Focus (on) Replication: Focus Processing in L1 and L2 English using the Fidelity, Refinement, and Exploratory Extension (FiREE) Replication Framework

Adam A. Bramlett[*] and Seth Wiener[*,†]

[*]Department of Languages, Cultures, and Applied Linguistics,

Carnegie Mellon University, 341 Posner Hall, Pittsburgh, PA 15213, USA

†Corresponding author: sethw1@cmu.edu

Declarations of Interests

The authors declare no financial support or conflicts of interest. Experimental materials, unidentifiable data, and code are openly available and shared through the Open Science Framework at: https://osf.io/wa4gv/?view_only=de113dbced6b46fab96ca8217b3c1ca6

Acknowledgments

The authors would like to thank the authors of Ge et al., 2021a for both sharing their materials and their original contribution to our understanding of L1 and L2 focus processing.

Abstract

A fundamental challenge in replication studies lies in balancing methodological fidelity with statistical conservatism while also leaving space for deeper theoretical exploration. This paper introduces a web-based replication framework designed to balance these sometimes competing interests: the Fidelity, Refinement, and Exploratory Extension (FiREE) Replication Framework. We demonstrate this framework by using entirely web-based methods to partially replicate an in-person English focus processing eye-tracking experiment (Ge et al., 2021a), which found L1 Dutch-L2 English learners showed delayed fixation patterns to focus alternatives compared to L1 English speakers. We selected this study due to the unexpected L1 Dutch effect, the study's recency and open materials, and its compatibility with individual difference extensions. We adhered closely to the fidelity of the original methodology and analysis. We recruited L1 English and L1 Dutch-L2 English participants via the internet and partially replicated previous findings by capturing efficient L1 processing and less efficient L2 processing. We then refined the analytical approach and treated time as a continuous, dynamic variable using Generalized Additive Models. Differences between L1 and L2 participants were found not to be static processing deficits, but rather dynamic changes in fixation patterns over time. We show that the L1 speaker advantage may be one of long-term consistency rather than early efficiency. In our exploratory extension, we further tested whether eye movements can be predicted using acoustic measurements of the stimuli and five individual difference measures. We found evidence of acoustics, participants' perceptual auditory sensitivity, and auditory motor reproduction abilities predicting eye fixations in line with interaction views of auditory processing. Our FiREE framework therefore ensures that replicability is not merely about reproducing a prior effect, but also about clarifying whether findings are meaningful, interpretable, and generalizable across different statistical, methodological, and theoretical contexts.

Introduction

Fidelity, Refinement, and Exploratory Extension (FiREE) Replication Framework

A researcher faces at least two challenges when attempting to replicate previously published research: staying faithful to the original design while also ensuring that the replication is analytically robust and theoretically informative. Contemporary discussions in applied linguistics emphasize that replication is not simply about reproducing identical methods, but rather about maintaining core features of the original design while also correcting limitations and expanding interpretability (Marsden et al., 2018; McManus, 2024; Porte & McManus, 2018). This paper contributes to that view by introducing the Fidelity, Refinement, and Exploratory Extension (FiREE) replication framework. FiREE integrates high-fidelity replication with methodological refinement and theory-driven exploratory analysis, providing a structured way to reproduce prior findings while also clarifying their robustness and generalizability.

Fidelity ensures the adherence to the original study design, using the same procedures and statistical models to assess whether the original findings hold under comparable conditions. Maintaining methodological consistency facilitates direct comparisons and contributes to broader efforts to establish replicability across fields.

Refinement strengthens the analytical framework by incorporating improved statistical practices to address issues such as multiple comparisons, overfitting, and model complexity. This step ensures that any observed effects were not artifacts of less conservative analytical approaches. Moreover, this step ensures that effects were not overlooked due to less robust modeling approaches. Researchers must distinguish between effects that are method-dependent and those that hold across multiple different analytical choices.

Finally, exploratory extensions allow for the systematic examination of theoretical assumptions that may not have been explicitly tested in the original study. Exploratory extensions provide a structured, often theory-motivated examination of factors, such as individual differences, which may contribute to the results. This allows researchers to test established explanatory framework while generating new hypotheses to account for the variation.

Our FiREE Replication approach encourages researchers to move beyond a binary success-or-failure replication framework (Nosek & Errington, 2020). Instead, see it as an opportunity to evaluate methodological robustness, refine statistical practices, and uncover theoretical insights that were not explicitly addressed in the original research. This perspective frames replication as an active tool for scientific progress, reinforcing that replicability is not merely about reproducing a prior effect, but also about ensuring that findings are meaningful, interpretable, and generalizable across different statistical, methodological, and theoretical contexts.

While existing classifications such as exact, close, and conceptual replication describe the degree of similarity between original and replication studies (Marsden et al., 2018; McManus, 2024; Porte & McManus, 2018), the FiREE framework is concerned with the structure of replication efforts—how researchers can combine fidelity, refinement, and exploratory extension in a single study. In this sense, FiREE is not a new replication type, but rather an orthogonal framework that encourages layered replication analysis design: a faithful reproduction of the original analysis (fidelity), followed by analytical improvements (refinement), and theoretically motivated additions (exploratory extension). That is, the FiREE framework could be applied to any replication regardless of degree of similarity. We view this as fully aligned with current thinking in applied psycholinguistics, particularly calls for transparency, theory-relevance, and statistical rigor in replication research. Moreover, as we demonstrate, the FiREE framework is a natural fit for behavioral and eye-tracking web-based studies.

In what follows, we first discuss the linguistic concept of focus and the findings of Ge et al. (2021a)—the focus processing eye-tracking study that we set out to partially replicate. We then briefly review the acoustics of focus and five areas in which individuals differ in their language processing behavior. These acoustic cues and individual differences measures serve as predictors in the exploratory extension. We make all our methods, materials, code, and data freely available on the Open Science Framework. Using our Ge et al. (2021a) replication results, we demonstrate our FiREE framework and connect our replication and extension findings to larger theories of

psycholinguistics and second language acquisition.

Focus in English *only*-sentences

Information in languages must be organized and presented in a manner that clearly and effectively conveys the speaker's intention. The study of how information is structured in language encompasses syntax, semantics, pragmatics, and prosody (see Breen et al., 2010; Lambrecht, 1994; Roberts, 2012). An important area of information structure is focus or the information that is considered most important, relevant, new or contrastive (Kiss, 1998). Focus is believed to be a linguistic universal (Comrie, 1989). How focus is implemented, however, varies across languages and can be constrained by a language's phonology and morphosyntax (Kiss, 1998; Lambrecht, 1994). That is, focus involves multiple levels of linguistic knowledge or interfaces. How speakers process focus in sentences is a rich area of psycholinguistics research (e.g., Cutler & Fodor, 1979; Filik et al., 2005; Wang et al., 2011).

English realizes focus primarily through prosodic prominence—typically, a pitch accent on the focused constituent. This is especially clear in English sentences containing the focus-sensitive particle *only*, which plays a central role in theoretical and experimental work on focus processing. In such sentences, the semantic scope of *only* is often ambiguous and must be resolved through prosodic cues. As an example, take the sentence, "Obama only vetoed the bill." The scope of the focus particle *only* can be associated with the verb 'vetoed' or the object 'the bill.' Semantic parsing, however, depends on which word(s) carries prosodic prominence. If 'the bill' carries prosodic prominence, the listener will understand that Obama vetoed nothing else but the bill. In contrast, if 'vetoed' carries prosodic prominence, the listener will understand that Obama did nothing else to the bill other than veto it.

In the present study, we "focus" on English preverbal *only*-sentences with varying positions of prosodic prominence because they offer a highly controlled paradigm for testing prosodic focus processing. Moreover, the ambiguity inherent in these sentences allows us to manipulate focus structure independently of word order or lexical content, providing a clean test case for examining interface processing—how syntax, semantics, and prosody are integrated in real time. This is

particularly valuable in second language (L2) research, where learners may struggle to integrate multiple cues across linguistic domains.

Ge et al. (2021a)—the study we set out to replicate—examined how L1 and L2 English speakers process *only*-sentences in real time. The authors used the "look-and-listen" visual world paradigm in which a participant looks at images on screen while listening to spoken sentences. Importantly, the images on the screen represented the intended target of the focus or the alternative focus (i.e., a competitor). For example, each experimental sentence stimulus contained *only* with prosodic prominence on either the verb or the object, creating two conditions as in "The dinosaur is only CARRYING the bucket, not throwing the bucket" (verb condition; capital letters denote prominence) or "The dinosaur is only carrying the BUCKET, not carrying the suitcase" (object condition).

Ge et al. (2021a) tested L1 English speakers and L2 English learners whose L1 was either Cantonese or Dutch. Dutch, like English, uses prosodic prominence to realize focus through an expanded F0 range, increased amplitude, and longer durations (Dimitrova et al., 2010). Dutch *only*-sentences (or *alleen*-sentences, the Dutch equivalent) can pattern like English *only*-sentences as in "De dinosaurus draagt alleen De EMMER" (The dinosaur is only carrying the BUCKET). Importantly, Dutch *alleen*-sentences can also place *only* after the object as in "De dinosaurus DRAAGT De emmer alleen" (The dinosaur is only CARRYING the bucket). In contrast, Cantonese *only*-sentences are considerably different from those in English (and Dutch). Cantonese has a number of different focus particles, which makes prosody somewhat optional for realizing focus (Fung, 2000; Ge et al., 2024; Lee, 2019; Wu & Xu, 2010).

The authors predicted that the presence of *only* would prompt participants to search for the picture that depicts a focus alternative. That is, participants' looks to the on-screen images would diverge once focus prosodic information was integrated with semantic and syntactic information. For example, upon hearing object-focused trials (e.g., "...only carrying the BUCKET, not carrying the suitcase"), participants would first look to the object (i.e., the bucket) and then look to the alternative (i.e., the suitcase) upon hearing *not*. Ge et al. (2021a) found L1 English speakers

considered the alternative of focus at an early stage (generally before hearing “not” in sentences). L2 speakers showed delayed eye movements to the alternative of focus (generally while or after hearing “not” in sentences), with the L1 Dutch speakers showing even more delayed behavior than the L1 Cantonese speakers. The authors interpreted these differences as evidence for problematic integration of multiple interfaces (e.g., syntax-semantics, syntax-pragmatics) in real time. They connected their findings to the Prosodic-Learning Interference Hypothesis (Tremblay et al., 2016, 2021), which states that L2 learning of prosodic cues is more difficult when the L1 and L2 use similar prosodic cues as in Dutch-English and less difficult when the L1 and L2 use different prosodic cues as in Cantonese-English, which also uses spoken particles as cues (see also Ge et al., 2021b).

This finding was partially replicated and extended by Jansen et al. (2023) who tested a new set of L1 Dutch-L2 English speakers and found that L2 learners with a stronger musical pitch perception ability were more likely to fixate on the target and less likely to fixate on the competitor. Thus, not only does the L1 affect L2 prosody acquisition, but also individuals’ perceptual abilities can play a role in L2 prosody acquisition.

We set out to replicate Ge et al. (2021a) using web-based eye-tracking and open materials and code. We collected L1 English and L1 Dutch data but not L1 Cantonese data given current geopolitical constraints. We also measure and test the acoustics of the stimuli and multiple individual differences to determine what acoustic cues and behavioral measures, if any, serve as reliable predictors of focus processing in an L1 and L2. We selected Ge et al. (2021a) for replication due to the recency and theoretical relevance of its findings, particularly the unexpected delays observed in Dutch L2 English learners. The study also offered a strong foundation for exploring individual differences, making it well-suited for extension through acoustic validation and learner-level predictors within the FiREE framework.

The acoustics of focus

In order for a word to be prominent, it must be realized with one or more acoustic correlates that increase or enhance its perceptibility. A speaker generally does this through an F0, amplitude,

duration, i.e., nuclear pitch accent on the focal element(s) (Gussenhoven, 1983). Speakers reliably mark focused words with a higher mean and max F0 (pitch), greater intensity (loudness), and longer durations than words not focused (Breen et al., 2010). With respect to the stimuli Ge et al. (2021a) created, the authors reported two significant duration differences: verb duration is longer in verb-focused condition than in object-focused condition; object duration is longer in object-focused condition than in verb-focused condition. No F0 or amplitude differences were reported. Presumably these acoustic cues contribute to eye movements. In our exploratory extension, we place an emphasis on linking variable input to eye movements to strengthen theory (Magnuson, 2019).

For example, Figure 1 plots the F0 contours of the stimuli by condition (color) with time on the x-axis. For verb-focused sentences (indicated in orange), we see an increase in F0 at the "verb1" time bin and for object-focused sentences (indicated in blue), we see an increase in F0 at the "object1" time bin. F0 information is dynamic and unfolds over time. We assume all listeners are sensitive to this information and therefore ask (EE1) how do acoustic cues predict participants' eye movements? We expect some findings given what we know about the acoustics of information structure (Breen et al., 2010).

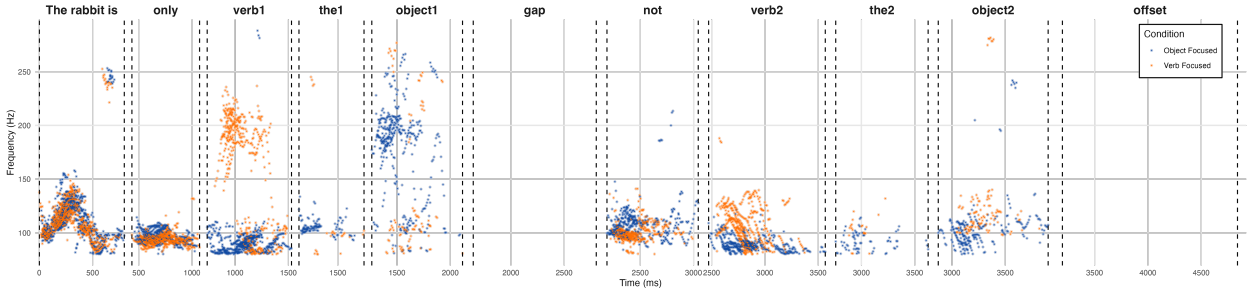

Figure 1

Fundamental frequency over time by object-focused (blue) and verb-focused (orange) sentences.

Individuals differ in their language processing behavior

Current experimental evidence suggests that there is a complex interplay between cognitive abilities, auditory perceptual abilities, and motor reproduction abilities during speech processing

(Bakkouche & Saito, 2025; Bramlett et al., 2024; Kachlicka et al., 2019; Saito et al., 2022). Whereas we cannot review (and explore) every possible individual difference predictor, we briefly review five ways in which individuals differ as it relates to spoken language processing: working memory, cognitive control, lexical proficiency, auditory perception abilities, and auditory motor reproduction abilities. These dimensions are theoretically relevant to focus processing because successful integration of prosodic, syntactic, and semantic cues in real time requires flexible cognitive control, precise auditory perception, and potentially sensorimotor alignment for tracking prosodic contours. We therefore include these measures to explore whether individual variability in these domains helps explain differences in how listeners process prosodic focus structure.

Working memory

Working memory, or the ability to keep recent input in mind and later draw on it (see Baddeley, 2003; Carpenter & Just, 2013) can affect a wide range of linguistic processes. Based on findings related to working memory and prosody processing research (e.g., Bishop, 2021; Ferreira & Karimi, 2015; Traxler, 2009), the answer most likely depends on the population, task demands, and methods used. For example, working memory, as measured by the digit span task (the same task we use in the present study), was not predictive in terms of identifying English emotional prosody (Sinagra & Wiener, 2022) or Italian lexical stress (Bramlett & Wiener, 2025) in groups of L1 adults, presumably given the relatively simple task demands of identifying happy/sad prosody or a penultimate/antepenultimate stressed word. How working memory affects the processing of focus prosody in a look-and-listen task is unclear. We use longer sentences involving potential object or verb targets that will require listeners to store linguistic information and draw on it later. (EE2) Do participants with higher working memory show earlier fixations to the focus alternative? If there is an effect of working memory, we assume it will be found among all participants and not just within the L1 or L2 group.

Cognitive control

Whereas there are several cognitive tasks used in the psycholinguistics literature (Ness et al., 2023), we were interested in a task that captures participants' ability to resolve conflict

during processing. For this reason, we used the Flanker task (Eriksen & Eriksen, 1974) in which participants must focus their attention on congruent stimuli (e.g., >>>>>) while resisting attention on incongruent stimuli (e.g., >><>>). Congruency tasks such as Flanker have been found to predict behavior in many different linguistic tasks and populations, especially bilinguals who must suppress their L1 while processing their L2 (Blumenfeld & Marian, 2014; Luk et al., 2011) (though see also Hedge et al., 2018 for reliability concerns). (EE3) Do L2 speakers with better cognitive control as indexed by performance on the Flanker task show earlier fixations to the focus alternative? If there is an effect of cognitive control, we assume it will be found among only the L2 participants and not the L1 group given the L2 group must suppress their L1 while performing the task.

Lexical proficiency

L1 and L2 speakers differ in their lexical proficiency, which has unsurprisingly been shown to contribute to differences in language processing (Yap et al., 2012; Zareva et al., 2005). For example, higher lexical proficiency scores are correlated with the speed of activation of the target and the degree to which lexical competition is resolved (Sarrett et al., 2022). Here, we are interested in whether increased English lexical proficiency, as measured by the LexTALE task (Lemhöfer & Broersma, 2012), leads to earlier looks for the L2 English participants. (EE4) Do L2 speakers with greater English proficiency show earlier fixations to the focus alternative? If there is an effect of lexical proficiency, we assume it will be found among only the L2 participants and not the L1 group given the L1 group should have less variation in their English proficiency.

Perceptual auditory sensitivity

Detecting prosodic cues requires a certain level of sensitivity to the incoming acoustics. Here we look at four measures of what Saito (2023) calls "explicit acuity in L2 speech learning": how sensitive a listener is to temporal and spectral cues or dimensions (e.g., formant, pitch, duration, and intensity). These measures have proven to be reliable measures for a range of L2 speech learning tasks (Bakkouche & Saito, 2025; Bramlett et al., 2024; Kachlicka et al., 2019; Saito et al., 2024). We also know that musical perception ability, which aligns closely with our

pitch task, was somewhat predictive of eye fixations Jansen et al. (2023). (EE5) Do speakers with better perceptual auditory sensitivity show earlier fixations to the focus alternative? If there is an effect of auditory sensitivity, we assume it will be found among all participants given that the temporal and spectral cues that we test—formant, pitch, duration, and intensity—are used in both Dutch and English for focus realization.

Auditory motor reproduction

Recent work (Saito et al., 2024; Tierney & Kraus, 2014; Tierney et al., 2017) has argued that motor abilities are a crucial part of auditory processing and that the ability to integrate auditory input to motor actions helps explain various aspects of adult L2 acquisition. We examine whether participants' ability to reproduce target sound sequences for melodies (pitch) and rhythm (duration) is informative for understanding L1 and L2 prosodic processing. (EE6) Do speakers with better auditory motor reproduction show earlier fixations to the focus alternative? If there is an effect of motor reproductions, we assume it will be found among all participants given that this integration ability is necessary for all languages.

Motivation for Replicating Ge et al. (2021a)

We selected Ge et al. (2021a) for replication because the study reported an unexpected L1 Dutch effect in focus processing, where L1 Dutch–L2 English speakers showed sensitivity to prosodic focus that diverged from predictions. Replicating this result was important for testing whether it reflected a robust cross-linguistic effect or a study-specific outcome.

Second, the original study was reasonably well documented and thus replicable, though not fully transparent by current standards. The task design and stimuli were described in sufficient detail to permit replication (McManus, 2024). At the same time, materials and analysis code were not publicly available, which falls short of emerging norms for reproducible research more broadly (Goodman et al., 2016) and for eye-tracking specifically (Bramlett & Wiener, 2024; Godfroid et al., 2025). Because the design was relatively simple and clearly specified, it still provided a suitable baseline for the FiREE framework.

Third, the study exemplified the broader challenge of balancing fidelity with the need for more robust analyses and theoretical exploration. It lacked acoustic analyses and relied on coarse time-binned models, gaps that align with recent calls for more informative and theory-driven approaches (Xie et al., 2023). Our replication therefore tested the reliability of the original findings while using FiREE to guide refinement and extension.

Research Questions

This study follows the Fidelity, Refinement, and Exploratory Extension (FiREE) framework to structure a layered replication of Ge et al. (2021a). To guide each phase of the study, we pose the following research questions:

- **RQ1 (Fidelity):** To what extent do we replicate the key finding from Ge et al. (2021a) that L1 Dutch-L2 English speakers show delayed processing of focus alternatives relative to L1 English speakers?
- **RQ2 (Refinement):** To what extent do group differences in focus processing, when analyzed using continuous time-sensitive models (i.e., Generalized Additive Models), emerge as dynamic shifts in fixations over time rather than static processing deficits?
- **RQ3 (Exploratory Extension):** To what extent do acoustic features of the stimuli and individual difference measures (e.g., auditory sensitivity, auditory-motor integration, cognitive control) predict changes in target fixations over time during focus processing?

Methods

Experimental materials, unidentifiable data, and R code are openly available through the Open Science Framework.

Participants

A total of 130 participants were initially recruited for the study. We recruited participants according to their first language (L1), which we defined as the language acquired from birth and the language that the speaker considered most fluent. We defined the L2 as the language acquired

in childhood after the L1 was in place and the language that the speaker considers less fluent than their L1. Unlike Ge et al. (2021a), which recruited L1 English participants from study abroad students population in Hong Kong, our L1 English participants were recruited via Prolific (N = 25) (Prolific, 2024) or in-person at a North American university (N = 50). All L1 English participants reported being born in the United States. The L1 Dutch-L2 English speakers (N = 55) were fully recruited via Prolific and all reported being born in the Netherlands. To ensure consistency, participants completed a detailed language background questionnaire (Marian et al., 2007), confirming their age of L2 acquisition, language exposure history, and self-rated proficiency. No participants had significant early exposure to other languages at home, and all participants completed primary and secondary education in English or Dutch-speaking environments, respectively. From these 130 participants, we required that all participants provide informed consent, pass a basic hearing screening using a dichotic pitch task (Milne et al., 2021), and pass a 5-point eye-tracking calibration with sufficient lighting. This left a total of 105 participants (L1 English N = 74; L1 Dutch N = 31) who qualified and completed the study. These participants were compensated for their time. To ensure that participants were engaged in the online tasks, we further specified that participants scored within 3 median absolute deviations (MADs) (Leys et al., 2013) for all behavioral tasks. After removal, this left 61 L1 English speakers and 27 L1 Dutch speakers. All participants provided informed consent in accordance with IRB-approved procedures and were compensated for their time and effort.

Ge et al. (2021a) recruited university students in lab settings in the Netherlands (L1 Dutch–L2 English) and Hong Kong (L1 Cantonese–L2 English). Our participants were recruited online via Prolific and completed the study remotely. This broadened the demographic profile of both L1 and L2 groups but also introduced more heterogeneity in testing environments. We view this as an implementation-level adaptation of recruitment, not a design-level methodological change.

Materials

All materials used in the present study were freely available and taken from previously published research (for which we are very grateful). The forward digit span task, Flanker task, and LexTALE (Lemhöfer & Broersma, 2012) were taken from the Gorilla (Anwyl-Irvine et al., 2019) Open Materials repository (see our OSF page for more details). The perceptual auditory sensitivity and auditory motor reproduction tasks were provided by Kachlicka et al. (2019) and Saito et al. (2020). Auditory and visual stimuli for the eye-tracking task were provided by Ge et al. (2021a). While Ge et al. (2021a) did not conduct an independent acoustic analysis, we used Parselmouth (Jadoul et al., 2018) and the R 'reticulate' package (Ushey et al., 2022) we extracted acoustic measures from Ge et al. (2021a)'s stimuli. This yielded four normalized key parameters: pitch range (min-max pitch per word), amplitude (measured in dB), duration, and word stress (measured as spectral tilt in lower frequencies). All scaled values can be seen in Figure 2.

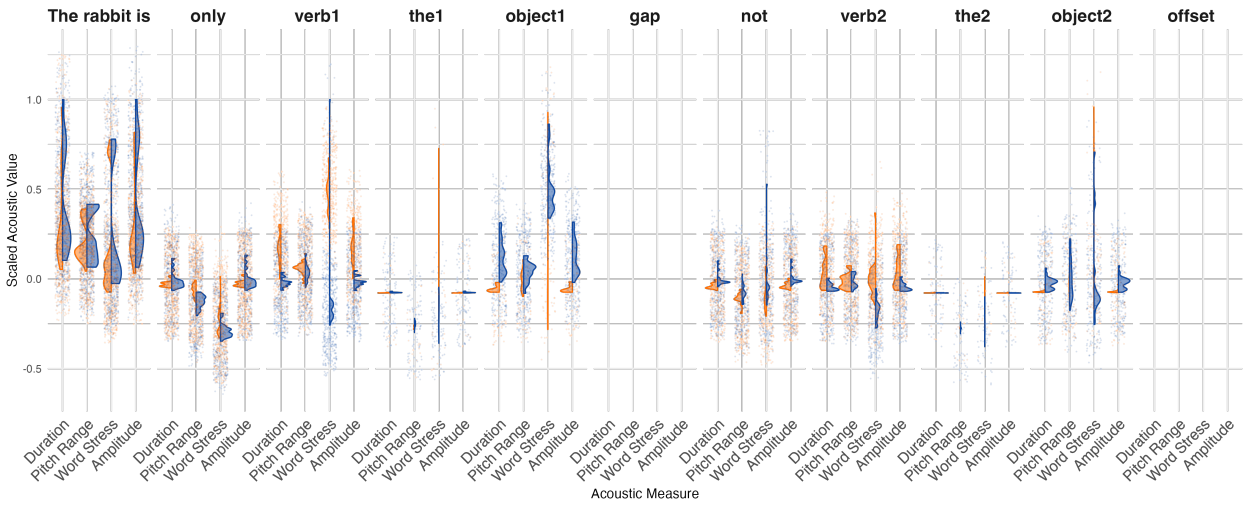

Figure 2

Scaled duration, fundamental frequency (pitch) range, stress prominence, and amplitude by time word for object-focused (blue) and verb-focused (orange) sentences.

Procedure

The experiment was hosted on Gorilla (Anwyl-Irvine et al., 2019) and distributed via Prolific, with participants completing the study on a personal computer in an environment that met the experiment's requirements, such as adequate lighting and minimal background noise. Participants started with the first three tasks in a fixed order: digit span, auditory sensitivity, and Flanker. The digit span task began with two digits, each on screen for 500 ms, followed by a fixation cross for 250 ms. Participants were required to use their mouse to click a visual number pad and enter the same digits in the same order. The adaptive, forward task increased by one digit after each correct response and decreased by one digit after each incorrect response. The task contained 13 trials and took approximately two minutes to complete. The internal consistency of the items as measured by Cronbach's alpha was .63.

The auditory sensitivity task included four same/different (AX) discrimination tasks across four cues: pitch, risetime (the speed at which a sound reaches its peak amplitude), duration, and formant contrasts. Each task consisted of 36 trials in which a sound was played, followed by a 250 ms fixation cross, and then a second sound. Participants responded by pressing the 'Z' or 'M' key to indicate whether the sounds were the same or different. Each task used a continuum of 50 stimuli taken from Kachlicka et al. (2019), with 10 stimuli selected at varying distances (e.g., 10 Hz or 15 Hz for pitch). The order of stimulus distances and same/different trials was randomized. Each battery task took approximately 90 seconds to complete. The internal consistency of these tasks, measured by Cronbach's alpha, was pitch = .75, risetime = .76, duration = .60, formants = .70. We note that both the auditory sensitivity and digit span tasks contained items of varying difficulty. For example, recalling two digits is inherently easier than recalling nine digits; discriminating between two sounds with a 50 Hz difference is easier than discriminating between two sounds with a 5 Hz difference. Given this variability, lower internal reliability, as measured by Cronbach (1951), is expected.

The Flanker task showed five arrows with the center arrow facing either the same (congruent) direction as the other four arrows or the opposite (incongruent) direction. Participants

were asked to use their keyboard and press 'Z' if the middle arrow was facing left and 'M' if it was facing right. There were 12 practice trials with feedback followed by 96 trials without feedback (48 congruent; 48 incongruent) spread over four blocks of 24 trials each. After each trial, a fixation cross with varying timing was shown ranging from 400 to 1,000 ms. The internal consistency of the items measured by Cronbach's alpha was .93.

Next, the eye-tracking task and the auditory-motor tasks were counterbalanced across participants to mitigate order effects and minimize potential biases related to task sequencing. For the eye-tracking task, gaze data were recorded using WebGazer.js (Papoutsaki et al., 2016), implemented within Gorilla (Anwyl-Irvine et al., 2019), which is the only implementation level difference between our task and Ge et al. (2021a). Participants began with four practice trials followed by 44 experimental trials (20 target and 24 filler). The same two counterbalanced lists used in Ge et al. (2021a) were used for the present study. In each trial, the 2x2 visual display and auditory stimulus were presented simultaneously followed by 1,000 ms of silence. These auditory stimuli were recorded at 44.1 kHz (16-bit resolution, mono) by a male native speaker of British English. Forty target sentences contained *only* with prosodic prominence on either the verb or the object, creating two experimental conditions: object-focus and verb-focus. An additional 48 fillers were created without the *only* focus particle. Like the targets, these fillers contained a *not*-fragment as in "The rabbit is licking the CANDY, not licking the ice cream." Figure 3 shows an example of a 2x2 visual stimulus. Each stimulus consisted of a target (e.g., the rabbit licking the candy), a competitor corresponding to the object alternative focus (e.g., the rabbit licking the ice cream), a competitor corresponding to the verb alternative focus (e.g., the rabbit throwing the candy) and a distractor (e.g., the rabbit throwing the ice cream). The non-target items were structured to ensure balanced looks across visual stimuli so that participants could not infer the target-competitor pairing by visual inspection alone. Participants were told to simply look at the visual display and listen to the audio. After each trial, a 1,000 ms slide with a different cartoon character at the center of the screen was shown before advancing to the next trial. No feedback was given during the task. The eye-tracking task took approximately 15 minutes to complete.

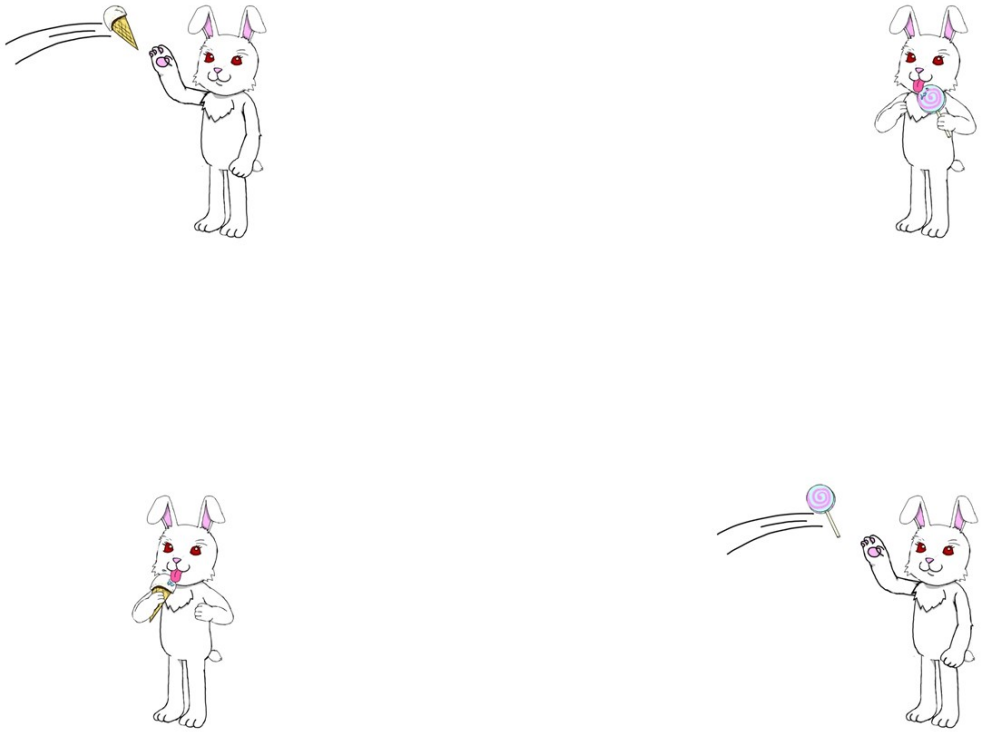

Figure 3

Example 2x2 visual display taken from Ge et al. (2021a).

The auditory-motor tasks consisted of two components: auditory-motor rhythm and auditory-motor melody, both adapted from Kachlicka et al. (2019). In the auditory-motor rhythm task, participants heard a rhythmic sequence consisting of 13 possible beat positions, played three times. They were then required to replicate the rhythm by pressing the space bar at the correct intervals. Each key press was time-stamped, and accuracy was assessed based on whether a beat was present (1) or absent (0). In the auditory-motor melody task, participants listened to a seven-note melody and were required to reproduce it using on-screen buttons corresponding to relative pitch levels. To maintain consistency, all melodies began with the middle pitch, ensuring a stable reference point for participants. Accuracy was determined by comparing the selected pitch to the correct response. The internal consistency of the items as measured by Cronbach's alpha

was .92 for the melody task and .88 for the rhythm task.

After completing the eye-tracking task and auditory-motor tasks, the participants completed a modified version of the English LexTALE task (Lemhöfer & Broersma, 2012). On each trial, participants were shown a fixation cross for 500 ms followed by a string of letters presented for 2000 ms. Participants used the 'J' and 'K' keys to indicate whether the displayed string was a real word or a non-word, within the given time limit. A total of 60 trials were completed, consisting of 40 words and 20 non-words presented in a randomized order. The internal consistency for the LexTALE task, measured by Cronbach's alpha, was .92. The English LexTALE and was adapted from publicly available Gorilla open materials; while the original versions of LexTALE did not impose time limits, our version required responses within 2000 ms. The task took approximately three minutes to complete. After completing the LexTALE, participants were asked to complete a language questionnaire that detailed their language experience (Marian et al., 2007).

Data analysis

Analyses were carried out in R (version 4.4.3; R Core Team, 2022) with a 0.05 alpha level for all null hypothesis significance tests. The individual difference measures were calculated so that each participant had a single score for each task. Working memory capacity, assessed using the forward digit span task, was determined by averaging the correctly recalled sequence lengths for each participant. One participant was removed for low performance in this task (outside 3 MAD). For the four auditory sensitivity tasks, all reaction times below 200 ms were removed across the four battery tasks. Following this, hits and false-alarms for each trial were scored and sensitivity to each contrast type (pitch, duration, formants, risetime) was then calculated using d-prime, a measure of sensitivity that accounts for bias. For all d' calculated measures a constant of .05 was added to both hits/false-alarms with a Haldane correction/log-linear correction for values of 0 or 1 (Hautus, 1995). One participant was removed for low performance in these tasks (outside 3 MAD).

The Flanker task was analyzed using drift rate estimates from a drift diffusion model (DDM) implemented in the brms package (Bürkner, 2017). Reaction times were first filtered to

remove extreme responses below 200 ms (anticipatory responses) and above 1750 ms (slow, inattentive responses). Trials were categorized as congruent (flankers match the target) or incongruent (flankers mismatch the target), and reaction time distributions were analyzed separately for each condition. A hierarchical Bayesian DDM was fit to the data using brms, with reaction time as the dependent variable and accuracy as the decision variable. Individual drift rate estimates were extracted for each participant from the model's random effects structure. These values were used as the final measure of cognitive control, aligning with previous work that treats drift rate as a sensitivity index for attentional selection in Flanker tasks (Poole et al., 2024).

For both auditory motor tasks, scores were computed so that each participant had a single score for each task. In the melody task, each individual note in a trial was scored as 1 for correct and 0 for incorrect, and these values were averaged within each trial to obtain a trial-level accuracy score. The final Melody Score was then calculated as the mean trial accuracy across all trials. In the rhythm task, each individual beat within a trial was also scored as 1 for correct and 0 for incorrect, with these values averaged within each trial to determine trial accuracy. Reaction times were normalized by subtracting an initial offset to align responses with the expected rhythmic structure. Trials in which participants produced double beats (multiple key presses within a single beat position) were adjusted by retaining only the first response. The final Rhythm Score was calculated as the mean trial accuracy across all trials. Three participants were removed for low performance (outside 3 MAD).

For eye-tracking data removal, eye-fixations outside the possible screen area were removed. As suggested in Bramlett and Wiener (2024), quadrants were defined by the origin to maximize signal retention. Fixations at the beginning of the trials were normally distributed from the center along both the x and y axes. Eye-fixations with face confirmation below 100% certainty were removed. We retained approximately 77.79% of eye-fixations. Our data removal is slightly higher the other recent web-based eye-tracking studies. because the task is a look-and-listen task and we had no other way to ensure that participants were attending to the task. We set a minimal frame rate for participants to 5 fps (Vos et al., 2022) and only five participants were removed for poor data

quality. See supplemental materials for detailed range across participants variable frame rates.

Like acoustic sensitivity measures, the scores for LexTALE exclude responses with RTs below 200 ms. While LexTALE is most commonly calculated by averaging the accuracy of non-words and words to control for bias (Lemhöfer & Broersma, 2012), we calculated LexTALE scores by using d′ after ignoring non-responses, which also controls for bias by using the difference between z-scored hits and false-alarms. For all d′ calculated, a constant of .05 was added to both hits/false-alarms with a Haldane (log linear) correction for values of 0 (Hautus, 1995).

For the Language Background Questionnaire, participants' first language and most fluent or dominant language was used to confirm their L1 status. Interestingly, 11 participants had language experience that mismatched their Prolific designation (e.g., their primary language they provided in our questionnaire did not match what they self-categorized within Prolific). However, the mismatch was not randomly distributed. Of the 11 participants that had mismatched language background information, 10 came from the English group and one from the Dutch group. The one Dutch speaker self-reported that English being their dominant language even though their L1 is Dutch. These participants were removed from further analyses.

Following replication reporting guidelines (McManus, 2024), we summarize in Table 1 the differences between our study and Ge et al. (2021a). We distinguish here between design-level changes, which are intentional experimental manipulations, and implementation-level adaptations, which are pragmatic accommodations made to reproduce a study as faithfully as possible with available tools. In our case, the contrasts largely reflect implementation-level adaptations associated with online testing, not design-level methodological changes. Analytical refinements and extensions were then layered on top of this replication.

Domain	Ge et al. (2021a)	This Study
Participants	L1 English (HK), L1 Dutch, L1 Cantonese	L1 English (US), L1 Dutch; no Cantonese
Recruitment	In-lab	Online (Prolific + US university)
Modality	Tobii eye-tracker (lab)	WebGazer.js (webcam)
Stimuli	Original only-sentence materials	Same stimuli; added acoustic analysis
Analysis	Time-binned LMMs	LMMs + GAMs + LASSO-GAMs

Table 1

Key differences between Ge et al. (2021a) and the present replication.

Results

Replication

Partial Replication: Testing prior findings

The foundation of any replication study is fidelity—ensuring that the methodological and analytical choices align as closely as possible with the original study. In this section, our primary goal is to faithfully replicate the methodological choices of Ge et al. (2021a), ensuring that our analyses align with the original study's design, statistical procedures, and reporting conventions. This includes using the same data transformations and inferential techniques to ensure that our replication remains as close as possible to the original study. Any unavoidable deviations, whether due to differences in implementation, data collection platform, or sample composition, are explicitly documented to clarify the degree to which methodological fidelity has been maintained.

Following Ge et al. (2021a), we analyzed fixation proportions to target objects, object-stressed competitors, verb-stressed competitors, and distractors across time, comparing conditions in which the object or verb was stressed. To maintain fidelity, we used a comparable time-binning approach, aligned with the beginning and end of each word in the sentence (e.g., "The rabbit is", "only", "verb1", "the1", "object1", "gap", "not", "verb2", "the2", "object2",

"offset"). This follows Ge et al. (2021a)'s critical word boundaries and allows for direct comparisons between studies with nine critical time bins. We calculated mean fixation proportions and standard errors for each interest area across time bins and conditions. Like Ge et al. (2021a), we plotted these fixation patterns using a time-series approach, with separate visualizations for each competitor type and experimental group. L1 English fixation proportions across areas of interest (AOI) for both our study and and estimated result of Ge et al. (2021a) can be found in Figure 4; L1 Dutch-L2 English fixation proportions can be found in Figure 5.

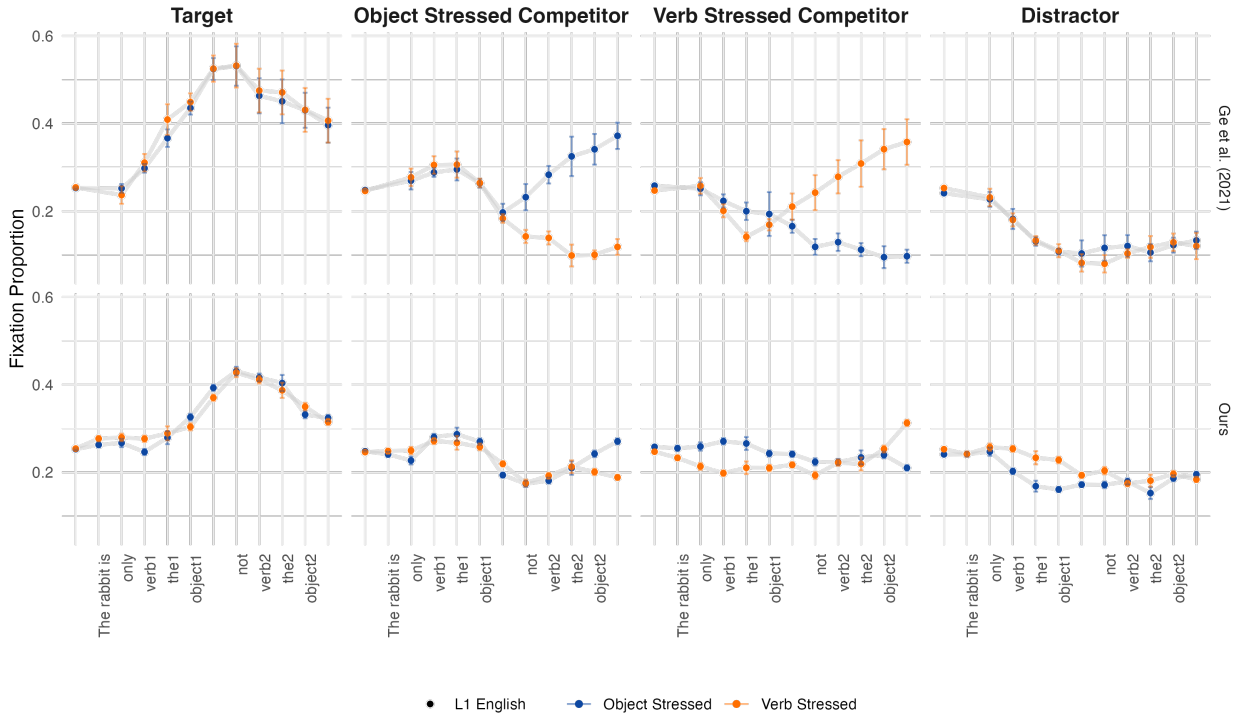

Figure 4

L1 English speakers' fixation proportions across time bins from Ge et al. (2021a) (top row) and our study (bottom row) by target, competitors, and distractor.

Following Ge et al. (2021a), we applied mixed-effects regression models to predict fixation proportions as a function of focus condition, time, and AOI for each time bin separately. We fit a series of linear mixed-effects models (LMMs) with random intercepts for participants and slopes for condition and AOI. The best-fitting models were selected based on lowest AIC. Following Ge

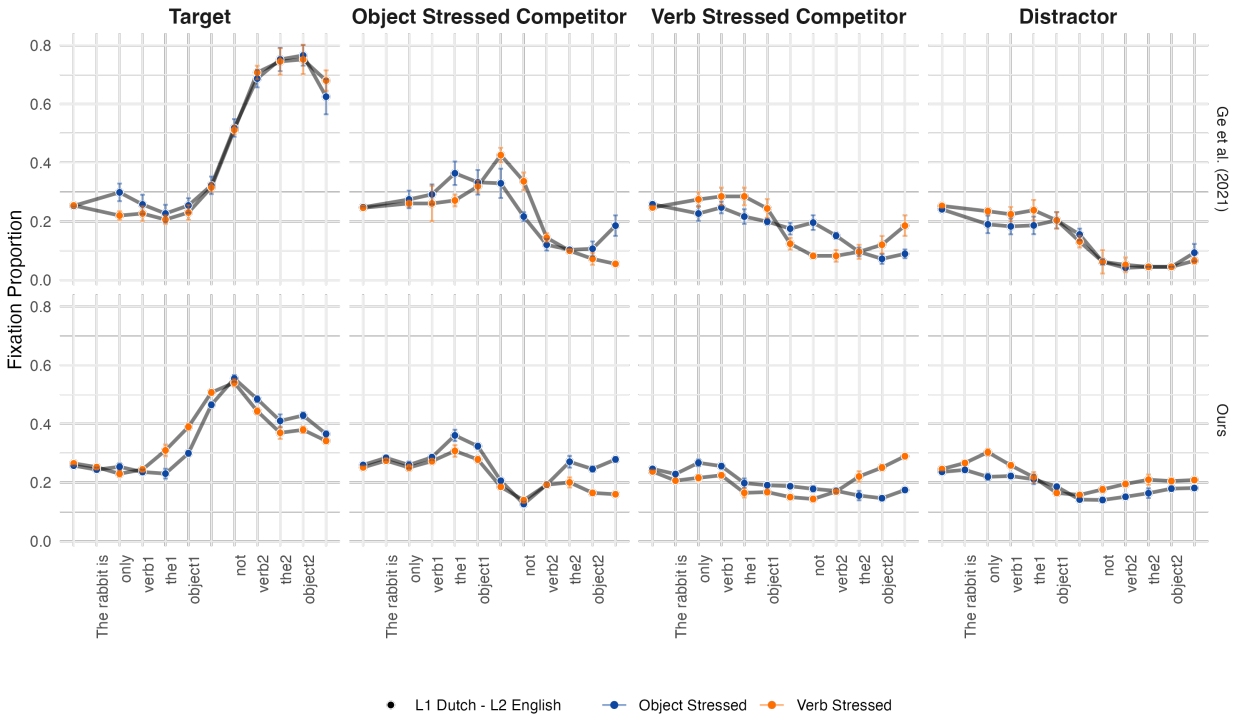

Figure 5

L1 Dutch-L2 English speakers' fixation proportions across time bins from Ge et al. (2021a) (top row) and our study (bottom row) by target, competitors, and distractor.

et al. (2021a), we used the baseline of object stressed competitor for condition and a baseline of object stressed competitor for AOI. This means that negative effects for either variable can be interpreted as object facing and positive effects are verb facing. Said another way, a positive effect for AOI means more looks toward verb competitor. Likewise, a positive effect for condition indicates more looks during verb-focused sentences.

Unlike Ge et al. (2021a), our model selection pipeline incorporated an automated model comparison approach, iteratively testing models with simpler effects structures. This change was necessary because many models did not converge. The full model included both AOI and condition, along with their interaction, as well as random intercepts for participants and slopes for items. If this failed to converge or created a singularity, then the model was reduced to remove individual slopes but still included the AOI × condition interaction. If that model failed to

converge, the AOI x condition interaction was removed, leaving only the main effects. The model was then further simplified by removing participant-specific effects entirely, keeping the overall effects of AOI and condition and participant random slopes. If none of these models converged or if the simplest model was the best, then the last step was a basic linear model without any participant-level adjustments, treating all data points as independent. This yielded 18 statistical model comparisons (9 critical time bins x 2 L1 groups). We present results in order from the beginning of the sentence to the end.

Starting with the L1 English models, three significant effects were found: A positive effect was found for AOI (β = 0.062, *SE* = 0.029, *t* = 2.17, *p* = 0.032), indicating more looks to verb competitors than object competitors during the "not" time bin. A positive interaction between AOI and condition was found (β = 0.175, *SE* = 0.056, *t* = 3.14, *p* = 0.0019) for the "offset" time bin, indicating that participants began to look more at the verb competitor during the verb-focused sentences. Additionally, in the same "offset" time bin, a negative effect was found for condition (β = -0.091, *SE* = 0.039, *t* = -2.30, *p* = 0.022) indicating more looks to competitors during object-focused sentences overall during the offset time bin. Our L1 English participants' results can be seen in comparison to Ge et al. (2021a) in Figure 6.

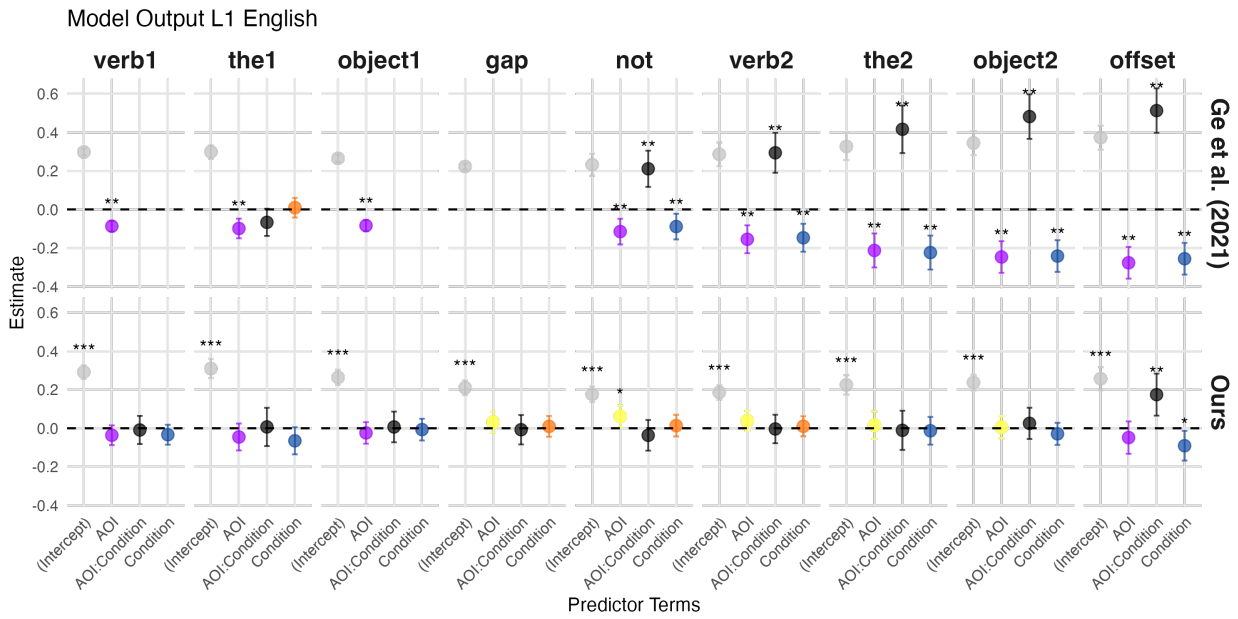

Figure 6

*Model outputs for L1 English participants across nine modeled time bins. Order of time bins appears in order across the top. Estimates of Ge et al. (2021a) data appears on top while our data appears below. Significance levels 0.05, 0.01, and 0.001 are indicated by *, **, and ***, respectively above the estimate. Purple and yellow references AOI, like that of figures 8 and 9. Blue and orange reference condition, like that of figures 4 and 5.*

For the L1 Dutch-L2 English models, nine significant effects were found: A negative effect of AOI (β = -0.099, *SE* = 0.041, *t* = -2.43, *p* = 0.017) was found for "the" time bin, indicating more looks to object competitor AOIs. Similarly, a negative effect for AOI was found for "object1" (β = -0.117, *SE* = 0.033, *t* = -3.59, $p < 0.001$), "the2" (β = -0.129, *SE* = 0.042, *t* = -3.08, *p* = 0.0025), and "object2" (β = -0.091, *SE* = 0.032, *t* = -2.83, *p* = 0.0055) time bins, all of which indicate more looks to the object competitors AOIs. Further, a negative effect of condition was found for both "object2" (β = 0.180, *SE* = 0.045, *t* = 3.98, $p < 0.001$) and "offset" (β = -0.130, *SE* = 0.050, *t* = -2.58, *p* = 0.011), indicating more looks to competitors during verb-focused sentences. Lastly, a positive interaction between AOI and condition was found during the "object2" (β = 0.180, *SE* = 0.045, *t* = 3.98, $p < 0.001$) and "offset" (β = 0.226, *SE* = 0.071, *t* = 3.17, *p* = 0.0019) time bins,

indicating more looks to object competitors during object-focused sentences. Our L1 Dutch-L2 English participant results can be seen in comparison to Ge et al. (2021a)'s results in Figure 6.

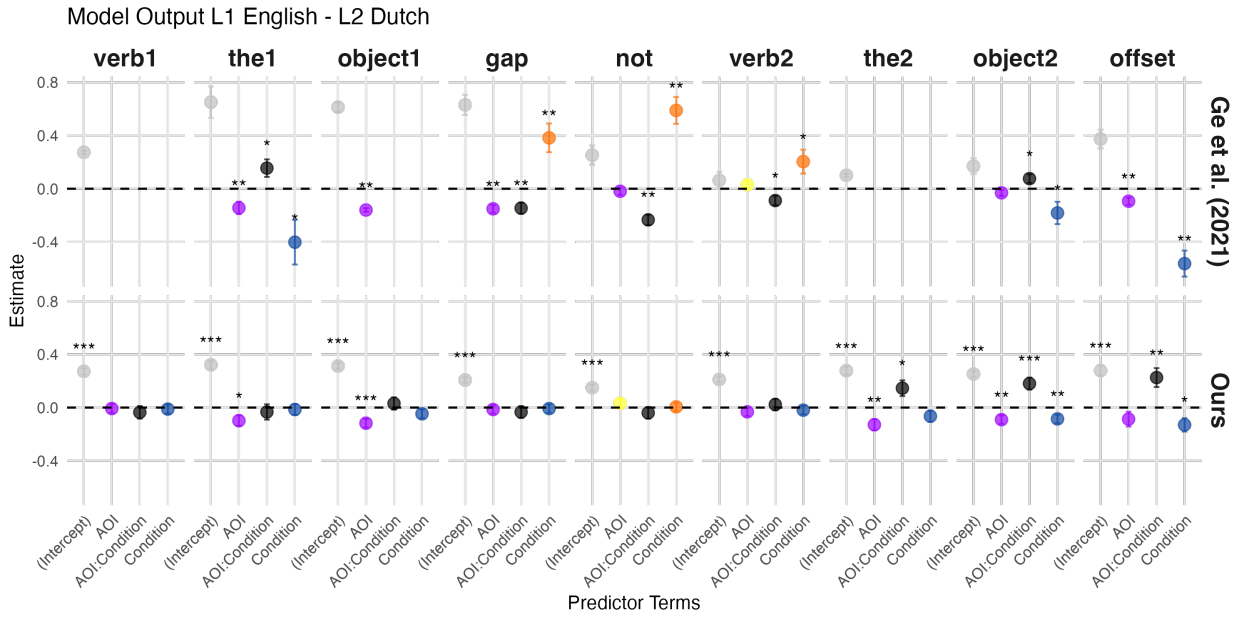

Figure 7

*Model outputs for L1 Dutch-L2 English participants across nine modeled time bins. Order of time bins appears in order across the top. Estimates of Ge et al. (2021a) data appears on top while our data appears below. Significance levels 0.05, 0.01, and 0.001 are indicated by *, **, and ***, respectively above the estimate. Like Figure 6, purple and yellow reference AOI, like that of figures 8 and 9. Blue and orange reference condition, like that of figures 4 and 5.*

Methodological refinement: A "focused" approach

Whereas fidelity in replication is essential, it is also necessary to evaluate whether the original analytical choices align with current best practices. The refinement phase of our analysis addresses potential limitations such as model specification, overfitting, multiple comparisons, and analytical transparency. Here, we implement refined statistical approaches that maintain the interpretability of the original findings while increasing statistical robustness. By comparing our refined results to both the original and replicated findings, we can assess whether methodological improvements impact the observed effects and whether the key conclusions of Ge et al. (2021a)

remain stable across different analytical approaches.

The first refinement we make is analyzing both binary competitor fixations and target fixations separately, rather than relying on aggregated measures. This approach provides a more comprehensive view of participants' behavior, capturing differences in how they allocate visual attention. Secondly, we combine the analysis from both language groups to be able to compare both within language by condition and across participant L1. Fig 8 shows a new way of plotting Figure 4, which allows the reader to compare when fixations to specific AOI deviate from each other and not just across conditions.

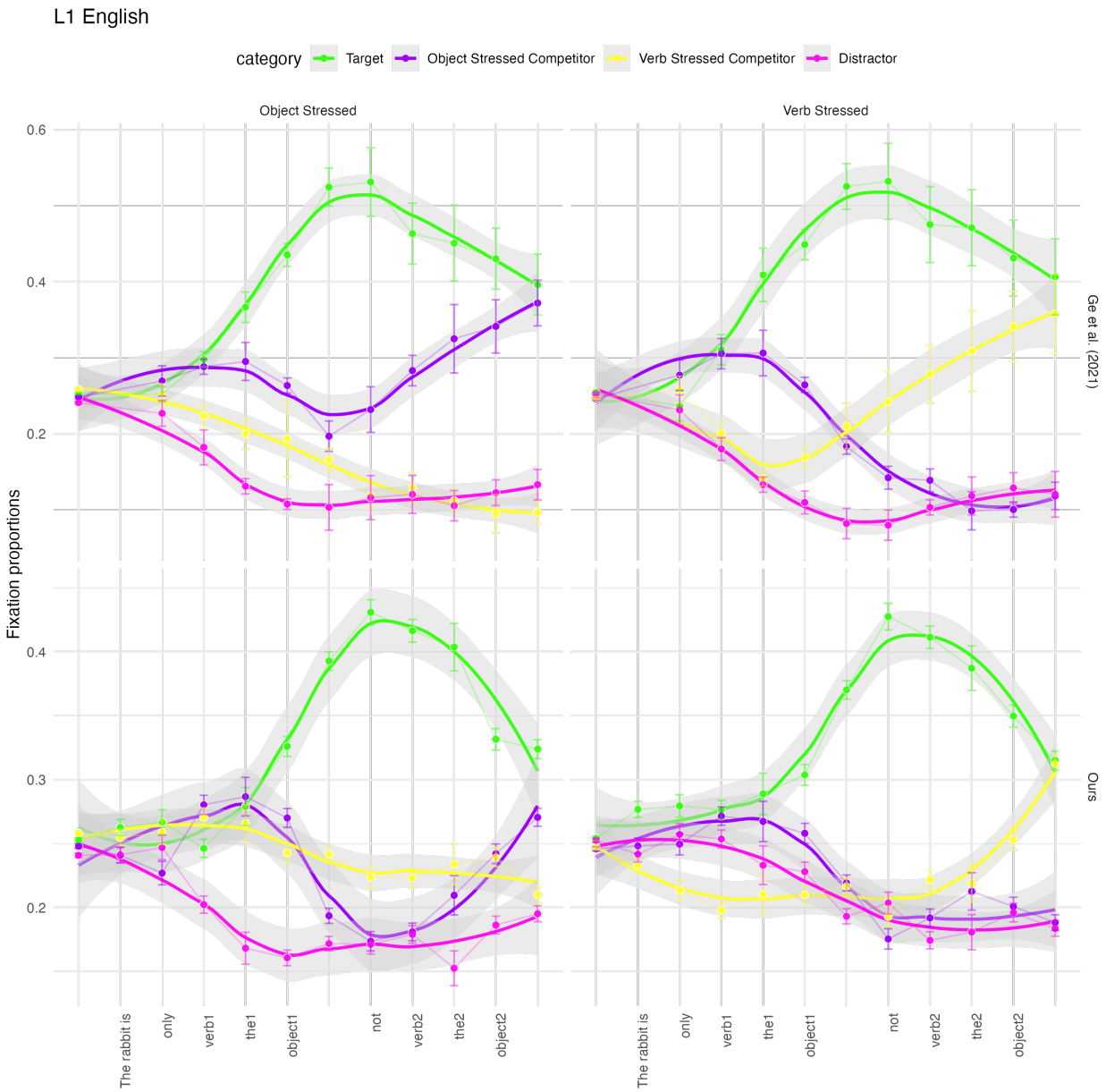

Figure 8

Fixation proportions for L1 English participants across object stressed (left) and verb stressed (right) sentences. Estimates of Ge et al. (2021a)'s data appears on top while our data appears below.

For these refined models, we present both target models and competitor models. That is, we compare the target fixations (green lines in Figures 8 and 9) in the target model and we are comparing the object competitor fixations (yellow lines in Figures 8 and 9) in the object-focused

sentences (left plots of Figures 8 and 9) as well as the verb competitor fixations (purple lines in Figures 8 and 9) during verb-focused sentences (right plots of Figures 8 and 9) for the competitor models.

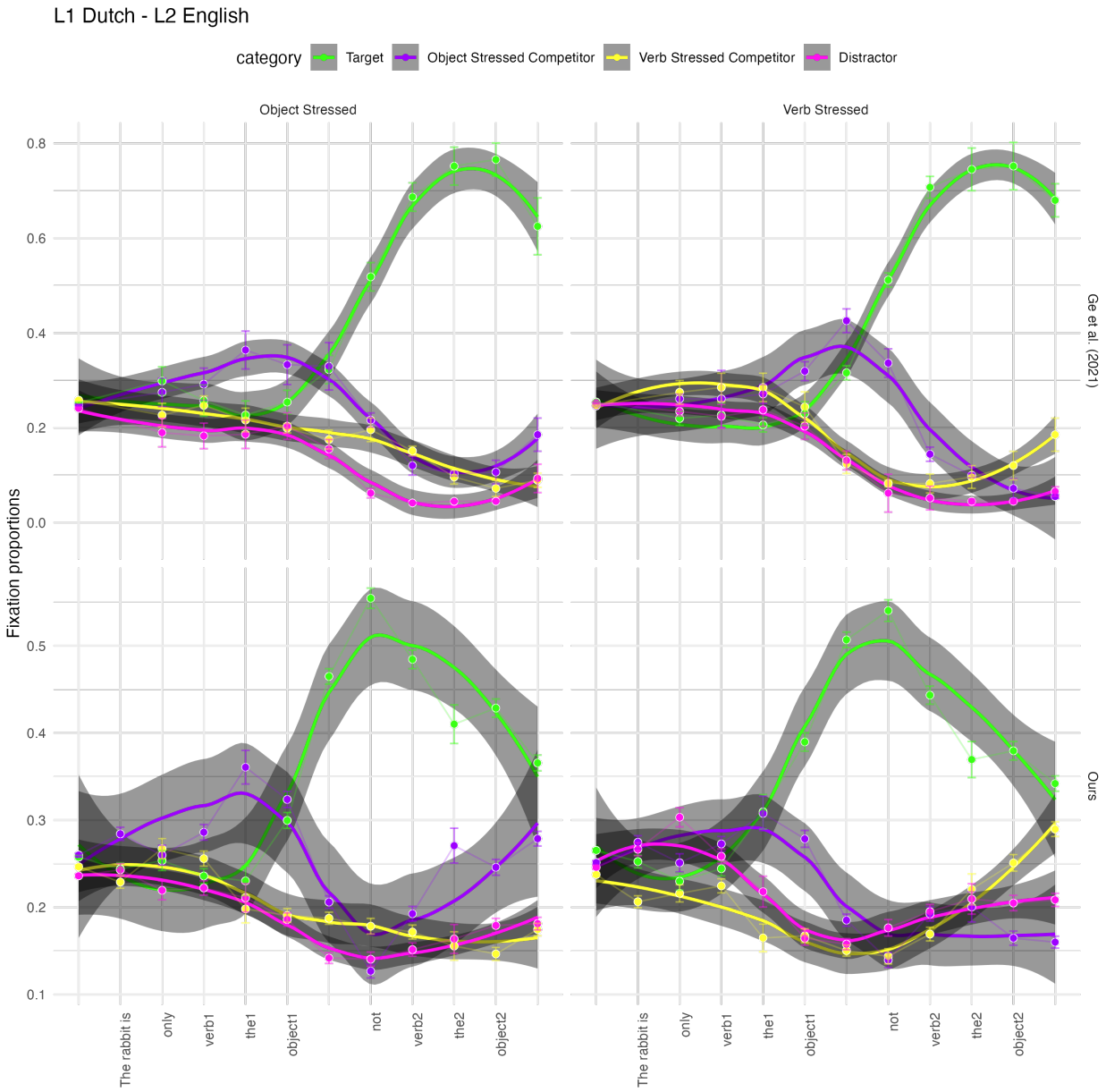

Figure 9

Fixation proportions for L1 Dutch-L2 English participants across object stressed (left) and verb stressed (right) sentences. Estimates of Ge et al., 2021a's data appears on top while our data appears below.

An additional difference in our refined model is that we do all time bins in two models: we split the sentence at the gap between the first phrase (e.g., "The rabbit only verb1 the1 object1) and second phrase (e.g., "not verb2 the2 object2). This yields four total models (target/competitor x phrases one/two).

Unlike the linear mixed-effects models (LMMs) used in the fidelity modeling, here we use Generalized Additive Models (GAMs) to better capture time-dependent changes in fixation proportions. Since eye-tracking data unfolds continuously over time, GAMs provide a more flexible way to model the non-linear dynamics of fixations that may not be well captured by traditional LMMs (Wood, 2017). In this approach, we compare two GAM variants: a full interaction model, which includes a three-way interaction between time, experiment group, and condition, and a main effects model, which assumes that changes in fixation proportions follow an additive pattern without time-dependent interactions. Both models include random smooth effects for participants, allowing for individual variability while capturing group-level fixation trends. Max models started with random items but were removed as none of the models converged.

For the first phrase of the target models, a positive main effect of stress was found (β = 0.155, *SE* = 0.053, *t* = 2.93, *p* = 0.003), indicating more target fixations during object focus sentences. A positive main effect of time was found (β = 0.067, *SE* = 0.012, *t* = 5.66, $p < 0.001$), indicating increased looks to targets over the first phrase in general. A negative two-way interaction between L1 and stress was found (β = -0.316, *SE* = 0.084, *t* = -3.76, $p < 0.001$), indicating fewer looks during verb stressed stimuli for L1 English participants. Additionally, we found a negative two-way interaction between time and stress in phrase 1 of the target model (β = -0.036, *SE* = 0.017, *t* = -2.12, *p* = 0.034), indicating more verb looks to targets later on in phrase one for verb-focused sentences. Lastly, the positive three-way interaction between time, L1, and stress (β = 0.136, *SE* = 0.027, *t* = 5.12, $p < 0.001$) indicates that over time, English participants exhibited an increasing trend in fixations to the target in verb-stressed sentences, in contrast to L1 Dutch-L2 English participants.

For target fixations during the second phrase, a main effect of L1 was found (β = 0.834, *SE*

= 0.237, t = 3.52, $p < 0.001$), indicating that English speakers had significantly more looks to targets during the second phrase. Additionally, a negative effect of time was found (β = -0.137, *SE* = 0.012, t = -11.36, $p < .001$), indicating fewer looks to targets over the second phrase. A negative two-way interaction between L1 and time was also found (β = -0.058, *SE* = 0.019, t = -3.09, p = 0.002), indicating more looks to targets for L1 Dutch-L2 English participants over time.

For the competitor first phrase model, a positive effect of time was found (β = 0.056, *SE* = 0.012, t = 4.63, $p < 0.001$), indicating more looks to competitors over time during the first phrase. A negative two-way interaction between time and stress was found (β = -0.095, *SE* = 0.018, t = -5.26, $p < 0.001$), indicating more looks to verb competitors as time increases. No significant effects were found for the second phrase competitor model. Lastly, a negative two-way interaction between L1 and stress was found (β = -0.178, *SE* = 0.086, t = -2.06, p = 0.040), indicating more looks to the verb competitor for L1 Dutch-L2 English participants.

For the second phrase competitor model, a positive effect of time was found (β = 0.161, *SE* = 0.014, t = 11.42, $p < 0.001$), indicating that more looks to competitors occurred over the duration of the second phrase. A positive interaction between time and L1 was also found (β = 0.055, *SE* = 0.022, t = 2.50, p = 0.012), indicating that L1 Dutch-L2 English speakers looked to competitors more as time went on over the second phrase.

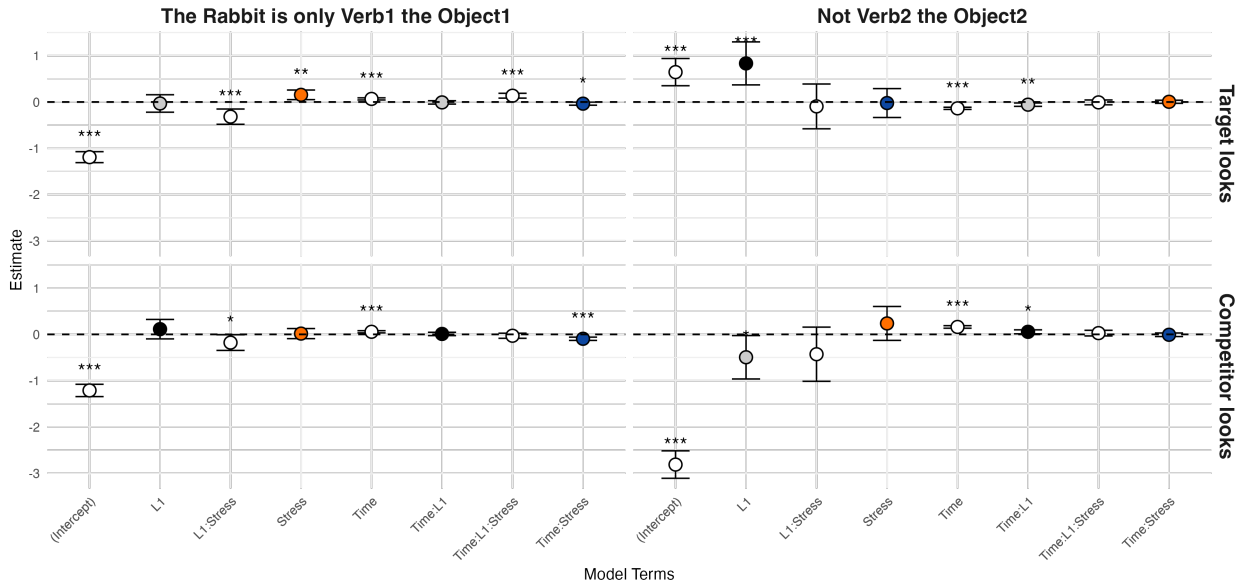

Figure 10

*Generalized additive model output across phrase one (left column) and phrase two (right column). Significance levels 0.05, 0.01, and 0.001 are indicated by *, **, and ***, respectively above the estimate. Like Figures 6 and 7, blue and orange reference condition based on reference levels in the model, like that of Figures 4 and 5, language is indicated by black (L1 Dutch-L2 English) and gray (L1 English), based on reference levels.*

Exploratory extension

Beyond replication and refinement, we extend the analysis to explore six extensions of Ge et al., 2021a. Specifically, we ask how fine-grained acoustic-phonetic properties of the stimuli affect participants' eye movements (EE1), and whether working memory (EE2), cognitive control (EE3), English lexical proficiency (EE4), perceptual auditory sensitivity (EE5), and auditory motor reproduction (EE6) affect participants' eye movements. The individual differences among our participants are visualized in Figure 11, which presents a comprehensive view of the multidimensional nature of these differences. Figure 2 presents the acoustic results.

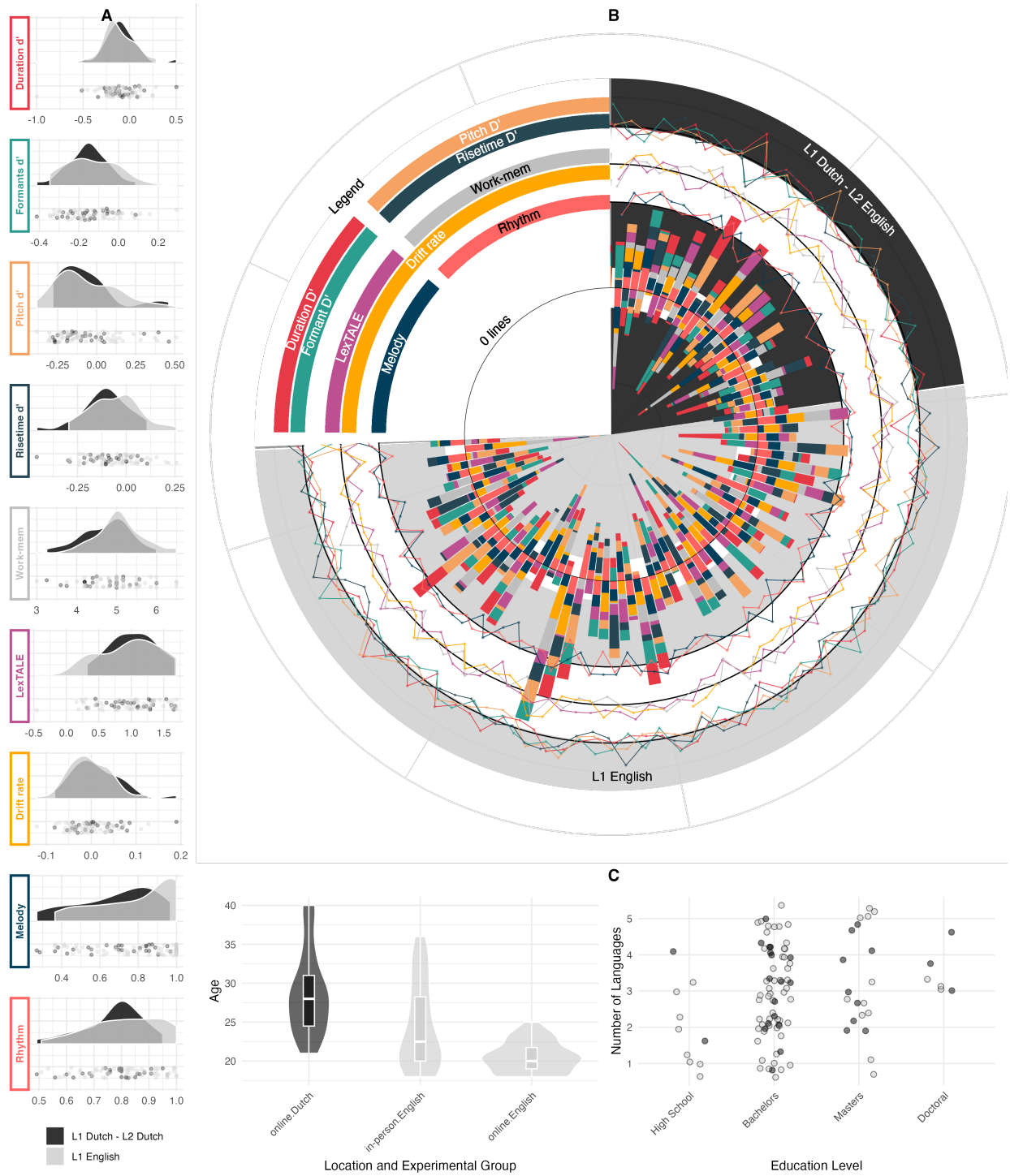

Figure 11

In this plot of individual differences, there are three sections A, B, and C. A- shows tasks across the L1 Dutch-L2 English and L1 English participants to give a distributional view of scores. B- provides individual difference vectors for each participant (stacked bar plots). The circular stacked bar plots are centered on a labeled 0 line so that negative and positive scores are stacked away from each other. Similarly, on the outer edge of the circular stacked bar plot the individual scores for tasks are shown in concentric circles by group (melodic and rhythmic scores, cognitive

We used our refined modeling approach as a baseline procedure. To optimize model selection while maintaining interpretability, we employed LASSO-GAM Feature Selection, combining LASSO regression (cv.glmnet(); (Friedman et al., 2010) with Generalized Additive Models (GAMs) (mgcv; (Wood, 2017). First, LASSO identified the most predictive variables from a set of individual differences (e.g., cognitive abilities, auditory perception, etc.), acoustic features (e.g., pitch, duration, etc.), and experimental conditions, while regularizing to prevent overfitting. Next, the selected predictors were incorporated into a GAM to model nonlinear time effects and random participant variability. This data-driven approach ensures that only the most informative factors contribute to the model, balancing flexibility, interpretability, and predictive accuracy.

The LASSO was applied in a generalized linear model (GLM) with a binomial link function, where the design matrix included individual differences, acoustic properties, and experimental conditions, along with key interaction terms. The individual difference measures included working memory capacity, drift rate (decision-making efficiency), lexical proficiency (LexTALE score), and motor reproduction. The acoustic features included pitch range (min-max pitch per word), duration, stress prominence (spectral tilt in lower frequencies), and amplitude (raw and dB-scaled). The experimental conditions included focus condition (object-focused vs. verb-focused) and participant group assignment (e.g., L1).

To account for potential interactive effects, we included key two-way and three-way interactions. These interactions tested whether perceptual, cognitive, and linguistic factors jointly influenced fixations. Specifically, we modeled interactions between individual differences and acoustic properties (e.g., pitch sensitivity × pitch range × duration) and interactions between experimental conditions and perceptual abilities (e.g., focus condition × working memory, L1 background × stress prominence). Additionally, we included a three-way interaction to examine whether the effect of working memory on lexical processing varied as a function of drift rate and lexical knowledge.

In the target model of phrase one, no main effects were found. However, a negative two-way interaction between duration d' and the duration of the word was found (β = -0.848, *SE* = 0.205, *t*

= -4.14, $p < 0.001$), indicating that fewer target looks occurred for those with lesser duration sensitivity during stimuli that have shorter durations. Similarly, a positive two-way interaction between formant d′ and word stress was found (β = 0.230, *SE* = 0.106, *t* = 2.16, *p* = 0.031), indicating more target fixations for those with better formant sensitivity for words with greater word stress. Lastly, a positive two-way interaction between pitch d′ and pitch range was found (β = 0.144, *SE* = 0.068, *t* = 2.11, *p* = 0.035), indicating that more target fixations occurred for individuals with greater pitch sensitivity for words with a larger pitch range.

For the second phrase target model, a negative effect of duration was found (β = -0.357, *SE* = 0.093, *t* = -3.85, *p* = 0.0001), indicating that shorter word acoustics lead to fewer target fixations. A negative effect of stress was also found (β = -0.060, *SE* = 0.022, *t* = -2.72, *p* = 0.0066), indicating that verb-focused sentences generally had fewer target fixations during the second phrase. A negative two-way interaction between duration d′ and word duration was also found (β = -0.396, *SE* = 0.125, *t* = -3.18, *p* = 0.0015), indicating fewer target fixations for those participants with lesser duration sensitivity during stimuli that have shorter duration. Finally, a negative interaction between pitch range and melody reproduction was found (β = -0.487, *SE* = 0.110, *t* = -4.42, $p < .00001$), indicating that fewer target fixations occurred during phrases with lower pitch for those with lower melodic reproduction abilities.

For the first phrase competitor model, a negative effect of duration was found (β = -0.477, *SE* = 0.195, *t* = -2.45, *p* = .014), indicating fewer looks to competitors for shorter duration word bins. A positive effect of L1 was also found (β = 0.325, *SE* = 0.120, *t* = 2.71, *p* = .007), indicating more looks to competitors during the first syllable for L1 Dutch-L2 English participants. A positive effect of word stress was also found (β = 0.197, *SE* = 0.097, *t* = 2.03, *p* = .042), indicating more competitor looks for words with higher word stress. In terms of two-way interactions, a positive interaction between duration d′ and duration of the word was found (β = 0.853, *SE* = 0.216, *t* = 3.96, $p < .001$), indicating more competitor looks when words have longer durations during the first phrase. The effect is mirrored in a positive interaction between formant d′ and word stress (β = 0.419, *SE* = 0.116, *t* = 3.61, $p < .001$), which indicates more competitor looks for

those with greater formant sensitivity when words have high word stress. A positive interaction between melody reproduction and pitch range (β = 0.239, *SE* = 0.072, *t* = 3.33, $p < .001$) further indicates that for participants with higher melodic reproduction abilities, words with greater pitch range lead to more competitor fixations during the first phrase. Finally, a negative interaction between stress and L1 (β = -0.267, *SE* = 0.047, *t* = -5.64, $p < .001$) indicates fewer competitor fixations for L1 English speakers during verb stressed first phrases.

Our last model is the second phrase competitor model, where a positive effect of duration was found (β = 0.362, *SE* = 0.091, *t* = 3.96, $p < .001$), indicating more competitor fixations for words with shorter duration during the second phrase. A negative effect of pitch d′ was also found (β = -0.393, *SE* = 0.194, *t* = -2.03, *p* = .043), indicating fewer competitor fixations for participants with lower pitch sensitivity. Likewise, a positive effect of stress was found (β = 0.355, *SE* = 0.077, *t* = 4.63, $p < .001$), indicating greater competitor fixations during the second phrase for object-focused phrases. A two-way negative interaction between L1 and stress was found (β = -0.188, *SE* = 0.052, *t* = -3.64, $p < 0.001$), indicating fewer competitor fixations for English speakers during verb focus sentences in the second phrase. Finally, a positive interaction between pitch d′ and pitch range was found (β = 0.664, *SE* = 0.131, *t* = 5.07, $p < 0.001$), indicating more competitor looks when a participant has more pitch sensitivity and the word has higher pitch range. All results for the exploratory extension analyses can be found in Figure 12. Table 2 summarizes the significant findings of the exploratory study.

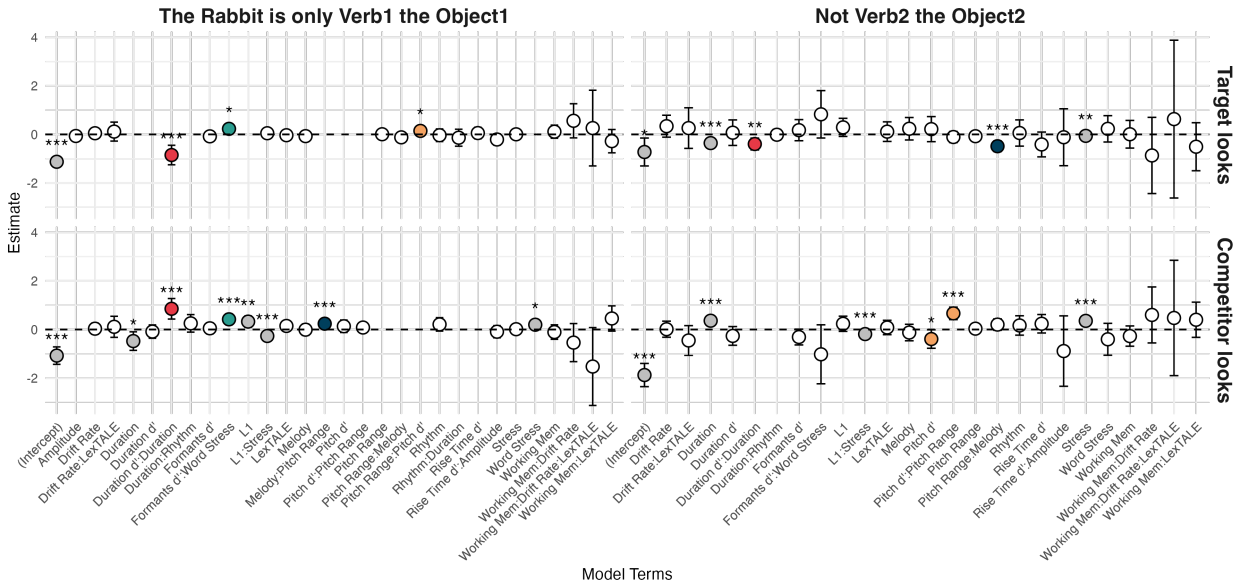

Figure 12

All non significant values remain white. Significant values are colorized by individual difference measure, like that of figure 11. Gray indicates significant but not an individual difference measure.

Discussion

Replication vs. Fidelity: a faithful reproduction with minimal changes

Ge et al. (2021a) investigated how L1 English speakers and L1 Dutch-L2 English speakers process focus in English *only*-sentences using the "look-and-listen" visual world paradigm. Their study reported two key findings that we set out to replicate: 1) L1 English speakers exhibit significantly earlier fixations to focus-alternative referents compared to L1 Dutch-L2 English speakers because L2 learners struggle to integrate multiple interfaces in real time. 2) L2 learners struggle to acquire focus prosody in the target language in accordance with the Prosodic-Learning Interference Hypothesis (Tremblay et al., 2016, 2021), which states that L2 learning of prosodic cues is more difficult when the L1 and L2 use similar prosodic cues as in Dutch-English.

Our fidelity-based replication found some limited evidence for these findings. Although we did not find the same effects in the same time bins that Ge et al. (2021a) reported, we did find that L1 speakers used prosodic information relatively more efficiently than L2 speakers. That is, we

AOI	Phrase	Effect Term	Effect	Effect Meaning
Target	1	Duration d' × Duration	Negative	Listeners with lower duration sensitivity fixated less on short-duration focus-marked words.
Target	1	Formants d' × Word Stress	Positive	Higher formant sensitivity led to increased fixations on stressed words.
Target	1	Pitch d' × Pitch Range	Positive	Listeners with greater pitch sensitivity fixated more when pitch range was wider.
Target	2	Stress	Negative	Listeners fixated less on targets in verb-focused sentences.
Target	2	Duration	Negative	Shorter words resulted in fewer fixations on the target.
Target	2	Duration d' × Duration	Negative	Fewer fixation on short-duration words for listeners with lower duration sensitivity.
Target	2	Pitch Range × Melody	Negative	Listeners with weaker melodic reproduction ability fixated less when pitch range was smaller.
Competitor	1	Duration	Negative	Fewer competitor fixations for words with shorter duration.
Competitor	1	L1	Positive	L1 Dutch-L2 English speakers showed greater fixations.
Competitor	1	Word Stress	Positive	Words with higher stress prominence attracted more competitor fixations.
Competitor	1	L1 × Stress	Negative	Fewer fixations by L1 Dutch-L2 English speakers for stressed words.
Competitor	1	Duration d' × Duration	Positive	More competitor fixations when words had longer duration.
Competitor	1	Formants d' × Word Stress	Positive	Higher formant sensitivity led to increased competitor fixations for stressed words.
Competitor	1	Melody × Pitch Range	Positive	Higher melodic ability led to increased fixations on competitors when pitch range was wider.
Competitor	2	Stress	Positive	Focus-marking increased competitor fixations.
Competitor	2	Duration	Positive	More competitor fixations for words with shorter duration.
Competitor	2	Pitch d'	Positive	Higher pitch sensitivity led to increased competitor fixations.
Competitor	2	Pitch d' × Pitch Range	Positive	Competitor fixations increased when pitch range was wider for listeners with higher pitch sensitivity.
Competitor	2	L1 × Stress	Negative	Fewer fixations for L1 English speakers during verb focus.

Table 2

Summary of fixation patterns across phrase models, organized by AOI, phrase, effect term, direction, and meaning.

observed looks to the focus competitor during "not" for L1 speakers but not for L2 speakers. We also observed for the L2 group relatively early competitor effects (before "not" or other prosodic information was available) and relatively late target looks. Thus, L1 Dutch-L2 English speakers seemed to struggle to integrate multiple interfaces in real time. One interpretation of our results–that align with Ge et al. (2021a)'s–is that the timing of L1 and L2 looks does, in fact, differ.

Our fidelity-based replication did not test for direct interactions between language groups. This was because Ge et al. (2021a) conducted separate statistical analyzes for each group, which means that group-level differences were inferred rather than explicitly tested via interaction

effects. The fact that our study did not replicate these key findings from Ge et al. (2021a) highlights the complexities of replication in psycholinguistics. This methodological limitation raises concerns about the robustness of previously reported L1-L2 differences, particularly given the potential for false positives when conducting multiple separate statistical tests. If group differences were found in one study but not another, this does not necessarily indicate a true underlying cognitive difference; rather, it could be the result of analytical choices or sample variability. This is particularly important when working with small effect sizes, where statistical significance may not always equate to meaningful or replicable findings (see Limitations below). Given these considerations, our fidelity replication highlights the importance of directly testing interactions, which we look at next in our refinement.

Replication vs. refinement: Balancing rigor and practicality

A key limitation of both our fidelity replication and the original study by Ge et al. (2021a) is the separation of L1 English and L1 Dutch participants into independent statistical models. Whereas this approach may have been necessary in Ge et al. (2021a) due to differences in eye-tracking sampling rates across participant groups, it prevented a direct statistical test of whether group differences were robust or whether they emerged as an artifact of separate analyses. To address this, we refined our analysis by adopting a time-sensitive model that accounts for gradual fixation changes across each phrase rather than treating time as discrete bins. Using Generalized Additive Models (GAMs), we assessed whether prosodic effects emerged immediately or developed over time. This refined approach allowed us to test whether L1 speakers truly showed an early fixation advantage and whether L2 speakers were uniformly delayed in processing prosodic focus cues.

The refined analysis produced three key findings that deepen our understanding of prosodic cue integration. First, our target model and competitor model for phrase one indicated that fixation patterns gradually increased over time, as shown by an effect of time with object focus leading to greater fixations. This means that fixations to the target did not shift immediately following prosodic cues, but rather developed dynamically over time as multiple levels of

linguistic information was integrated (see Lambrecht, 1994).

Second, rather than revealing a uniform L2 delay, our results show that L1 Dutch-L2 English speakers were not consistently slower than L1 English speakers in processing prosodic focus cues. In fact, the interaction between stress and L1 indicates that this difference is negative for both target and competitors models indicating fewer looks overall rather than an L1 advantage. A three-way interaction between time, L1, and stress further revealed that L1 English speakers gradually developed a fixation preference for the target in verb-stressed sentences whereas L2 speakers did not. This suggests that during the first phrase, L1 English speakers showed a more stable late-stage fixation pattern. In other words, the L1 speaker advantage may be one of long-term consistency rather than early efficiency.

Third, our refined approach revealed that group-level differences may have been previously overstated due to statistical modeling choices. In Ge et al. (2021a), L1 and L2 differences were inferred from separate statistical models rather than directly tested as an interaction. Our unified model revealed a negative interaction between L1 and stress in the first phrase competitor model, indicating that L1 English speakers showed fewer fixations to the competitor during verb-stressed sentences. At the same time, a negative interaction between time and stress showed that fixations to the competitor decreased more sharply for L1 English speakers as time progressed. These findings indicate that differences between L1 and L2 participants were not static processing deficits but instead emerged gradually through dynamic changes in fixation patterns over time.

Importantly, our refined analysis does not overturn the key findings of Ge et al. (2021a), rather it suggests that their interpretation may have been overly rigid. As opposed to a clear and immediate L1 advantage, our results indicate that prosodic cue integration is dynamic, shifting over time rather than appearing as a static group-level effect. By adopting an approach that accounts for these time-dependent changes, we provide a more nuanced understanding of how both L1 and L2 speakers process prosodic focus cues.

Exploratory extension: Where we go from here

Whereas methodological fidelity and statistical refinement allowed us to test for consistency with Ge et al. (2021a), we set out to explore how fine-grained acoustic-phonetic properties of the stimuli affect participants' eye movements (EE1), and whether working memory (EE2), cognitive control (EE3), English lexical proficiency (EE4), perceptual auditory sensitivity (EE5), and auditory motor reproduction (EE6) affect participants' eye movements.

Acoustics and acoustic sensitivity

As expected, the acoustic properties of the stimuli played a significant role in shaping fixation patterns (Magnuson, 2019). Across target and competitor analyses and during both the first and second phrase, we repeatedly found evidence for participants using duration, stress (spectral tilt), and pitch range information in real-time in line with previous acoustic findings Baumann and Winter (2018) and Breen et al. (2010). In addition, we found evidence that individual differences in acoustic sensitivity interacted with the acoustics of the stimuli to further influence eye movements (see Table 2). For example, in the first phrase competitor model, a positive interaction between duration d′ (individual sensitivity to duration contrasts) and word duration indicated that listeners with higher duration sensitivity were more likely to fixate on competitors when words had longer durations. This suggests that perception of prosodic prominence is contingent on both the acoustic properties of speech and the listener's ability to track duration-based prominence cues. Similarly, pitch variation also influenced fixation behavior. In the second phrase competitor model, a positive interaction between pitch d′ and pitch range indicated that listeners with greater pitch sensitivity were more likely to fixate on competitors when words had a larger pitch range. This suggests that prosodic focus effects were enhanced for listeners who could perceive fine-grained pitch variations, leading them to shift fixations toward competitors when pitch was exaggerated. These results are in line with a growing body of research that shows that prosodic processing is not driven solely by categorical acoustic marking; rather, the acoustic properties of speech interact with individual auditory sensitivity to shape fixation patterns dynamically (Bramlett & Wiener, 2025; Bramlett et al., 2024; Jansen et al., 2023;

Roy et al., 2017).

Acoustics and auditory motor reproduction abilities

Beyond basic acoustic sensitivity, listeners' melody reproduction abilities further influenced how they processed prosodic focus. These higher-level perceptual skills moderated the extent to which fixations aligned with prosodic prominence in the speech signal especially at the later parts of the sentence (second phrases). In the first phrase target model, a negative interaction between pitch range and melodic reproduction ability revealed that listeners with lower melodic reproduction abilities exhibited fewer fixations to the target when pitch range was compressed. This suggests that prosodic focus effects were stronger for listeners who could reproduce pitch variation with greater precision—when the pitch range was reduced, those with weaker melodic skills failed to fixate as reliably on focus-marked words. A similar trend was observed in the first phrase competitor model, where a positive interaction between melody reproduction and pitch range indicated that listeners with greater melodic ability fixated more on competitors when pitch range was exaggerated. This suggests that musically skilled listeners were more sensitive to prosodic variation, shifting their attention in response to subtle acoustic differences. This strongly supports Jansen et al. (2023), which found positive effects of music on perception of L2 focus prosody and, more generally, supports claims for domain-general auditory processing (Bakkouche & Saito, 2025; Bramlett et al., 2024; Kachlicka et al., 2019; Saito et al., 2022).

Cognitive factors: Working wemory, cognitive control, and LexTALE

Despite their hypothesized role in prosodic processing, working memory, cognitive control, and lexical proficiency (LexTALE) did not significantly predict fixation behavior in our study. This does not necessarily mean that these predictors do not contribute to focus processing. There are at least three explanations for our null results. First, we did not have a large enough sample. This is probably true for our L1 Dutch-L2 English participants (N = 27). Second, our specific population was at or near ceiling in many tasks, such as LexTALE. The L1 Dutch-L2 English group scored very high on the LexTALE task, with many participants actually outperforming L1 English participants (though it is interesting to note despite this high lexical proficiency, the L2

participants' eye movements still indicate less efficient real-time processing). It is not clear if a wider range of English proficiency may lead to more varying results. Third, the look-and-listen task may require fewer cognitive resources than a word recognition task that involves multiple choices and greater attention. We tentatively conclude that our look-and-listen results were shaped more by perceptual and motor reproduction abilities than by cognitive resources.

Moving beyond L1-L2 differences

In sum, our exploration findings extend our replication and refinement by demonstrating that prosodic focus processing is not solely driven by language background (L1 vs. L2) but is shaped by a combination of acoustic properties of stimuli and listener-specific traits. While Ge et al. (2021a) attributed differences in focus processing to L1 effects, our results suggest that domain-general, listener-specific auditory and motor abilities play a more central role than previously assumed (Bakkouche & Saito, 2025; Bramlett et al., 2024; Kachlicka et al., 2019; Saito et al., 2022). While previous studies suggested that L2 listeners exhibit a uniform delay, our results show that fixation patterns are better predicted by individual differences in auditory sensitivity tied to specific acoustic properties of the stimuli (Xie et al., 2023). In other words, prosodic processing is highly individualized, and L2 delays may reflect perceptual and cognitive variability rather than a fixed group-level effect.

Finally, our results highlight the importance of including individual differences in L1 and L2 speech perception research. Traditional L1/L2 comparisons often treat L1 monolinguals as a control (Rothman et al., 2023), attributing differences in speech processing to categorical group distinctions without specifying the underlying mechanisms. However, this approach lacks parsimony, as it assumes that the L1 itself is the explanatory factor rather than identifying the perceptual and cognitive mechanisms that drive these differences. Our findings suggest that focus processing is better explained through mechanistic factors such as individual variation in acoustic sensitivity (e.g., pitch d', duration d') rather than broad L1 effects. This more parsimonious framework accounts for why some L2 speakers approach L1-like processing while some L1 speakers do not consistently exhibit the expected pattern. Rather than treating L1 effects as static,

we show that they emerge from individual differences in sensitivity to speech cues, aligning with adaptive models of speech perception (Xie et al., 2023).

Limitations

There are at least three notable deviations from our study and Ge et al. (2021a): differences between data collection methods, false positives-negatives, and group-level differences.

Regarding the difference in data collection methods, Ge et al., 2021a conducted their study in a controlled lab setting, using high-precision eye-trackers with different frame rates depending on the participant group. L1 Dutch speakers were tested in the Netherlands using an eye-tracker with a 500 Hz sampling rate, whereas L1 English speakers were tested in Hong Kong with a 300 Hz sampling rate. The use of web-based eye-tracking introduces variability in gaze data due to differences in participant screen sizes, webcam qualities, and environmental conditions. Although we implemented stringent calibration procedures and data filtering, our web-based sample had variable frame rates ranging from 5 Hz to 60 Hz, as is common in web-based eye-tracking (Bramlett & Wiener, 2024; Vos et al., 2022). The lower sampling rate compared to lab-based studies may have affected the temporal precision of fixation patterns.

Secondly, the presence of false positives and false negatives in either our study or Ge et al. (2021a) could contribute to discrepancies between our results and theirs. While our sample sizes (Dutch = 27, English = 61) are comparable to those of Ge et al. (2021a) (Dutch = 35, English = 40), the number of statistical tests conducted per language (9) increases the likelihood of Type I errors. Given this, there is a 59.34% probability of obtaining at least one false positive in both studies, making it crucial to interpret significant findings with caution. The risk of false negatives is more complex to quantify. If prosodic effects exist but are small, our sample sizes may be underpowered to detect them, leading to Type II errors. However, estimating this risk is particularly challenging due to the lack of established effect sizes for prosody in focus processing. Since this field is still emerging, future research should aim to establish reliable effect size estimates to improve statistical power calculations and minimize both false positives and false negatives. To clarify, we are not claiming that all effects across the two studies are merely false

positives but rather that being able to separate the false positives from real effects is not feasible at this stage. Similarly, while our statistical approach using Generalized Additive Models and LASSO-based feature selection provided a more nuanced analysis of fixation dynamics, it also introduced complexity in model interpretation. The inclusion of multiple individual difference measures allowed for a richer understanding of variability in focus processing, but further replications with larger and more diverse samples are needed to determine the generalizability of these findings.

Third, although we followed a recruitment strategy similar to Ge et al. (2021a), differences in L2 proficiency, English exposure, or individual cognitive-perceptual abilities between participant samples may have contributed to the discrepancies in findings. It should be noted that Ge et al. (2021a) recruited their L1 English speakers in Hong Kong, which means that they may not be directly comparable to other L1 English speaking populations. Exposure to or experience with tone languages like Cantonese and Mandarin may have influenced English focus processing patterns. The greater reliance on lexical pitch in these languages could have heightened participants' attentional allocation to prosodic cues in English, potentially enhancing sensitivity to focus marking. Indeed, there is evidence that L1 tonal experience improves L2 English stress perception to behavior better than that of L1 listeners (Choi, 2021; Choi et al., 2019). By contrast, our L1 participants were recruited online through prolific rather than through university cohorts. Such differences broaden the demographic profile but also introduce more heterogeneity. This is common in replication studies and should be considered alongside modality and analytical factors when interpreting variation between the current study and Ge et al. (2021a).

Conclusion

Replication studies often blur the line between confirmatory and exploratory analyses, complicating efforts to distinguish between fidelity and refinement (Yanai & Lercher, 2020). This lack of transparency raises an important question: should replications adhere to the original analyses, even if statistically flawed, or refine the analyses at the risk of introducing unintended differences (McManus, 2022)? Our FiREE framework addresses this challenge by stressing

fidelity and refinement while also encouraging exploratory extensions into theoretically relevant areas. Our FiREE replication demonstrated that prosodic focus processing is more variable than previously assumed and that strict replication alone may not capture the full picture. Rather than a binary success-or-failure, replications should be seen as an iterative process—one that strengthens methodological rigor while uncovering new theoretical insights.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Baddeley, A. (2003). Working memory and language: An overview. *Journal of communication disorders*, *36*(3), 189–208.

Bakkouche, L., & Saito, K. (2025). Effects of auditory processing, memory, and experience on early and later stages of second language speech learning. *Second Language Research*, 02676583251317909.

Baumann, S., & Winter, B. (2018). What makes a word prominent? predicting untrained german listeners' perceptual judgments. *Journal of Phonetics*, *70*, 20–38.

Bishop, J. (2021). Exploring the similarity between implicit and explicit prosody: Prosodic phrasing and individual differences. *Language and speech*, *64*(4), 873–899.

Blumenfeld, H. K., & Marian, V. (2014). Cognitive control in bilinguals: Advantages in stimulus–stimulus inhibition. *Bilingualism: Language and Cognition*, *17*(3), 610–629.

Bramlett, A. A., Brown, B., Dueck, J., & Wiener, S. (2024). Measuring music and prosody: Accounting for variation in non-native speech discrimination with working memory, specialized music skills, and music background [proceedings], 773–776. https://doi.org/10.21437/SpeechProsody.2022-157

Bramlett, A. A., & Wiener, S. (2024). The art of wrangling: Best practices for reporting web-based eye-tracking data in language research. *Linguistic Approaches to Bilingualism*, *1*, e2205v1. https://doi.org/https://doi.org/10.1075/lab.23071.bra

Bramlett, A. A., & Wiener, S. (2025). Individual differences modulate prediction of italian words based on lexical stress: A close replication and lasso extension of sulpizio and mcqueen (2012). *Journal of Cultural Cognitive Science*, 1–27.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes*, *25*(7-9), 1044–1098. https://doi.org/10.1080/01690965.2010.504378

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Carpenter, P. A., & Just, M. A. (2013). The role of working memory in language comprehension. In *Complex information processing* (pp. 51–88). Psychology Press.

Choi, W. (2021). Cantonese advantage on English stress perception: Constraints and neural underpinnings. *Neuropsychologia*, *158*, 107888.

Choi, W., Tong, X., & Samuel, A. G. (2019). Better than native: Tone language experience enhances english lexical stress discrimination in cantonese-english bilingual listeners. *Cognition*, *189*, 188–192.

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology*. University of Chicago Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cutler, A., & Fodor, J. A. (1979). Semantic focus and sentence comprehension. *Cognition*, *7*(1), 49–59. https://doi.org/10.1016/0010-0277(79)90010-6

Dimitrova, D. V., Stowe, L. A., Redeker, G., & Hoeks, J. C. (2010). Focus particles and prosody processing in dutch: Evidence from erps. *Speech Prosody*, *100979*, 1–4.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics*, *16*(1), 143–149.

Ferreira, F., & Karimi, H. (2015). Prosody, performance, and cognitive skill: Evidence from individual differences. *Explicit and implicit prosody in sentence processing: Studies in honor of Janet Dean Fodor*, 119–132.

Filik, R., Paterson, K. B., & Liversedge, S. P. (2005). Parsing with focus particles in context: Eye movements during the processing of relative clause ambiguities. *Journal of Memory and Language*, *53*(4), 473–495. https://doi.org/10.1016/j.jml.2005.07.002

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://www.jstatsoft.org/article/view/v033i01

Fung, R. S.-Y. (2000). *Final particles in standard cantonese: Semantic extension and pragmatic inference*. The Ohio State University.

Ge, H., Chen, A., & Yip, V. (2021a). Comprehension of focus-to-accentuation mapping in sentences with *only* by advanced cantonese learners and dutch learners of english. *Studies in Second Language Acquisition*, *43*(1), 25–49. https://doi.org/10.1017/S0272263120000140

Ge, H., Mulders, I., Kang, X., Chen, A., & Yip, V. (2021b). Processing focus in native and non-native speakers of english: An eye-tracking study in the visual world paradigm. *Applied Psycholinguistics*, *42*, 1057–1088. https://doi.org/10.1017/S0142716421000230

Ge, H., Lee, A. K. L., Yuen, H. K., Liu, F., & Yip, V. (2024). Bilingual exposure might enhance l1 development in cantonese–english bilingual autistic children: Evidence from the production of focus. *Autism*, *28*(7), 1795–1808.

Godfroid, A., Finch, B., & Koh, J. (2025). Reporting eye-tracking research in second language acquisition and bilingualism: A synthesis and field-specific guidelines. *Language Learning*, *75*(1), 250–294. https://doi.org/10.1111/lang.12664

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Gussenhoven, C. (1983). Focus, mode and the nucleus. *Journal of Linguistics*, *19*(2), 377–417. https://doi.org/10.1017/S0022226700007799

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods*, *50*, 1166–1186.

Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, *71*, 1–15.

Jansen, N., Loerts, H., Harding, E. E., Başkent, D., & Lowie, W. (2023). The influence of musical abilities on the processing of l2 focus prosody: An eye-tracking study. *20th International Congress of Phonetic Sciences*, 599–603.

Kachlicka, M., Saito, K., & Tierney, A. (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language*, *192*, 15–24. https://doi.org/10.1016/j.bandl.2019.02.004

Kiss, K. E. (1998). Identificational focus versus information focus. *Language*, *74*(2), 245–273. https://doi.org/10.2307/417867

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents* (Vol. 71). Cambridge University Press.

Lee, P. P.-l. (2019). *Focus manifestation in mandarin chinese and cantonese: A comparative perspective*. Routledge.

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–343. https://doi.org/10.3758/s13428-011-0146-0

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.*, *49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Luk, G., De Sa, E., & Bialystok, E. (2011). Is there a relation between onset age of bilingualism and enhancement of cognitive control? *Bilingualism: Language and cognition*, *14*(4), 588–595.

Magnuson, J. S. (2019). Fixations in the visual world paradigm: Where, when, why? *Journal of Cultural Cognitive Science*, *3*(2), 113–139.

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (leap-q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067)

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field [Open Access, Systematic Review Article]. *Language Learning*, *68*(2), 321–391. https://doi.org/10.1111/lang.12286

McManus, K. (2022). Replication research in instructed SLA. In L. Gurzynski-Weiss & Y. Kim (Eds.), *Instructed second language acquisition research methods* (pp. 104–122). John Benjamins Publishing Company. https://doi.org/10.1075/rmal.3.05mcm

McManus, K. (2024). The future of replication in applied linguistics: Toward a standard for replication studies [First View, Open Access]. *Annual Review of Applied Linguistics*, 1–17. https://doi.org/10.1017/S0267190524000011

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Ness, T., Langlois, V. J., Kim, A. E., & Novick, J. M. (2023). The state of cognitive control in language processing. *Perspectives on Psychological Science*, 17456916231197122.

Nosek, B. A., & Errington, T. M. (2020). The best time to argue about what a replication means? before you do it. *Nature*, *583*(7817), 518–520. https://doi.org/10.1038/d41586-020-02142-6

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). Webgazer: Scalable webcam eye tracking using user interactions. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 3839–3845. https://doi.org/10.1145/2702613.2702627

Poole, D., Grange, J. A., & Milne, E. (2024). Putting the spotlight back onto the flanker task in autism: Autistic adults show increased interference from foils compared with non-autistic adults. *Journal of cognition*, *7*(1), 46.

Porte, G., & McManus, K. (2018). *Doing replication research in applied linguistics* (1st) [eBook, 190 pages]. Routledge. https://doi.org/10.4324/9781315621395

Prolific. (2024). Prolific [Accessed: 2025-09-09].

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*, 6–1. https://doi.org/10.3765/sp.5.6

Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Duñabeitia, J. A., Gharibi, K., Hao, J., Kolb, N., Kubota, M., Kupisch, T., et al. (2023). Monolingual comparative normativity in bilingualism research is out of "control": Arguments and alternatives. *Applied Psycholinguistics*, *44*(3), 316–329.

Roy, J., Cole, J., & Mahrt, T. (2017). Individual differences and patterns of convergence in prosody perception. *Laboratory Phonology*, *8*(1).

Saito, K. (2023). How does having a good ear promote successful second language speech acquisition in adulthood? introducing auditory precision hypothesis-l2. *Language Teaching*, *56*(4), 522–538.

Saito, K., Cui, H., Suzukida, Y., Dardon, D. E., Suzuki, Y., Jeong, H., Révész, A., Sugiura, M., & Tierney, A. (2022). Does domain-general auditory processing uniquely explain the outcomes of second language speech acquisition, even once cognitive and demographic variables are accounted for? *Bilingualism: Language and Cognition*, *25*(5), 856–868.

Saito, K., Kachlicka, M., Sun, H., & Tierney, A. (2020). Domain-general auditory processing as an anchor of post-pubertal second language pronunciation learning: Behavioural and neurophysiological investigations of perceptual acuity, age, experience, development, and attainment. *Journal of Memory and Language*, *115*, 104168.

Saito, K., Kachlicka, M., Suzukida, Y., Mora-Plaza, I., Ruan, Y., & Tierney, A. (2024). Auditory processing as perceptual, cognitive, and motoric abilities underlying successful second language acquisition: Interaction model. *Journal of Experimental Psychology: Human Perception and Performance*, *50*(1), 119.

Sarrett, M. E., Shea, C., & McMurray, B. (2022). Within-and between-language competition in adult second language learners: Implications for language proficiency. *Language, Cognition and Neuroscience*, *37*(2), 165–181.

Sinagra, C., & Wiener, S. (2022). The perception of intonational and emotional speech prosody produced with and without a face mask: An exploratory individual differences study. *Cognitive Research: Principles and Implications*, *7*(1), 89.

Tierney, A., & Kraus, N. (2014). Auditory-motor entrainment and phonological skills: Precise auditory timing hypothesis (path). *Frontiers in human neuroscience*, *8*, 949.

Tierney, A., White-Schwoch, T., MacLean, J., & Kraus, N. (2017). Individual differences in rhythm skills: Links with neural consistency and linguistic ability. *Journal of cognitive neuroscience*, *29*(5), 855–868.

Traxler, M. J. (2009). A hierarchical linear modeling analysis of working memory and implicit prosody in the resolution of adjunct attachment ambiguity. *Journal of psycholinguistic research*, *38*, 491–509.

Tremblay, A., Broersma, M., Coughlin, C. E., & Choi, J. (2016). Effects of the native language on the learning of fundamental frequency in second-language speech segmentation. *Frontiers in psychology*, *7*, 985.

Tremblay, A., Kim, S., Shin, S., & Cho, T. (2021). Re-examining the effect of phonological similarity between the native-and second-language intonational systems in

second-language speech segmentation. *Bilingualism: Language and cognition*, *24*(2), 401–413.

Ushey, K., Allaire, J., & Tang, Y. (2022). *Reticulate: Interface to 'python'* [R package version 1.26]. https://CRAN.R-project.org/package=reticulate

Vos, M., Minor, S., & Ramchand, G. C. (2022). Comparing infrared and webcam eye tracking in the Visual World Paradigm. *Glossa Psycholinguistics*, *1*(1), 1–37. https://doi.org/10.5070/g6011131

Wang, L., Bastiaansen, M., Yang, Y., & Hagoort, P. (2011). The influence of information structure on the depth of semantic processing: How focus and pitch accent determine the size of the n400 effect. *Neuropsychologia*, *49*(5), 813–820. https://doi.org/10.1016/j.neuropsychologia.2010.12.035

Wood, S. (2017). *Generalized additive models: An introduction with r* (2nd ed.). Chapman; Hall/CRC.

Wu, W. L., & Xu, Y. (2010). Prosodic focus in hong kong cantonese without post-focus compression. *Speech prosody*, *2010*, 1–4.

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex*, *166*, 377–424. https://doi.org/10.1016/j.cortex.2023.05.003

Yanai, I., & Lercher, M. (2020). A hypothesis is a liability. *Statistical Modeling, Causal Inference, and Social Science*, (231), 1–5. https://doi.org/10.1186/s13059-020-02133-w

Yap, M. J., Balota, D. A., Sibley, D. E., & Ratcliff, R. (2012). Individual differences in visual word recognition: Insights from the English Lexicon Project. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 53. https://doi.org/10.1037/a0024177

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, *27*(4), 567–595.