Controlled and Automatic Processing in the Acceptability Judgment Task: An Eye-Tracking Study

Ryo Maie and Aline Godfroid

Michigan State University

Abstract

We conducted an eye-tracking study of the acceptability judgment task (AJT) by drawing on the dual process theory of controlled and automatic processing. We conceptually replicated the work of Godfroid et al. (2015) and then extended it in two respects: (a) we analyzed both late and early measures of eye movement to differentiate between the effect of time pressure on controlled and on automatic processes, and (b) we examined how the automaticity of participants' lexical processing moderated the effect of time pressure. Under timed and untimed conditions, 31 L1 and 40 L2 English speakers performed the AJT while their eye movements were recorded. Through statistical modeling of the eye-tracking data, we demonstrated that (a) time pressure inhibits not only controlled processing but also automatic processing and (b) the time pressure effect is most pronounced for the late eye-movement measures of L2 speakers with slow lexical decoding skills. We explain that time pressure may not work as theoretically predicted by L2 researchers (i.e., to suppress controlled processes associated with explicit knowledge) and that its effect is not uniform across different L2 speakers.

Keywords controlled processing; automatic processing; acceptability judgment task; eye tracking

Introduction

The acceptability judgment task (AJT) has a long tradition in the field of second language acquisition (SLA) as a means for probing what learners know of the target language, that is, their developing interlanguage system (Plonsky, Marsden, Crowther, Gass, & Spinner, 2019; Spinner & Gass, 2019). A synthesis by Plonsky et al. (2019) reported that the use of AJTs in second language (L2) research has steadily increased since the 1970s, finding a total of 302 studies that have used some type of judgment data to access the acceptability of target language sentences. The wide popularity of AJTs, however, is at odds with its disputed construct validity because SLA researchers have yet to agree on what AJTs actually measure. Often, validity studies have focused on the constructs of explicit and implicit knowledge, and they have predominantly concentrated on L2 knowledge as a learned product—a static outcome. In this line of research, the biggest controversy has concerned whether imposing time pressure on an AJT renders the task a (relatively) pure measure of implicit knowledge (DeKeyser, 2003; Ellis, 2005) in contrast to an untimed AJT, which has been deemed to (be more likely to) measure (more) explicit knowledge in L2 research.

In this paper, we have reported a study in which we analyzed first language (L1) and L2 speakers' performance on AJTs from a processing perspective. We adopted the dual process theory of controlled/automatic processing (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977) and investigated whether controlled and/or automatic processes can be affected by time pressure. In doing so, we conceptually replicated and extended a previous L2 processing study of AJTs conducted by Godfroid et al. (2015), in which the authors tracked online eye movements of L1 and L2 speakers while the speakers made acceptability judgments under timed and untimed conditions. Whereas the original study for English solely focused on regressions, that is, right-to-left eye movements, which yields a late measure that indexes controlled processing, we analyzed additional eye-movement measures of both early and late processing in an attempt to differentiate between the effect of time pressure on controlled and on automatic processes. Moreover, drawing on insights from recent psycholinguistic research (Hopp, 2014; McDonald, 2006) and reading models (Koda, 2005; LaBerge & Samuels, 1974; Reichle, Warren, & McConnell, 2009), we examined whether the automaticity of participants' lexical processing moderated the effect of time pressure.

Background Literature

Controlled and Automatic Processes in the Acceptability Judgment Task

The dual process theory of human information processing originated in the seminal work of Schneider and Shiffrin (1977) and Shiffrin and Schneider (1977), who investigated how normal healthy adults develop fluency in performing visual search tasks. The theory contends that attention is a fundamental mechanism of learning and contrasts two types of cognitive processing—controlled and automatic—as opposite ends of the same developmental continuum (see also Schneider & Chein, 2003, for a review). Controlled processing is defined as a cognitive activity that is "activated under control of, and through attention, by the subject," whereas automatic processing is processing "without the necessity of active control or attention" (Schneider & Shiffrin, 1977, p. 20). Functionally, controlled processes are conceptualized to be slow, serial, and effortful and to expend an individual's attentional resources. Automatic processes, however, are fast, parallel, and effortless and require minimal cognitive effort. Learning, in this light, can be considered a gradual transition from controlled to automatic processing, ultimately culminating in automaticity. Although many cognitive psychologists have agreed that automaticity is a multifaceted construct, they have continued to disagree on the features with which to define automaticity (see DeKeyser, 2001; Moors & De Houwer, 2006). The dual process theory, in particular, highlights the role of attention in the process of automatization and conceives of automatization as a gradual withdrawal of attention from a task (or the components thereof). Although other accounts of automaticity exist (e.g., Anderson, 1982; Logan, 1988), the distinction between controlled and automatic processing that constitutes the core of dual process theory lends itself well to operationalization with an eye-tracking methodology. As such, we adopted dual process theory as a guiding theoretical framework for the current study.

In the L2 literature, controlled and automatic processing have often been researched in connection with monitoring, a cover term for psycholinguistic processes during which one checks that messages that are planned, uttered, or comprehended correspond to one's L2 system (or communicative intentions). For beginning-level learners, monitoring is a resource-consuming, and hence, controlled process because these learners have many (sub)processes of production and comprehension to which to attend. Although monitoring is typically associated with explicit knowledge and controlled processing (e.g., Krashen, 1982), research on speech production has indicated that L1 speakers and advanced L2 speakers also monitor their speech, which is evidenced in self-corrections, under performance conditions that are more automatic (Kormos, 2006; Levelt, 1989). To this point, Krashen noted that "self-correction can also be done using the acquired system alone, with one's 'feel' for correctness" (Krashen, 1982, p. 108). There may thus be more than one route that leads to monitoring behavior, termed "monitoring by rule" and "monitoring by feel" (DeKeyser, 2010, p. 90) or, in the context of the current study, monitoring by controlled processing versus monitoring by automatic processing.

In the context of AJTs, one can monitor the accuracy of a sentence by either controlled or automatic processing or by using some combination of both. Monitoring through controlled processing may occur specifically when learners, due to the nature of AJTs, scan a sentence with controlled, and potentially metalinguistic, strategies from the outset (i.e., top-down task effects). So-called monitoring by feel may occur when learners detect a grammatical error in an ungrammatical sentence (i.e., bottom up) and the detection of the error causes them to become conscious and revert to controlled processing again. Seen in this light, monitoring by feel also entails some low level of attention from learners for them to process the input. Given that L1 and advanced L2 speakers (who were the target populations of this study) most often rely on highly automatized knowledge and skills to monitor their speech (DeKeyser, 2009; Kormos, 2006), we could expect that their monitoring in AJTs should also proceed automatically, following the bottom-up route. This prediction, however, needed to be approached with caution because the AJT, by its nature, orients test-takers to focus on the accuracy and form of the (usually) written language and thus invites a more controlled type of monitoring than is typically expected (Maie & DeKeyser, 2020).

The Effect of Time Pressure on Controlled and Automatic Processing

Research in cognitive psychology has revealed that automatic processes are relatively robust to stressors (e.g., fatigue, stress, and lack of vigilance), whereas controlled processes are more sensitive to stressors (e.g., Heuer, Spijkers, Kiesswetter, & Schmidtke, 1998). Given the functional properties of the two processes (e.g., faster vs. slower), this distinction is unsurprising. In consequence, controlled processes should be more affected by time pressure, but automatic processes should be more likely to be impervious to the effect. SLA researchers have posited similar claims, especially those researchers interested in the construct validity of AJTs (Ellis, 2005, 2015; Loewen, 2009). Ellis (2005), for instance, claimed that imposing time pressure renders the AJT a better measure of implicit knowledge than of explicit knowledge. This is because time pressure limits L2 speakers' monitoring by controlled processing (which is slow and effortful), and the use of explicit knowledge likely entails controlled processing. Recently, this pairing of explicit knowledge and controlled processing has been challenged by some (DeKeyser, 2003; Vafaee, Suzuki, & Kachinske, 2017) who argue that the use of explicit knowledge can be automatized through practice and hence become available to automatic processing; yet, for our purposes, the rationale provided by Ellis and others has remained valid.

Conceptually speaking, it has made sense to researchers to hypothesize that time pressure selectively inhibits controlled processes. The available evidence, however, has been too limited for researchers to draw any definitive conclusions. Hulstijn and Hulstijn (1984), for instance, reported that time pressure equally affected the speech performance of those who monitored by either explicit (controlled) or implicit (automatic) knowledge. If this is true, if indeed it is the case that even speech monitoring on the basis of highly automatized knowledge can be affected by time pressure, then the same effect when reading AJTs might be observed as well. The answer to this question, again, has yet to be revealed because very few researchers have to date directly focused on what learners actually do while they make acceptability judgments. This is evident in the fact that most previous studies have relied on accuracy of judgments as the sole dependent variable. One exception, however, was Godfroid et al. (2015), the study that we conceptually replicated and extended.

Godfroid et al. (2015) conducted an eye-tracking study to assess how imposing time pressure can influence the ways in which L1 and L2 speakers process target AJT sentences. They specifically focused on regressions, that is, right-to-left movements, and examined how the rate of regression changed (i.e., decreased) depending on the introduction of time pressure. Godfroid et al. argued that regressions likely index the controlled processing of stimuli (i.e., structural reanalysis; Von der Malsburg & Vasishth, 2011), whereas straight-pass reading with no regression, that is, straight left-to-right reading, represents fluent and automatic processing (Reichle et al., 2009). Godfroid et al.'s (2015) analysis of participants' reading paths revealed that three so-called scanpath categories best characterized the data: (a) no regression, (b) unfinished reading of a sentence with one or more regressions, and (c) finished reading with regression(s). They tallied the number of items in each category and examined how the proportions of these items were affected by three independent variables of interest: (a) L1 versus L2, (b) timed versus untimed, and (c) grammatical versus ungrammatical. The results showed that, although L1 speakers were not affected by time pressure, the regression rate of L2 speakers was significantly suppressed in the timed condition. For instance, L2 speakers' rate of finished reading with regressions decreased from 80% on the untimed AJT to 57% on the timed AJT, whereas that of L1 speakers remained constant—68% on the timed and 69% on the untimed AJT.

In sum, although there is little direct evidence for the effect of time pressure, Godfroid et al. (2015) seemed to provide the first evidence that time pressure does indeed restrict monitoring by controlled processing. This finding only spoke to part of the picture, however, because Godfroid et al. did not examine automatic processes. In the present study, we extended their experiment by including both early and late measures of eye movements to distinguish between the effect of time pressure on controlled and automatic processing. Furthermore, we also added participants' lexical processing automaticity as a moderating variable. This was done because psycholinguistic research has provided evidence that lexical processing skills may be an important covariate when investigating phenomena concerning L2 syntactic or sentence-level processing (Hopp, 2014; McDonald, 2006). By extension, we argue below that lexical processing automaticity may also play a role in mitigating the effect of time pressure in AJTs.

The Role of Lexical Processing Automaticity in Acceptability Judgment Tasks

Psycholinguistic research on syntactic processing has demonstrated that L2 speakers' automaticity in decoding and recognizing individual words interacts with the extent to which they can successfully parse sentences under time pressure (Hopp, 2014; McDonald, 2006; see also Godfroid, 2020). Hopp (2014), for instance, identified automaticity of lexical decoding skills to be of primary importance, in that "L2 readers with less automatic lexical access . . . do not reach the syntactic structure building stage in online comprehension" (p. 272). Theoretically speaking, this can be explained and expected by various models of visual reading. In SLA research, the dominant approach to researching reading skills has been the component skills approach (Carr & Levy, 1990) that holds that reading is "the product of a complex information-processing system, involving a constellation of mental operations" (Koda, 2005, p.19). These mental operations are arguably organized hierarchically (Koda, 2005, 2007; Reichle et al., 2009; Stanovich, 2000), from orthographic decoding to word recognition to syntactic parsing and integration. LaBerge and Samuels (1974) proposed a highly influential model of reading that described the transition of written stimuli to meanings as a sequence of stages of information processing (e.g., perceptual and graphemic processing, lexical processing, and syntactic processing). They highlighted the role of attention and automaticity as important variables for understanding the process of reading and argued that the automaticity of lower-level processes such as word recognition is critical to successful reading; otherwise, higher-level processes such as syntactic parsing would be impossible because the entire operation would exceed one's capacity of attention before it reached the higher-level stage. In the L2 context, Koda (2007) also made a similar claim that “[b]ecause inefficient decoding is resource demanding, it severely restricts readers’ involvement in higher order comprehension operations” (p.23). In this light, slower lexical processing means that L2 learners require more time and effort to process individual words. When under time pressure, L2 learners cannot compensate for a lack in processing efficiency by taking more time. Hence, readers who do not have automatized lexical recognition will have less time/attentional capacity available when they reach the syntactic stage, or worse, they may run out of such resources before even getting there. From a validity perspective, this means that one cannot simply expect time pressure to affect different L2 speakers equally; lexical processing automaticity must be also be considered when timed AJTs are used to measure any kind of linguistic knowledge.

Last, another type of evidence supporting the role of lexical processing automaticity comes directly from research on the monitoring of L2 speech (Kormos, 2000, 2006). Studies have shown that the degree of self-repair (as an overt indication of monitoring) concerning grammatical features diminishes as learners attain automaticity of speech de-/encoding mechanisms. This means that for these advanced speakers, controlled monitoring may not be a necessary process even under task conditions that allow for it (i.e., without time pressure, when focusing on the accuracy of the language, as in untimed AJTs). They rely primarily on automatic processing even when task conditions do not demand it, possibly because this is more efficient. Using lexical processing as an index of overall processing automaticity (e.g., Lim & Godfroid, 2015; Suzuki & Sunada, 2018), one may then predict that time pressure will primarily impact those who possess (comparatively) less automatic processing skills. Accordingly, our prediction, which distinguished between the effect of time pressure on controlled and automatic processing, can be summarized in Figure 1. Although the degree of automatic monitoring should be minimally affected by time pressure, that of controlled monitoring must be significantly suppressed in order to meet the time constraints of the test. Participants should differ in the degree to which they rely on controlled processes in the untimed condition, and hence they should differ in how time pressure affects them. Hence, the effect of time pressure on controlled processing should be seen most clearly in those participants who do not possess (relatively) automatized lexical processing skills.

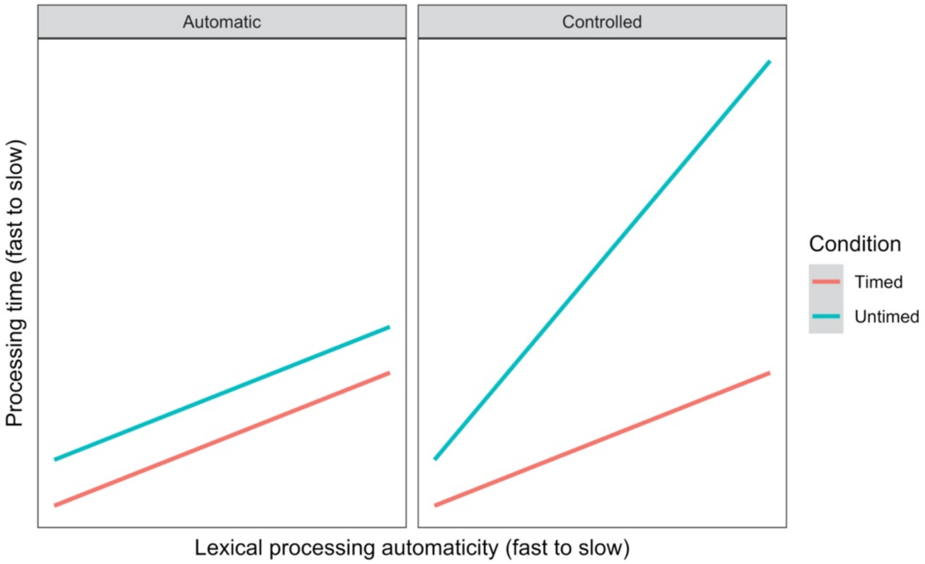

Figure 1 Expected effect of time pressure on controlled and automatic processing

The Present Study

We conducted an eye-tracking study of L1 and L2 speaker participants' processing of AJTs to test whether time pressure suppressed their eye fixations and regressions, the suppression of which would indicate the degree of controlled and automatic processing during the judgment process. In doing so, we conceptually replicated Godfroid et al. (2015) and extended their study in two ways: first by analyzing both early and late measures of eye-movement registration and second by examining how lexical processing automaticity moderated the effect of time pressure. Our research questions were:

1. In what way does time pressure suppress monitoring by controlled and/or automatic processing in AJTs as operationalized by late and early measures of eye-movement registration, respectively?
2. To what extent does lexical processing automaticity moderate the effect of time pressure on monitoring by controlled and/or automatic processing in AJTs?

We predicted that time pressure would selectively inhibit controlled processes only (see Figure 1). In operational terms, this meant that the early measures would not differ across timed and untimed conditions, whereas the late measures would be significantly diminished in the timed condition. We also expected this effect on the late measures to be moderated by the participants' lexical processing automaticity (for the discussion of the theoretical rationale for this effect in reading, see the section The Role of Lexical Processing Automaticity in Acceptability Judgment Tasks). Indeed, if time pressure primarily affects less fluent readers, who draw on more controlled processes during reading, the effect of lexical processing automaticity would provide further converging evidence in favor of our prediction that time pressure primarily affects controlled processing.

Methods

Participants

We recruited 31 L1 English speakers and 40 advanced L2 English speakers via recruitment flyers distributed to a research participant pool at a midwestern university in the United States as well as through word of mouth. Most of the L1 participants were undergraduate students pursuing a degree in fields other than linguistics or language-related disciplines. For the

L2 participants, we applied the following three inclusion criteria: (a) they had to be late learners of L2 English with age of arrival in an English-speaking country at or later than 12 years, (b) they had to have a minimum length of residence in an English-speaking country of one year, and (c) they had to be advanced L2 speakers of English with a TOEFL iBT (Internet-based Test; ETS, 2021 score of 90 or higher. Their education level varied from a high school diploma to a doctoral degree. Two participants were excluded, one L1 speaker and one L2 speaker because their performance on the semantic classification task was unusually slow or they provided the same response to all items in the timed AJT. Thus, the final sample consisted of 69 participants. Table 1 summarizes the demographic information for the participants. The L2 participants came from various L1 backgrounds (see Appendix S1 in the online Supporting Information). To facilitate comparisons between the current study and Godfroid et al. (2015), we have also summarized in Table 2 how the two studies compare to each other with respect to their major sampling and methodological processes.

Table 1 Background information for the L1 and L2 speakers participating in the study

	L1 (*n* = 30)				L2 (*n* = 39)			
Variable	*M*	*SD*	Min.	Max.	*M*	*SD*	Min.	Max.
Age at testing (years)	23.56	6.62	19	48	29.79	5.95	21	49
AoA (years)					26.71	11.96	15	34
LoR (months)					48.64	39.64	12	216
TOEFL iBT					103.76	8.13	91	116

Note. AoA = age of arrival; LoR = length of residence; TOEFL iBT = TOEFL Internet-based Test.

Table 2 Comparisons of key methodological features of Godfroid et al. (2015) and the current study

Features	Godfroid et al.	This study
Participants	L1: 20 undergraduates L2: 40 undergraduates of mixed proficiency in English	L1: 30 graduates and undergraduates L2: 39 graduate and undergraduate students with advanced proficiency in English
Research site	A midwestern university in the United States	A midwestern university in the United States
Stimuli	English sentences (*k* = 68) from Ellis (2005) covering 17 grammatical structures	English sentences (*k* = 96) from Godfroid and Kim (2021) covering six grammatical structures
Time limit	1.2 × median response time in pilot L1 speaker data	1.2 × median response time in pilot L1 speaker data[a]
Task order	T-AJT, U-AJT	T-AJT, semantic classification task, U-AJT
Eye-tracker	EyeLink 1000 Plus	EyeLink 1000 Plus
Dependent variables	Regressions (i.e., three types)	Regressions (i.e., binary) First-pass reading time Rereading time Mean fixation count

Note. T-AJT = written timed acceptability judgment task; U-AJT = written untimed acceptability judgment task.

[a] Our resulting time limits were different from those of Godfroid et al. (2015) because we adopted different test stimuli and a different L1 pilot sample.

Materials

Acceptability Judgment Task. The participants performed two types of AJTs: a written timed AJT (T-AJT) and a written untimed AJT (U-AJT). The T-AJT required the participants to make acceptability judgments under time pressure, whereas the participants had unlimited time to perform in the U-AJT. We adopted 96 test stimuli from Godfroid and Kim (2021) that covered three morphological and three syntactic structures of English: (a) third-person singular -*s*, (b) mass versus count nouns, (c) passives, (d) verb complements, (e) embedded questions, and (f) comparatives. One-half of the sentences (*k* = 48) were grammatical, but the other one-half were ungrammatical. The length of each sentence was controlled within each structure (*M* = 10.68 words, min. = 8, max. = 13). We created two lists of the same sentence stimuli and counterbalanced their grammaticality so that grammatical sentences in List A were ungrammatical in List B and ungrammatical sentences in List A were grammatical in List B (see Appendix S2 in the online Supporting Information). The full list of acceptability judgement stimuli (Maie & Godfroid 2021a) is available on IRIS (http://www.iris-database.org). We randomly assigned the participants to either one of the lists at the beginning of the experiment, taking List A for the T-AJT and List B for the U-AJT, or vice versa.

To determine an appropriate time limit for each item, we conducted a pilot study with 10 L1 speakers. The participants in the pilot study had experience teaching English as a L2 or foreign language or were pursuing a master's or doctoral degree in a language-related discipline. They were told to make judgments in the untimed condition and to respond at their natural speed. Following previous studies, we set the pilot study participants' median reaction time, which was calculated for each sentence individually, multiplied by 1.2 as an appropriate time limit for L2 speaker participants (Ellis, 2005; Godfroid et al., 2015). These time limits ranged from 2,172 ms for a nine-word sentence to 7,850 ms for an 11-word sentence. Both the T-AJT and the U-AJT were programmed with Experiment Builder in the EyeLink 1000 Plus eye-tracking system (SR Research, 2021), which recorded participants' eye-movement patterns and the accuracy of their judgments during the AJTs. Internal consistency based on Kuder-Richardson Formula 20 was .79 and .79 for the L1 participants, and .87 and .88 for L2 participants, on the T-AJT and U-AJT respectively.

Semantic Classification Task. We operationalized lexical processing automaticity as the coefficient of variation (CV) of reaction time (RT) in a semantic classification task (Segalowitz & Segalowitz, 1993; see also Lim & Godfroid, 2015).[1] In this task, the participants judged whether a series of English words that they were viewing were animate or inanimate. The full list of semantic classification stimuli (Maie & Godfroid, 2021b) is available on IRIS (http://www.iris-database.org). Segalowitz and others (e.g., Segalowitz, 2010; Segalowitz & Frenkiel-Fishman, 2005) have argued that animacy judgment tasks provide a more genuine measure of word recognition skills than do lexical decision tasks because they entail a deeper and stronger processing of meaning than do lexical decision tasks. The task included 25 animate and 25 inanimate words taken from the stimuli in the AJTs. Thus, the measure of automaticity of participants' lexical knowledge was germane to the task at hand. All of the target words were highly frequent, occurring with a mean frequency of 181.49 per million (*SD* = 208.99) in the British National Corpus (BNC Consortium, 2007; see Appendix S3 in the online Supporting Information for the stimuli).

Each trial started with an asterisk presented at the center of the screen for 500 ms that subsequently turned into a target word to be judged. Participants responded as quickly and accurately as possible by pressing either the YES-key (i.e., animate: Right-Shift) or NO-key (i.e., inanimate: Left-Shift). The order of presentation was randomized, and the entire task was programmed in DMDX (Forster & Forster, 2003). The CV was calculated for each participant by dividing the standard deviation of each individual's RT with their corresponding mean RT for responding to all 50 words. Thus, by taking into account an individual's general speed, smaller CVs are understood to reflect more stable lexical processing and larger CVs more variable processing. Due to space constraints, we have reported our data cleaning procedure for the task in Appendix S8 in the online Supporting Information. The internal consistency of the semantic classification based on Cronbach's alpha (using RTs) was .94 for L1 participants and .92 for L2 participants. Positive correlations between the mean RT and the CV indicated that faster participants also had more stable lexical processing skills, supporting the interpretation of the CV as an index of automaticity (Segalowitz & Segalowitz, 1993): L1 participants, *r*(30) = .63, 95% CI [.35, .81], *p* < .001; L2 participants, *r*(39) = .68, 95% CI [.47, .82], *p* < .001.

Procedure

After meeting the researcher, the participants signed a consent form to indicate their willingness to participate in the study. They then filled out a demographic questionnaire, the completion of which led to the main AJT tasks. The order of the tasks was kept consistent, with the T-AJT completed first, followed by the U-AJT. The semantic classification task was assigned between the two judgment tasks. For the T-AJT, the participants made judgments as quickly and accurately as possible, whereas they were allowed as much time as they needed for the U-AJT (see Appendix S6 in the online Supporting Information for the instruction sheet). After completing the study, the participants received $10 as compensation for their time and effort. The entire experiment took approximately one hour to complete.

While the participants performed the AJTs, their eye movements were recorded with the EyeLink 1000 Plus (SR Research, 2021) with a sampling rate of 1000 Hz. The recording was monocular and from the participants' right eye. The participants sat in front of a computer monitor (Dell, 1920 × 1080, with a 60 Hz refresh rate) with their head placed against a headrest that was adjusted to position their face at the center of the screen. The display screen was connected to a host computer from which the researcher administered the test stimuli. Sentences were presented individually in randomized order at the center of the screen in black (RGB: 0, 0, 0) 24-point Consolas font against a light-gray background (RGB: 204, 204, 204). Because sentences with different grammatical structures varied in length, they were left-aligned, but a one-inch margin was inserted around the edges of the screen to minimize spatial recording error (Godfroid, 2020). Both the T-AJT and the U-AJT began with an instruction page that repeated the oral instructions that the participants had been given and were followed by the camera calibration. Each task (96 sentences) was divided into four blocks of 24 sentences. The recording camera was calibrated, and the calibration was validated before every block (i.e., four times in total) with a 9-point calibration procedure. Each judgment trial began with a black dot at the top-left corner of the screen that served as a drift check (i.e., a check for spatial recording error). Upon confirming that the participants were fixating on the dot, the researcher pressed a button to reveal a test sentence. The participants then read the sentence quietly and made a judgment by pressing either the YES-key (Right-Shift) when the sentence was grammatical or the NO-key (Left-Shift) when the sentence was ungrammatical. The researcher monitored the recording quality during the experiment and performed additional recalibrations of the camera when they were needed.

Eye-Movement Measures

We analyzed four eye-movement measures, both early and late, to operationalize the degree of monitoring by controlled or automatic processing. Because we were interested in how the introduction of time pressure changed the amount of time it took for the participants to read each sentence, we mostly focused on reading behaviors at the sentence level. To this end, we treated every word in the sentences as an interest area and extracted individual fixations from each word. We then calculated the different eye-movement measures (defined below) for each word in the sentence and averaged the results across all the words in the sentence. This sentence-level analysis stands in contrast to many sentence processing studies with eye-movement registration in which fixations and regressions are extracted from only some interest areas (e.g., some words or phrases) in the sentence.

Regression. To identify regressions, we compartmentalized target sentences into four regions in relation to the primary interest area that contained the ungrammatical element in an ungrammatical sentence or the corresponding grammatical element in the grammatical counterpart of the sentence (see Figure 2). The first region spanned from the beginning of the sentence to just before the (un)grammatical element; the second and third regions consisted of the primary interest area (the ungrammatical or equivalent grammatical element) and the spill-over area; and the last region included the rest of the sentence. The spill-over region was defined as the area up to the first content word following the primary interest area (where a content word was defined as any noun, verb, or adjective). Regressions thus referred to *regressions out*, defined for English as a right-to-left saccade that was launched from the primary interest area, the spill-over area, or the sentence-final region to any earlier region. In Panel B of Figure 2, a saccade initiated from the word *enjoy* to the word *in* is one example of a regression. Although we adopted the same scheme of scanpath categories as had Godfroid et al. (2015)—(a) no regression, (b) regression(s) without finished reading of the sentence, and (c) regression(s) with finished reading—we found that regression without finished reading was quite rare in our data (see Table S5.3 in Appendix S5 in the online Supporting Information). We thus compressed the second and third categories, and performed further statistical analyses based on a binary outcome, that is, whether or not readers regressed during reading. As regressions signal an interruption to the default (forward) reading process, they can be considered a later measure of eye-movement registration, indexing controlled processing.[2]

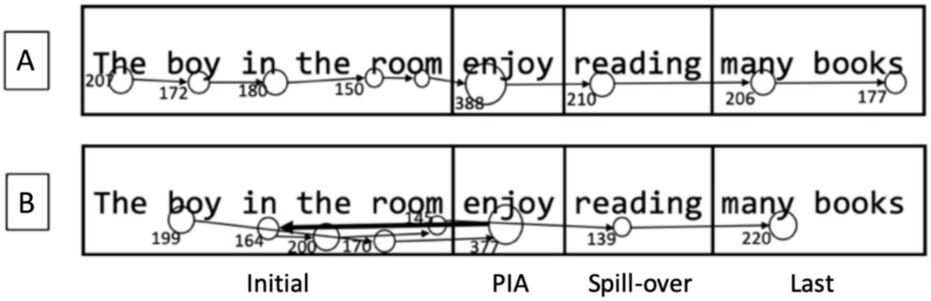

Figure 2 Two examples of fixation paths illustrating reading.
Note. Panel A: without regression; Panel B: with regression. PIA = primary interest area

First-Pass Reading Time and Rereading Time. First-pass reading time (FPT) was defined as the sum of gaze durations (i.e., length of fixations on a word during first-pass reading) across the sentence. FPT is one of the early measures of eye-movement registration. It is thought to index cognitive processes in the initial stages of sentence processing such as word recognition and lexical access (Conklin, Pellicer-Sánchez, & Carrol, 2018; Godfroid, 2020). Rereading time (RRT) was defined as the sum of all non-first-pass fixations. RRT is a late measure of eye-movement registration that, it has been argued, reflects effortful and nonautomatic processes (Conklin et al., 2018; Godfroid, 2020).

Mean Fixation Count. Mean fixation count (MFC) was the total number of fixations on a sentence (e.g., eight fixations in Figure 2, Panel B). Accounting for the rate of regression and refixation, the MFC captures how frequently the participants fixated on sentences and therefore in how much detail they processed the information. A higher MFC is arguably tied to more intense cognitive processing (Rayner, 1998). As an aggregate measure, MFC subsumes both early and late processing (Godfroid, 2020). However, our categorization of MFC as a measure of controlled processing primarily rested on the assumption that it is particularly late processes during reading (e.g., refixations following a regression) that cause the number of fixations to increase, and hence the MFC to increase. Thus in this study, MFC was considered a late eye-movement measure that is tied to controlled processing (Conklin et al., 2018).

Analysis

First, we descriptively analyzed all variables of interest, that is, judgment accuracy, CV, and eye-tracking measures, and examined the correlations among them. We computed the Pearson correlations based on the participant means for each variable. It should be noted that the eye-movement measures were extracted from correct responses on AJT sentences only so as to follow the same practice as in Godfroid et al.'s (2015) study and thereby to facilitate comparisons of our results with theirs.

Second, we modeled the four eye-movement measures for each sentence from individual participants by fitting generalized linear mixed models using Bayesian inference. We used the software R (Version 4.0.3; R Core Team, 2020) and the R package brms (Version 2.12.0; Büerkner, 2017), which provided an interface to fit the Bayesian models using Stan (Version 2.18.0; Stan Development Team, 2018), which is a probabilistic programming language for performing full Bayesian optimization. In Bayesian analysis, prior knowledge in the form of probability distributions is combined with observed data to produce posterior distributions (see Norouzian, Miranda, & Plonsky, 2018, for a review in L2 contexts). Computationally, posterior distributions are derived as the precision-weighted average of the prior and the observed data. We deemed Bayesian analysis to be an optimal approach because our study as a replication could capitalize on information from Godfroid et al. (2015) by treating their results as our prior. We have detailed the workflow of our statistical analysis in Appendix S9 in the online Supporting Information, including the specification of prior distributions, the development of statistical models, and model checking (e.g., posterior predictive checks, sensitivity analysis). We refer readers to Gelman et al. (2013) and Kruschke (2015) for comprehensive reviews of Bayesian data analysis and its workflow. Our dataset and R scripts (Maie & Godfroid 2021c) are available on both IRIS (http://www.iris-database.org) and the Open Science Foundation (https://osf.io/x2az6/) so that interested parties can recreate our analysis.[3]

Table 3 includes a summary of the generalized linear mixed models for each eye-movement measure. We modeled, for instance, regressions with a binomial distribution. The probability of regression, *p*, was predicted by a combination of predictor variables (i.e., fixed effects) and *p* was transformed by the logit link function so that the predictors corresponded to the logit-transformed regressions (i.e., the log odds of regressions) in a linear manner. In choosing fixed and random effects, we opted to develop our models using a theory- and empirically-driven approach rather than to engage in model selection (see Whittingham, Stephens, Bradbury, & Freckleton, 2006, for a criticism of the model selection approach).

Table 3 Summary of statistical models fit to eye-movement measures

Measure	Error distribution	Parameter	Link function	Comment
Regression	Binomial	*n, p*	Logit	*p* of regression predicted by linear predictors
First-pass reading time	Normal	μ, σ^2	Identity	μ predicted by linear predictors with σ^2 estimated from data. This variable was log10-transformed.
Rereading time	Normal	μ, σ^2	Identity	μ predicted by linear predictors with σ^2 estimated from data. This variable was log10-transformed
Mean fixation count	Conway-Maxwell-Poisson[a]	λ, ν	Logarithm	λ predicted by linear predictors and ν estimated from data.

Note. n = number of items, p = probability, μ = mean, σ^2 = variance, λ = mean of Poisson distribution, ν = overdispersion or underdispersion.

[a]The Conway-Maxwell-Poisson distribution is a generalization of Poisson distribution. Poisson assumes its mean and variance to be equal and does not often approximate overdispersed or underdispersed count data, which was true for our mean fixation count. Conway-Maxwell-Poisson overcomes this issue by incorporating a parameter ν, in addition to λ. See Appendix S9 for the entire model.

For fixed effects, we first included variables that were statistically significant in Godfroid et al. (2015, p. 285): the main effect of group (L1 and L2), condition (untimed and timed), grammaticality (grammatical and ungrammatical), the two-way interaction of group and condition, and the two-way interaction of condition and grammaticality. Although the main effects of group and grammaticality were not significant themselves, we included them for the two-way interactions in which they took part. To model the effect of CV, we also added the main effect of the CV, the two-way interaction of group and the CV, the two-way interaction of condition and the CV, and the three-way interaction of group, condition, and the CV. The CV was standardized within each group in the form of *z* scores. We coded the categorical fixed effects with contrast coding (i.e., –0.5 or 0.5) to avoid multicollinearity of the fixed effects, especially multicollinearity between a main effect and its interaction terms. The L1 participants, the untimed condition, and the grammatical sentences were coded with –0.5. We also included sentence length as a covariate to control for any influence of sentence length on the eye-movement measures.

For random effects, we first included by-participant and by-item random intercepts to model any variation that was due to unobserved variables specific to individual participants and items. To this, we added by-participant and by-item random slopes for condition because we expected that participants and items would differ in terms of how much they could be affected by time pressure (Hopp, 2014; McDonald, 2006). We also modeled the by-participant random slope for grammaticality and its interaction with condition because the participants could differ in how sensitive they were to ungrammaticality in general, as well as how much this sensitivity could vary across the two conditions (Vafaee et al., 2017). The fixed and random effects in a concatenated form were specified as follows in R:

Y ~ Group + Condition + Grammaticality + CV + Group:Condition + Group:CV + Condition:Grammaticality + Condition:CV + Group:Condition:CV + Length + (1 + Condition*Grammaticality | Participant) + (1 + Condition | Item)

For prior distributions, we incorporated the results from Godfroid et al. (2015) into our binomial model for the regression rate. Although Godfroid et al. analyzed regressions both with and without finished reading, we selected the results for regressions with finished reading because trials with unfinished reading were rare in our current data (see Table S5.3 in Appendix S5 in the online Supporting Information). For all the fixed effects taken from Godfroid et al.'s (2015) study, we assigned a prior distribution as a normal distribution, N(μ,σ^2), with the mean equal to the estimate of the model coefficient they reported, and the variance equal to the corresponding (squared) standard error. For all other fixed effects as well as models for the other eye-tracking measures, we set a weakly informative prior based on the pilot data from L1 speakers and on knowledge that applied to a general class of problems (e.g., a probability must be bounded between 0 and 1, fixations longer than 1,000 ms are unlikely, and so forth). This type of prior can be distinguished from noninformative (or diffuse) priors that assume complete ignorance about the given subject matter. As Gelman et al. (2013) noted, there is always some information available for almost every real-world problem, at least in the form of common sense. In our binomial model, for instance, a prior of $N\,(0,1)$ on condition expected the effect of time pressure on the logit scale to be variable from −1 to +1 within one standard deviation unit (i.e., −73.10–73.10%) when all other predictors were held constant.[4] From our review of Godfroid et al.'s (2015) study, this range seemed more than sufficient. To provide more detailed information on assigning priors, we have summarized all of our priors in numerical terms in Appendix S9 in the online Supporting Information, where we have also examined the sensitivity of the posterior distributions to the priors.

We estimated (or approximated) the posterior distributions by a Markov chain Monte Carlo simulation from four chains of 10,000 iterations each, with a warm-up period of 5,000 iterations and the amount of thinning set to 2 to (somewhat) reduce autocorrelation of the posterior samples. Stan employs a No-U-Turn Sampler as a Markov chain Monte Carlo algorithm, which is an extension of Hamiltonian Monte Carlo (Hoffman & Gelman, 2014). To determine whether the chain converged on model parameters with a stationary distribution, we monitored whether the value of R hat ($\hat{R}$) associated with each parameter (as a convergence index) was within the range of $1 \leq \hat{R} \geq 1.1$ (Gelman & Rubin, 1992). We adopted expected a posteriori (i.e., the mean of the posteriori distributions) and 95% credible intervals (CrI; i.e., highest posterior density intervals) as the estimates of model coefficients.

Results

Preliminary Analyses

Descriptive Statistics. Figure 3 summarizes the means and 95% confidence intervals for judgment accuracy as well as for the four eye-movement measures. Numeric summaries of these variables are provided in Appendix S5 in the online Supporting Information. Overall, the L1 participants made more accurate judgments than did the L2 participants. Both groups were more accurate in U-AJT than in T-AJT. Moreover, both groups were more accurate in accepting grammatical sentences than in rejecting ungrammatical ones. Focusing on the eye-tracking measures, the L2 participants were slower than the L1 participants on all the measures. These effects were most pronounced in the U-AJT, for which dramatic differences appeared in two of the late measures (RRT, Figure 3, Panel D, and MFC, Figure 3, Panel E). The L2 participants took much longer than did the L1 participants on the untimed test and their data were far more variable than were the L1 participant data.

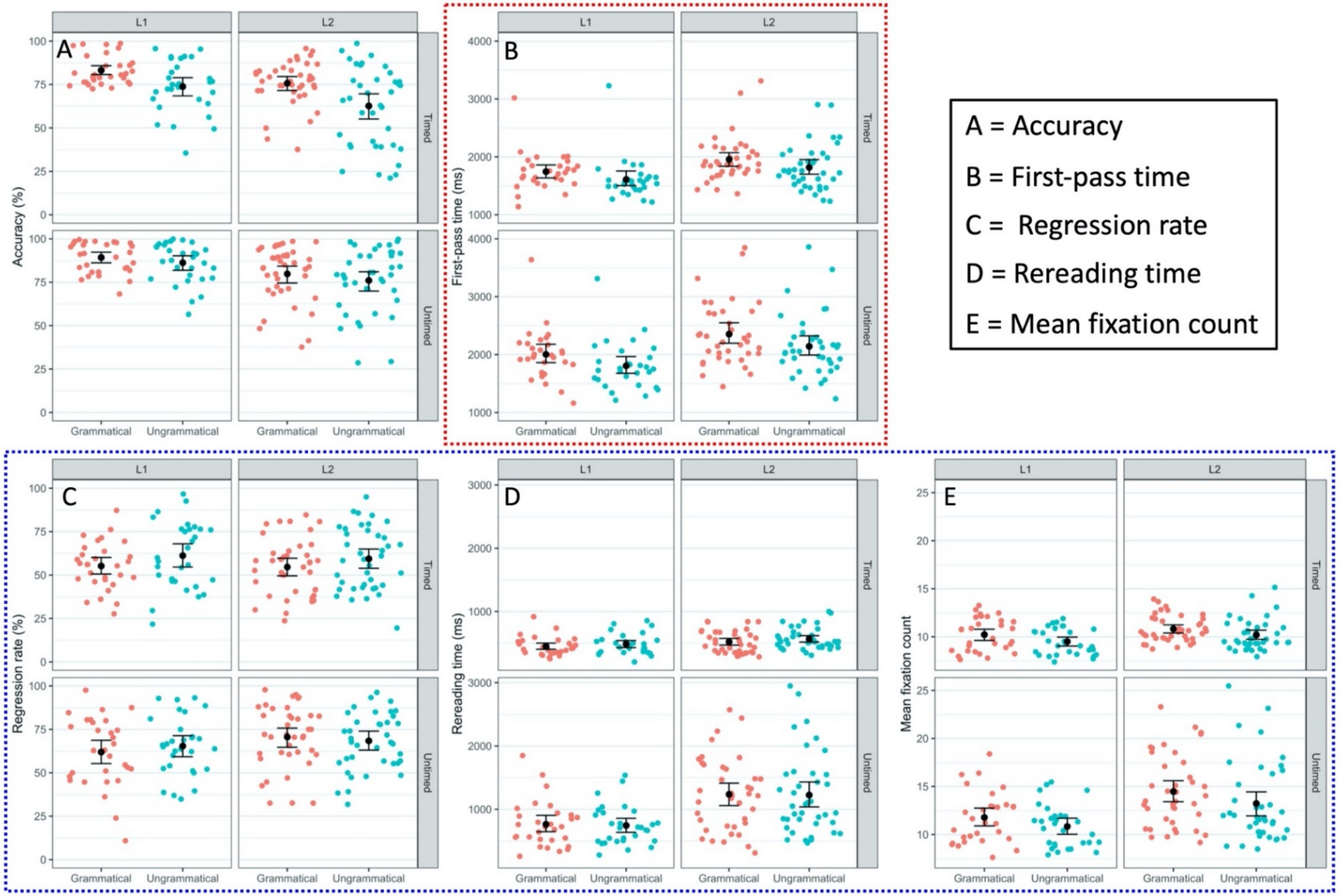

Figure 3 Descriptive statistics of accuracy rates and early/later eye-movement measures grouped in red for early measures and in blue for late measures.

Note. The points are means and the error bars correspond to 95% confidence intervals

Correlational Analyses. Figure 4 presents correlation matrices for the L1 and L2 participants' accuracy, CV, and eye-movement measures. The values on the right side of the diagonal show the results of the grammatical sentences; the values on the left side of the diagonal show the results of the ungrammatical sentences. Because Figure 4 conveys complex information (i.e., 120 correlations), we refer only to general patterns in the data. First, the late measures (regression, RRT, and MFC) positively correlated with each other in most cases, which reinforced our interpretation of their being late measures. The early measure, that is, FPT, was also positively related to the late measures, but the associations were not so strong as those among the late measures. Second, eye-movement measures were almost always negatively related to accuracy of judgments (but for a few cases). This indicated that those L1 and L2 participants who took longer to read the sentence also had lower accuracy scores. Third, notable positive correlations between the CV and eye-movement measures were only found for the L2 participants in U-AJT, and this was further restricted to the late measures only. Larger CV values here indicated more variable (and hence slower) lexical processing skills. This meant that those L2 speakers with less stable processing skills engaged in more controlled processing in U-AJT, which lent support to our decision to incorporate lexical processing automaticity as a moderating variable.

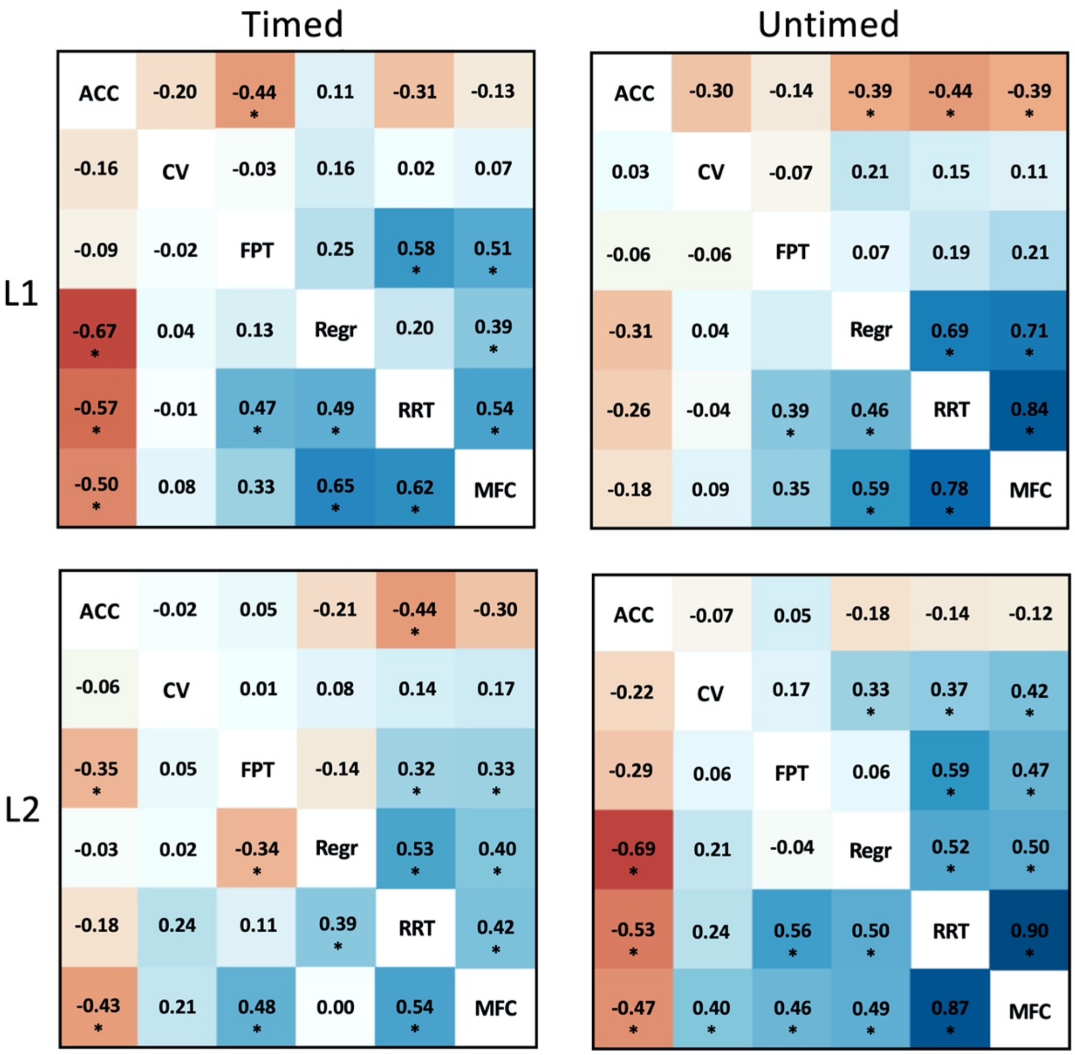

Figure 4 Correlation matrices of variables of interests.
Note. The values on the right side of the diagonal show the results of the grammatical sentences; the values on the left side of the diagonal show the results of the ungrammatical sentences. Acc = accuracy; FPT = first-pass reading time; Regr = regression; RRT = rereading time; MFC = mean fixation count. **p* < .05.

Statistical Modeling

In this section, we present the results from our statistical models. Due to the sheer complexity and multiplicity of the models, we have organized our results according to the distinction between early and late measures. In Figure 5, we present a summary of the estimates of regression coefficients and their 95% credible intervals. From these parameter estimates, we produced model-based predictions of every participant's response on each item and summarized these values with respect to the effect of time pressure and how it was moderated by lexical processing automaticity (i.e., CV). Figure 6 summarizes the results. In the remainder of the section, we restrict our discussion to regression parameters that pertained to the research questions (their estimates, standard errors, 95% credible intervals, and posterior probability are reported). Detailed numerical and graphical summaries of the generalized linear mixed models can be found in Appendix S4 and S7 in the online Supporting Information, including the posterior probability of each regression coefficient's being in a given direction (i.e., positive or negative).

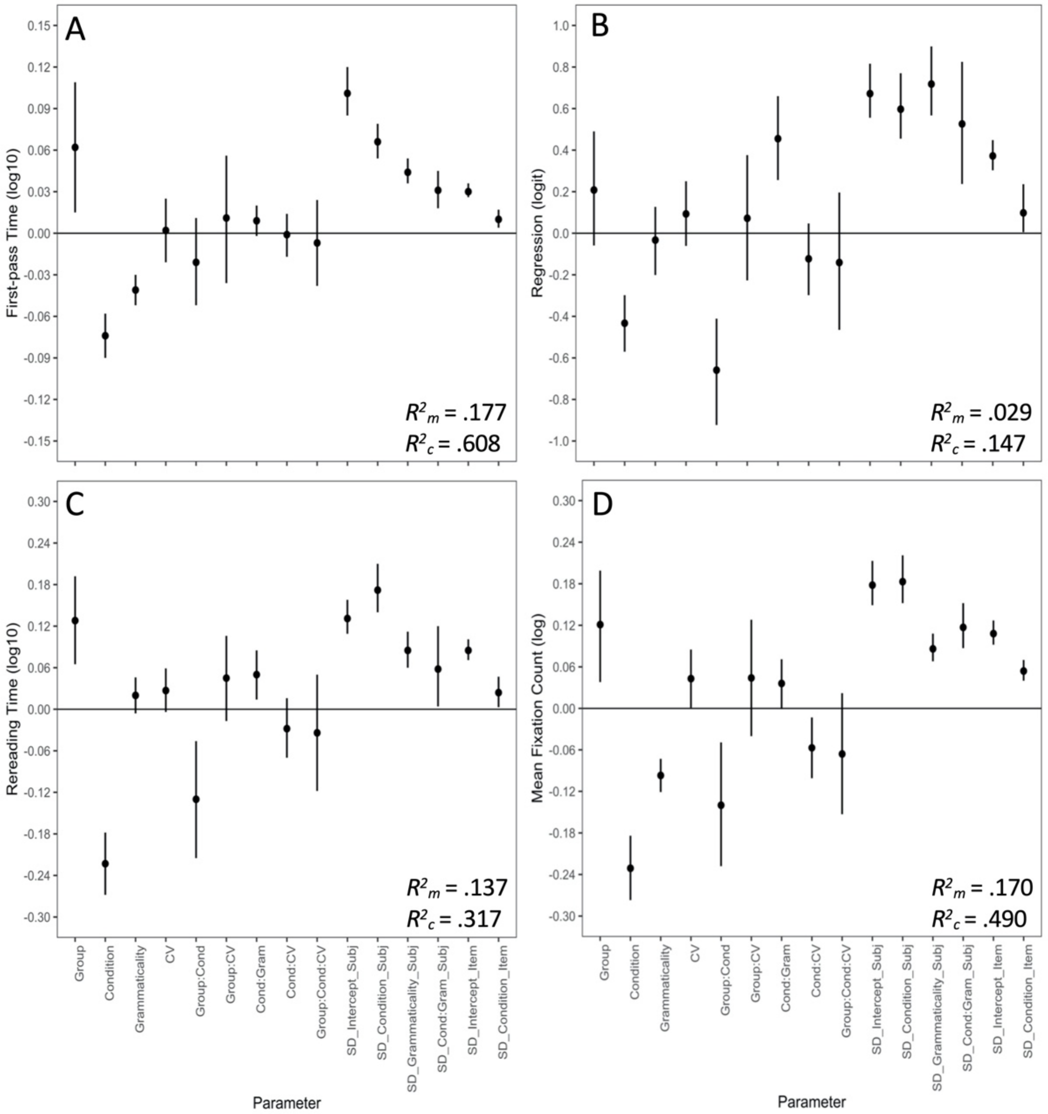

Figure 5 Summary of the generalized linear mixed models fitted to the eye-movement measures. *Note*. The points present point estimates (expected a posteriori) of the posterior distributions of the parameters and the error bars are 95% credible (higher posterior density) intervals. Panel A: First-pass reading time; Panel B: Regression; Panel C: Rereading time; Panel D: Mean fixation count. $R^2{}_m$ = Marginal R^2 (fixed effects); $R^2{}_c$ = Conditional R^2 (fixed and random effects); SD_Intercept_Subject = random intercepts for participants; SD_Condition_Subj = varying slopes of Condition among participants; SD_Condition_Subj = varying slopes of Grammaticality among participants; SD_Cond:Gram_Subj = varying slopes of the two-way interaction of Condition and Grammaticality among participants; SD_Intercept_Item = random intercepts for individual items; SD_Condition_Item = varying slopes of Condition among items.

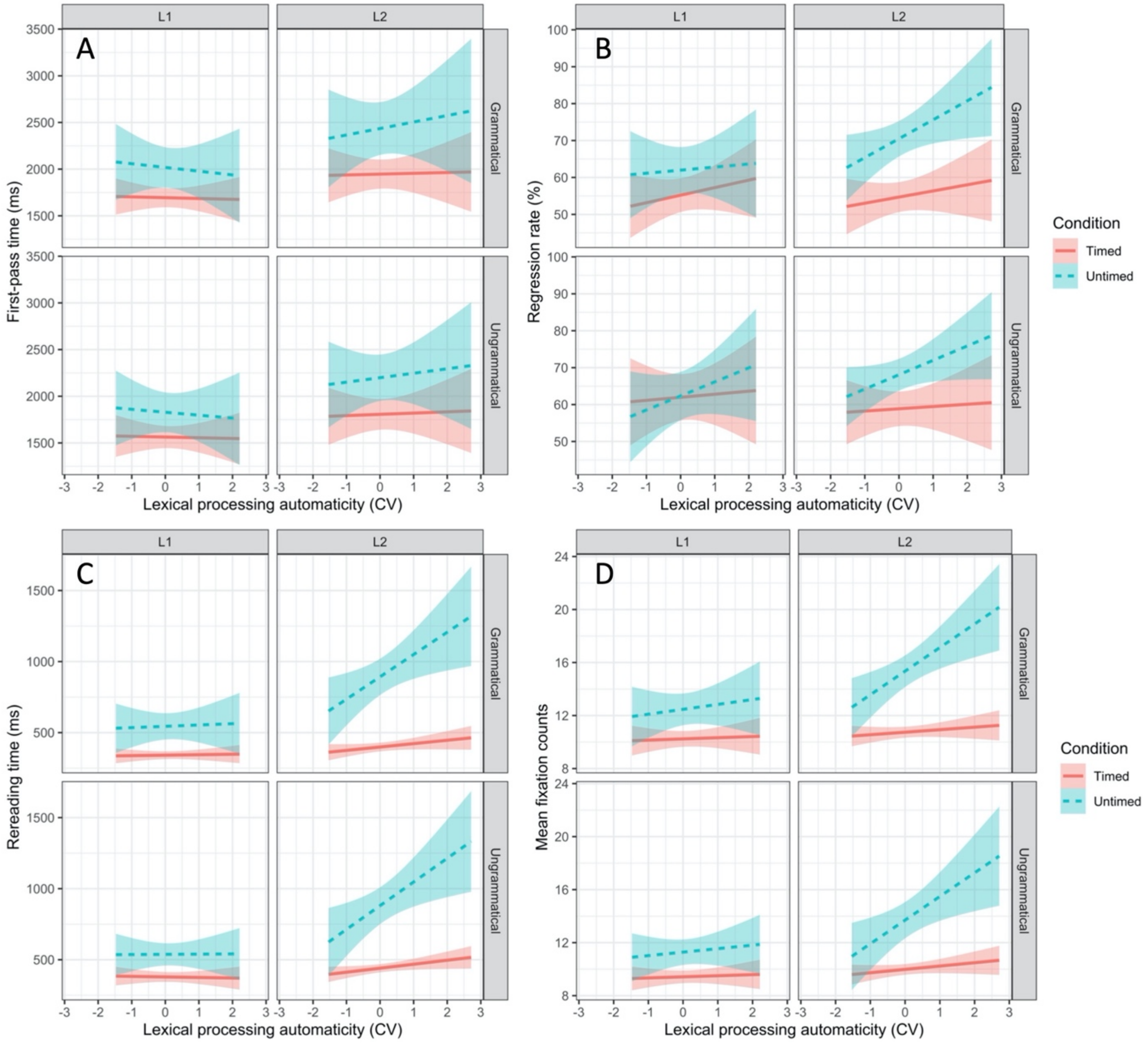

Figure 6 Model-based predictions of the effect of time pressure as a function of the coefficient of variation (CV).
Note. Panel A: First-pass reading time; Panel B: Regression; Panel C: Rereading time; Panel D: Mean fixation count. The error bands correspond to 95% credible intervals.

Early Measure. Against our expectation that monitoring by automatic processing should not be affected by time pressure, we observed an effect of time pressure on FPT, $b = -0.074$, $SE = 0.008$, 95% CrI [–0.090, –0.058], $P(b < 0) \approx 1.000$. There was also a tendency for this effect to be larger for the L2 participants than for the L1 participants, estimated through the two-way interaction of group and condition, $b = -0.021$, $SE = 0.016$, 95% CrI [–0.052, 0.011], $P(b < 0)$

= .901. We call this a tendency here because the 95% credible intervals of the interaction term spanned the value 0 and hence conveyed some level of uncertainty regarding the interaction effect. In Figure 6, Panel A, the interaction can be seen in the fact that the difference between T-AJT and U-AJT (i.e., the red and blue lines, respectively) was larger for the L2 participants. With the effect of grammaticality and CV aggregated, the difference in FPT between U-AJT and T-AJT was 232 ms for the L1 participants and 377 ms for the L2 participants.[5]

In Figure 6, Panel A shows that lexical processing automaticity (CV) did not mitigate the effect of time pressure on the early measure: the relationship between FPT and CV was similar (parallel slopes) in the timed and the untimed test. Although not directly related to the research questions, we also found an effect of group, $b = 0.062$, $SE = 0.024$, 95% CrI [0.015, 0.109], $P(b > 0) = .995$, and of grammaticality, $b = –0.041$, $SE = 0.006$, 95% CrI [–0.052, –0.030], $P(b < 0) \approx 1.000$. The L2 participants thus spent more time monitoring sentences using automatic processing than did the L1 participants. Both participant groups also had longer FPTs when a sentence was grammatical.

Late Measures. As we had predicted, there was an effect of time pressure on all three of the late measures: regression, $b = –0.433$, $SE = 0.069$, 95% CrI [–0.570, –0.298], $P(b < 0) \approx 1.000$; RRT, $b = –0.233$, $SE = 0.023$, 95% CrI [–0.268, –0.178], $P(b < 0) \approx 1.000$; MFC, $b = –0.231$, $SE = 0.024$, 95% CrI [–0.277, –0.184], $P(b < 0) \approx 1.000$. The L2 participants, moreover, were more severely affected by time pressure, as the interaction of group and condition showed: regression, $b = –0.659$, $SE = 0.132$, 95% CrI [–0.923, –0.411], $P(b < 0) \approx 1.000$; RRT, $b = –0.130$, $SE = 0.043$, 95% CrI [–0.215, –0.046], $P(b < 0) = .998$; MFC, $b = –0.140$, $SE = 0.045$, 95% CrI [–0.228, –0.052], $P(b < 0) = .998$. In Figure 6, Panels B to D, this interaction effect is clearly visible in that the difference in the reading times between T-AJT and U-AJT was far more prominent for the L2 participants, although this difference seemed to be contingent upon participants' lexical processing automaticity. When the effect of the other predictor variables was aggregated, the difference between U-AJT and T-AJT was 6% (regression), 190 ms (RRT), and two fixations (MFC) for the L1 participants and 20%, 363ms, and three fixations for the L2 participants. In addition, there was also an effect of grammaticality for RRT, $b = 0.020$, $SE = 0.013$, 95% CrI [−0.006, 0.046], $P(b > 0) = .938$, and for MFC, $b = -0.097$, $SE = 0.012$, 95% CrI [−0.121, −0.073], $P(b < 0) \approx 1.000$, and of the two-way interaction of condition and grammaticality for all three: regression, $b = 0.455$, $SE = 0.105$, 95% CrI [0.256, 0.666], $P(b > 0) \approx 1.000$; RRT, $b = 0.050$, $SE = 0.018$, 95% CrI [0.014, 0.085], $P(b > 0) = .996$; MFC, $b = 0.036$, $SE = 0.019$, 95% CrI [0.000, 0.072], $P(b > 0) = .973$. Both participant groups, thus, regressed more frequently and took longer to reread in ungrammatical than grammatical sentences, but this effect was more pronounced in the timed condition. For the number of fixations, however, the participants fixated more frequently on grammatical sentences than ungrammatical sentences, and this difference became larger in the untimed condition. Figure 3, Panels B to D, validates these interpretations with the descriptive summary of raw data.

Unlike for the early measure, lexical processing automaticity seemed to moderate the effect of time pressure on the late measures. Figure 6, Panels B to D, shows that the difference between T-AJT and U-AJT became larger with increasing values of the CV (i.e., with less stable lexical processing). This effect, furthermore, was only true for the L2 participants. In our statistical models, this difference in the moderation effect between the L1 and the L2 participants corresponded to the three-way interaction of group, condition, and CV: regression, $b = -0.141$, $SE= 0.167$, 95% CrI [−0.465, 0.196], $P(b < 0) = .806$; RRT, $b = -0.034$, $SE = 0.043$, 95% CrI [−0.118, 0.050], $P(b < 0) = .785$; MFC, $b = -0.066$, $SE = 0.044$, 95% CrI [−0.153, 0.021], $P(b <$

0) = .935 (see R interaction expression Group:Cond:CV in Figure 5, Panels B to D). It must be noted, however, that there was some uncertainty over the exact magnitude of the moderation effect because the 95% credible intervals extended widely from negative values to slightly positive values. To isolate the moderation effect and make it more visually noticeable, we plotted the difference in the late measures between T-AJT and U-AJT as a function of lexical processing automaticity (see Figure 7). The negative slope shown in Figure 7 means that time pressure more severely suppressed the late eye-movement measures of those participants with slower lexical processing automaticity (indicated by larger CV values). It was clear that this was the case for the L2 participants but not for the L1 participants.

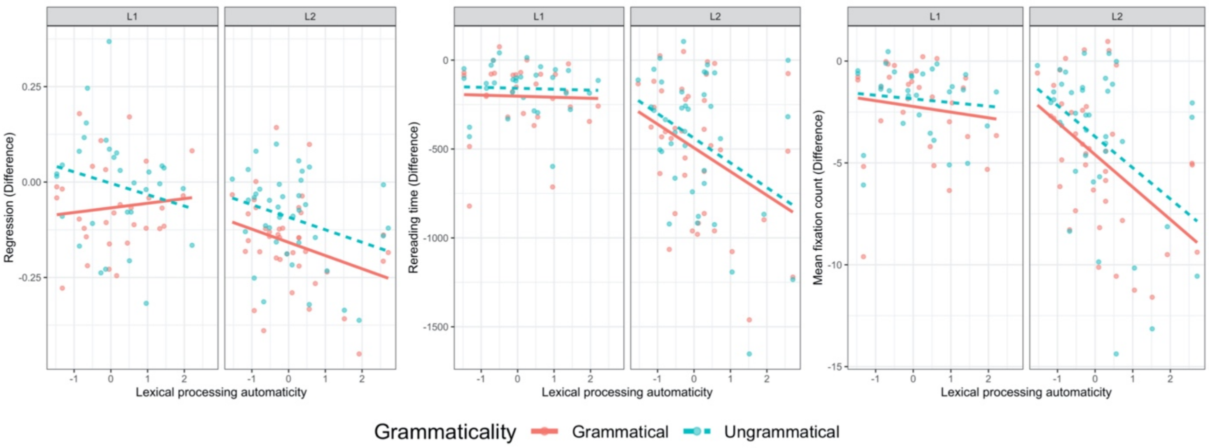

Figure 7 The difference between timed acceptability task and untimed acceptability task as a function coefficient of variation.
Note. Negative slopes indicate that time pressure more severely suppressed the eye-movement measures of those with slower lexical processing automaticity.

Discussion

We examined the role of time pressure in shifting the balance of automatic and controlled processing during linguistic (acceptability judgement) test performance by L1 and L2 speakers. Time pressure exerted robust effects on both the L1 and the L2 participants' performance, causing them to suppress fixations and regressions in the T-AJT. This effect was found to be stronger for the L2 participants, and this was especially true on the late measures of eye-movement registration (see in Figure 5, Panels B to D, the model parameter Group:Cond, the R expression for the interaction of group with condition). One unexpected but interesting finding was that time pressure also suppressed FPT, the early measure, which provided an index of the degree of monitoring by automatic processing. The fact that time pressure affected both early and late processing suggests that there may have been multiple effects at play.

First, time pressure may have forced the participants to perform the task at a faster rate than they normally would do. From a theoretical perspective, the reduction in FPT was somewhat unexpected given the strong link between lexical access (specifically an initial stage of lexical access known as a familiarity check or L_1) and saccade planning in the E-Z Reader model of eye-movement control (e.g., Reichle et al., 2009). The E-Z Reader model does not account for task effects in reading, and it therefore remains unclear how adding time pressure might speed up the process of lexical access that is believed to drive the eyes forward during reading, and hence reduce FPT. The main competitor model SWIFT (Engbert, Nuthmann, Richter, & Kliegl, 2005) does not account for task effects in reading either, although the model could more readily be extended to account for our results. In SWIFT, saccades are generated at more-or-less fixed time intervals by a mechanism described as an “autonomous timer” (Engbert et al., 2005, p. 781). According to the model, the autonomous timer is susceptible to variables such as word frequency and, importantly, it can be overruled when necessary. In the current context, adding time pressure could thus be considered to have dialed up the pace of the internal timer, which affected the generation of all the saccades and accordingly the length and number of fixations during all_stages of reading (i.e., fixations in both first-pass/early reading and later ones). This interpretation is consistent with recent extensions of the model such as that by Lewis, Shvartsman, and Singh (2013) that consider the effect of a range of variables (including task demands) that jointly and adaptively optimize eye-movement control. We must admit, however, that this interpretation is speculative at this time and testing why controlled and/or automatic processing are affected by time pressure remains an interesting topic for further research.

Another potential explanation of the result, raised by an anonymous reviewer, is that FPT may not have been process pure in that it could have indexed not only automatic but also controlled processing. Following this account, it was only the portion of controlled processing that was affected by time pressure and hence, automatic processing was in fact unaffected. Although we acknowledge that this is an interesting possibility, the current study was not designed to test this possibility. If indeed FPT indexed controlled processing, we believe this may have been due to the unique nature of the AJT being an accuracy- and form-focused task. We thus envision a follow-up study that will manipulate time pressure in a natural, comprehension-based reading task that is free from the controlled and metalinguistic characteristics inherent to the AJT. Such a study could also be used to pit competing hypotheses from theoretical models about eye-movement control in reading against each other.

Second, consistent with our predictions, we found that time pressure suppressed nonautomatic, strategic monitoring of the test sentences as manifested in the late eye-movement measures. In fact, the stronger time pressure effect on the L2 participants was more clearly evident for these late measures (see Figure 6 and, in Figure 5, Panels B to D, the two-way interaction of group with condition that was expressed in R as Group:Condition). This may be, in part, because L2 processing is slower than L1 processing (Godfroid et al., 2015; Hopp, 2014; McDonald, 2006), but more importantly, because the L2 participants may have read the sentences in a qualitatively different way, engaging more deliberate and controlled strategies to monitor the test sentences (DeKeyser, 2009; Ellis, 2005; Vafaee et al., 2017). As Bialystok (1978) originally argued, this is characteristic of the fact that in the AJT, L1 speakers tend to monitor sentences with implicit, automatized knowledge, whereas L2 speakers use explicit, controlled knowledge.

Regarding regressions during reading, which was the late measure that this study and Godfroid et al. (2015) had in common, we found that both participants groups made fewer regressions in the T-AJT than in the U-AJT (compare the lower and upper portions of Panel C in Figure 3). The results of the two studies converged in that time pressure had a stronger impact on L2 participants than on L1 participants. In the original study, however, the L1 participants' processing was unaffected by time pressure. Although it is hard to pinpoint exact reasons why both participating speaker groups were affected in the current study, one plausible explanation is to attribute these findings to particularities in participants and materials within each study (e.g., different L1 speaker samples for the pilot phase, different test stimuli, and time limits). Hence, our results not only replicated but also expanded on the findings of the original study by showing that under some test conditions L1 processing can also be affected.

In summary, time pressure indeed restricted monitoring by controlled processing. This effect, furthermore, was found to be stronger for the L2 participants than for the L1 participants, and especially so for those L2 participants with less automatic lexical processing skills. These results thus confirm the widely held view in L2 research that time pressure restricts the involvement of controlled processing when L2 speakers make acceptability judgments. What was unexpected, however, was that time pressure would also affect monitoring on the basis of automatic processing by inhibiting some degree of attention that is fundamentally required for speakers to perform the task. If we apply terminologies pertaining to explicit and implicit knowledge to our results, this meant that time pressure was instrumental in suppressing the use of explicit knowledge, yet it also degraded the application of implicit knowledge (or automatized explicit knowledge). This is in fact consistent with Hulstijn and Hulstijn (1984), in which two groups of L2 speakers with and without explicit knowledge were found to be equally affected by time pressure on a story retelling task.

The Role of Lexical Processing Automaticity

We included participants' lexical processing automaticity (operationalized with the CV) as a variable that could interact with the effect of time pressure. Indeed, the CV did moderate the time pressure effect, but this was only the case for less fluent L2 participants (as indicated by high CV values) and only on the late measures of eye movements (see three-way interaction of group, condition, and CV in Figure 5, Panels B to D). This is consistent with the view that under time pressure, word decoding and recognition must be sufficiently fast so that readers have sufficient time and resources to transition into higher-level processes such as syntactic parsing and sentence integration (Koda, 2007; LaBerge & Samuels, 1974). In Figures 4 and 6, L2 lexical automaticity was positively related to the late eye-movement measures, but only for the untimed condition. Put differently, it was in the U-AJT that the L2 participants could exhibit the full extent of their controlled processing. In the T-AJT, time pressure equalized reading behavior for all the participants regardless of their CV (see especially Figure 6). It seems that those L2 participants with higher lexical processing automaticity performed very similarly regardless of time pressure.

From a different angle, however, it can also be seen that those participants with less processing automaticity were the ones who were most vulnerable to the time pressure effect. Correlational matrices in Figure 4 show that the CV correlated positively with the late eye-movement measures, and these, in turn, correlated negatively with accuracy of judgments, yet there was no direct relationship between the CV and accuracy. This points to a model of AJT performance in which the less fluent L2 participants spent more time rereading sentences in the U-AJT, but this was detrimental to their overall accuracy on the task. Compared to the more lexically fluent participants, who engaged in less controlled processing of the U-AJT sentences, the less lexically fluent participants performed more poorly. Hence, our results highlight that the effect of time pressure is not monotonic across different L2 speakers, and lexical processing automaticity, among others, is a moderating variable. Viewed differently, our results also suggest that for those with more automatic (or more stable) lexical decoding skills, AJTs can elicit the use of automatized knowledge regardless of time pressure. For those with less automatic (or less stable) lexical processing skills, however, time pressure can limit the involvement of their controlled processes, but it can also degrade their performance based on automatized knowledge (as seen in the time pressure effect on the early measure, discussed previously). In this light, we concur with Hopp (2014) that slow lexical decoders may simply not be able to cope with time pressure. Again, there is room for one to be cautious about the task validity and reliability of T-AJTs, or, at least, be cognizant about the nature of the knowledge and processing that they are likely to elicit.

The Construct Validity of Timed and Untimed Acceptability Judgment Tasks

As a synthesis of the effects of time pressure that we have discussed so far, Table 4 provides an overall summary of the findings with respect to which process—controlled or automatic—dominated the L2 participants' performance in the T-AJT and U-AJT. It is clear that the traditional association of T-AJTs with automatic processing, characteristic of implicit (or automatized explicit) knowledge, and of U-AJTs with controlled processing, characteristic of explicit knowledge, is overly simplistic when (a) the effect of time pressure on both controlled and automatic processing is examined and (b) L2 speakers' lexical processing automaticity is considered. Rather, the construct validity of the tests depends in important ways on individual differences in the test takers' profile. U-AJTs, for instance, can evoke monitoring by controlled processing, but this is likely to be restricted to those with less automatic (or less stable) lexical decoding skills. When L2 speakers possess more automatized (or more stable) skills, they tend to monitor more smoothly via automatic processing regardless of whether the task is timed or not. Taking one step further back, this means that U-AJTs can elicit both controlled and automatic processing and hence measure both explicit and implicit (or automatized explicit) knowledge.

Table 4 Summary of findings in terms of controlled and automatic processes

	Participants' lexical processing automaticity	
Measure	High	Low
T-AJT	Automatic	Automatic (?)
U-AJT	Automatic	Controlled

Note. T-AJT = timed acceptability judgment task; U-AJT = untimed acceptability task.

We remain agnostic, however, as to whether adding time pressure to a test can truly push test takers to use their automatic processing skills, and this is especially the case for those with less automatic (or less stable) lexical processing skills (who already may not possess a lot of automaticity). As Hopp (2014) argued, having less automatized lexical skills (Koda, 2005; LaBerge & Samuels, 1974) can be detrimental during reading because slow lexical decoders are simply not capable of successfully parsing sentences under time pressure. Given our result that the effect of time pressure also extends to automatic processing (i.e., the early eye-movement measures), these readers with less lexical processing automaticity may simply break down during reading without reaching the syntactic structure building stage.[6] Future research is certainly needed to resolve the issue. Taken together, the findings of the current study reveal that time pressure may not work as theoretically predicted by some SLA researchers and more fine-grained experimental research, including individual differences research, is necessary before the use of time pressure can be deemed a valid measure of automatic processing skills.

Limitations and Future Directions

To this point, we have not addressed any limitations in our research and methodological design that are expected to be further investigated by future studies. First, we did not conduct any by-structure analyses for the effect of time pressure. It is likely that different grammatical structures produce different patterns of eye movements and hence react differently to the effect of time pressure (McDonald, 2006). Examples might be the difference between two of the structures that we chose to highlight in this study: indirect questions (e.g., The new students asked Joe where they could/*could they go to buy food) and mass/count nouns (e.g., They eat a lot of beef/*beefs in the evening). Clearly, the word order error in the question stems from a long-distance dependency that may invite more regressions and rereading than the local error on the noun, and hence these two structures may respond to the effects of time pressure differently. Unfortunately, our data only contained eight sentences of each grammatical structure in each condition (i.e., grammaticality) and this made it impossible to carry out any reliable statistical analyses. Second, our individual difference variables were restricted to lexical processing automaticity only. It would be interesting to see what other learner-internal variables can moderate the effect of time pressure. Working memory may be one such variable (as demonstrated by McDonald, 2006). Last, no study itself is complete when investigating phenomena as complex as language learning and processing. By means of conceptual replication/extension and the use of Bayesian inference, we updated our prior belief about the effect of time pressure in AJTs. We expect that future studies will now incorporate our results as their prior. As long as the effects that we observed are true and that materials and procedures are well-designed, future studies will not only replicate our results but also make our inference more certain.

Conclusion

We conceptually replicated and extended Godfroid et al. (2015) by conducting fine-grained analyses of L1 and L2 readers' eye movements in AJTs to investigate what happens while they perform the task. We found that time pressure successfully inhibited the use of controlled processing (as measured by later eye-movements) while the participants made acceptability judgments. This effect worked more strongly for the L2 participants, and especially for those participants with comparatively less lexical processing automaticity. Although these results confirmed what has been often assumed in L2 research (i.e., time pressure suppresses controlled processing), we also found that the effect of time pressure extended to monitoring by *automatic* processing (as measured by early eye-movements), such that time pressure tended to deprive the participants of the necessary time and effort required to perform the task.

Notes

1. Although we acknowledge that the CV merely quantifies processing stability (one aspect of automaticity), the construct of automaticity is multifaceted and cognitive psychologists disagree on features with which to define automaticity (see Moors & De Houwer, 2006). Rather than exhausting all of the features to adequately characterize the construct, we preferred to retain a common practice in L2 research, that is, to use the CV as a proxy for the whole construct of automaticity (rather than processing stability).

2. Regressions can be categorized further based on whether they occur during first-pass reading or are temporally and/or spatially delayed (see Godfroid, 2020, for further details). As such, regressions differ in how late in the reading process they occur, which is why we chose to refer to them as *later* rather than *late* measures of eye-movement registration.

3. It should be noted that Bayesian analysis uses simulations to approximate and estimate posterior distributions and, hence, our results will not be exactly reproducible even with the same data and with the same code.

4. $\text{Logit}(p) = \log\left(\frac{p}{1-p}\right)$. $\log\left(\frac{p}{1-p}\right)$ = 1, for example, can be exponentiated as $\frac{p}{1-p} = 2.718$, hence $p = 0.7310$.

5. We computed these difference scores by (a) entering corresponding values of predictor variables and multiplying them with the value of regression coefficients (and this is for the whole linear predictors: *a* + *bx*), (b) then transforming the resultant value back onto the original scale, and (c) subtracting the result in one condition from that in the other condition while marginalizing (averaging) the effect of other nuisance variables.

6. Although this account is plausible, it does not provide a full explanation of our results either because L2 participants with less automatic (or less stable) lexical processing skills did not

necessarily have lower accuracy scores. Specifically, in the T-AJT, the L2 participants' CV did not correlate with their accuracy of judgments (*r* = −.02 for grammatical sentences and −.06 for ungrammatical sentences; see Figure 4).

References

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*, 369–406. https://www.jstor.org/stable/1423026

Bialystok, E. (1979). Explicit and implicit judgements of L2 grammaticality. *Language Learning*, *29*, 81–103. https://doi.org/10.1111/j.1467-1770.1979.tb01053.x

BNC Consortium. (2007). The British national corpus (Version 3; BNC XML edition). http://www.natcorp.ox.ac.uk/

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

Carr, T. H., & Levy, B. A. E. (1990). *Reading and its development: Component skills approaches*. San Diego, CA: Academic Press.

Conklin, K., Pellicer-Sánchez, A., & Carrol, G. (2018). *Eye-tracking: A guide for applied linguistic research*. Cambridge, UK: Cambridge University Press.

DeKeyser, R. M. (2001). Automaticity and automatization. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 125–151). New York, NY: Cambridge University Press.

DeKeyser, R. M. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 313–348). Oxford, UK: Wiley-Blackwell.

DeKeyser, R. M. (2009). Cognitive-psychological processes in second language learning. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 119–138). Oxford, UK: Wiley-Blackwell.

DeKeyser, R. M. (2010). Monitoring processes in Spanish as a second language during a study abroad program. *Foreign Language Annals*, *43*, 80–92. https://doi.org/10.1111/j.1944-9720.2010.01061.x

Education Testing Service. (2021). *TOEFL iBT Test*. Test of English as a Foreign Language. https://www.ets.org/toefl/test-takers/ibt.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172. https://doi.org/10.1017/S0272263105050096

Ellis, R. (2015). Form-focused instruction and the measurement of implicit and explicit L2 knowledge. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 417–442). Amsterdam, The Netherlands: John Benjamins.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamic model of saccade generation during reading. *Psychological Review*, *112*, 777–813. https://doi.org/10.1037/0033-295x.112.4.777

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, *35*, 116–124. https://doi.org/10.3758/bf03195503

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Boca Raton, FL: CRC Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457–511. https://doi.org/10.1214/ss/1177011136

Godfroid, A. (2020). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. New York, NY: Routledge.

Godfroid, A., & Kim, K. M. (2021). The contributions of implicit-statistical learning aptitude to implicit second-language knowledge. *Studies in Second Language Acquisition*, 1–29. https://doi.org/10.1017/S0272263121000085

Godfroid, A., Loewen, S., Jung, S., Park, J. H., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge. *Studies in Second Language Acquisition*, *37*, 269–297. https://doi.org/10.1017/S0272263114000850

Heuer, H., Spijkers, W., Kiesswetter, E., & Schmidtke, V. (1998). Effects of sleep loss, time of day, and extended mental work on implicit and explicit learning of sequences. *Journal of Experimental Psychology: Applied*, *4*, 139–162. https://doi.org/10.1037//1076-898x.4.2.139

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1351–1381.

Hopp, H. (2014). Working memory effects on the L2 processing of ambiguous relative clauses. *Language Acquisition*, *21*, 250–278. https://doi.org/10.1080/10489223.2014.892943

Hulstijn, J. H., & Hulstijn, W. (1984). Grammatical errors as a function of processing constraints and explicit knowledge. *Language Learning*, *34*, 23–43. https://doi.org/10.1111/j.1467-1770.1984.tb00994.x

Koda, K. (2005). *Insights into second language reading*. New York, NY: Cambridge University Press.

Koda, K. (2007). Reading and language learning: Crosslinguistic constraints on second language reading development. *Language Learning*, *57*, 1–44. https://doi.org/10.1111/0023-8333.101997010-i1

Kormos, J. (2000). The role of attention in monitoring second language speech production. *Language Learning*, *50*, 343–382. https://doi.org/10.1111/0023-8333.00120

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

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Englewood Cliffs, NJ: Prentice Hall.

Kruschke, J. (2015). *Doing Bayesian data analysis: A tutorial introduction with R, JAGS, and Stan* (2nd ed.). Cambridge, MA: Academic Press.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*, 293–323. https://doi.org/10.1016/0010-0285(74)90015-2

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.

Lewis, R. L., Shvartsman, M., & Singh, S. (2013). The adaptive nature of eye movements in linguistic tasks: How payoff and architecture shape speed-accuracy trade-offs. *Topics in Cognitive Science*, *5*, 581–610. https://doi.org/10.1111/tops.12032

Lim, H., & Godfroid, A. (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied Psycholinguistics*, *36*, 1247–1282. https://doi.org/10.1017/S0142716414000137

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*, 492–527. https://doi.org/10.1037/0033-295X.95.4.492

Loewen, S. (2009). Grammaticality judgment tests and the measurement of implicit and explicit L2 knowledge. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning* (pp. 94–112). Bristol, UK: Multilingual Matters.

Maie, R., & DeKeyser, R. M. (2020). Conflicting evidence of explicit and implicit knowledge from objective and subjective measures. *Studies in Second Language Acquisition*, *42*, 359–382. https://doi.org/10.1017/S0272263119000615

Maie, R., & Godfroid, A. (2021a). *Acceptability judgment test. Materials from "Controlled and automatic processing in the acceptability judgment task: An eye-tracking study".* [Language test]. IRIS Database, University of York, UK. https://doi.org/10.48316/fmyk-1020

Maie, R., & Godfroid, A. (2021b). *Semantic classification stimuli. Materials from "Controlled and automatic processing in the acceptability judgment task: An eye-tracking study".* [Language test]. IRIS Database, University of York, UK. https://doi.org/10.48316/vcsb-4198

Maie, R., & Godfroid, A. (2021c). *Eye-tracking data and R code. Datasets from "Controlled and automatic processing in the acceptability judgment task: An eye-tracking study".* [Dataset]. IRIS Database, University of York, UK. https://doi.org/10.48316/41dw-ep98

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, *55*, 381–401. https://doi.org/10.1016/j.jml.2006.06.006

Moors, A., & De Houwer, J. (2006). Automaticity: A theoretical and conceptual analysis. *Psychological Bulletin*, *132*, 297–326. https://doi.org/10.1037/0033-2909.132.2.297

Norouzian, R., de Miranda, M. A., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning*, *68*, 1032–1075. https://doi.org/10.1111/lang.12310

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2019). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research*, 1–39. https://doi.org/10.1177%2F0267658319828413

R Core Team (2020). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

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

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

Schneider, W., & Chein, J. M. (2003). Controlled & automatic processing: Behavior, theory, and biological mechanisms. *Cognitive Science*, *27*, 525–559. https://doi.org/10.1207/s15516709cog2703_8

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*, 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge.

Segalowitz, N., & Frenkiel-Fishman, S. (2005). Attention control and ability level in a complex cognitive skill: Attention shifting and second-language proficiency. *Memory & Cognition*, *33*, 644–653. https://doi.org/10.3758/BF03195331

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics*, *14*, 369–385. https://doi.org/10.1017/S0142716400010845

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*, 127–190. https://psycnet.apa.org/doi/10.1037/0033-295X.84.2.127

Spinner, P., & Gass, S. M. (2019). *Using judgments in second language acquisition research*. New York, NY: Routledge.

SR Research (2021). EyeLink 1000 Plus [Apparatus and software]. Ottawa, Ontario, Canada: SR Research. https://www.sr-research.com/eyelink-1000-plus/

Stan Development Team (2018). Stan: A C++ library for programming and sampling [Computer software]. http://mc-stan.org

Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York, NY: Guilford Press.

Suzuki, Y., & Sunada, M. (2018). Automatization in second language sentence processing: Relationship between elicited imitation and maze tasks. *Bilingualism: Language and Cognition*, *21*, 32–46. https://doi.org/10.1017/S1366728916000857

Vafaee, P., Suzuki, Y., & Kachinske, I. (2017). Validating grammaticality judgment tests: Evidence from two new psycholinguistic measures. *Studies in Second Language Acquisition*, *39*, 59–95. https://doi.org/10.1017/S0272263115000455

Von der Malsburg, T., & Vasishth, S. (2011). What is the scanpath signature of syntactic reanalysis? *Journal of Memory and Language*, *65*, 109–127. https://doi.org/10.1016/j.jml.2011.02.004

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*, 1182–1189. https://doi.org/10.1111/j.1365-2656.2006.01141.x

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1: L2 participants' L1 Backgrounds (*n* = 39).

Appendix S2: Sentence Stimuli for the Acceptability Judgment Task.

Appendix S3: Word Stimuli for Semantic Classification Task.

Appendix S4: Summaries of Statistical Models.

Appendix S5: Descriptive Statistics.

Appendix S6: The Instruction Sheet.

Appendix S7: Posterior Distributions of the Fixed Effects Parameters.

Appendix S8: Processing of Semantic Classification Data.

Appendix S9: Bayesian Data Analysis.