigue in hearing-impaired adults. *Ear Hear*, *38*, e39–e48.

Alhanbali, S., Dawes, P., Millman, R. E., Munro, K. J. (2019). Measures of listening effort are multidimensional. *Ear Hear*, *40*, 1084–1097.

Ayasse, N. D., Hodson, A. J., Wingfield, A. (2021). The principle of least effort and comprehension of spoken sentences by younger and older adults. *Front Psychol*, *12*, 629464.

Baese-Berk, M. M., Levi, S. V., Van Engen, K. J. (2023). Intelligibility as a measure of speech perception: Current approaches, challenges, and recommendations. *J Acoust Soc Am*, *153*, 68–76.

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J Stat Softw*, *67*, 1–48.

Blumenfeld, H. K., & Marian, V. (2013). Parallel language activation and cognitive control during spoken word recognition in bilinguals. *J Cogn Psychol*, *25*, 547–567.

Bobb, S. C., Von Holzen, K., Mayor, J., Mani, N., Carreiras, M. (2020). Co-activation of the L2 during L1 auditory processing: An ERP cross-modal priming study. *Brain Lang*, *203*, 104739.

Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer (version 5.1.13).

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Front Neurosci*, *12*, 152.

Borghini, G., & Hazan, V. (2020). Effects of acoustic and semantic cues on listening effort during native and non-native speech perception. *J Acoust Soc Am*, *147*, 3783–3794.

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *J Acoust Soc Am*, *121*, 2339–2349.

Brännström, K. J., Rudner, M., Carlie, J., Sahlén, B., Gulz, A., Andersson, K., Johansson, R. (2021). Listening effort and fatigue in native and non-native primary school children. *J Exp Child Psychol*, *210*, 105203.

Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System*, *36*, 22–34.

Brown, V. A. (2021). An introduction to linear mixed-effects modeling in R. *Adv Methods Pract Psychol Sci*, *4*, 2515245920960351.

Brown, V. A., McLaughlin, D. J., Strand, J. F., Van Engen, K. J. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Q J Exp Psychol (Hove)*, *73*, 1431–1443.

Brustad, K., & Zuniga, E. (2019). Levantine Arabic. In J. Huehnergard & N. Pat-El (Eds.), *The Semitic Languages* (pp. 403–432). Routledge.

Bsharat-Maalouf, D., Degani, T., Karawani, H. (2023). The involvement of listening effort in explaining bilingual listening under adverse listening conditions. *Trends Hear*, *27*, 23312165231205107.

Bsharat-Maalouf, D., & Karawani, H. (2022a). Bilinguals' speech perception in noise: Perceptual and neural associations. *PLoS One*, *17*, e0264282.

Bsharat-Maalouf, D., & Karawani, H. (2022b). Learning and bilingualism in challenging listening conditions: How challenging can it be? *Cognition*, *222*, 105018.

Bunta, F., Douglas, M., Dickson, H., Cantu, A., Wickesberg, J., Gifford, R. H. (2016). Dual language versus English-only support for bilingual children with hearing loss who use cochlear implants and hearing aids. *Int J Lang Commun Disord*, *51*, 460–472.

Chen, P., Bobb, S. C., Hoshino, N., Marian, V. (2017). Neural signatures of language co-activation and control in bilingual spoken word comprehension. *Brain Res*, *1665*, 50–64.

Corps, R. E., & Rabagliati, H. (2020). How top-down processing enhances comprehension of noise-vocoded speech: Predictions about meaning are more important than predictions about form. *J Mem Lang*, *113*, 104114.

Cowan, T., Paroby, C., Leibold, L. J., Buss, E., Rodriguez, B., Calandruccio, L. (2022). Masked-speech recognition for linguistically diverse populations: A focused review and suggestions for the future. *J Speech Lang Hear Res*, *65*, 3195–3216.

De Houwer, A. (2023). The danger of bilingual–monolingual comparisons in applied psycholinguistic research. *Appl Psycholinguist*, *44*, 343–357.

De Rosario-Martinez, H., Fox, J., Team, R. C., De Rosario-Martinez, M. H. (2015). Package "phia." *CRAN Repos*. Retrieved 1, 2015. https://CRAN.R-project.org/package=phia.

Degani, T., Prior, A., Hajajra, W. (2018). Cross-language semantic influences in different script bilinguals. *Biling Lang Cogn*, *21*, 782–804.

Desjardins, J. L., Barraza, E. G., Orozco, J. A. (2019). Age-related changes in speech recognition performance in Spanish–English bilinguals' first and second languages. *J Speech Lang Hear Res*, *62*, 2553–2563.

Desjardins, J. L., & Doherty, K. A. (2014). The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. *Ear Hear*, *35*, 600–610.

Douglas, M. (2011). Spoken language assessment considerations for children with hearing impairment when the home language is not English. *Perspect Hear Hear Disord Child*, *21*, 4–19.

Francis, A. L., Tigchelaar, L. J., Zhang, R., Zekveld, A. A. (2018). Effects of second language proficiency and linguistic uncertainty on recognition of speech in native and nonnative competing speech. *J Speech Lang Hear Res*, *61*, 1815–1830.

Gagne, J.-P., Besser, J., Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends Hear*, *21*, 2331216516687287.

Garcia, D. L., & Gollan, T. H. (2022). The MINT Sprint: Exploring a fast administration procedure with an expanded multilingual naming test. *J Int Neuropsychol Soc*, *28*, 845–861.

Garcia Lecumberri, M. L., Cooke, M., Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Commun*, *52*, 864–886.

Giuliani, N. P., Brown, C. J., Wu, Y.-H. (2021). Comparisons of the sensitivity and reliability of multiple measures of listening effort. *Ear Hear*, *42*, 465–474.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *J Exp Psychol Learn Mem Cogn*, *22*, 1166–1183.

Gollan, T. H., Montoya, R. I., Cera, C., Sandoval, T. C. (2008). More use almost always means a smaller frequency effect: Aging, bilingualism, and the weaker links hypothesis. *J Mem Lang*, *58*, 787–814.

Gollan, T. H., Montoya, R. I., Fennema-Notestine, C., Morris, S. K. (2005). Bilingualism affects picture naming but not picture classification. *Mem Cogn*, *33*, 1220–1234.

Gollan, T. H., Montoya, R. I., Werner, G. A. (2002). Semantic and letter fluency in Spanish-English bilinguals. *Neuropsychology*, *16*, 562–576.

Gollan, T. H., Slattery, T. J., Goldenberg, D., Van Assche, E., Duyck, W., Rayner, K. (2011). Frequency drives lexical access in reading but not in speaking: The frequency-lag hypothesis. *J Exp Psychol Gen*, *140*, 186–209.

Gollan, T. H., Starr, J., Ferreira, V. S. (2015). More than use it or lose it: The number-of-speakers effect on heritage language proficiency. *Psychon Bull Rev*, *22*, 147–155.

Grosjean, F. (2008). *Studying Bilinguals*. Oxford University Press.

Grosjean, F. (2010). *Bilingual: Life and Reality*. Harvard University Press.

Herrmann, B., & Johnsrude, I. S. (2020). A Model of Listening Engagement (MoLE). *Hear Res*, *397*, 108016.

Hilbe, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.

Hisagi, M., Barragan, B., Diaz, A., White, K., Winter, M. (2024). Auditory discrimination in aging bilinguals vs. monolinguals with and without hearing loss. *Front Aging*, *4*, 1302050.

Holmes, E., Folkeard, P., Johnsrude, I. S., Scollie, S. (2018). Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment. *Int J Audiol*, *57*, 483–492.

Hornsby, B. W., Naylor, G., Bess, F. H. (2016). A taxonomy of fatigue concepts and their relation to hearing loss. *Ear Hear*, *37*(Suppl 1), 136S.

Hunter, C. R., & Humes, L. E. (2022). Predictive sentence context reduces listening effort in older adults with and without hearing loss and with high and low working memory capacity. *Ear Hear*, *43*, 1164–1177.

Hyönä, J., Tommola, J., Alaja, A.-M. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *Q J Exp Psychol A*, *48*, 598–612.

Jain, S., & Nataraja, N. P. (2019). The effect of fatigue on working memory and auditory perceptual abilities in trained musicians. *Am J Audiol*, *28*(2S), 483–494.

Johnson, J., Xu, J., Cox, R., Pendergraft, P. (2015). A comparison of two methods for measuring listening effort as part of an audiologic test battery. *Am J Audiol*, *24*, 419–431.

Kaplan Neeman, R., Roziner, I., Muchnik, C. (2022). A clinical paradigm for listening effort assessment in middle-aged listeners. *Front Psychol*, *13*, 820227.

Kavé, G. (2005). Phonemic fluency, semantic fluency, and difference scores: Normative data for adult Hebrew speakers. *J Clin Exp Neuropsychol*, *27*, 690–699.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*, 7–36.

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

Kilman, L., Zekveld, A. A., Hällgren, M., Rönnberg, J. (2015). Subjective ratings of masker disturbance during the perception of native and non-native speech. *Front Psychol*, *6*, 1065.

Klatte, M., Lachmann, T., Meis, M. (2010). Effects of noise and reverberation on speech perception and listening comprehension of children and adults in a classroom-like setting. *Noise Health*, *12*, 270.

Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., Zekveld, A. A., Kramer, S. E. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hear Res*, *323*, 81–90.

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

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

Koelewijn, T., Zekveld, A. A., Lunner, T., Kramer, S. E. (2018). The effect of reward on listening effort as reflected by the pupil dilation response. *Hear Res*, *367*, 106–112.

Kousaie, S., Baum, S., Phillips, N. A., Gracco, V., Titone, D., Chen, J.-K., Chai, X. J., Klein, D. (2019). Language learning experience and mastering the challenges of perceiving speech in noise. *Brain Lang*, *196*, 104645.

Krizman, J., Bradlow, A. R., Lam, S. S.-Y., Kraus, N. (2017). How bilinguals listen in noise: Linguistic and non-linguistic factors. *Biling Lang Cogn*, *20*, 834–843.

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *J Stat Softw*, *82*, 1–26.

Lam, A., Hodgson, M., Prodi, N., Visentin, C. (2018). Effects of classroom acoustics on speech intelligibility and response time: A comparison between native and non-native listeners. *Build Acoust*, *25*, 35–42.

Lau, M. K., Hicks, C., Kroll, T., Zupancic, S. (2019). Effect of auditory task type on physiological and subjective measures of listening effort in individuals with normal hearing. *J Speech Lang Hear Res*, *62*, 1549–1560.

Marian, V., Blumenfeld, H. K., Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *J Speech Lang Hear Res*, *50*, 940–967.

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within-and between-language competition. *Biling Lang Cogn*, *6*, 97–115.

Mathôt, S., & Vilotijević, A. (2022). Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behav Res Methods*, *55*, 3055–3077.

Mattys, S. L., Davis, M. H., Bradlow, A. R., Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Lang Cogn Process*, *27*, 953–978.

Mayo, L. H., Florentine, M., Buus, S. (1997). Age of second-language acquisition and perception of speech in noise. *J Speech Lang Hear Res*, *40*, 686–693.

McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupil response for accurately recognized accented speech. *J Acoust Soc Am*, *147*, EL151–EL156.

Meador, D., Flege, J. E., MacKay, I. R. (2000). Factors affecting the recognition of words in a second language. *Biling Lang Cogn*, *3*, 55–67.

Micula, A., Rönnberg, J., Fiedler, L., Wendt, D., Jørgensen, M. C., Larsen, D. K., Ng, E. H. N. (2021). The effects of task difficulty predictability and noise reduction on recall performance and pupil dilation responses. *Ear Hear*, *42*, 1668–1679.

Micula, A., Rönnberg, J., Książek, P., Murmu Nielsen, R., Wendt, D., Fiedler, L., Ng, E. H. N. (2022). A glimpse of memory through the eyes: Pupillary responses measured during encoding reflect the likelihood of subsequent memory recall in an auditory free recall test. *Trends Hear*, *26*, 23312165221130581.

Modiano, M. (2023). The vicissitudes of bilingualism and plurilingualism in the European Union. *J Eur Stud*, *53*, 53–69.

Mor, B., & Prior, A. (2022). Frequency and predictability effects in first and second language of different script bilinguals. *J Exp Psychol Learn Mem Cogn*, *48*, 1363.

Neagu, M.-B., Kressner, A. A., Relaño-Iborra, H., Bækgaard, P., Dau, T., Wendt, D. (2023). Investigating the reliability of pupillometry as a measure of individualized listening effort. *Trends Hear*, *27*, 23312165231153288.

Norman, T., & Degani, T. (2024). Context effects in the L2: Evidence for compensatory mechanisms. *Int J Biling*, *28*, 279–315.

Ohlenforst, B., Wendt, D., Kramer, S. E., Naylor, G., Zekveld, A. A., Lunner, T. (2018). Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. *Hear Res*, *365*, 90–99.

Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N. J., Kramer, S. E. (2017). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hear Res*, *351*, 68–79.

Oosthuizen, I., Picou, E. M., Pottas, L., Myburgh, H. C., Swanepoel, D. W. (2020). Listening effort in native and nonnative English-speaking children using low linguistic single-and dual-task paradigms. *J Speech Lang Hear Res*, *63*, 1979–1989.

Peng, Z. E., & Wang, L. M. (2019). Listening effort by native and nonnative listeners due to noise, reverberation, and talker foreign accent during English speech perception. *J Speech Lang Hear Res*, *62*, 1068–1081.

Pichora-Fuller, M. K. (2016). How social psychological factors may modulate auditory and cognitive functioning during listening. *Ear Hear*, *37*, 92S–100S.

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

Picou, E. M., & Ricketts, T. A. (2014). The effect of changing the secondary task in dual-task paradigms for measuring listening effort. *Ear Hear*, *35*, 611–622.

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

Pielage, H., Zekveld, A. A., Saunders, G. H., Versfeld, N. J., Lunner, T., Kramer, S. E. (2021). The presence of another individual influences listening effort, but not performance. *Ear Hear*, *42*, 1577–1589.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Version 4.2.2. Retrieved from https://www.R-project.org/.

Regalado, D., Kong, J., Buss, E., Calandruccio, L. (2019). Effects of language history on sentence recognition in noise or two-talker speech: Monolingual, early bilingual, and late bilingual speakers of English. *Am J Audiol*, *28*, 935–946.

Relaño-Iborra, H., Wendt, D., Neagu, M. B., Kressner, A. A., Dau, T., Bækgaard, P. (2022). Baseline pupil size encodes task-related information and modulates the task-evoked response in a speech-in-noise task. *Trends Hear*, *26*, 1–17.

Richer, F., & Beatty, J. (1985). Pupillary dilations in movement preparation and execution. *Psychophysiology*, *22*, 204–207.

Rimikis, S., Smiljanic, R., Calandruccio, L. (2013). Nonnative English speaker performance on the Basic English Lexicon (BEL) sentences. *J Speech Lang Hear Res*, *56*, 792–804.

Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., Abrams, H. B. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Appl Psycholinguist*, *27*, 465–485.

Rönnberg, J., Holmer, E., Rudner, M. (2019). Cognitive hearing science and ease of language understanding. *Int J Audiol*, *58*, 247–261.

Rönnberg, J., Holmer, E., Rudner, M. (2021). Cognitive hearing science: Three memory systems, two approaches, and the ease of language understanding model. *J Speech Lang Hear Res*, *64*, 359–370.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, O., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., Rudner, M. (2013). The Ease of Language Understanding (ELU) model: Theoretical, empirical, and clinical advances. *Front Syst Neurosci*, *7*, 31.

Rönnberg, J., Rudner, M., Foo, C., Lunner, T. (2008). Cognition counts: A working memory system for Ease of Language Understanding (ELU). *Int J Audiol*, *47*(Suppl 2), S99–105.

Rosenhouse, J., Haik, L., Kishon-Rabin, L. (2006). Speech perception in adverse listening conditions in Arabic-Hebrew bilinguals. *Int J Biling*, *10*, 119–135.

Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Dunabeitia, J. A., Gharibi, K., Hao, J., Kolb, N., Kubota, M., Kupisch, T., Laméris, T., Luque, A., van Osch, B., Pereira Soares, S. M., Prystauka, Y., Tat, D., Tomić, A., Voits, T., Wulff, S. (2023). Monolingual comparative normativity in bilingualism research is out of "control": Arguments and alternatives. *Appl Psycholinguist*, *44*, 316–329.

Rovetti, J., Goy, H., Zara, M., Russo, F. A. (2022). Reduced semantic context and signal-to-noise ratio increase listening effort as measured using functional near-infrared spectroscopy. *Ear Hear*, *43*, 836–848.

Scharenborg, O., Coumans, J. M., van Hout, R. (2018). The effect of background noise on the word activation process in nonnative spoken-word recognition. *J Exp Psychol Learn Mem Cogn*, *44*, 233.

Scharenborg, O., & van Os, M. (2019). Why listening in background noise is harder in a non-native language than in a native language: A review. *Speech Commun*, *108*, 53–64.

Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: Evidence from pupillometry. *Front Psychol*, *5*, 137.

Schmidtke, J. (2016). The bilingual disadvantage in speech understanding in noise is likely a frequency effect related to reduced language exposure. *Front Psychol*, *7*, 678.

Schmidtke, J. (2018). Pupillometry in linguistic research: An introduction and review for second language researchers. *Stud Second Lang Acquis*, *40*, 529–549.

Schwartz, A. I., & Kroll, J. F. (2006). Bilingual lexical activation in sentence context. *J Mem Lang*, *55*, 197–212.

Sebastián-Gallés, N., Echeverría, S., Bosch, L. (2005). The influence of initial exposure on lexical representation: Comparing early and simultaneous bilinguals. *J Mem Lang*, *52*, 240–255.

Shechter, A., & Share, D. L. (2021). Keeping an eye on effort: A pupillometric investigation of effort and effortlessness in visual word recognition. *Psychol Sci*, *32*, 80–95.

Shi, L.-F. (2010). Perception of acoustically degraded sentences in bilingual listeners who differ in age of English acquisition. *J Speech Lang Hear Res*, *53*, 821–835.

Shi, L.-F. (2012). Contribution of linguistic variables to bilingual listeners' perception of degraded English sentences. *J Speech Lang Hear Res*, *55*, 219–234.

Shi, L.-F. (2015). How "proficient" is proficient? Bilingual listeners' recognition of English words in noise. *Am J Audiol*, *24*, 53–65.

Shi, L.-F., & Sanchez, D. (2010). Spanish/English bilingual listeners on clinical word recognition tests: What to expect and how to predict. *J Speech Lang Hear Res*, *53*, 1096–1110.

Shimizu, T., Makishima, K., Yoshida, M., Yamagishi, H. (2002). Effect of background noise on perception of English speech for Japanese listeners. *Auris Nasus Larynx*, *29*, 121–125.

Shook, A., & Marian, V. (2012). Bimodal bilinguals co-activate both languages during spoken comprehension. *Cognition*, *124*, 314–324.

Shook, A., & Marian, V. (2013). The bilingual language interaction network for comprehension of speech. *Biling Lang Cogn*, *16*, 304–324.

Siegle, G. J., Ichikawa, N., Steinhauer, S. (2008). Blink before and after you think: Blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology*, *45*, 679–687.

Skoe, E., & Karayanidi, K. (2019). Bilingualism and speech understanding in noise: Auditory and linguistic factors. *J Am Acad Audiol*, *30*, 115–130.

Tabri, D., Chacra, K. M. S. A., Pring, T. (2015). Speech perception in noise by monolingual, bilingual and trilingual listeners. *Int J Lang Commun Disord*, *46*, 101005020050084–101005020050012.

Van Assche, E., Duyck, W., Hartsuiker, R. J. (2016). Context effects in bilingual sentence processing: Task specificity. In R. R. Heredia, J. Altarriba, A. B. Cieślicka (Eds.), *Methods in Bilingual Reading Comprehension Research* (*Vol. 1*, pp. 11–31). The Bilingual Mind and Brain Book Series.

Van Engen, K. J., & McLaughlin, D. J. (2018). Eyes and ears: Using eye tracking and pupillometry to understand challenges to speech recognition. *Hear Res*, *369*, 56–66.

Van Steenbergen, H., & Band, G. P. (2013). Pupil dilation in the Simon task as a marker of conflict processing. *Front Hum Neurosci*, *7*, 215.

Van Wijngaarden, S. J., Steeneken, H. J., Houtgast, T. (2002). Quantifying the intelligibility of speech in noise for non-native listeners. *J Acoust Soc Am*, *111*, 1906–1916.

Visentin, C., Prodi, N., Cappelletti, F., Torresin, S., Gasparella, A. (2019). Speech intelligibility and listening effort in university classrooms for native and non-native Italian listeners. *Build Acoust*, *26*, 275–291.

Visentin, C., Valzolgher, C., Pellegatti, M., Potente, P., Pavani, F., Prodi, N. (2022). A comparison of simultaneously-obtained measures of listening effort: Pupil dilation, verbal response time and self-rating. *Int J Audiol*, *61*, 561–573.

Voeten, C. C. (2019). buildmer: Stepwise elimination and term reordering for mixed-effects regression. [Computer software]. Retrieved from https://CRAN.R-project.org/package=buildmer.

Von Hapsburg, D., Champlin, C. A., Shetty, S. R. (2004). Reception thresholds for sentences in bilingual (Spanish/English) and monolingual (English) listeners. *J Am Acad Audiol*, *15*, 088–098.

Wang, Y., Naylor, G., Kramer, S. E., Zekveld, A. A., Wendt, D., Ohlenforst, B., Lunner, T. (2018). Relations between self-reported daily-life fatigue, hearing status, and pupil dilation during a speech perception in noise task. *Ear Hear*, *39*, 573–582.

Warzybok, A., Brand, T., Wagener, K. C., Kollmeier, B. (2015). How much does language proficiency by non-native listeners influence speech audiometric tests in noise? *Int J Audiol*, *54*(Suppl 2), 88–99.

Weber, A., & Cutler, A. (2004). Lexical competition in non-native spoken-word recognition. *J Mem Lang*, *50*, 1–25.

Weiss, D., & Dempsey, J. J. (2008). Performance of bilingual speakers on the English and Spanish versions of the Hearing in Noise Test (HINT). *J Am Acad Audiol*, *19*, 005–017.

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

Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hear Res*, *369*, 67–78.

Wilschut, T., & Mathôt, S. (2022). Interactions between visual working memory, attention, and color categories: A pupillometry study. *J Cogn*, *5*, 16.

Winn, M. B. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends Hear*, *20*, 2331216516669723.

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

Winn, M. B., & Teece, K. H. (2021). Listening effort is not the same as speech intelligibility score. *Trends Hear*, *25*, 1–26.

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

Wu, Y.-H., Stangl, E., Zhang, X., Perkins, J., Eilers, E. (2016). Psychometric functions of dual-task paradigms for measuring listening effort. *Ear Hear*, *37*, 660–670.

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

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

Zekveld, A. A., Koelewijn, T., Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends Hear*, *22*, 2331216518777174.

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

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

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

Zhang, X., & Samuel, A. G. (2018). Is speech recognition automatic? Lexical competition, but not initial lexical access, requires cognitive resources. *J Mem Lang*, *100*, 32–50.