Studies in Second Language Acquisition (2024), 1–27
doi:10.1017/S0272263124000421

METHODS FORUM

Task-generated processes in second language speech production: *Exploring the neural correlates of task complexity during silent pauses*

Andrea Révész[1], Hyeonjeong Jeong[2], Shungo Suzuki[3], Haining Cui[4], Shunsui Matsuura[5], Kazuya Saito[6] and Motoaki Sugiura[7]

[1]University College London; [2]Tohoku University; [3]Waseda University; [4]McGill University; [5]Kyoto University; [6]University College London and [7]Tohoku University
Corresponding author: Andrea Révész; Email: a.revesz@ucl.ac.uk

(Received 07 January 2024; Revised 04 April 2024; Accepted 20 May 2024)

Abstract

The last three decades have seen significant development in understanding and describing the effects of task complexity on learner internal processes. However, researchers have primarily employed behavioral methods to investigate task-generated cognitive load. Being the first to adopt neuroimaging to study second language (L2) task effects, we aimed to provide novel insights into the neural correlates of task-related variation in L2 oral production. To advance research methodology, we also tested the utility of a neuroimaging technique, functional magnetic resonance imaging (fMRI), in examining the impact of task-related variables on L2 speech production when combined with cognitive–behavioral tools (speech analysis, expert and learner judgments). Our research focus was the effects of task complexity on silent pausing. Twenty-four Japanese learners of English completed eight simple and complex versions of decision-making tasks, half in their first language and half in their L2. The dataset for the present study included the L2 speech and fMRI data, expert judgments, and participants' difficulty ratings of the L1 and L2 tasks they completed. Based on our findings, we concluded that brain imaging and L1 task difficulty ratings were more sensitive to detecting task complexity effects than L2 self-ratings and pausing measures. These results point to the benefits of triangulating cognitive and neural data to study task-based neurocognitive processes.

Keywords: fluency; L2 production; neuroimaging; pausing; task complexity

Introduction

In the past decade, second language (L2) researchers have shown an increased interest in identifying methods suitable for investigating task effects on L2 performance (see Révész, 2021a, for a review). This enhanced focus on techniques to examine task-based production has stemmed from several sources. On the theoretical front, there is a growing recognition among L2 researchers that, to test theoretical frameworks of task-based performance and learning (e.g., Skehan, 1998; Robinson, 2001a), it is essential to provide validity evidence for all constructs invoked in them, including the independent and dependent variables and the causal processes posited to mediate links between them (Kane, 2006; Messick, 1995; Norris & Ortega, 2003; Révész, 2014). Tapping these various aspects of task-based models calls for a careful selection of research methods. The thorough and valid testing of task effects is also imperative from the perspective of pedagogy, as the resulting research outcomes are intended to credibly inform L2 teaching. For example, building our knowledge base about the impact of task variables on L2 performance may assist in discovering sources of task-related difficulty and in assessing the potential of manipulating tasks to generate learning opportunities assumed to foster L2 learning. Finally, in the spirit of the "methodological turn" in L2 research (Byrnes, 2013, p. 825), the study of methodological issues merits attention in its own right to help increase rigor in our field, thereby raising the validity of our research.

Against this background, this study had two primary aims. First, being the first to adopt neuroimaging to study task effects on L2 spontaneous oral production, we intended to provide novel insights into the neural correlates of task-related variation in L2 oral production. Second, to advance L2 research methodology, our goal was to test the utility of a neuroimaging technique (fMRI) in examining the impact of task-related variables on L2 speech production when combined with cognitive–behavioral tools (speech analysis, expert and learner judgments). The focus of our research was task effects on silent pausing (pausing henceforth), a well-studied and common phenomenon of L2 speech. In particular, we investigated how pausing behaviors and associated neural processes might vary as a function of task complexity (i.e., the inherent cognitive demands of tasks), motivated by previous research on pausing and cognitive models of speech production and task-based performance.

Background

Models of speech production and task-based performance

Psycholinguistic models of speech production (Kormos, 2006, Levelt, 1999) typically see speaking as involving several different but incremental stages. The first stage, conceptualization, entails generating a preverbal plan through macro- and microplanning. During macroplanning, the speaker decides on the information to be presented and the order in which it will be expressed. As part of microplanning, they further elaborate the preverbal plan by specifying the informational perspective, including the focus, the argument structure and semantic relations, and the mood of the message. While macroplanning is assumed to be language general, microplanning is presumed to entail language-specific information, as the conceptual features to be encoded are dependent on language (e.g., tense) (DeBot, 1992; Kormos, 2006). The next stage, formulation, begins with lexical encoding, which involves pairing the conceptual specifications with the appropriate lemmas from the mental lexicon. There is an ongoing debate regarding lexical selection, depending on the theoretical assumptions of speech production. Some modular models presume that language cues are assigned to each concept of the preverbal message so that the subsequent lemma selection process is achieved with all the necessary conceptual information for identifying matching lexical entries (Kormos, 2006; Pouliesse & Bogaert, 1994). Others, however,

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

also include an intermediary stage, the verbalizer, between conceptualization and formulation to assist with mapping conceptual features and lemmas, assuming that the preverbal message is not language-specific and the intended language for lexical entries is specified as a result of activation in the mental lexicon by the conceptual features (De Bot & Schreuder, 1993). Once the appropriate lemmas have been identified, their grammatical properties are retrieved for morphosyntactic encoding, that is, building the surface structure of the message drawing on syntactic and morphological rules. Formulation ends with encoding the message into a plan of articulatory movements corresponding to the phonological representation of the message. In the final stage, articulation, the speaker executes the planned articulatory gestures to produce the overt speech. Throughout all three stages, the speaker engages in self-monitoring, checking whether the output of each stage (e.g., preverbal message, overt speech) matches their communicative intention. While the macroplanning stage of conceptualization is expected to pose similar demands on speakers with varied proficiency, lower-proficiency speakers will likely struggle more with microplanning and formulation, given their more limited L2 knowledge and partially automated processing skills (Suzuki & Kormos, 2023).

Partly drawing on models of speech production, two cognitive–interactionist frameworks, Robinson's (2001a, 2011) Cognition Hypothesis and Skehan's (1998, 2009) Limited Capacity Model, have been proposed to theorize task effects on L2 oral performance and learning. The primary independent variable in each model is cognitive task demands, which is referred to as cognitive complexity by Skehan (1998) and as task complexity by Robinson (2001a). Skehan (2009) postulates that manipulating task-related features may create differential pressures on conceptualization and/or formulation processes during speech production, and the quality of the resulting linguistic performance depends on the extent to which the conceptualizer and/or formulator can deal with the cognitive demands of the task against attentional or working memory constraints. Inspired by a multiple-resources account of attention besides models of speech production, Robinson (2011) posits that increasing the cognitive complexity of tasks will not only affect speech production processes predictably but also interactional patterns and the processing and retention of task-relevant input. In both frameworks, the principal dependent variable is the linguistic outcome of task-based performance described in terms of linguistic complexity, accuracy, and fluency. For fluency, both models hypothesize that when the cognitive demands of a task increase, the fluency of oral performance will decrease because L2 learners are likely to engage with controlled processing. Thus, the specific prediction for pausing, a breakdown feature of fluency (Skehan, 2003; Tavakoli & Skehan, 2005), is that more cognitively demanding tasks will lead to longer and more frequent pausing.

Based on previous empirical work on pausing, Skehan and Robinson's prediction might be further refined by taking pause location into account. Researchers (de Jong, 2016; Field, 2011; Lambert et al., 2017; Suzuki & Kormos, 2023; Tavakoli et al., 2017, 2020) have proposed that, depending on the location at which pauses occur, they are more or less likely to reflect certain speech production processes. In particular, there seems to be a greater likelihood that mid-clause pausing is associated with formulation processes, such as lexical encoding, syntactic, and phonological encoding, whereas end-clause pausing relates to conceptualization. Thus, following Robinson (2001a) and Skehan (2009), it might be hypothesized that task manipulations that increase pressure on the conceptualizer will lead to more frequent and longer end-clause pauses. On the

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

other hand, task demands exerting enhanced strain on the formulator will result in greater incidence and length of mid-clause pausing.

To date, only two empirical studies, Wang (2014) and Lambert et al. (2017) provide information about the validity of these predictions about the relationship between task manipulations and pausing behavior, both investigating task repetition effects focusing on pause length and frequency respectively. In Wang's (2014) research, participants repeated an oral task twice, whereas Lambert et al. (2017) involved L2 learners in repeating a series of tasks six times. Both studies found that end-clause pausing decreased from the first to the second task performance, but Lambert et al. (2017) discovered no change in the incidence of end-clause pauses for further repetitions. For mid-clause pausing, neither Wang's (2014) nor Lambert et al.'s (2017) research yielded an effect for repeating a task once. Lambert et al. (2017), however, observed a reduced rate of mid-clause pauses when comparing participants' first and third task performance. Drawing on models of speech production (Kormos, 2006; Levelt, 1989), Lambert et al. (2017) interpreted these findings as suggesting that the first repetition helped ease pressure on conceptualization processes, whereas the second task repetition assisted learners in carrying out more efficient linguistic encoding. While these conclusions seem logical, neither of these studies has provided direct evidence for the speech production processes assumed to explain the effects of task repetition on pausing patterns.

Validity considerations in assessing links between task factors and L2 performance

In general, previous research investigating the impact of task-related variables on L2 performance has dedicated relatively little attention to the causal processes that mediate the relationship between task manipulations and linguistic measures of task performance. However, as Norris and Ortega (2003) highlighted and other models of validation (Kane, 2006; Messick, 1995) also imply, if researchers would like to reach solid and valid conclusions about theoretical predictions regarding task effects on L2 performance, it is crucial to obtain validity evidence for measurement of every construct involved in the predictions, including the task variable in focus (independent variable), the indices of task performance used to assess the impact of the task variable (dependent variable), and the causal processes hypothesized to mediate links between the task variable studied and the performance measures (mediator).

Recently, L2 researchers have dedicated a lot of effort to aligning task-based research with this methodological recommendation. Most research attention has been allocated to identifying valid ways of selecting measures of linguistic complexity, accuracy, and fluency to assess task-based predictions (e.g., Bulté & Housen, 2012; Housen & Kuiken, 2009; Housen, Kuiken & Vedder, 2012; Norris & Ortega, 2009). Of particular relevance to the current research is previous validation work on measures of fluency, especially research focusing on the previously mentioned distinction between mid- and end-clause pausing. To date, most validation studies of fluency have drawn on Segalowitz's (2010) fluency framework, which describes fluency as including three different but interrelated subconstructs: cognitive fluency has to do with how efficiently the cognitive mechanisms underlying speech performance operate; utterance fluency refers to the observable aspects of oral performance including pausing, speed, and hesitation; and perceived fluency captures the listener's judgments about the speaker's cognitive fluency.

Conceptualized in terms of this framework, previous research has yielded at least three types of validity evidence that support the value of distinguishing between mid- and end-clause pausing. First, empirical studies of utterance fluency found that L2

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

speakers, as compared to first language (L1) speakers (de Jong, 2016; Duran-Karaoz & Tavakoli, 2020; Felker et al., 2019; Kahng, 2014; Riazantseva, 2001; Skehan & Foster, 2007; Tavakoli, 2011) and to more proficient L2 speakers (Duran-Karaoz & Tavakoli, 2020; Tavakoli et a., 2017, 2020), pause more *often* in the middle of clauses, but show similar patterns in terms of end-clause pausing. Second, researchers exploring the contribution of L2 cognitive fluency to L2 utterance fluency concluded that the frequency of mid-clause pausing is the strongest representative of the construct of L2 breakdown fluency (Kahng, 2020; Suzuki & Kormos, 2023). Finally, extant research on links between L2 perceived and utterance fluency has revealed a key role for mid-clause (but not end-clause) pausing in capturing L2 fluency (Kahng, 2018; Saito et al., 2018; Suzuki & Kormos, 2020; Suzuki et al., 2021). Overall, these results are in line with the theoretical assumption that mid-clause pausing mirrors the efficiency of L2 formulation and end-clause pauses reflect the operations of the conceptualizer (de Jong, 2016; Field, 2011; Lambert et al., 2017; Suzuki & Kormos, 2023; Tavakoli et al., 2017, 2020). In line with these empirical observations, L2 speakers are expected to experience more difficulty with formulation due to their developing proficiency, whereas conceptualization processes are predicted to be less influenced by L2 skills. Taken together, past work on fluency suggests that adopting mid- and end-clause pausing patterns as dependent variables may enable gaining valid insights about the potential effects of task complexity on L2 performance.

Besides finding valid methods to gauge linguistic performance, a growing amount of task-based research has been concerned with the issue of supplying valid evidence for the task manipulation(s) under scrutiny (Norris, 2010; Norris & Ortega, 2003; Révész, 2014). For example, studies of task complexity more and more frequently employ independent measures of mental effort or cognitive load to gauge the validity of the task manipulation they investigate. Some researchers have used subjective methods to assess the amount of mental effort learners exerted during task performance, eliciting learner self-reports (e.g., Robinson, 2001b) or expert judgments of task difficulty (e.g., Révész et al., 2014; Révész et al., 2016). A small number of studies (e.g., Lee, 2019; Révész et al., 2016; Sasayama, 2016, Xu et al., 2022) have additionally used objective tools, which involved observing learners' behaviors during task performance (e.g., dual-task methodology, eye-tracking). Notably, reflecting an increased concern with methodological issues in task-based research (e.g., Mackey, 2020; Norris, 2010; Révész, 2014, 2021a, 2021b), a few studies specifically defined their goal as to assess the usefulness of different techniques, alone or in combination, to capture the mental effort or cognitive load imposed by task demands. Like the current research, Révész et al. (2016) and Sasayama (2016) focused on oral production, triangulating data collected through the dual-task methodology and self-ratings of mental effort and task difficulty. Révész et al. (2016) additionally elicited expert judgments and Sasayama's (2016) time estimations to assess task complexity effects on cognitive load during oral task performance. In both projects, dual-task methodology and the various subjective methods yielded converging results overall, indicating that the task versions designed to be more complex were indeed more cognitively demanding. Given the parallel results generated by the objective and subjective methods in these validation studies, it appears justified to use subjective tools to assess task difficulty, as they are easier to administer and are nonobtrusive.

To date, as compared with validation research on measures of task complexity and linguistic performance, relatively few studies have focused on methods to assess the causal processes assumed to mediate relationships between task complexity and linguistic output (Robinson, 2001a; Skehan, 2009). Similar to related work on validating

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

task complexity manipulations, the small amount of research available on task processes has assessed the utility of various subjective and objective methods (see Révész, 2021b, for a review). Studies on oral production, in particular, have utilized questionnaires (Révész, 2009; Sasayama & Norris, 2019), interviews (Ortega, 2005; Pang & Skehan, 2014), and stimulated recall protocols (Kim et al., 2015; Révész, Kourtali et al., 2017; Torres, 2018) to obtain information about learners' subjective experiences during task performance. To gain more objective insights, researchers have relied on dual-task methodology (Révész et al., 2016; Sasayama, 2016) and eye-tracking (Révész et al., 2014) to investigate speech production processes. More recently, neuroimaging has additionally been suggested as an objective tool that could be useful for investigating task-based processing (Révész, 2021b). However, its utility for this purpose has not yet been evaluated. A principal aim of this study was to begin exploring the capacity of neuroimaging to provide insights into task-generated processes and to complement and extend existing insights that have been gained through behavioral methods.

Neuroimaging as a potential way to tap task-generated cognitive processes

The specific neuroimaging technique we intended to explore is functional magnetic resonance imaging (fMRI). Simply put, fMRI captures when there are increases in blood flow resulting from heightened brain activity. When a greater amount of blood is supplied to a certain area of the brain, this neural activity is detected by the fMRI scanner. In the past, researchers have predominantly employed fMRI to investigate language user's neural activity during input processing in tightly controlled experiments. Few studies have used fMRI to investigate neural processes involved in naturalistic language use, with even fewer studies focusing on the cortical mechanisms called upon during spontaneous oral language production.

Among these studies, two previous L1 experiments are worth highlighting here, as their speech elicitation technique was similar to the one employed in our research. Based on the same dataset, Morales et al. (2022) and Wu et al. (2022) set out to compare brain activation patterns during the processing and production of L1 naturalistic discourse, to identify the neural correlates of psycholinguistic characteristics of speech (e.g., coherence, lexical complexity, emotional content). In an fMRI machine, participants were asked to orally respond to common topics (e.g., describe how to make a coffee) and listen to speech samples on comparable themes. The researchers revealed that, during both production and comprehension, brain activation patterns were associated with several psycholinguistic properties of the language produced or listened to by the participants. For example, Morales et al. (2022) found that, in speaking and listening alike, certain areas involved in the theory of mind network (e.g., medial prefrontal cortex [mPFC], precuneus, anterior temporal, and lateral parietal cortex) showed greater activation when the discourse was less coherent. The researchers, however, also discovered modality-specific effects depending on the properties of semantic information. For instance, Wu et al. (2022) observed that action-related content was processed in the sensory–motor area in both speech production and comprehension, but for emotional content, there was more increased activation in emotion-related areas (e.g., anterior cingulate cortex and insula) during speech production as compared with comprehension.

The neural correlates of L2 spontaneous speech production are even less explored than those involved in L1 speech. Among the few L2 studies (Jeong et al., 2011; Jeong et al., 2016), Jeong et al.'s (2016) study is closest in focus to this research. The purpose of this study was to identify brain activation patterns associated with communicative as compared with noncommunicative oral production. The experiment involved participants in watching short videos, in which an actor interacted with an object (e.g., played a guitar). In the communicative condition, participants were instructed to talk to the actor in the video, whereas, under the noncommunicative condition, their task was to describe the actor's situation. The fMRI analyses compared brain activation patterns when participants produced L1 and L2 across the communicative and descriptive speech conditions, as a function of language status (L1 versus L2) and L2 oral proficiency. The researchers found that both L1 and L2 speech production enhanced activation in certain brain areas only during communicative activities. These included regions involved in the theory-of-mind system (e.g., mPFC, precuneus, posterior superior temporal sulcus [pSTS]), retrieval and integration of concepts (left angular gyrus [AG]), and semantic retrieval (e.g., left middle temporal gyrus [MTG]). Notably, the left posterior supramarginal gyrus (SMG), an area associated with the planning of speech acts, was activated only during L2 communicative production. As expected, the study also yielded L2 oral proficiency effects for brain areas related to lexical and semantic retrieval (e.g., left MTG) during L2 communicative production. In addition, during L2 production, as expected, greater activation was observed in areas associated with syntactic and phonological processing (e.g., left inferior frontal gyrus [IFG]) than during L1 speech production irrespective of condition.

In sum, previous research suggests that neuroimaging has the potential to yield insights into the type of processing in which speakers primarily engage during speech production. Specifically, it appears that certain brain areas can be linked to processes associated with conceptualization (e.g., theory-of-mind network) and other regions to linguistic encoding processes, even during short sentence production. If so, we would expect that task manipulations that pose increased demands on the conceptualizer and formulator will enhance brain activity in areas related to conceptualization and language, respectively. A principal aim of this study was to explore if, as predicted, fMRI scans are indeed sensitive to task complexity manipulations.

Research questions and hypotheses

We formed two research questions to investigate the impact of task complexity on L2 pausing behaviors and associated neural processes:

RQ1: To what extent does task complexity influence silent pause frequency and length (mid- versus end-clause) during L2 speech production?

RQ2: To what extent does task complexity influence neural processes during silent pauses (mid- versus end-clause) during L2 speech production?

Inspired by models of speech production (Kormos, 2006; Levelt, 1999) and task-based performance (Robinson, 2001a; Skehan, 2009), we assumed that more complex tasks would put greater pressure on conceptualization processes, due to the increased conceptual demands they pose. In turn, we expected that the enhanced conceptual demands during more complex tasks would result in fewer attentional resources available for linguistic encoding processes, leading to increased pressure on the formulator. As mid-clause and end-clause pauses have been associated with greater engagement in formulation and conceptualization respectively (de Jong, 2016; Field, 2011; Lambert et al., 2017; Suzuki & Kormos, 2023; Tavakoli et al., 2017, 2020), we

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

hypothesized that these presumed task complexity effects would affect the behavioral and neural correlates of mid- and end-clause pausing as follows:

H1: More complex tasks will elicit longer and more mid-clause and end-clause pauses.

H2: During more complex tasks, there will be greater activation in conceptualization (i.e., theory-of-mind network) and language-related brain areas (e.g., left IFG) during pauses, with stronger effects for end-clause and mid-clause pauses respectively.

We also formed a methodology-related research question and hypothesis, exploring the extent to which our subjective measures of task complexity (expert judgments, speaker ratings) would yield converging results with those obtained through objective behavioral (pause length and frequency) and neuroimaging measures (brain activation):

RQ3: To what extent do subjective ratings of task complexity relate to the frequency and length of pausing and the neural correlates of pausing?

H3: Higher task complexity ratings will relate to greater pause length and frequency and greater activation in conceptualization- and language-related brain areas during pauses.

Methodology

Design

The data for this study comes from a larger dataset. The participants were 26 Japanese users of L2 English. All 26 participants carried out eight oral tasks altogether in an fMRI scanner. Each participant completed the low-complexity version of four of the tasks and the high-complexity version of the other four tasks, half of them in L2 English and the other half in L1 Japanese. Task complexity and language were counterbalanced across the participants. We could only include data for 24 L2 performances and 21 L1 performances in our analyses here; we had to exclude the rest of the data due to the poor quality of some speech recordings and excessive head movement during scanning. Immediately after carrying out a task, participants were asked to provide task difficulty and mental effort ratings on a 9-point Likert scale. Two experts also gave judgments about the anticipated difficulty and mental effort posed by the tasks. Prior to performing the experimental tasks, participants completed a background questionnaire and the listening part of the Oxford Placement Test (Allen, 2004).

The main focus of the current study is the participants' performance on the four tasks they carried out in their L2 and the self-ratings they provided for these ($n = 96$). In addition, as a subjective measure of task complexity, we also included participants' difficulty and mental effort ratings for the four tasks they completed in their L1 ($n = 84$). Given the counterbalanced design, participants did not complete the same four tasks in L2 English and L1 Japanese. Nevertheless, each of the eight tasks is equally represented in the L2 and L1 datasets, with any potential task effects controlled for through counterbalancing.

Participants

All the participants were undergraduate students at a Japanese university. The mean age was 20.33 years for both the L2 and L1 performances (L2: $SD = 1.43$, L1: $SD = 1.46$)

with a range of 19–24. Participants were nearly equally distributed in terms of gender (L2: 10 female, 14 male; L1: 11 female, 10 male). All L2 participants had been studying English as a first foreign language in formal school settings for an average of 9.66 years (*SD* = 2.82). Out of the 24 L2 participants, only three had studied abroad for one month after turning 18. The rest of the students had no study-abroad experience. The English proficiency levels of the L2 participants were in the B1–B2 bands on the Common European Framework of Reference (CEFR) scale, as determined by the listening component of the Oxford Placement Test (*M* = 81.08, *SD* = 6.43).

Instruments and Procedures

Tasks and Task Complexity Manipulations

The eight experimental tasks took the form of decision-making monologic tasks. Half of the eight tasks required participants to select four essential items from a list of eight to take with them in critical situations: having to swim to a desert island when their boat was sinking, to walk to the nearest emergency shelter after surviving an earthquake, to drive to an emergency accommodation upon receiving a flood alert, and to get to the closest camp when surviving a plane crash. The remaining four tasks required participants to select five people from a set of eight as part of further disaster situations: deciding who should receive a potentially life-saving vaccination, who should take the parachutes from a plane about to crash, who to save first from a building on fire, and who to select as government advisors in a health emergency (see the Supporting Information online for each task). Two applied linguistics experts, also experienced language teachers in the Japanese context, assisted with making the tasks and choice of items/people culturally appropriate for the participants. We also piloted the tasks with participants similar in demographics to the actual participants.

For each task, we designed simple and complex versions. In the simple versions of the tasks, the decisions among items or people were designed to be more straightforward. For example, a bottle of wine could be more easily eliminated than a bottle of water or a smoking advertiser than a doctor. Two highly experienced task-complexity researchers were asked to judge whether the task versions designed to be more complex indeed involved more complex decisions. Both experts evaluated all the task versions intended to be more complex as more cognitively demanding than their simple counterparts, resulting in 100% agreement between the two experts.

Self-rating scales

The self-rating scales evaluated participants' perceptions of (a) the mental effort required by the task and (b) the difficulty of the task. They were instructed to judge each statement on a 9-point Likert scale immediately after carrying out a task version. The questionnaire items were presented to the participants in English for the L2 performances and in Japanese for the L1 performances. The items were worded as follows:

This task required no mental effort at all.	1	2	3	4	5	6	7	8	9	This task required extreme mental effort.
This task was not difficult at all.	1	2	3	4	5	6	7	8	9	This task was extremely difficult.

Data collection

The participants took part in one individual session. First, we obtained informed consent, followed by the administration of a paper-and-pencil background questionnaire (10 min) and the Oxford Placement Listening Test (15 min max). The rest of the experiment took place in the fMRI scanner. First, participants were introduced to the task instructions and experimental procedures. As part of this practice phase, they read the task instructions, listened to a sample practice task performance, carried out the practice task, and completed the task perception questionnaire. Participants were also given detailed guidance on how to reduce head movements while undergoing the fMRI scans. The practice trial within the MRI machine was also aimed at familiarizing participants with how to control head movements while speaking. To further restrict head motion, participants' heads were secured with a combination of a foam pad and a restraint belt. Participants were encouraged to ask any questions they had regarding the procedures.

Next, participants moved on to completing the two experimental sessions, English and Japanese, inside the MRI. In each session, they carried out two simple and two complex tasks, presented in a counterbalanced order across participants. During each trial, participants had 1 minute to review the task instructions, 2 minutes to carry out their oral performance, and 10 seconds to complete the self-rating scales. Participants were asked to verbally provide their ratings. There was a 15-second rest between trials. The total time for the fMRI experiment was 834 seconds for each session. Participants' spoken responses were captured using an MRI-compatible noise-cancelling microphone (Optoacoustics Ltd., Moshav Mazor, Israel).

Scanning was performed using a 3T Philips Achieva dStream scanner. Functional images were acquired using gradient-echo planer image sequences with the following parameters: echo time=30 ms, flip angle=80°, slice thickness=3 mm, field of view=192 mm, 64×64 matrix. Thirty-two axial slices spanning the entire brain were obtained every 2 seconds. After excluding three dummy scans performed due to the T1 saturation effect, 417 volumes were obtained for each participant and session. T1-weighted anatomical images were also acquired from each participant to serve as a reference for anatomical correlates. The following preprocessing procedures were performed using Statistical Parametric Mapping (SPM12) software (Wellcome Centre for Human Neuroimaging, London, UK) and MATLAB (MathWorks, Natick, MA, USA): adjustment of acquisition timing across slices, correction for head motion, coregistration to the anatomical image, spatial normalization using the anatomical image and the MNI template, and smoothing using a Gaussian kernel with a full width at a half maximum (FWHM) of 8 mm.

Data analyses

Behavioral analyses

All 96 L2 speech performances were transcribed and then annotated for pauses by the fourth author. Given that only a few filled pauses were identified, our further analyses focused on silent pausing. Silent pauses were identified manually, given that the audio data obtained from the fMRI scanner were noisy to allow for automatic detection of pauses. Following previous research (Goldman-Eisler, 1968), the threshold for silent pauses was defined as 250 ms. We annotated the data for clause boundaries in TextGrid files using the Praat software (Boersma & Weenink, 2022). Then, we coded the pauses according to whether they appeared within or between clauses. To check reliability, we randomly selected three participants (12.5% of the data), and the first author also coded their speech samples. The intercoder agreement was high for all coding categories (pause identification:

100%, clause boundary: 99.5%, and pause location: 99.6%). Once we completed the coding, we divided the number of pauses by the total number of clauses for each task and used the resulting proportions in further analyses involving pause frequency.

fMRI analyses

In our study, we conducted a detailed fMRI analysis using SPM12, employing conventional within-participant (first-level) and between-participant (second-level) analyses. Starting at the first level, we performed a voxel-by-voxel multiple regression analysis in the time courses to estimate brain activation for each participant. We focused on the hemodynamic response during L2 task performances, hypothesizing variations in brain activity related to pause locations (mid-clause and end-clause) during simple and complex tasks. To quantify these variations, we constructed a design matrix incorporating four regressors corresponding to different task conditions: simple mid-clause (SM), simple end-clause (SE), complex mid-clause (CM), and complex end-clause (CE). These regressors were defined by the timing of silent pause onsets and duration of pauses, information derived from behavioral speech analysis for each participant. Additionally, six movement parameters (three translations and three rotations) were included as noninterest regressors. For each participant, contrast images were generated to compare the effect of pause location ([SE+CE > CM +SM] and [CM+SM > SE+CE]), task complexity ([CE+CM > SE+SM]), and their interactions ([CE -CM > SE – SM]) and ([CM- CE > [SM – SE]). These contrast images distilled the essence of brain activity differences for our specific hypotheses, including the main effects and interactions within our 2×2 factorial design.

Transitioning to the second level of analysis, we applied a one-sample t-test to these contrast images across all participants. This crucial step was aimed at identifying whether the patterns of brain activity changes we observed were consistent and significantly different from zero across the group. We employed a random effects model to conduct statistical inference on the contrasts of parameter estimates, ensuring that our findings could be generalized to the broader population. This structured approach, from individual-level analyses to group-level statistical inferences, allowed us to comprehensively investigate the neural correlates of task complexity and pause locations in L2 speech tasks.

To address the challenge of multiple comparisons inherent in fMRI data, we set a statistical threshold of p <.05, implementing whole-brain cluster size correction as recommended by Slotnick (2017). A Monte Carlo simulation with 10,000 iterations on a 64×64×32 whole-brain grid, smoothed with an 8-mm FWHM Gaussian kernel, established a voxel threshold of p <.001. This threshold was corrected to p <.05 with a cluster extent threshold of 45 voxels, ensuring the robustness of our findings. Activation peak coordinates were reported in the Montreal Neurological Institute space and identified using the automated anatomical labeling atlas in SPM12. The Marsbar toolbox (Brett et al., 2002) was used to extract parameter estimates in the four conditions for each participant to illustrate the activation profile in the observed brain area.

Results

Preliminary analyses

Task complexity and self-ratings

Table 1 provides the descriptive statistics for the mental effort and task difficulty self-ratings given by participants across simple and complex task versions and L2 and L1

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

Table 1. Ratings of mental effort and task difficulty.

				95% *CI*	
Measure	*N*	*M*	*SD*	Lower	Upper
L2 English					
Mental effort					
Simple tasks	48	5.56	1.93	4.75	6.38
Complex tasks	48	5.38	2.09	4.49	6.26
Task difficulty					
Simple tasks	48	5.54	2.04	4.68	6.40
Complex tasks	48	5.81	1.73	5.08	6.54
L1 Japanese					
Mental effort					
Simple tasks	42	3.57	2.55	2.80	4.34
Complex tasks	42	3.83	2.22	3.16	4.50
Task difficulty					
Simple tasks	42	3.21	2.09	2.58	3.85
Complex tasks	42	3.67	2.15	3.02	4.32

performances. We built a series of linear mixed-effects models, for the L1 and L2 data separately, to investigate the extent to which task complexity affected participants' self-ratings. The fixed effect in our models was task complexity, and participants and task prompt served as random effects. The dependent variable was the self-rating of mental effort or task difficulty in the models. As shown in Table 2, task complexity did not emerge as a significant predictor of either mental effort or task difficulty self-ratings for the L2 task performances. However, the L1 data yielded a significant difference in task difficulty ratings for the simple and complex task versions. In sum, the participants did not perceive the simple and complex tasks as requiring differential mental effort or as different in difficulty when they completed them in their L2. However, they perceived the complex task versions as more difficult (but not requiring more mental effort) when completing them in their L1. Task complexity explained approximately 2% of the variance in L1 self-ratings of task difficulty.

Table 2. Results for models examining the effects of task complexity on ratings of mental effort and difficulty.

	Fixed effects				Random effects		Model effect size	
Predictor	*Est*	*SE*	*t*	*p*	Factor	*SD*	R^2m	R^2c
L2 English								
Mental effort								
Task complexity	0.19	0.30	0.63	0.53	Part (Int) Task(Int)	1.24 0.59	<.01	0.48
Task difficulty								
Task complexity	–0.26	0.25	–1.04	0.30	Part (Int) Task(Int)	1.40 0.25	<.01	0.58
L1 Japanese								
Mental effort								
Task complexity	–0.35	0.29	–1.22	0.23	Part (Int) Task(Int)	1.94 0.53	<.01	0.71
Task difficulty								
Task complexity	–0.55	0.26	–2.06	0.04	Part (Int) Task(Int)	1.65 0.59	.02	0.69

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

Table 3. Word count for simple and complex tasks.

				95% *CI*	
Measure	*N*	*M*	*SD*	Lower	Upper
Simple tasks	48	140.31	34.15	130.65	149.97
Complex tasks	48	141.81	36.70	131.43	152.19

Table 4. Results for model examining the effects of task complexity on speech length.

	Fixed effects				Random effects		Model effect size	
Predictor	*Est*	*SE*	*t*	*p*	Factor	*SD*	R^2m	R^2c
Task complexity	–1.38	2.16	–0.64	0.52	Part (Int) Task (Int)	31.23 12.51	<.01	0.91

Speech length by task complexity

Table 3 provides the unpruned word count for participants' L2 performances across the simple and complex task conditions. We constructed a linear mixed-effect model, with word count as the dependent variable, task complexity as a fixed effect, and participants and task prompt as random effects. As shown in Table 4, participants' length of speech did not vary by the intended task complexity manipulation, accounting for less than 1% of the variance in word count.

Research question 1: effects of task complexity on pause frequency and length

Table 5 provides the descriptive statistics for the frequency and length of silent pauses by location, that is, whether the pauses were observed mid-clause or end-clause. The table also demonstrates pausing patterns across the two task complexity conditions.

To address our first research question, we constructed a linear mixed-effects model to examine the extent to which task complexity affected pausing behaviors. The fixed effects in our models were task complexity, pause location, and their interaction; and the random effects included participants and task prompt. The dependent variable was the frequency of silent pauses. As shown in Table 6, the analysis yielded a significant

Table 5. Silent pause frequency and length by task complexity and pause location.

					95% *CI*	
Measure	Pause location	*N*	*M*	*SD*	Lower	Upper
Pause frequency						
Simple tasks	Mid–clause	24	1.65	0.57	1.41	1.89
	End–clause	24	0.65	0.17	0.58	0.73
Complex tasks	Mid–clause	24	1.68	0.75	1.36	1.99
	End–clause	24	0.68	0.18	0.60	0.76
Pause length						
Simple tasks	Mid–clause	24	0.84	0.20	0.75	0.92
	End–clause	24	1.07	0.41	0.89	1.24
Complex tasks	Mid–clause	24	0.83	0.18	0.76	0.91
	End–clause	24	0.99	0.32	0.85	1.13

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

Table 6. Results for models examining the effects of task complexity, pause location, and their interaction on pausing behaviors.

	Fixed effects				Random effects		Model effect size	
Predictor	*Est*	*SE*	*t*	*p*	Factor	*SD*	R^2m	R^2c
Pause frequency								
Complexity	–0.03	0.08	–0.34	0.73	Part (Int)	.28	.52	.69
Pause location	1.00	0.08	12.74	<0.01	Task (Int)	.08		
Comp: pause location	0.00	0.11	–0.04	0.97				
Pause length							.02	.11
Complexity	0.05	0.03	1.44	0.15	Part (Int)	.19		
Pause location	–0.17	0.03	–5.91	<0.01	Task (Int)	.07		
Comp: pause location	–0.04	0.04	–1.12	0.26				

effect for pause location only, with participants pausing less often but longer in end-clause than mid-clause positions. The fixed effects explained substantially more variance in the model for pause frequency than pause length, accounting for 52% and 2% of the variation respectively. This indicates that pause location, mid- versus end-clause, affected pause frequency to a greater extent than pause length.

Research question 2: effects of task complexity on neural processes during pausing

Unlike our behavioral analyses, the fMRI analyses found a main effect for task complexity, but no interaction between task complexity and pause location. In complex tasks as to compared simple tasks, pauses (regardless of location) elicited greater activation in broad brain areas related to theory-of-mind activities, conceptualization, and preparation of speech. These areas included the bilateral precentral gyri, right putamen, and left cerebellum for speech planning and monitoring (Hervais-Adelman et al., 2015; Runnqvist et al., 2021; Silva et al., 2022); and left angular gyrus (AG), precuneus, and mPFC from the theory-of-mind network for conceptualization (Ferstl et al., 2008; Morales et al., 2022; Sassa et al., 2007) (see Table 7 and Figure 1). Conversely, simple tasks did not yield higher activation for pauses than complex tasks.

Table 7. Brain areas showing greater activation associated with pauses during complex than simple tasks (complex > simple).

Brain areas	x, y, z	*t*	cluster size
Complex > simple			
Right precentral gyrus	46, –8, 28	6.28	845
	56, 0, 26	5.80	
Left precentral gyrus	–50, –8, 26	5.40	579
	–58, –2, 24	4.86	
Left angular gyrus	–28, –64, 42	5.17	726
Left precuneus	–14, –56, 48		
Right precuneus	14, –64, 40	4.70	381
Left cerebellum	–16, –78, –24	5.13	720
Medial prefrontal cortex	–8, 48, 28	4.95	138
Right putamen	30, 10, –2	4.90	218

Notes. For each area, the coordinates (*x*, *y*, *z*) of the activation peak in MNI space, peak *t*-value, and size of the activated cluster in number (*k*) of voxels (2×2×2 mm^3) are shown for all subjects (n = 24). The threshold was set at p <.05 FWE correction with the cluster level.

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

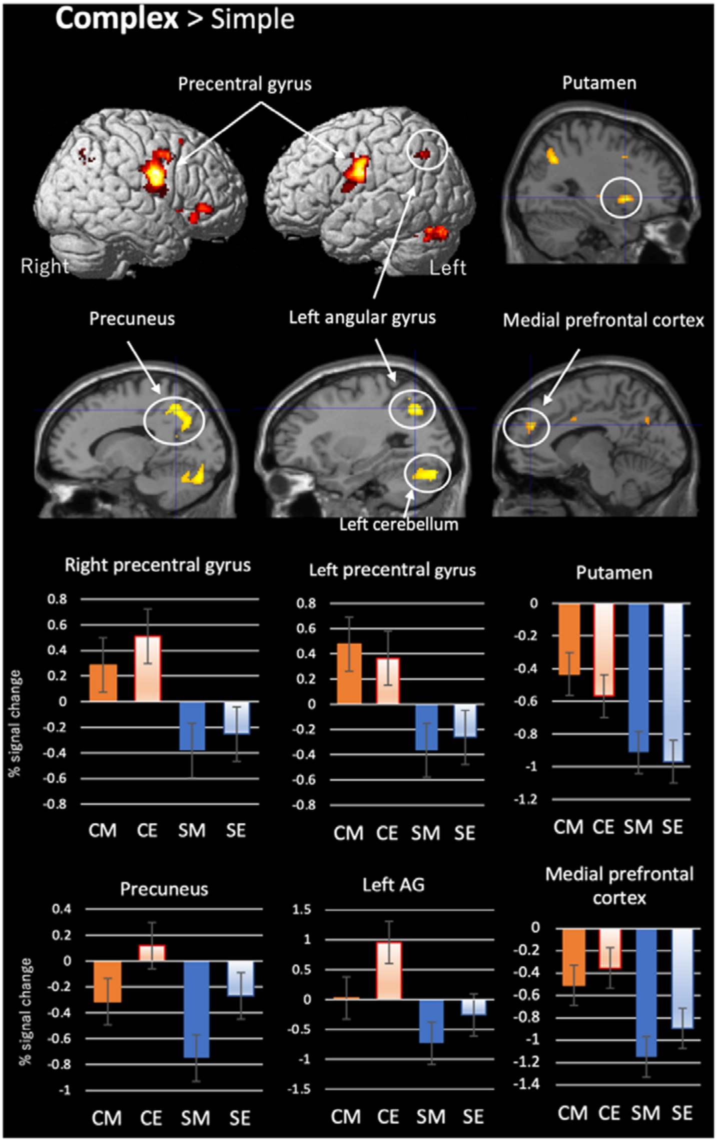

Figure 1. Brain areas showing greater activation associated with pauses during complex than simple tasks (complex > simple).
Note: The activation profile represents the mean percent signal change of each condition; CM: complex mid-clause, CE: complex end-clause, SM: simple mid-clause, SE: simple end-clause. Error bars indicate the standard error of the mean (SEM).

Similar to the behavioral results, the neuroimaging data also identified a pause location effect. As shown in Table 8, pauses at end-clause locations showed increased activation in the bilateral precuneus, extending to the posterior cingulate cortex and the left angular gyrus. These brain areas are associated with theory-of-mind activities, regulating internal thought and conceptualization (Ferstl et al., 2008; Sassa et al., 2007;

Table 8. Brain areas exhibiting differential activation between pauses at end- and mid-clause pause locations.

Brain areas	x, y, z	*t*	cluster size
End–clause pause > Mid–clause pause			
Left precuneus	–4, –66, 58	8.46	3520
Right precuneus	4, –72, 46	9.26	
Posterior cingulate cortex	0, –34, 28	6.21	
Left angular gyrus	–44, –56, 50	6.90	124
	–40, –72, 44		
Left cerebellum	–22, –86, –20	7.41	2180
Right cerebellum	46, –64, –24	6.83	345
Mid–clause pause > End–clause pause			
Left triangular part of inferior frontal gyrus	–44, 32, 4	4.47	47
Right insular	30, 26, 0	3.97	60

Notes. For each area, the coordinates (*x*, *y*, *z*) of the activation peak in MNI space, peak *t*-value, and size of the activated cluster in number (*k*) of voxels (2×2×2 mm^3) are shown for all subjects ($n = 24$). The threshold was set at $p < .05$ family-wise correction (FWE) with the cluster level.

Smallwood et al., 2013), and bilateral cerebellum, which is involved in speech planning (Runnqvist et al., 2021). In contrast, mid-clause pauses, when compared to end-clauses, led to greater activation in the left triangular part of the inferior frontal gyrus (IFG), a key language area (Friederici, 2011), and the right insula, which is associated with motor speech control (Oh et al., 2014).

In sum, no interaction effect was detected between task complexity and pause location in the whole brain analysis, but we detected a main effect for task complexity and pause location. However, greater activation was found in the precuneus and the left angular gyrus during complex as compared with simple tasks (both mid- and end-clause locations) and at end-pause locations as compared with mid-clause locations (during both simple and complex tasks), reflecting greater cognitive demands on conceptualization.

Research question 3: relationships between self-ratings of mental effort and task difficulty, pausing patterns, and neural correlates of pausing

Relationships of mental effort and task difficulty self-ratings to pause frequency and length

To address our third research question, we constructed a series of linear mixed-effects regression models to examine the extent to which participants' L2 ratings of mental effort and task difficulty predicted pause frequency and pause length. The fixed effects in our models were participants' self-ratings of mental effort or difficulty, pause location, and their interaction; and the random effects included participants and task prompt. The dependent variable was the frequency or length of silent pauses.

As shown in Table 9, the models for pause frequency yielded a significant interaction between mental effort and pause location and between task difficulty and pause location. As Figure 3 illustrates, the more often participants paused within clauses, the more effortful and more difficult they perceived the task to be. The fixed effects in both models explained 55% of the variance in the frequency of pausing.

Table 9 also shows that the model involving pause length and mental effort identified mental effort as a significant predictor of pause length. Figure 3 demonstrates that participants who paused longer reported exerting greater mental effort. Similarly, the model including pause length and task difficulty found that task difficulty predicted the

Table 9. Results for models examining self-ratings, pause location, and their interaction as predictors of pausing behaviors.

	Fixed effects				Random effects		Model effect size	
Predictor	*Est*	*SE*	*t*	*p*	Factor	*SD*	R^2m	R^2c
Pause frequency								
Mental effort	0.01	0.02	0.65	0.51	Part (Int)	.27	.55	.70
Pause location	0.63	0.16	4.02	<0.01	Task(Int)	.03		
Men effort: pause loc	0.07	0.03	2.50	0.01				
Pause frequency								
Task difficulty	0.00	0.03	–0.04	0.97	Part (Int)	.27	.55	.71
Pause location	0.53	0.17	3.09	<0.01	Task(Int)	.06		
Task diff: pause loc	0.08	0.03	2.85	<0.01				
Pause length								
Mental effort	0.03	0.01	2.68	<0.01	Part (Int)	.18	.02	.10
Pause location	–0.11	0.06	–1.83	0.07	Task(Int)	.06		
Men effort: pause loc	–0.01	0.01	–1.47	0.14				
Pause length								
Task difficulty	0.03	0.01	2.70	<0.01	Part (Int)	.17	.02	.09
Pause location	–0.16	0.06	–2.49	<0.01	Task(Int)	.07		
Task diff: pause loc	–0.01	0.01	–0.49	0.62				

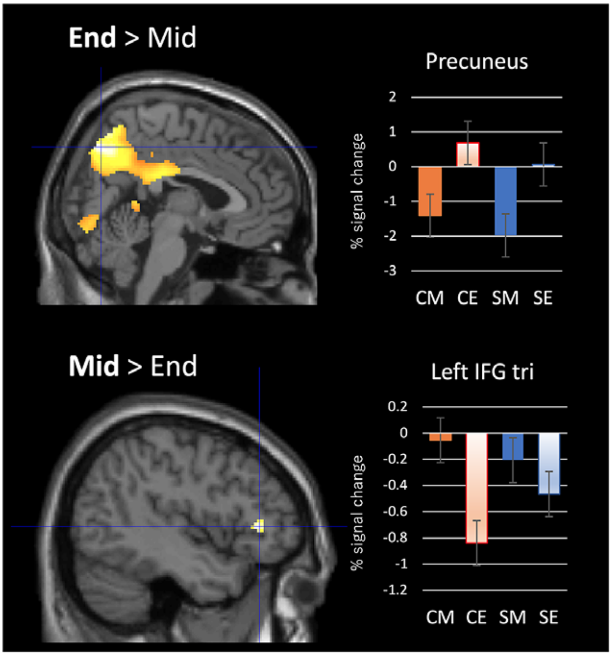

Figure 2. Brain area activation at end- and mid-clause pause locations.
Note: The activation profile represents the mean percent signal change for each condition; CM: complex mid-clause, CE: complex end-clause, SM: simple mid-clause, SE: simple end-clause. Error bars indicate SEM. Left IFG tri: left triangular part of inferior frontal gyrus.

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

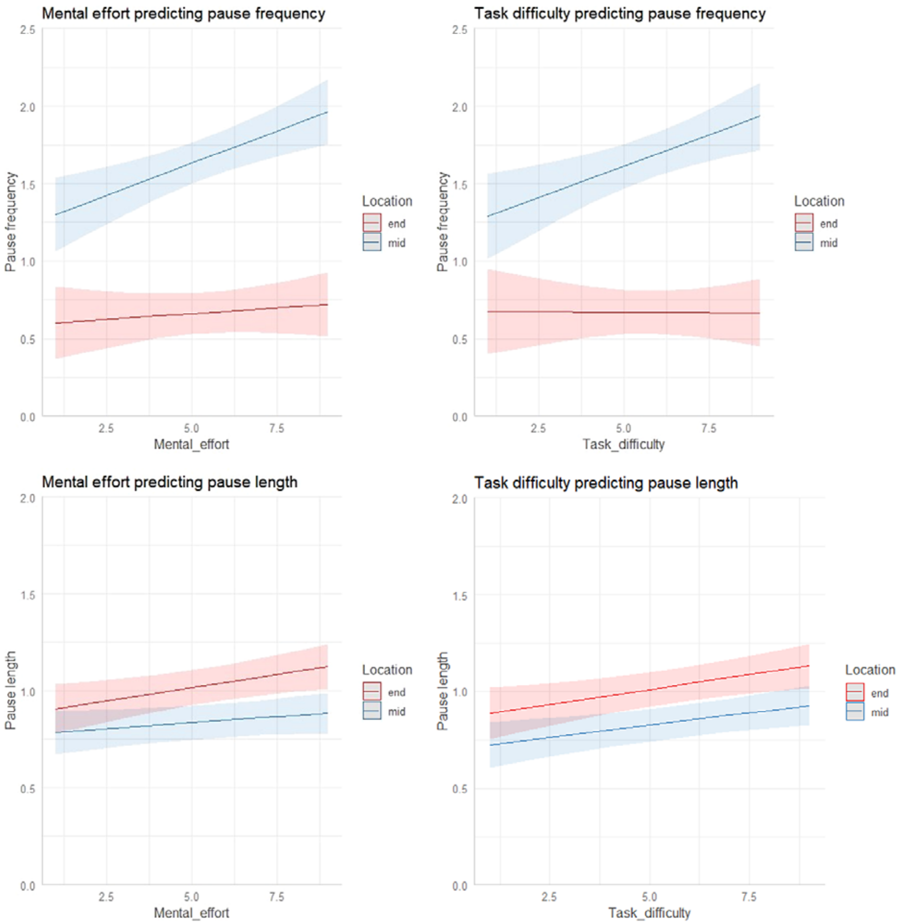

Figure 3. Self-ratings of mental effort and task difficulty predicting pausing behaviors.

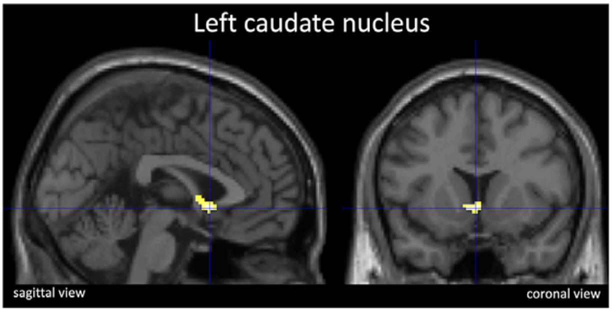

Figure 4. Perceived mental effort effect in mid-clause positions.

length of pausing. As shown in Figure 3, the longer participants paused, the more difficulty they felt the task posed. The analysis also yielded a main effect for pause location, indicating that end-clause pauses were significantly longer than mid-clause pauses during participants' task performance.

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

Relationships of mental effort and task difficulty self-ratings to brain activity during pauses

To elucidate the relationship between participants' L2 self-ratings of mental effort and task difficulty and their actual brain activity at mid-clause and end-clause pause locations, we performed a parametric modulation analysis using SPM12, separately for the mental effort or task difficulty data. In the first-level analysis, we added the mental effort/task difficulty rating for each task performance as a parametric modulator to the main regressor for mid- and end-clause locations. Then, in the second-level analysis, we tested the parametric regressors of mental/task difficulty with a one-sample *t*-test. While we found no significant effect for task difficulty, the left caudate nucleus (x, y, z coordinates = –4, 14, –4, *t*=4.90, 104 voxels) showed increased activation as perceived mental effort increased at mid-clause pause locations (see Figure 4). However, no such effect was observed at end-clause pause locations.

Discussion

Our first two research questions were concerned with the effects of task complexity on observable pausing patterns and associated neural processes. Building on models of speech production (Kormos, 2006; Levelt, 1999) and task-based performance (Robinson, 2001a; Skehan, 2009), we hypothesized that increased task complexity would lead to greater pressure on the conceptualizer, given the enhanced conceptual demands more complex tasks exert. Due to the greater strain on conceptualization processes, we also anticipated that participants would have fewer attentional resources to allocate to linguistic encoding, resulting in increased pressure on the formulator. In turn, we assumed that more complex tasks would lead to an increase in pause frequency and length at both pause locations, as pausing at mid-clause and end-clause locations have been related to involvement in the formulation and conceptualization processes respectively. In parallel, we also expected that more complex tasks would activate conceptualization- and speech-related brain areas to a greater extent during pauses, with more pronounced effects observed for end-clause and mid-clause pauses respectively. Our results have provided no support for our behavioral predictions, yielding no task complexity effects for either pause length or frequency. The neural data, however, largely confirmed our hypotheses. Although we found no interaction effect between task complexity and pause location, we detected greater activation in theory-of-mind-, conceptualization-related brain areas (precuneus and angular gyrus) at end-clause positions and in speech planning and monitoring areas (bilateral precentral gyri and right putamen) at both pause locations during more complex task performance. It is also important to highlight that, while the expert judgments of task complexity and L1 difficulty ratings provided evidence in support of the validity of our task manipulation, participants' L2 self-ratings of mental effort and task difficulty yielded no significant difference between tasks designed to be more and less complex. Interestingly, however, in addressing our third research question, we found that the more effortful and more difficult participants perceived the task to be, the more often and longer participants paused mid-clause. In line with this, the fMRI data, during mid-clause pause positions, revealed that higher mental ratings were correlated with increased activation in the left caudate nucleus. This region functions as a language control area, as indicated by Crinion et al. (2006).

This is an intriguing set of results, yielding several points for discussion. One issue concerns the discrepancy between our behavioral and neural findings. A possible

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

explanation for the observed task complexity effects on the neural data but the lack of those on the behavioral indices might be that the difference in complexity across the two task versions was not sufficiently large to be detected through the behavioral measures employed in the present study. Although in previous research task-related variables were found to influence pausing (Lambert et al., 2017; Wang, 2014), our task manipulation and its impact on participants might not have been robust enough to affect observable pausing patterns. Notably, the mental effort and task difficulty ratings did not identify any impact of task complexity either, indicating an alignment between participants' pausing behaviors and task perceptions. The fact, however, that the neural measures did yield task complexity effects affirms that, consistent with the experts' judgments and the L1 difficulty ratings, the simple and complex task versions did actually instigate processing differences. In other words, it appears that the neural data were more sensitive to detecting the influence of task complexity than the pausing patterns we observed. A lack of convergence between our behavioral and neural measures is not a unique finding. Researchers investigating first language writing processes, for example, found that similar observable behaviors do not necessarily implicate the same brain mechanisms, that is, solely relying on behavioral-level analyses may mask processing differences (Richards et al., 2017).

Another interesting, methods-related finding concerns the link we observed between participants' self-ratings and the pause-related behavioral and neural data. While participants' L2 self-ratings of mental effort and task difficulty did not yield any difference for our intended task complexity manipulation contrary to our expectation, we found that higher self-ratings of mental effort and task difficulty were associated with increased mid-clause pause length and frequency. We also found that higher self-ratings of mental effort were related to greater activation in a language-control brain area (left caudate nucleus) at mid-clause pause locations. Given that mid-clause pausing is assumed to indicate difficulty with linguistic encoding processes (de Jong, 2016; Field, 2011; Lambert et al., 2017; Suzuki & Kormos, 2023; Tavakoli et al., 2017, 2020), these results could be interpreted as suggesting that participants largely based their self-ratings on the linguistic rather than the conceptual difficulty that they had experienced during task performance. This interpretation is also aligned with the findings for L1 self-ratings of task difficulty, which revealed that participants perceived the complex task versions as more difficult. When participants carried out the tasks in their L1, they were unlikely to experience linguistic difficulty, making it more probable that they would judge the difficulty of the tasks based on the conceptual demands they posed (Lee, 2019).

An alternative and/or additional explanation could be that the linguistic difficulty posed by the L2 tasks was perceived by participants to be considerably larger than the issues they might have encountered conceptualizing their message. Given the context of the study, this might not be surprising; in Japan people are often exposed to disaster preparedness and response training, naturally decreasing the conceptual demands imposed by disaster-related tasks. This issue could be disentangled in future research by asking participants to provide separate task difficulty and mental effort ratings for linguistic and conceptual task demands. This more refined self-perception data could also assist with obtaining a fuller picture of perceived speech production processes during task-based work, with potential theoretical implications for task-based models and practical implications for task-based teaching.

Another potential methodological implication emerging from our study concerns the use of L1 self-ratings when establishing task complexity. In line with Lee's (2019) proposal, our results suggest that, if the researchers' aim is to establish the conceptual

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

demands of tasks, obtaining L1 rather than L2 self-ratings of task difficulty might yield more fine-tuned results. As we discussed earlier, during L1 performance speakers are less likely to encounter difficulty with linguistic encoding, increasing the likelihood that self-ratings provide an accurate judgment of the conceptual difficulty posed by the task. This conclusion is also supported by our fMRI data, which along with the L1 self-ratings, detected an effect for task complexity, contrary to the L2 behavioral data we collected (self-ratings and linguistic performance measures).

It is also worth highlighting that our results suggest that L2 task complexity has neural correlates, providing a new type of evidence for the validity of the construct of task complexity. In particular, the greater activation we observed in areas associated with the theory-of-mind network and conceptualization during complex tasks is aligned with our intended task design manipulation that the more complex task versions would lead to greater conceptual demands. The theory-of-mind system underlines humans' ability to understand others' beliefs, desires, and intentions, rendering its successful use crucial for effective social interaction and verbal communication. As the participants in the current study were in imaginary rather than real-life situations, they probably needed to infer what someone would do in these critical situations, therefore making them rely on the theory-of-mind system. If so, it seems logical that the theory-of-mind network would be more involved when participants were conceptualizing their message under more complex conditions, given that these tasks were designed with the intention that participants would imagine and decide how to act in more complex critical situations. Similar to our findings, theory-of-mind-related areas were found to display higher brain activation during written text comprehension activities that require inferencing and interpretation (Ferstl et al., 2008), such as reasoning about the intentions and mental states of narrative characters (Mason & Just, 2009). This finding is particularly relevant to our study, as, according to Levelt's (1989) speech production model, self-monitoring is part of one's comprehension rather than the production system. Thus, the effect of task complexity could also be interpreted as evidence of participants' engagement with self-monitoring, involving the theory-of-mind system through the evaluation of how the evolving message would be perceived by potential interlocutors. As discussed earlier, studies of speech production also showed greater neural activation of the theory-of-mind network when participants engaged in communication with others as compared to when they completed description activities without the need to communicate (Jeong et al., 2016; Sassa et al., 2007). Our study, however, is among the first to observe theory-of-mind effects on neural processes during spontaneous L2 speech production.

The results obtained here are also interesting to consider in relation to Robinson's (2001a, 2011) Cognition Hypothesis and Skehan's (1998, 2009) Limited Capacity Model. If we take the observed differences in neural processes across the simple and complex task versions as proof that the task manipulation worked in the present study, we can conclude that the results did not confirm the predictions of the models for the dependent variable of fluency, as the amount of pausing, contrary to the prediction of the models, would not rise as the cognitive demands of the tasks increased. However, the data did provide evidence in support of the task frameworks in that the task complexity manipulations, the independent variable in the models, affected speech production processes, a presumed causal variable in the Cognition Hypothesis and Limited Capacity Model. In particular, the task versions designed to be more complex exerted greater conceptual demands, as reflected in the neural data and predicted by the models.

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

Besides yielding evidence for the construct of task complexity and related task-based models, our findings reinforce the value of distinguishing between mid-clause and end-clause pausing. As discussed in the literature review, researchers have identified various types of behavioral evidence in favor of the assumption that mid-clause and end-clause pauses are associated with L2 formulation and conceptualization processes, respectively (e.g., de Jong, 2016; Field, 2011; Lambert et al., 2017; Suzuki & Kormos, 2023; Tavakoli et al., 2017, 2020). The neural data obtained in the current study are aligned with this assumption, given that the fMRI scans found increased brain activity in conceptualization-related brain areas at end-clause but not in mid-clause positions during tasks designed to be more complex. Also, self-ratings were related only to mid-clause pause frequency and brain activity but not to end-clause pausing. In other words, expanding on existing behavioral evidence, the present research has generated novel neural evidence for the merit of distinguishing between pauses based on location.

Limitations and future directions

This study has several limitations that are important to acknowledge and consider in future research. First, although our results suggest that it is highly profitable to carry out fMRI studies to gain more insights into L2 speech production processes and their links to task complexity, this neuroimaging technique, at least for now, yields results that are relatively low in ecological validity. Performing oral tasks in an fMRI scanner differs considerably from carrying L2 oral tasks in real-life settings, which limits the generalizability of our findings. Second, we exclusively employed neuroimaging to tap the link between task complexity and the speech production processes in which participants engaged. It would be interesting to complement neuroimaging techniques with objective behavioral tools such as dual-task methodology and eye-tracking, given that these behavioral methods could provide complementary insights into how task complexity may influence speech production processes. A further limitation has to do with the absence of more comprehensive information about participants' thoughts during pauses and their perceptions about mental effort and task difficulty. Besides using more fine-grained self-rating tools as suggested earlier, future studies would benefit from employing introspective methods (e.g., stimulated recall) to achieve a more in-depth understanding of participants' conscious thought processes and perceptions during tasks of varied complexity. Another limitation of our research lies in its relatively small sample size. While a sample size of 24 allowed us to detect medium effect sizes ($f = .32$) given the repeated-measures design and within-subjects independent variable according to G-power, it was not sufficiently large to reveal small-size effects. Thus, replicating the study with a larger number of participants would be a worthwhile endeavor in the future. A further useful research direction would be to make more detailed distinctions among pause locations and examine corresponding cognitive and neural processes as a function of task manipulations. For example, it would be worthwhile to distinguish pause locations in terms of more specific syntactic constituents (e.g., different types of phrases). In future research, it would also be interesting to examine how the intended task complexity manipulation affected other linguistic areas, such as the linguistic complexity, accuracy, and functional adequacy of learners' production. Finally, the current study only included Japanese users of L2 English who carried out disaster-related decision-making tasks, future studies are warranted to investigate whether the results we obtained would transfer to different first-language groups, second languages, and task types.

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

Conclusion

Being among the first to use neuroimaging to investigate task-related effects on L2 spontaneous oral production, the primary aim of this study was to provide novel insights into the neural correlates of task complexity. Our results identified differences in brain activation patterns across simple and complex versions of decision-making tasks, providing neural evidence in support of the construct of task complexity and the validity of the task manipulation in the present study. In contrast, however, participants' L2 subjective self-ratings of task difficulty, contrary to their L1 self-ratings, did not yield an effect for task difficulty. Our second goal was to explore the potential value of using neuroimaging to examine the impact of task-related variables on spontaneous oral production. While our L2 behavioral measures identified no influence of task complexity, the fMRI scans revealed that brain activation patterns varied as a function of task complexity consistent with participants' L1 self-ratings.

Overall, we interpreted these findings as suggesting that brain imaging was more sensitive to detecting small-size task complexity effects than the more traditional L2 self-ratings and pausing measures, thus confirming the value of triangulating behavioral measures with neuroimaging data. Importantly, this is not to suggest that task complexity researchers and practitioners should move away from using subjective self-ratings to establish task difficulty and from linguistic performance measures to examine task effects on L2 production. A large body of research indicates that these tools are likely to detect task complexity differences, probably larger in size than the ones observed here, and thus likely to generate meaningful results from a practical perspective. Importantly, however, our findings, if replicated, do imply that researchers might benefit from greater use of L1-ratings and expert judgments to establish task complexity. As compared to L2 self-ratings, these tools appear to yield more sensitive results aligned with neural measures, while being equally practical.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263124000421.

Data availability statement. The experiment in this article earned the Open Materials badge for transparent practices. The materials are available at https://www.iris-database.org/details/gsCp3-6k9Ej.

Acknowledgements. This research was supported by the UCL-Tohoku University Strategic Partner Funds; the Designated National University Core Research Cluster of Disaster Science, Tohoku University, Japan; JSPS KAKENHI Grant Number 23K21946; and the ESRC-AHRC UK-Japan SSH Connections Grants (ES/S013024/1).

References

Allan, D. (2004). *Oxford placement test (1 or 2)*. Oxford University Press.

Boersma, P., & Weenink, D. (2022). Praat: doing phonetics by computer [Computer program]. Version 6.4.13, retrieved 10 June 2022 from http://www.praat.org/

Brett, M., Anton J. L., Valabregue R., & Poline J. B. (2002). Region of interest analysis using the MarsBar toolbox for SPM 99. *Neuroimage*, *16*(2), S497.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA (pp. 21–46). John Benjamins.

Byrnes, H. (2013). Notes from the editor. *Modern Language Journal*, *97*(4), 825–827. https://doi.org/10.1111/j.1540-4781.2013.12051.x

Crinion, J., Turner, R., Grogan, A., Hanakawa, T., Noppeney, U., Devlin, J., Aso, T., Urayama, S., Fukuyama, H., Stockton, K., Usui, K., Green, D., & Price, C. (2006). Language control in the bilingual brain. *Science*, *312*(5779), 1537–1540. https://doi.org/10.1126/science.1127761

De Bot, K. (1992). A bilingual production model: Levelt's speaking model adapted. *Applied Linguistics*, *13*(1), 1–24. https://doi.org/10.1093/applin/13.1.1

De Bot, K., & Schreuder, R. (1993). Word production and the bilingual lexicon. In R. Schreuder & B. Weltens (Eds.), *The bilingual lexicon* (pp. 191–214). John Benjamins.

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: the effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching (IRAL)*, *54*(2), 113–-132. https://doi.org/10.1515/iral-2016-9993

Duran-Karaoz, Z., & Tavakoli, P. (2020). Predicting L2 fluency from L1 fluency behavior: The case of L1 Turkish and L2 English speakers. *Studies in Second Language Acquisition*, *42*(4), 671–695. https://doi.org/10.1017/S0272263119000755

Felker, E., Klockmann, H., & De Jong, N. (2019). How conceptualizing influences fluency in first and second language speech production. *Applied Psycholinguistics*, *40*(1), 111–136. https://doi.org/10.1017/S0142716418000474

Ferstl, E. C., Neumann, J., Bogler, C., & von Cramon, D. Y. (2008). The extended language network: a meta-analysis of neuroimaging studies on text comprehension. *Human Brain Mapping*, *29*(5), 581–593. https://doi.org/10.1002/hbm.20422

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking: Research and practice in assessing second language speaking* (pp. 65–111). Cambridge University Press.

Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Hervais-Adelman, A., Moser-Mercer, B., Michel, C. M., & Golestani, N. (2015). fMRI of simultaneous interpretation reveals the neural basis of extreme language control. *Cerebral Cortex*, *25*(12), 4727–4739. https://doi.org/10.1093/cercor/bhu158

Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (Eds.) (2012). Dimensions of L2 performance and proficiency. Complexity, accuracy and fluency in SLA. John Benjamins.

Jeong, H., Hashizume, H., Sugiura, M., Sassa, Y., Yokoyama, S., Shiozaki, S., & Kawashima, R. (2011). Testing second language oral proficiency in direct and semidirect settings: A social-cognitive neuroscience perspective. *Language Learning*, *61*(3), 675–699. https://doi.org/10.1111/j.1467-9922.2011.00635.x

Jeong, H., Sugiura, M., Suzuki, W., Sassa, Y., Hashizume, H., & Kawashima, R. (2016). Neural correlates of second-language communication and the effect of language anxiety. *Neuropsychologia*, *84*, 2–12. https://doi.org/10.1016/j.neuropsychologia.2014.11.013

Kane, M. T. (2006). Validation. In Brennan R. (Ed.), *Educational measurement* (4th ed., pp. 17–64), Westport, CT: American Council on Education and Praeger.

Khang, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, *64*(4), 809–854. https://doi.org/10.1111/lang.12084

Khang, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*(3), 569–91. https://doi.org/10.1017/S0142716417000534

Kahng, J. (2020). Explaining second language utterance fluency: Contribution of cognitive fluency and first language utterance fluency. *Applied Psycholinguistics*, *41*(2), 457–480. https://doi.org/10.1017/S0142716420000065

Kim, Y., Payant, C., & Pearson, P. (2015). The intersection of task-based interaction, task complexity, and working memory: L2 question development through recasts in a laboratory setting. *Studies in Second Language Acquisition*, *37*(3), 549–581. https://doi.org/10.1017/S0272263114000618

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum.

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition*, *39*(1), 167–196. https://doi.org/10.1017/S0272263116000085

Lee, J. (2019). Task complexity, cognitive load, and L1 speech. *Applied Linguistics*, *39*(3), 506–539. https://doi.org/10.1093/applin/amy011

Levelt, W. J. M. (1989). *Speaking from intention to articulation*. MIT Press.

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

Levelt, W. J. M. (1999). Producing spoken language: A blueprint of the speaker. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 83–122). Oxford University Press.

Mackey, A. (2020). *Interaction, feedback and task research in second language learning: Methods and design*. Cambridge University Press.

Mason, R., & Just, M. (2009). The role of the theory-of-mind cortical network in the comprehension of narratives. *Language and Linguistics Compass*, *3*(1), 157–174. https://doi.org/10.1111/j.1749-818X.2008.00122.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Morales, M., Patel, T., Tamm, A., Pickering, M. J., & Hoffman, P. (2022). Similar neural networks respond to coherence during comprehension and production of discourse. *Cerebral Cortex*, *32*(19), 4317–4330. https://doi.org/10.1093/cercor/bhab485

Norris, J. M. (2010). *Understanding instructed SLA: Constructs, contexts, and consequences* [Plenary talk]. European Second Language Association (EUROSLA) conference, Reggio Emilia, Italy.

Norris, J. M., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 717–761). Blackwell.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555–578. https://doi.org/10.1093/applin/amp044

Oh, A., Duerden, E. G., & Pang, E. W. (2014). The role of the insula in speech and language processing. *Brain and Language*, *135*, 96–103. https://doi.org/10.1016/j.bandl.2014.06.003

Ortega, L. (2005). What do learners plan? Learner-driven attention to form during pre-task planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 77–109). John Benjamins. https://doi.org/10.1075/lllt.11.07ort

Pang, F., & Skehan, P. (2014). Self-reported planning behaviour and second language performance in narrative retelling. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 95–128). John Benjamins. https://doi.org/10.1075/tblt.5.04pan

Poulisse, N., & Bongaerts, T. (1994). First language use in second language production. *Applied Linguistics*, *15*, 36–57.

Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition*, *31*(3), 437–470. https://doi.org/10.1017/S0272263109090366

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics*, *35*(1), 87–92. https://doi.org/10.1093/applin/amt039

Révész, A. (2021a). Methodological approaches to investigating task-based language teaching: Advances and challenges. In M. J. Ahmadian & M. H. Long (Eds.), *The Cambridge handbook of task-based language teaching* (pp. 605–627). Cambridge University Press.

Révész, A. (2021b). Exploring task-based cognitive processes: Methodological advances and challenges. *TASK*, *1*(2), 266–288. https://doi.org/10.1075/task.21017.rev

Révész, A., Kourtali, N., & Mazgutova, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning*, *67*(1), 208–241. https://doi.org/10.1111/lang.12205

Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual task methodology, subjective self-ratings, and expert judgments: A validation study. *Studies in Second Language Acquisition*, *38*(4), 703–737. https://doi.org/10.1017/S0272263115000339

Révész, A., Sachs, R., & Hama, M. (2014). The effects of task complexity and input frequency on the acquisition of the past counterfactual construction through recasts. *Language Learning*, *64*(3), 615–650. https://doi.org/10.1111/lang.12061

Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition*, *23*(4), 497–526. https://doi.org/10.1017/S027226310100403X

Richards, T. L., Berninger, V. W., Yagle, K. J., Abbott, R. D., & Peterson, D. J. (2017). Changes in DTI diffusivity and fMRI connectivity cluster coefficients for students with and without specific learning disabilities in written language: Brain's response to writing instruction. *Journal of Nature and Science*, *3*(4), e350. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5488805/

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: A triadic framework for investigating task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287–318). Cambridge University Press.

Robinson, P. (2001b). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics*, *22*(1), 27–57. https://doi.org/10.1093/applin/22.1.27

Robinson, P. (2011). Task-based language learning: A review of issues. *Language Learning*, *61*(S1), 1–36. https://doi.org/10.1111/j.1467-9922.2011.00641.x

Runnqvist, E., Chanoine, V., Strijkers, K., Pattamadilok, C., Bonnard, M., Nazarian, B., Sein, J., Anton, J.-L., Dorokhova, L., Belin, P., & Alario, F.-X. (2021). Cerebellar and cortical correlates of internal and external speech error monitoring. *Cerebral Cortex Communications, 2*(2), Article tgab038. https://doi.org/10.1093/texcom/tgab038

Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Applied Psycholinguistics*, *39*(3), 593–617. https://doi.org/10.1017/S0142716417000571

Sasayama, S. (2016). 'complex' task really complex? Validating the assumption of cognitive task complexity. *The Modern Language Journal*, *100*(1), 231–254. https://doi.org/10.1111/modl.12313

Sasayama, S., & Norris, J. (2019). Unravelling cognitive task complexity: Learning from learners' perspectives on task characteristics and second language performance. In Z. E. Wen & M. J. Ahmadian (Eds.), *Researching L2 task performance and performance: In honour of Peter Skehan* (pp. 95–132). John Benjamins. https://doi.org/10.1075/tblt.13.06sas

Sassa, Y., Sugiura, M., Jeong, H., Horie, K., Sato, S., & Kawashima, R. (2007). Cortical mechanism of communicative speech production. *NeuroImage*, *37*(3), 985–992. https://doi.org/10.1016/j.neuroimage.2007.05.059

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Silva, A. B., Liu, J. R., Zhao, L., Levy, D. F., Scott, T. L., & Chang, E. F. (2022). A neurosurgical functional dissection of the middle precentral gyrus during speech production. *The Journal of Neuroscience*, *42*(45), 8416–8426. https://doi.org/10.1523/jneurosci.1614-22.2022

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Skehan, P. (2003). Task-based instruction. *Language Teaching*, *36*(1), 1–14. https://doi.org/10.1017/S026144480200188X

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4). 510–532. https://doi.org/10.1093/applin/amp047

Skehan, P., & Foster, P. (2007). Complexity, accuracy, fluency and lexis in task-based performance: a meta-analysis of the Ealing research. In Van Daele, P., Housen, A., Kuiken, F., Pierrard, M., & Vedder, I. (Eds.), Complexity, fluency and accuracy and fluency in second language use, learning and teaching (pp. 207–226). University of Brussels Press.

Slotnick, S. D. (2017). Cluster success: fMRI inferences for spatial extent have acceptable false-positive rates. *Cognitive Neuroscience*, *8*(3), 150–155. https://doi.org/10.1080/17588928.2017.1319350

Smallwood, J., Gorgolewski, K. J., Golchert, J., Ruby, F. J. M., Engen, H., Baird, B., Vinski, M. T., Schooler, J. W., & Margulies, D. S. (2013). The default modes of reading: modulation of posterior cingulate and medial prefrontal cortex connectivity associated with comprehension and task focus while reading. *Frontiers in Human Neuroscience*, *7*, 734. https://doi.org/10.3389/fnhum.2013.00734

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition*, *42*(1), 143–167. https://doi.org/10.1017/S0272263119000421

Suzuki, S., & Kormos, J. (2023). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, *45*(1), 38–64. https://doi.org/10.1017/S0272263121000899

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal*, *105*(2), 435–463. https://doi.org/10.1111/modl.12706

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal*, *65* (1), 71–79. https://doi.org/10.1093/elt/ccq020

Tavakoli, P., Nakatsuhara, F., & Hunter, A-M. (2017). *Scoring validity of the APTIS speaking test: Investigating fluency across tasks and levels of proficiency*. British Council.

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press

Tavakoli, P., Nakatsuhara, F., & Hunter A-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *Modern Language Journal*, *104*(1), 169–191. https://doi.org/10.1111/modl.12620

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–277). John Benjamins. https://doi.org/10.1075/lllt.11.15tav

Torres, J. (2018). The effects of task complexity on heritage and L2 Spanish development. *Canadian Modern Language Review*, *74*(1), 128–152. https://doi.org/10.3138/cmlr.3770

Wang, Z. (2014). On-line time pressure manipulations: L2 speaking performance under five types of planning and repetition conditions. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 27–62). John Benjamins. https://doi.org/10.1075/tblt.5.02wan

Wu, W., Morales, M., Patel, T., Pickering, M. J., & Hoffman, P. (2022). Modulation of brain activity by psycholinguistic information during naturalistic speech comprehension and production. *Cortex*, *155*, 287–306. https://doi.org/10.1016/j.cortex.2022.08.002

Xu, T. S., Zhang, L. J., & Gaffney, J. S. (2022). Examining the relative effects of task complexity and cognitive demands on students' writing in a second language. *Studies in Second Language Acquisition*, *44*(2), 483–506. https://doi.org/10.1017/S0272263121000310

Cite this article: Révész, A., Jeong, H., Suzuki, S., Cui, H., Matsuura, S., Saito, K., & Sugiura, M. (2024). Task-generated processes in second language speech production: *Exploring the neural correlates of task complexity during silent pauses. Studies in Second Language Acquisition,* 1–27. https://doi.org/10.1017/S0272263124000421

https://doi.org/10.1017/S0272263124000421 Published online by Cambridge University Press