Claudia Felser | Ian Cunnings | Claire Batterham | Harald Clahsen

The timing of island effects in nonnative sentence processing

Suggested citation referring to the original publication:
Studies in Second Language Acquisition 34 (2012) pp. 67–98
DOI https://doi.org/ 10.1017/S0272263111000507
ISSN (print) 0272-2631
ISSN (online) 1470-1545

Studies in Second Language Acquisition, 2012, **34**, 67–98.
doi:10.1017/S0272263111000507

THE TIMING OF ISLAND EFFECTS IN NONNATIVE SENTENCE PROCESSING

Claudia Felser, Ian Cunnings, Claire Batterham, and Harald Clahsen
University of Essex

Using the eye-movement monitoring technique in two reading comprehension experiments, this study investigated the timing of constraints on wh-dependencies (so-called island constraints) in first- and second-language (L1 and L2) sentence processing. The results show that both L1 and L2 speakers of English are sensitive to extraction islands during processing, suggesting that memory storage limitations affect L1 and L2 comprehenders in essentially the same way. Furthermore, these results show that the timing of island effects in L1 compared to L2 sentence comprehension is affected differently by the type of cue (semantic fit versus filled gaps) signaling whether dependency formation is possible at a potential gap site. Even though L1 English speakers showed immediate sensitivity to filled gaps but not to lack of semantic fit, proficient German-speaking learners of English as a L2 showed the opposite sensitivity pattern. This indicates that initial wh-dependency formation in L2 processing is based on semantic feature matching rather than being structurally mediated as in L1 comprehension.

The authors gratefully acknowledge an Economic and Social Research Council grant (RES-000-22-2508) to the first author, which supported the research carried out for this study.

Address correspondence to Claudia Felser, Potsdam Research Institute for Multilingualism, University of Potsdam, Karl-Liebknecht-Strasse 24-25, 14476 Potsdam, Germany; e-mail: felser@uni-potsdam.de.

A number of recent second-language (L2) processing studies have investigated how postchildhood language learners process unbounded or filler-gap dependencies such as (1) (see Dallas & Kaan, 2008, for a review).

(1) ***Which magazine*** *did the old lady say that she read ___ with great pleasure?*

When processing sentences such as (1), the fronted constituent *which magazine* (the filler) needs to be temporarily stored in working memory and associated with its subcategorizer, the verb *read*, when this is encountered. The two kinds of computational processes involved here—memory storage on the one hand and filler integration on the other (Gibson, 1998)—are each subject to different types of constraint. There are few published studies that have investigated the role of constraints on wh-extraction, known as island constraints (Ross, 1967), in L2 processing, and the way filler integration is accomplished in L2 sentence comprehension is still poorly understood.

Although there is evidence that L2 comprehenders, like first language (L1) speakers, are able to link a fronted wh-element to its lexical licenser during processing and can quickly evaluate its semantic fit (Williams, 2006; Williams, Möbius, & Kim, 2001), the results from other studies indicate that the processing of wh-dependencies in a L2 is not mediated by structural information to the same extent as in L1 comprehension (Felser & Roberts, 2007; Marinis, Roberts, Felser, & Clahsen, 2005). The present study builds on and extends previous research on the processing of filler-gap dependencies by examining whether L2 comprehenders are sensitive to constraints on wh-extraction, and by further investigating the process of filler integration in L2 compared to L1 processing.

L1 PROCESSING OF FILLER-GAP DEPENDENCIES

Results from a large body of L1 processing studies have shown that, in sentence comprehension, the formation of filler-gap dependencies as in (1) is constrained by processing-capacity limitations on the one hand and lexical and phrase structure information on the other. Longer dependencies tend to be computationally more costly than shorter ones, and the need to maintain a filler in working memory across structurally or referentially complex intervening material may lead to processing overload (see Gibson, 1998, 2000).

Processing-capacity limitations have also been argued to account for so-called island effects such as the unacceptability and uninterpretability of sentences like (2), which involves illicit extraction of a wh-phrase from a relative clause (RC).

(2) ****Which magazine*** *did the old lady* [$_{RC}$ *who read* ___] *laugh out loud?*

According to Kluender (2004), many island phenomena can be explained by the increased referential processing and memory load at island clause boundaries, a proposal that challenges previous grammatical accounts for islands (e.g., Chomsky, 1973). In example (2), for instance, local processing overload will result in the original gap search being abandoned when the relative pronoun *who* is encountered, preventing the parser from linking *the magazine* to the embedded verb *read* (for similar proposals and further discussion, see Alexopoulou & Keller, 2007; Hofmeister & Sag, 2010; Kluender & Kutas, 1993). It is still far from clear, however, whether processing-capacity limitations are able to account for all types of extraction islands (see Saah and Goodluck, 1995, and Wagers & Phillips, 2009, for evidence against this hypothesis). Several L1 processing studies have found evidence for the parser's sensitivity to extraction islands during online comprehension (see Phillips, 2006, for a review).

Using a plausibility manipulation as a diagnostic for dependency formation in both island and nonisland environments, Traxler and Pickering (1996), for example, showed that English L1 speakers respect relative clause islands during parsing. The analysis of participants' eye-movement patterns during their processing of sentences such as (3) revealed that a temporary dependency between *the book* or *the city* and the embedded verb *write* was formed immediately after the verb was first encountered, but only in the absence of relative clause islands as in (3a).

(3) a. *We like the book* (*city*) *that the author wrote unceasingly and with great dedication about while waiting for a contract.*
b. *We like the book* (*city*) *that the author who wrote unceasingly and with great dedication saw while waiting for a contract.*

No plausibility effects—that is, elevated reading times for *the city . . . wrote* compared to *the book . . . wrote*—were observed at the verb region in island environments such as (3b), which suggested that participants did not attempt to link the wh-filler to a potential gap inside an extraction island.

Although cognitive resource limitations no doubt play an important part in constraining the length and acceptability of filler-gap dependencies, there is ample evidence that dependency formation in L1 processing is also guided by lexical-semantic information as well as structural information such as verb subcategorization requirements and configurational properties of the emerging phrase structure representation.

According to Pickering and Barry (1991), filler integration involves lexically based association of the filler with its subcategorizer when this

is encountered. However, associating a displaced element with its lexical licenser may itself be a complex process that involves different kinds of subprocesses. Linguistic theories usually distinguish between the requirement for semantic and pragmatic compatibility between a verb and its arguments on the one hand and the saturation of valency or subcategorization frames on the other. Generative grammar, for example, has traditionally distinguished s(emantic)-selection from subcategorization or c(ategory)-selection. In head-driven phrase structure grammar, semantic information is specified as part of a lexical item's CONTENT feature, whereas subcategorization information is provided in the shape of SUBJ and COMPS lists (formerly conflated under the SUBCAT feature), with elements in these argument lists being cancelled when corresponding syntactic constituents are encountered (Pollard & Sag, 1994).

Filler integration during L1 sentence comprehension appears to involve (at least) two mental processes that correspond to this linguistic distinction, one based on subcategorization and constituent structure information—that is, gap filling in the strict sense of the term—and the other semantic in nature. In head-initial languages such as English, effects of gap filling observed at or immediately after the verb in sentences containing direct object gaps may reflect a semantic goodness-of-fit evaluation (as witnessed, for example, by semantic plausibility effects as in Traxler & Pickering, 1996), structure-based gap filling (giving rise to filled-gap effects as in Stowe, 1986), or most likely, both (Nicol, 1993). In head-final languages such as German or Japanese, semantic and structure-based gap filling can be more easily empirically dissociated (see Clahsen & Featherston, 1999; Nakano, Felser, & Clahsen, 2002).

Native speakers have also been found to link displaced constituents to structurally defined gaps whose presence is contingent on the hierarchical structural representations built during processing, calling into question Pickering and Barry's (1991) hypothesis that gap filling is purely lexically driven (see, among others, Gibson & Warren, 2004; Lee, 2004; Marinis et al., 2005; Nicol, 1993; Roberts, Marinis, Felser & Clahsen, 2007). For dependencies spanning more than one clause, sufficiently elaborate phrase structure representations will provide intermediate structural gap sites that help break long dependencies up into a series of smaller ones, thus allowing for distant fillers to undergo cyclic memory refreshing—for instance, example (1), which according to generative-transformational theories of grammar involves successive-cyclic wh-movement (Chomsky, 1973). In (4), the fronted wh-phrase *which magazine* originates as the direct object of *read* (i.e., at the position marked e_i) and moves to main clause initial position in (at least) two steps, with the unfilled specifier of the embedded complementizer phrase (CP) providing an intermediate landing site (e_i').

(4) [*Which magazine*]$_i$ *did the old lady say* [$_{CP}$ e_i' *that she read* e_i *with great pleasure*]*?*

Evidence from processing studies suggests that L1 speakers do indeed postulate such intermediate gaps during online comprehension, and that breaking up long dependencies into a set of shorter ones facilitates filler integration at the ultimate gap site (Gibson & Warren, 2004; Marinis et al., 2005). In sentences containing relative clause islands such as (2), however, reactivation of the filler at embedded clause boundaries is precluded, as the specifier of CP in this case is filled by another wh-element, potentially leading to processing overload and rendering this kind of sentence unacceptable.

In sum, establishing filler-gap dependencies in L1 processing is known to be subject to the following types of constraint: (a) complexity- or working-memory-based constraints, which limit the distance between filler and gap; (b) semantic or pragmatic constraints on filler integration such as goodness-of-fit; and (c) structural constraints on gap filling such as the availability of an unfilled argument slot. The aim of the present study is twofold: to investigate and compare the timing of island effects in native and nonnative language processing and to gain a better understanding of the way filler integration is accomplished in L2 comprehension.

DEPENDENCY FORMATION IN L2 PROCESSING

Both native and nonnative comprehenders have been shown to employ an active filler strategy (Frazier & Clifton, 1989); that is, they seek to minimize the length of filler-gap dependencies by attempting to link a fronted constituent to the earliest potential subcategorizer or other lexical licenser encountered in the input (see Williams, 2006; Williams et al., 2001).

Williams et al. (2001) used an online word-by-word plausibility judgment (or stop-making-sense) task to examine L2 learners' processing of sentences such as *Which machine (friend) did the mechanic fix the motorbike with two weeks ago?* The rationale was that if participants tried to link the wh-filler to the potential gap following the verb *fix*, then the plausibility of the wh-phrase as a direct object of *fix* should affect the number of "stop" responses at or around this verb as well as the size of the filled-gap effect at the noun phrase (NP) that follows (*the motorbike*)—that is, the degree of processing disruption caused by finding a potential direct object gap filled by an overt NP.[1] Proficient learners of English from both wh-movement and wh-in-situ backgrounds performed similarly to the L1 English controls in this study in making more stop decisions for sentences that contained an implausible filler (*which*

friend) compared to those that contained a plausible one (*which machine*), suggesting that both L1 and L2 comprehenders initially attempted to form a wh-dependency at the verb *fix*. The opposite pattern was seen at the head of the following NP (*motorbike*); that is, the number of stop decisions was smaller when the initial dependency was implausible compared to when it was plausible. The analysis of participants' reading times, however, revealed a subtle L1-L2 difference such that the filled-gap effect was already visible at the determiner in the L1 group but only seen at the following noun in the L2 group.

To further explore the timing of filled-gap effects in L2 processing, Williams (2006) replicated this study using slightly modified materials that contained longer postverbal noun phrases (e.g., *the very noisy motorbike*). In Experiment 1, a similar pattern of stop-making-sense decisions was observed as in the earlier study, with both L1 and L2 speakers providing more stop decisions in the implausible compared to the plausible condition at the verb *fix*, and a reversal of this pattern was seen in the region following the verb. This indicates that both L1 and L2 comprehenders immediately evaluated the plausibility of the filler as a potential theme argument of the verb when the verb was encountered. It is important to point out, however, that no corresponding plausibility effects were seen in participants' reading times at the verb region, except for Romance-speaking learners in the by-participant analysis. All groups showed evidence of a filled-gap effect modulated by plausibility during their processing of the postverbal NP, however.

In Experiment 2, in which the participants' task was changed to a memory task that did not explicitly require them to monitor the incoming sentence for semantic and pragmatic coherence, the reading time analyses yielded no reliable plausibility effects for the participant group as a whole. Further analyses revealed that only participants who had scored highly in the memory task showed elevated reading times for sentences containing initially plausible compared to implausible object NPs during their processing of the postverbal NP. However, whereas high-memory L1 speakers showed filled-gap effects modulated by plausibility from the determiner onward, plausibility effects were delayed until the preposition following the postverbal NP in high-memory L2 speakers. Taken together, the results from the two experiments led Williams (2006) to conclude that both L1 and L2 speakers immediately postulate direct object gaps when encountering a potential subcategorizer, but that L2 speakers' use of plausibility information may be delayed in tasks that do not explicitly require any online semantic evaluation.

It is necessary to note, however, that in both Williams et al.'s (2001) and Williams's (2006) studies, it was the (plausibility-modulated) filled-gap effects—rather than learners' stop-making-sense decisions—that

were found to be delayed in L2 processing. Plausibility and filled gaps are in fact two rather different types of diagnostic for dependency formation: Recognizing implausible fillers requires semantic goodness-of-fit evaluation between the filler and the verb, and filled gaps require participants to recognize that a postverbal argument position to which the current filler could potentially be linked is occupied already. This means that, unlike semantic goodness-of-fit evaluation, recognizing a filled gap requires not only access to verb argument structure information but also the ability to map arguments onto appropriate syntactic positions in the emerging structural sentence representation. Evidence from monolingual processing studies using event-related potentials shows that filled direct object gaps in English elicit brain responses that have been associated with early syntactic structure building (Hestvik, Maxfield, Schwartz, & Shafer, 2007), whereas implausible fillers elicit brain responses thought to index semantic processing (Garnsey, Tanenhaus, & Chapman, 1989). These findings indicate that filled gaps and lack of semantic fit each trigger qualitatively different subprocesses of gap filling. Combining both diagnostics in the same experiment, as in Williams and colleagues' studies, makes it difficult to dissociate semantic plausibility effects from effects based on argument competition, or to tell whether initial dependency formation involved both structurally mediated gap filling and semantic goodness-of-fit evaluation, or semantic evaluation only.

Evidence that learners may not postulate purely structurally defined gaps during L2 processing comes from studies by Marinis et al. (2005) and Felser and Roberts (2007). Marinis et al., for example, found no evidence for intermediate syntactic gaps in proficient L2 learners' online representations of sentences such as (5a).

(5) a. *The actress* [$_{CP}$ ***who***$_i$ *the journalist suggested* [$_{CP}$ e_i' *that the talented writer had inspired* e_i]] *will go on stage tonight.*
b. *The actress* [$_{CP}$ ***who***$_i$ *the journalist's suggestion about the talented writer had inspired* e_i] *will go on stage tonight.*

Marinis et al.'s results showed that the potential availability of a structurally defined intermediate gap (marked e_i') for the wh-filler *who*, referring to *the actress*, in (5a) led to significantly shorter reading times at the subcategorizing verb (e.g., *inspired*) in (5a) compared to (5b), in which no intermediate gap is available, only for L1 English speakers but not for L1 German, Greek, Japanese, or Chinese speakers. In other words, whereas for L1 English speakers filler integration at the ultimate gap site was facilitated by the possibility of breaking up the long wh-dependency into smaller steps—compare also example (4)—no such facilitation or intermediate gap effect was observed for L2 learners from either wh-movement or wh-in-situ backgrounds.

In a similar vein, Felser and Roberts (2007) found evidence for the presence of indirect object gaps in L1 but not in L2 listeners using the cross modal priming paradigm. When listening to sentences that contained fronted indirect objects such as *John saw the peacock* [*to which*]$_i$ *the small penguin gave the nice birthday present* e_i *in the garden last weekend*, only L1 speakers, but not advanced Greek-speaking learners of English, were found to mentally reactivate the referent of the indirect object phrase, *the peacock*, at the purported structural gap site (i.e., at the point marked e_i). Like the results from Marinis et al.'s (2005) study, this finding suggests that filler integration in L2 language processing may not be mediated by purely structurally defined gaps.

These findings can be accounted for by the shallow structure hypothesis for L2 processing (Clahsen & Felser, 2006a, 2006b), according to which late L2 learners are less sensitive than L1 speakers to structural cues in the input and have difficulty computing detailed hierarchical phrase structure representations in real time. Like other models of L2 processing (e.g., MacWhinney, 2005), however, Clahsen and Felser's model currently lacks any assumptions about when during real-time processing different types of information become available.

It is important to note that the absence of configurational gap effects in L2 processing (Felser & Roberts, 2007; Marinis et al., 2005) only provides indirect support for the hypothesis that filler integration in L2 comprehension may be semantically rather than structurally driven, and more research is needed that probes more directly into the nature of dependency formation in L2 processing. Comparing the time course of gap filling in L1 versus L2 comprehension might advance one step closer toward developing an empirically founded model of the time course of real-time L2 processing.

SENSITIVITY TO ISLANDS IN L2 COMPREHENSION

It is possible to assume, in view of the evidence presented here, that L2 learners are nativelike in trying to link a filler to the closest potential lexical licenser. For the language comprehension system to adopt an active filler strategy makes sense considering that processing and memory resources are limited, and these are likely to be drained more easily in L2 than in L1 processing. The parser's desire to keep dependencies short is seemingly obviated by island constraints, however, which prevent dependency formation under certain conditions.

Compared to the large number of L1 processing studies on the role of island constraints on parsing, very little is still known about what happens in L2 sentence comprehension when a potential subcategorizer or

gap is encountered in the input but dependency formation is prohibited by an island constraint. In a pioneering study, Juffs and Harrington (1995) found that Chinese-speaking learners correctly rejected sentences containing island violations such as *What did Sam see the man who stole?* around 90% of the time on average in an online grammaticality judgment task, a figure that was slightly lower if subject-controlled word-by-word stimulus presentation was used. In a later replication study by Juffs (2005), also using word-by-word presentation, island violations were judged as ungrammatical somewhat less frequently (around 70%) overall, even by L1 English speakers. These differences might have been due to differences in the experimental materials used or in participants' instructions, or to general processing demands being greater in word-by-word compared to whole-sentence presentation as used by Juffs and Harrington with island violations more likely to go unnoticed in word-by-word presentation.

Using a speeded grammaticality judgment task, Sato (2007) found that proficient Japanese-speaking learners of English were significantly faster and better at detecting relative clause island violations in sentences such as *It was the toy that the babysitter saw the children who enjoyed* compared to semantic violations in sentences such as *It was the toy that the babysitter said the children had excited.* Although the learners' sensitivity to extraction islands (as measured by A′ scores) was lower than that of the L1 controls, the learners' well-above-chance-level score of .85 suggests that they were generally able to detect island violations even under processing pressure.[2] It is necessary to note, however, that metalinguistic end-of-sentence judgments cannot provide any conclusive information about learners' online sensitivity to extraction islands, as they do not tap directly into ongoing parsing or comprehension processes.

The current study investigates learners' sensitivity to extraction islands during online comprehension using eye-movement recording during reading, a technique that provides a detailed millisecond-by-millisecond record of participants' reading times throughout a sentence, and that has also been shown to be suitable for the study of L2 processing (Frenck-Mestre, 2005). Examining both early and later eye-movement measures reveals how processing unfolds over time (Clifton, Staub, & Rayner, 2007; Staub & Rayner, 2007). Eye-movement measures such as first fixations or first-pass reading times are thought to index the earliest stages of processing including and immediately following lexical access. Regression path duration (also known as go-past time) is the sum of all fixations on a region until it is first exited to the right and thus may also include regressive eye movements to earlier sentence regions. Regressive eye movements are likely to reflect processing difficulty or disruption that occurred during participants' initial reading of a given region, whereas regression path duration may also index slightly later processes such

as the integration of the words or phrases in the current interest region with the preceding text (Rayner, Warren, Juhasz, & Liversedge, 2004). Second-pass or rereading time, in contrast, includes additional fixations within a region only after the eyes have already moved away from it and is thought to index later processes such as reanalysis or discourse integration. In short, whereas early measures such as first-pass reading times are likely to reflect automatic (first-pass) processing, later measures such as rereading times can be taken to reflect later (second-pass) processing, with regression path duration possibly reflecting aspects of both (Clifton et al., 2007).

To get a clearer picture of the relative timing of island effects in L1 versus L2 comprehension, it would seem prudent not only to use a highly time-course sensitive experimental technique such as eye-movement monitoring but also to examine learners whose general reading speed is comparable to that of native speakers. Comparing the timing of island effects using both semantic (Experiment 1) and structural (Experiment 2) diagnostics may provide information about the nature of dependency formation in L1 compared to L2 processing. If learners postulate structurally defined gaps immediately on coming across a potential gap site during L2 comprehension, whereas their use of plausibility information is potentially delayed, as has been suggested by Williams (2006), then island effects may be observed earlier during processing in the second experiment, which used filled gaps as a diagnostic, compared to the first experiment, which used a plausibility diagnostic. However, if online filler integration in L2 comprehension is semantics driven and not initially guided by subcategorization or constituent structure information—a possible extension of Clahsen and Felser's (2006a, 2006b) shallow structure hypothesis—then filled-gap effects, but not plausibility effects, may be found to be delayed in L2 sentence processing.

EXPERIMENT 1

To investigate and compare island sensitivity in L1 and L2 comprehenders, an eye-movement monitoring experiment was first carried out, which used plausibility as a diagnostic for dependency formation inside extraction islands like in Traxler and Pickering's (1996) original experiment. To minimize the possibility of the results being affected by factors such as poor L2 grammar proficiency, typological L1-L2 distance, or differences in writing scripts that might have led to a potentially confounding slowdown of the learners' general reading or processing speed, the L2 group was comprised of proficient German-speaking learners of L2 English. Note that wh-extraction from relative clause islands is also unacceptable in German (as has been confirmed experimentally by Alexopoulou & Keller, 2007).

Method

Participants. Twenty-four adult German-speaking learners (7 males and 17 females, mean age = 22.6) of L2 English and 39 L1 English-speaking controls (11 males and 28 females, mean age = 23.7), all recruited from the University of Essex community, participated in the first experiment. The L2 participants had first been exposed to English between ages 7 and 14 (M = 11.1, SD = 1.6) in a formal school setting and had spent an average of 3.9 years (SD = 4.9) immersed in English at the time of testing. The learners scored between 60 and 100% (M = 85.04%) in the Quick Placement Test (Oxford University Press, 2001), which indicates that their general level of English proficiency ranged from upper intermediate to upper advanced. All participants had normal or corrected to normal vision and were naïve as to the ultimate purpose of the experiment. They received a small fee for their participation.[3]

Materials. The experiment had a 2 × 2 design with the materials modeled after those of Traxler and Pickering (1996). Twenty-eight short paragraphs were created consisting of a lead-in sentence followed by a second (critical) sentence in four experimental conditions as illustrated in (6a–d).[4]

(6) *The new shampoo was featured in the popular magazine.*
   a. No constraint, plausible
      Everyone liked the magazine that the hairdresser read extensively and with such enormous enthusiasm about before going to the salon.
   b. No constraint, implausible
      Everyone liked the shampoo that the hairdresser read extensively and with such enormous enthusiasm about before going to the salon.
   c. Island constraint, plausible
      Everyone liked the magazine that the hairdresser who read extensively and with such enormous enthusiasm bought before going to the salon.
   d. Island constraint, implausible
      Everyone liked the shampoo that the hairdresser who read extensively and with such enormous enthusiasm bought before going to the salon.

The critical second sentences in (6a–d) all contained a transitive main verb whose direct object was modified by a relative clause introduced by the relative complementizer *that*; additionally, all of them were grammatical. The two island constraint conditions (6c, d) contained a further level of embedding in the form of a second relative clause introduced by the relative pronoun *who*, whose presence renders this clause an island for extraction.

In the two no-constraint conditions (6a, b), the earliest potential gap site was at the embedded verb *read.* The plausibility of the wh-filler's

referent, the direct object NP *the magazine* or *the shampoo*, as an object of the embedded verb was manipulated as a diagnostic for whether a dependency would be formed at this point. In other words, if participants applied an active filler strategy and initially tried to link the filler to *read*, then the implausibility of *the shampoo* as a direct object of *read* should lead to a slowdown in processing at or following the verb *read* in (6b) compared to sentences that contain initially plausible direct object fillers such as *the magazine* in (6a). In the two island constraint conditions, in contrast, dependency formation at the verb *read* should be blocked by the wh-pronoun *who* (see Traxler & Pickering, 1996). It is important to note that the ultimately correct gaps were located further downstream, at the preposition *about* in (6a, b) and at the verb *bought* in (6c, d), and that all sentences were globally plausible.[5]

The lead-in sentences always mentioned both the plausible and implausible manipulated NPs, with the relative ordering of the two NPs counterbalanced, such that half contained the implausible NP followed by the plausible NP, and half contained the plausible NP followed by the implausible NP. The manipulated nouns in the plausible and implausible conditions were matched for length and word form frequency using the CELEX database (Baayen, Piepenbrock, & van Rijn, 1993). The mean length of these nouns was 5.00 characters in the plausible condition and 5.04 characters in the implausible condition. Plausible nouns had a mean word form frequency of 75.04 per million and implausible ones a mean of 82.96. *T* tests showed that neither of these differences were significant, either for length, $t(54) = 0.090$, $p = 0.929$, or frequency, $t(54) = 0.314$, $p = 0.753$. The subcategorization biases of the verbs at the first potential gap site were also assessed using a sentence completion task (Trueswell, Tanenhaus, & Kello, 1993), to ensure that all critical verbs readily admitted direct object NPs. To this end, 16 L1 English speakers were given a list of sentence fragments consisting of a proper noun followed by a potentially transitive verb (e.g., *John hunted*) and asked to complete the fragment with the first suitable continuation that came to their mind. Only verbs that elicited 50% or more ($M = 69\%$) direct object continuations were used in the experimental materials.

The strength of the plausibility manipulation was further pretested by asking 10 L1 English speakers to rate the plausibility of 72 short sentences such as *The hairdresser read the magazine/shampoo* on a scale from 1 (*plausible*) to 5 (*implausible*). Two counterbalanced presentation lists were created to ensure that each participant only saw one member of each plausible-implausible sentence pair. On the basis of these scores, those NP pairs that showed the greatest difference in plausibility ratings were selected. For the 28 pairs selected, the mean ratings were 1.21 for plausible and 4.49 for implausible NPs, a difference that proved highly significant, $t_1(9) = 17.05$, $p < .001$; $t_2(27) = 29.69$, $p < .001$.

The experimental items were distributed across four presentation lists using a Latin square design, mixed with 32 fillers and pseudorandomized. Ten of the filler items were structurally similar to the experimental ones, 5 of which were globally plausible, and 5 mildly implausible. A further 5 of the remaining 24 fillers were also mildly implausible globally. The purpose of including globally implausible fillers was to help prevent participants from developing a strategy of ignoring critical sentences that were initially implausible on the assumption that all sentences would ultimately prove globally plausible. To ensure that participants read the experimental sentences properly for meaning, two thirds of all trials were followed by a yes/no comprehension question, half of which required a *yes* and half a *no* response.

Readers' sensitivity to extraction islands should be reflected statistically in an interaction between the factors plausibility and constraint. Given Traxler and Pickering's (1996) findings, the L1 speakers were expected to show plausibility effects at or around the verb *read* in nonisland environments only, in the shape of longer reading times for locally implausible compared to plausible sentences in the no-constraint pair (6a, b). Following previous findings that suggest that L1 speakers are sensitive to extraction islands from early on during processing (e.g., Traxler & Pickering), the predicted interaction might already be visible in early eye-movement measures—that is, during the L1 speakers' initial reading of the critical verb region. If L2 learners are also sensitive to plausibility information and respect extraction islands during processing, plausibility effects should be restricted to the no-constraint conditions in the learner data as well. A lack of sensitivity to island constraints, in contrast, should be reflected in a main effect of plausibility that is not modulated by the factor constraint. Moreover, given Williams's (2006) suggestion that sensitivity to plausibility information in L2 processing may be delayed, main effects of, or interactions with, the factor plausibility might be expected to be restricted to later eye-movement measures in the L2 group, or to be visible only at later sentence regions.

Procedure. All participants were tested individually in a dedicated, quiet laboratory room. The experiment began with the presentation of five practice items to familiarize participants with the procedure. The experimental items were presented in Courier New font in black letters against a white background on a computer screen and were displayed across three lines of text.

Participants' eye movements were recorded as they read through the experimental paragraphs presented on the screen using the head-mounted EyeLink II system. The system records eye movements through two cameras that are mounted on a headband and held in place with a cradle on the participants' head. Although participants read binocularly,

only information from the right eye was recorded, at a sample rate of 500 Hz. At the beginning of an experimental session, the eye tracker was calibrated on a 9-point grid, and calibration was checked again before each new trial. Participants were asked to read the experimental paragraphs silently for comprehension at their normal reading speed, and to press a button on the control pad when they had finished. The end-of-trial comprehension questions required a binary yes/no push-button response.

The L1-English speakers completed the experiment in a single session lasting about 30–40 min. The L2 participants were tested in two separate sessions of similar length, with the eye-movement experiment and a brief vocabulary test administered in the first session and the proficiency test in the second. The vocabulary test consisted of a checklist containing all critical vocabulary items, including the manipulated NPs and critical verbs, and the learners were asked to read through the list carefully and circle any words that were unfamiliar.

Data Analysis. To examine the presence and timing of island effects, reading times were analyzed for two regions of text: the critical region, consisting of the verb at the first potential gap site and the following word—for example, *read extensively* in (6) above—and the spillover region, consisting of the following words up until the end of the line (e.g., *and with*). Three reading time measures will be reported for these regions. First-pass reading time is the summed duration of all initial fixations on a region until that region is exited to either the left or right. Regression path duration is the sum of all fixations on a region until this region is first exited to the right, and rereading time is the summed duration of all fixations on a region after it first exited to either the left or right.

Short fixations of 80 ms or below within one degree of visual arc of another fixation were automatically merged, and any other extremely short ($\leq$ 80 ms) or long ($>$ 800 ms) fixations removed before any further analysis. Individual outlier data points beyond 2.5 *SD*s from a participant's mean for each measure at each region were also removed prior to the statistical analysis. Reading times for trials in which track loss occurred or in which a region was initially skipped were treated as missing data, and trials in which a region was not fixated again following the first pass contributed a value of zero to the calculation of average rereading times.

To establish whether the two groups' reading time patterns at the critical and spillover regions were statistically different, a series of preliminary mixed ANOVAs were carried out with plausibility (plausible, implausible) and constraint (no constraint, island constraint) as within-subjects factors, and group (L1 speakers, L2 learners) as a between-subjects factor, for each of the two interest regions. For regions in which interactions with the

factor group were observed, the reading time data from the L1 and L2 speakers were analyzed separately. In the absence of any significant interactions with group at a given region, no separate per-group analyses were conducted.

Results

The L2 participants answered 84% of the end-of-trial comprehension questions correctly, and the L1 speakers 86% overall, indicating that both groups paid attention to the task and read the stimulus items for meaning. Track loss accounted for 0.7% of the L1 and 0.9% of the L2 data. Items that the L2 participants had indicated contained unknown vocabulary items were also removed, which affected a further 1.9% of the L2 data. Skipping rates for the two reported regions were less than 6.8% in both groups, and the removal of outliers led to the loss of no more than 4.2% of the L2 and 4.7% of the L1 speakers' remaining data per measure and region.

Summaries of participants' reading times and the results from the preliminary ANOVAs are provided in Tables 1 and 2, respectively.

Critical Region. Table 1 shows that the L2 group's reading time patterns across the four experimental conditions differed from those of the L1 controls in the critical region, and most notably so in first-pass reading times. Results from the preliminary ANOVAs revealed significant main effects of constraint in first-pass and regression path times, with reading times generally being longer in the no-constraint conditions. Main effects of group, which indicate that the L2 participants tended to read the experimental items more slowly than the L1 participants, were significant for first-pass times by both subjects and items, and by items only in the rereading times. Most important, there was a three-way interaction between constraint, plausibility, and group in the first-pass times, and a Constraint × Group interaction in the regression path times in the analysis by items. Given these interactions with the factor group at this region, the reading time data from each participant group were analyzed separately.

The L1 speakers showed a significant main effect of constraint in first-pass times, $F_1(1, 38) = 14.30$, $p < .01$; $F_2(1, 27) = 16.81$, $p < .001$, reflecting the fact that this region was read faster in the island constraint than in the no-constraint conditions, but there was no interaction. In regression path times, only a hint was found of a Plausibility × Constraint interaction, which was not statistically reliable, $F_1(1, 38) = 3.33$, $p = .076$; $F_2(1, 27) = 3.06$, $p = .091$. The analysis of the L1 speakers' rereading times showed a main effect of plausibility, $F_1(1, 38) = 13.42$, $p < .01$; $F_2(1, 27) = 12.93$, $p < .01$, which was qualified by a significant Plausibility × Constraint interaction,

Table 1. Reading times in ms for three eye-movement measures at the critical and spillover regions in Experiment 1

	L1 speakers (n = 39)			L2 learners (n = 24)		
Condition	First-pass reading time	Regression path time	Rereading time	First-pass reading time	Regression path time	Rereading time
Critical region						
No constraint–plausible	564 (144)	689 (191)	263 (238)	595 (208)	780 (270)	452 (408)
No constraint–implausible	533 (134)	722 (206)	412 (375)	683 (239)	819 (279)	467 (437)
Island constraint–plausible	471 (157)	703 (221)	339 (307)	573 (167)	706 (265)	501 (483)
Island constraint–implausible	484 (136)	655 (187)	344 (235)	554 (173)	705 (230)	487 (400)
Spillover region						
No constraint–plausible	416 (133)	542 (169)	132 (145)	488 (236)	624 (228)	213 (176)
No constraint–implausible	394 (125)	727 (267)	179 (160)	476 (204)	754 (407)	272 (314)
Island constraint–plausible	367 (121)	483 (176)	132 (130)	459 (160)	573 (240)	246 (219)
Island constraint–implausible	372 (112)	518 (203)	159 (159)	452 (175)	533 (214)	248 (253)

Note. *SDs* are given in parentheses.

Table 2. Summary of results of preliminary ANOVAs at the critical and spillover regions in Experiment 1

		Critical region			Spillover region		
Factor	F1-F2	First-pass reading time	Regression path time	Rereading time	First-pass reading time	Regression path time	Rereading time
Constraint	F1 (1, 61)	20.12[c]	7.15[c]	0.78	5.70[b]	27.48[c]	0.03
	F2 (1, 27)	21.23[c]	10.45[c]	0.49	6.50[b]	32.55[c]	0.10
Plausibility	F1 (1, 61)	0.01	0.08	2.53	0.58	11.64[c]	4.53[b]
	F2 (1, 27)	0.02	0.09	3.96[a]	0.28	7.83[b]	6.74[b]
Group	F1 (1, 61)	6.37[b]	1.70	2.74	5.57[b]	1.28	5.05[b]
	F2 (1, 27)	11.92[b]	3.16[a]	25.41[c]	15.82[c]	1.60	20.57[c]
Constraint × Plausibility	F1 (1, 61)	1.30	2.37	3.92[a]	0.60	15.49[c]	2.12
	F2 (1, 27)	1.34	2.79	3.25[a]	0.53	5.78[b]	7.21[b]
Constraint × Group	F1 (1, 61)	0.01	2.25	0.48	0.11	0.00	0.20
	F2 (1, 27)	0.30	5.34[b]	0.58	0.04	0.00	0.38
Plausibility × Group	F1 (1, 61)	2.56	0.44	2.45	0.00	2.03	0.04
	F2 (1, 27)	0.99	0.68	2.11	0.06	3.63[a]	0.02
Constraint × Plausibility × Group	F1 (1, 61)	7.61[b]	0.26	1.71	0.34	0.06	0.51
	F2 (1, 27)	6.02[b]	0.25	1.07	0.19	0.10	1.36

[a] $p < .1$.
[b] $p < .05$.
[c] $p < .01$.

$F_1(1, 38) = 6.94$, $p < .05$; $F_2(1, 27) = 9.72$, $p < .01$. To follow up the observed interaction, planned paired sample *t* tests were carried out that showed a significant difference between the no-constraint pair only, $t_1(38) = 4.29$, $p < .001$; $t_2(27) = 5.03$, $p < .001$, which confirmed that the critical region was reread more slowly in the implausible compared to the plausible condition. No such difference was found between the two island constraint conditions, $t_1(38) = 0.13$, $p = .897$; $t_2(27) = 0.38$, $p = .707$.

The L2 learners patterned differently from the L1 speaker controls in their first-pass reading times in showing a main effect of constraint, $F_1(1, 23) = 7.11$, $p < .05$; $F_2(1, 27) = 10.70$, $p < .01$, which was modulated by a significant Plausibility × Constraint interaction, $F_1(1, 23) = 5.08$, $p < .05$; $F_2(1, 27) = 6.03$, $p < .05$. *T* tests confirmed that the learners' first-pass reading times in the no-constraint conditions were longer for implausible than for plausible sentences, $t_1(23) = 2.23$, $p < .05$; $t_2(27) = 1.98$, $p = .058$, with no differences between the two island constraint conditions, $t_1(23) = 0.62$, $p = .542$; $t_2(27) = 0.33$, $p = .741$. There were no further reliable effects or interactions other than a main effect of constraint in regression path times, $F_1(1, 23) = 10.31$, $p < .01$; $F_2(1, 27) = 12.95$, $p < .01$, again reflecting shorter reading times for the island constraint compared to the no-constraint conditions.[6]

Spillover Region. The preliminary analyses for this region revealed significant main effects of constraint in the first-pass and regression path times, main effects of plausibility in regression path and rereading times, and main effects of group in first-pass times and rereading times. There were no significant interactions with group in any measure, but the main effects of plausibility in the regression path and rereading times were qualified by Constraint × Plausibility interactions in both measures (reliable by subjects and items in the regression path times, and by items only in the rereading times). The absence of any reliable interactions with the factor group suggests that the native and nonnative participants behaved similarly at this region, and as such subsequent per-group ANOVAs did not seem warranted. *T* tests were conducted on the regression path and rereading time data for the participant group as a whole to examine the Constraint × Plausibility interactions in these measures. These revealed longer reading times for implausible than plausible sentences in the no-constraint conditions in both the regression path times, $t_1(62) = 4.73$, $p < .001$; $t_2(27) = 3.07$, $p < .01$, and rereading times, $t_1(62) = 2.53$, $p < .05$; $t_2(27) = 3.29$, $p < .01$, whereas no differences were observed in either measure between the island constraint conditions—regression path times: $t_1(62) = 0.25$, $p = .801$; $t_2(27) = 0.06$, $p = .956$; rereading times: $t_1(62) = 0.91$, $p = .365$; $t_2(27) = 0.63$, $p = .534$.

The Plausibility × Constraint interactions found for regression path and rereading times match those found at the critical region and indicate

that participants were sensitive to both plausibility information and relative clause islands during their reading of the spillover region.

Discussion

Although the analysis of participants' reading times indicated that neither L1 nor L2 comprehenders attempted to link a wh-filler to a potential gap inside a relative clause island during processing, it also revealed some subtle differences between the two participant groups in the timing of island effects. Whereas the L2 group already showed the predicted interaction—a plausibility effect restricted to the no-constraint pair—during their initial inspection of the critical region (i.e., in their first-pass reading times), this effect was delayed slightly in the L1 group. In other words, according to the experimental diagnostic, the L2 speakers showed evidence of being sensitive to relative clause islands slightly earlier during processing than the L1 controls.

The L1 speakers initially showed a main effect of constraint only that was not modulated by plausibility, with the predicted interaction between the two factors only significant in their rereading times at the critical region, and at the spillover region. The main effect of constraint in the L1 speakers' first-pass reading times could potentially be due to the presence of an extra clause boundary, signaled by the pronoun *who*, in the island constraint conditions, which may have helped the parser segment the overall sentence into chunks and thus facilitated the processing of the critical embedded verb. It is also conceivable that the L1 speakers' processing of the critical verb region was slowed in the no-constraint conditions because they initially tried to form a wh-dependency at this point regardless of the filler's plausibility as a direct object. However, according to the design of Experiment 1, only an interaction between the factors constraint and plausibility would provide a clear indication of gap filling in the absence of an island constraint.

Taking into account the results from other L1 processing studies that found semantic anomalies to affect early processing measures (e.g., Murray & Rowan, 1998; Rayner et al., 2004; Traxler & Pickering, 1996), any conclusion to the effect that L1 speakers' use of plausibility information in gap filling should be generally delayed does not seem warranted here. The marginal Plausibility × Constraint interaction seen in the L1 group's regression path durations at the critical region, with a 33 ms advantage for plausible compared to implausible direct objects in the no-constraint conditions, seems to suggest that the L1 speakers were also sensitive to the experimental diagnostic fairly early during processing. Rayner et al. hypothesized that the timing of plausibility

effects in L1 processing may be related to the violation's severity, which suggests that the slightly later appearance of plausibility effects in this L1 group's data than in Traxler and Pickering's study could be due to differences in the severity of the local semantic incongruence between their materials and the ones of this study.[7] Considering other reading time evidence that indicates greater sensitivity to plausibility violations in L2 compared to L1 sentence processing (Roberts & Felser, 2011), the statistical L1-L2 differences in the timing of plausibility effects observed in Experiment 1 might indeed reflect differences between the two participant groups in the degree of the perceived severity of the violation.

The finding that the L2 speakers showed clear evidence of immediate sensitivity to plausibility information argues against Williams's (2006) suggestion that the use of plausibility information might be delayed in tasks that do not explicitly require any semantic evaluation. Instead, these results show that initial dependency formation in L2 language comprehension involves semantic goodness-of-fit evaluation. The second experiment should help in determining whether initial gap filling in L2 processing is also structurally mediated.

EXPERIMENT 2

The results from Experiment 1 showed that both L1 and L2 comprehenders are sensitive to relative clause islands during processing. The purpose of Experiment 2 was to replicate this finding using a different experimental diagnostic, and to examine whether the choice of diagnostic affects the relative timing of island effects in nonnative compared to native speakers.

Method

Participants. Participants included 26 German-speaking learners (7 males and 19 females, mean age 24.8) of L2 English and 28 L1 English-speaking controls (16 males and 12 females, mean age 22.1) recruited from the University of Essex community, who were offered a small fee for their participation. The learners were comparable to those who participated in Experiment 1 in terms of their age, English learning history, and general level of L2 proficiency. They had first started learning English between ages 7 and 13 (M = 10.4, SD = 1.35) at school and at the time of testing had spent an average of 3.3 years (SD = 5.0) in an English-speaking environment. Their scores in the Quick Placement Test ranged from 62 to 100% (M = 84.65%), placing them in the upper intermediate proficiency bracket or above. All participants had normal or corrected

to normal vision and were naïve with regard to the ultimate purpose of the experiment.

Materials. The design and materials for this experiment were similar to those used in Experiment 1, but instead of manipulating the plausibility of the filler as a direct object of the embedded verb, filled gaps were used as a diagnostic for dependency formation. A total of 24 sentence quadruplets were constructed as shown in (7).

(7) *There are all sorts of magazines on the market.*
a. No constraint, gap
Everyone liked the magazine that the hairdresser read quickly and yet extremely thoroughly about before going to the beauty salon.
b. No constraint, filled gap
Everyone liked the magazine that the hairdresser read articles with such strong conclusions about before going to the beauty salon.
c. Island constraint, gap
Everyone liked the magazine that the hairdresser who read quickly and yet extremely thoroughly bought before going to the beauty salon.
d. Island constraint, filled gap
Everyone liked the magazine that the hairdresser who read articles with such strong conclusions bought before going to the beauty salon.

The experimental conditions differed only in the region following the verb *read*, which in the filled-gap conditions (7b, d) was followed by a NP and in the gap conditions (7a, c) by an adverbial. All the sentences were globally grammatical. If participants tried to link the filler *the magazine* to the potential object gap following the embedded verb *read,* processing would be expected to slow down in cases in which the hypothesized gap was found to be filled by an overt direct object NP (7b, d), compared to those conditions in which an initial gap analysis was locally possible (7a, c).

The critical words (*articles*, *quickly*) were matched for length, $t(46) = 0.162$, $p = .729$; frequency, $t(46) = 0.671$, $p = .929$; and number of syllables, $t(46) = 0.755$, $p = .674$, according to the CELEX database, and were matched in terms of mean lexical decision latency, $t(46) = 0.165$, $p = .687$, according to the norms provided by Balota et al. (2007), and the rest of the sentence was matched for length. As before, only verbs that were optionally transitive were chosen, based on the results of the sentence completion task described previously.

The predictions are parallel to those for Experiment 1; that is, sensitivity to islands should be reflected in a Gap × Constraint interaction, with reading times at the critical region being shorter for (7a) than for (7b), and with no difference between the two constraint conditions (7c, d). If L2 learners are like L1 speakers in that they will attempt to link fillers to structural gaps (as has been argued by Williams, 2006), the

predicted interaction should be visible from early eye-movement measures onward in both participant groups.

In addition to the 24 critical items, a total of 56 filler items were created, including 5 pseudofillers with similar structures to those of the critical items. As before, two-thirds of all trials were followed by yes/no comprehension questions, balanced across all conditions. The materials were again distributed across four presentation lists, mixed with the fillers and pseudorandomized, so that each participant saw a total of 80 items.

Procedures. The experimental data cleaning and data analysis procedures were the same as in Experiment 1.

Results

Overall comprehension accuracy was high, with both participant groups answering 90% of the end-of-trial questions correctly. Track loss accounted for 0.3% of the L1 data and 0.8% of the L2 data. A further 1.44% of the L2 data were excluded on the basis of participants not knowing critical vocabulary items (the words manipulated in the filled-gap diagnostic) in a vocabulary list.

As before, statistical analyses are reported for two sentence regions: the critical region, containing the word following the potential direct object gap—for example, *articles* and *quickly* in (7)—and the spillover region, containing the next three words (e.g., *with such strong* or *and yet extremely*). Skipping rates for both groups in both reported regions were below 8.43%, and outlier removal resulted in the loss of no more than 4.02% of the L2 and 3.16% of the L1 data for each measure at each region. Tables 3 and 4 provide overviews of both the reading time data and the preliminary ANOVA results for Experiment 2, respectively.

Critical Region. Preliminary mixed ANOVAs with gap (unfilled, filled) and constraint (no constraint, island constraint) as within-subjects factors and group (L1 speakers, L2 learners) as a between-subjects factor revealed significant main effects of constraint in the first-pass and regression path times, main effects of gap in the rereading times, and main effects of group in the first-pass and rereading times. These were qualified by a Constraint × Group interaction in the regression path times, and by a three-way interaction between constraint, gap, and group that was significant by subjects and items in the first-pass times, and reliable by subjects and marginal by items in the regression path times. Given the observed interactions with group, separate per-group analyses were conducted.

Table 3. Reading times in ms for three eye-movement measures at the critical and spillover regions in Experiment 2

	L1 speakers (n = 28)			L2 learners (n = 26)		
Condition	First-pass reading time	Regression path time	Rereading time	First-pass reading time	Regression path time	Rereading time
Critical region						
No constraint–gap	283 (64)	394 (138)	176 (127)	357 (106)	408 (126)	247 (250)
No constraint–filled gap	301 (79)	461 (175)	336 (213)	332 (86)	379 (114)	372 (587)
Island constraint–gap	281 (50)	366 (122)	228 (183)	315 (77)	397 (122)	229 (239)
Island constraint–filled gap	268 (61)	324 (77)	235 (146)	325 (101)	414 (164)	309 (331)
Spillover region						
No constraint–gap	531 (233)	787 (328)	304 (223)	622 (155)	890 (547)	372 (366)
No constraint–filled gap	500 (138)	1,016 (514)	431 (347)	614 (152)	935 (378)	539 (613)
Island constraint–gap	518 (175)	873 (323)	369 (335)	654 (193)	763 (233)	408 (433)
Island constraint–filled gap	500 (218)	783 (227)	357 (222)	598 (159)	776 (303)	407 (408)

Note. *SDs* are given in parentheses.

Table 4. Summary of results of preliminary ANOVAs at the critical and spillover regions in Experiment 2

		Critical region			Spillover region		
Factor	F1-F2	First-pass reading time	Regression path time	Rereading time	First-pass reading time	Regression path time	Rereading time
Constraint	F1 (1, 52)	6.52[b]	4.39[b]	1.76	0.01	4.74[b]	1.25
	F2 (1, 23)	5.18[b]	2.22	1.45	0.03	6.45[b]	0.38
Gap	F1 (1, 52)	0.11	0.05	16.28[c]	4.17[b]	1.11	8.11[b]
	F2 (1, 23)	0.33	0.03	16.80[c]	0.99	1.22	1.99
Group	F1 (1, 52)	7.92[b]	0.30	0.47	6.73[b]	0.12	0.49
	F2 (1, 23)	28.95[c]	0.95	10.01[b]	40.67[c]	0.10	7.43[b]
Constraint × Gap	F1 (1, 52)	0.02	1.29	3.30[a]	0.26	7.05[b]	6.90[b]
	F2 (1, 23)	0.00	0.25	2.15	0.41	2.06	2.19
Constraint × Group	F1 (1, 52)	0.18	7.79[b]	0.10	0.33	0.48	0.86
	F2 (1, 23)	0.20	14.45[b]	0.02	0.16	1.57	0.42
Gap × Group	F1 (1, 52)	0.51	0.43	0.16	0.06	0.19	0.29
	F2 (1, 23)	0.20	0.11	0.33	0.02	0.50	0.75
Constraint × Gap × Group	F1 (1, 52)	6.37[b]	7.72[b]	1.00	0.82	4.73[b]	0.06
	F2 (1, 23)	5.00[b]	3.39[a]	0.33	1.34	3.70[a]	0.11

[a] $p < .1$.
[b] $p < .05$.
[c] $p < .01$.

For the native speakers, a 2 × 2 repeated measures ANOVA showed the expected Gap × Constraint interaction in first-pass reading times, albeit marginal by items, $F_1(1, 27) = 5.005$, $p < .05$; $F_2(1, 23) = 3.101$, $p = .092$. Although the filled-gap condition elicited higher reading times than the gap condition in the no-constraint conditions but not in the island constraint condition as predicted, subsequent *t* tests revealed no reliable differences between conditions for either the no-constraint pair or the island constraint pair—no-constraint: $t_1(27) = 1.32$, $p = .198$; $t_2(23) = 1.29$, $p = .212$; island constraint: $t_1(27) = 1.08$, $p = .289$; $t_2(23) = 1.21$, $p = .238$. For regression path durations, a parallel ANOVA showed a main effect of constraint, $F_1(1, 27) = 13.172$, $p < .01$; $F_2(1, 23) = 10.887$, $p < .01$, and a significant Gap × Constraint interaction in the analysis by participants, $F_1(1, 27) = 7.381$, $p < .05$; $F_2(1, 23) = 2.599$, $p = .121$. There was a trend in the by-participants analysis in the regression path times for longer reading times in filled-gap compared to unfilled-gap sentences in the no-constraint conditions, $t_1(27) = 1.81$, $p = .082$; $t_2(23) = 1.11$, $p = .278$, with no significant differences between the constraint conditions, $t_1(27) = 1.68$, $p = .104$; $t_2(23) = 1.49$, $p = .150$. The analysis of the L1 speakers' rereading times yielded a main effect of gap, $F_1(1, 14) = 17.790$, $p < .001$; $F_2(1, 23) = 11.319$, $p < .01$, that was modulated by an interaction with the factor constraint, $F_1(1, 27) = 7.394$, $p < .05$; $F_2(1, 23) = 10.091$, $p < .01$. Pairwise comparisons showed that the difference between the two gap conditions was significant for the no-constraint pair only, with a postverbal noun indicating a filled direct object gap eliciting higher reading times than a postverbal adverb, $t_1(27) = 4.23$, $p < .001$; $t_2(23) = 5.09$, $p < .001$.

The analysis of the L2 group's reading times yielded a rather different picture. There was a main effect of constraint in the first-pass times, marginal by participants, $F_1(1, 25) = 3.93$, $p = .058$; $F_2(1, 23) = 4.45$, $p < .05$, which reflected the fact that sentences in the no-constraint conditions were read slightly faster than in the island constraint conditions overall. A significant main effect of gap was found in rereading times, $F_1(1, 25) = 5.71$, $p < .05$; $F_2(1, 23) = 9.56$, $p < .01$, with the learners showing shorter reading times for postverbal adverbs compared to postverbal nouns across both constraint conditions. No interactions between the two factors were found, however, for any eye-movement measure.[8]

Spillover Region. At the three words following the critical region, preliminary between-groups ANOVAs revealed significant main effects of constraint in the regression path times, main effects of gap by subjects in first-pass and rereading times, and a significant main effect of group in the first-pass times. Constraint × Gap interactions were found in the subjects analyses of the regression path and rereading times, and there was a reliable three-way interaction in the subjects analysis that was marginal by items for the regression path times.

The analyses of the L1 speakers' reading times showed a significant Gap × Constraint interaction in their regression path durations, $F_1(1, 27) = 10.37$, $p < .01$; $F_2(1, 23) = 4.63$, $p < .05$. Pairwise comparisons revealed a reliable difference between the no-constraint pair only, whereby the filled-gap condition elicited significantly higher reading times than the gap condition, $t_1(27) = 2.76$, $p < .05$; $t_2(23) = 2.20$, $p < .05$, in line with the filled-gap effects that were seen at the critical region for rereading times. There was no difference between the island constraint pair, $t_1(27) = 1.39$, $p = .176$; $t_2(23) = 0.98$, $p = .338$. No further main effects or interactions were found in this group for this region.

The L2 group showed a main effect of constraint in regression path times that was marginal by participants, $F_1(1, 25) = 3.58$, $p = .070$; $F_2(1, 23) = 6.56$, $p < .05$, and a main effect of gap in rereading times in the participant analysis, $F_1(1, 25) = 6.32$, $p < .05$; $F_2(1, 23) = 2.09$, $p = .162$. A significant interaction between the two factors was found only in rereading times, in the analysis by participants, $F_1(1, 25) = 4.34$, $p < .05$; $F_2(1, 23) = 1.08$, $p = .309$. Pairwise comparisons confirmed that there were longer reading times in the analysis by participants, marginal by items, for the filled-gap condition in comparison to the gap condition for the no-constraint pair, $t_1(25) = 2.55$, $p < .05$; $t_2(23) = 1.74$, $p = .096$, with no reliable difference between the constraint conditions, $t_1(25) = 0.01$, $p = .989$; $t_2(23) = 0.01$, $p = .989$.

In sum, although the selective filled-gap effects indicative of island sensitivity were seen from the critical region onward in the L1 group, the predicted interaction of gap and constraint was delayed until the postcritical region in the L2 learners' reading times.

GENERAL DISCUSSION

Taken together, the results from Experiments 1 and 2 indicate that both L1 and L2 speakers are sensitive to relative clause islands during real-time comprehension. L1-L2 differences were observed, in contrast, with regard to the relative timing of plausibility versus filled-gap effects. When a semantic diagnostic for dependency formation was used (Experiment 1), effects indicative of island sensitivity were visible slightly earlier in the L2 group than in the L1 group. Even though the L2 group showed the expected interaction of constraint and plausibility during the initial reading of the critical verb region, in the L1 group this interaction proved reliable only for rereading times. In contrast, when filled gaps were used as an experimental diagnostic (Experiment 2), it was the L1 speakers who showed the expected interaction from relatively early on during processing, whereas in the learners this was delayed until the spillover region, in which it was found in rereading times only.

Sensitivity to Islands in L2 Processing

There was no evidence in either of these experiments that participants attempted to link a wh-filler to a gap inside an extraction island, in line with earlier findings by Traxler and Pickering (1996) and others for L1 speakers. These results thus confirm and extend previous findings by Cunnings, Batterham, Felser, & Clahsen (2010), Juffs (2005), Juffs and Harrington (1995), and Sato (2007) indicating that learners of English from typologically different L1 backgrounds (including German, Chinese, Japanese, and Spanish) are sensitive to island constraints in processing tasks.

Clear evidence for learners' immediate sensitivity to islands during processing is provided by the results from Experiment 1, in the shape of a Plausibility × Constraint interaction in their first-pass reading times. Effects of island sensitivity in L2 processing were also present but comparatively delayed in Experiment 2, which used filled gaps as an experimental diagnostic. The possible reasons for this difference in timing will be discussed further in the Timing of Island Effects and the Nature of Dependency Formation section. It is important to note that because the presence or absence of a potential gap was signaled by different lexical items in Experiment 2, the main effect of gap that the learner group showed in their rereading times at the critical region does not provide any evidence that the learners temporarily violated the island constraint during second-pass processing. Instead, this could simply have reflected lexical processing differences between the different words (e.g., *articles* vs. *quickly*) occurring in postverbal position in the filled-gap versus unfilled-gap conditions.

That L2 speakers should be sensitive to extraction islands in processing tasks would be unsurprising from the point of view of performance-based accounts for islands (e.g., Kluender, 2004). Processing complex sentences and having to maintain a fronted constituent in memory is bound to be at least as resource demanding in a L2 as it is in a L1; that is, both L1 and L2 comprehenders are likely to temporarily push the original wh-filler far down the memory stack when encountering a second wh-filler at a relative clause island boundary, as a result of the increased referential processing load and memory burden at this point. Thus, the original gap search will be suspended or abandoned, which will keep the parser from postulating a direct object gap inside island clauses of the kind under investigation.

The Timing of Island Effects and the Nature of Dependency Formation

The results from Experiment 1 confirm previous findings showing that L2 learners, like L1 speakers, adopt an active filler strategy and are able

to link a filler to its lexical licenser as soon as this is encountered (e.g., Williams et al., 2001; Williams, 2006). Moreover, unlike previous L2 processing studies, the observed L1-L2 differences in the timing of island effects can provide more information about the nature of dependency formation in L1 versus L2 processing. A rather striking difference was seen between the learners' reading time patterns in Experiment 1, which used plausibility as a diagnostic for dependency formation, and their reading time patterns in Experiment 2, which used a structural diagnostic.

It is necessary to recall that gap filling in L1 language processing seems to involve at least two different mental subprocesses, one involving semantic evaluation or feature matching (e.g., determining whether *a magazine* or *shampoo* is readable) and the other sensitive to constituent structure, with the filler being linked to an unfilled argument position (corresponding to distinctions made in linguistic theories such as head-driven phrase structure grammar [HPSG]; Pollard & Sag, 1994). The finding that the L1 participants showed relatively early sensitivity to the filled-gap diagnostic used in Experiment 2 indicates that initial dependency formation here involved the attempt to link the filler to a potential argument slot within the emerging verb phrase in the no-constraint conditions. In Experiment 1, statistically reliable evidence for the L1 group's sensitivity to plausibility information during their processing of the critical region was found only in rereading times. There is no evidence in these L1 speaker data, then, to suggest that a wh-filler's goodness-of-fit evaluation temporally preceded the postulation of a structural gap.

A rather different picture emerges from the L2 participant groups' results, however. Although the learners' reading time patterns in Experiment 1 suggest that the filler's semantic plausibility as a participant in the event denoted by the embedded verb was evaluated immediately, their sensitivity to the filled-gap diagnostic in Experiment 2 was clearly delayed. In this experiment, the main effect of constraint seen in the learners' first-pass reading times at the critical postverbal region suggests that they were aware of the presence of the additional clause boundary (signaled by the relative pronoun *who*) in the two island constraint conditions compared to the two no-constraint conditions, but this effect was not modulated by the availability of a structural gap after the verb. The predicted Gap × Constraint interaction was only seen during the learners' processing of the spillover region, in a relatively late eye-movement measure.

Taken together, the results from Experiment 1 and 2 fail to support Williams's (2006) claim that structural gaps are postulated immediately in both L1 and L2 processing whereas learners' use of plausibility information may be slightly delayed. Instead, the results of this study indicate that initial wh-dependency formation in L2 processing involves semantic goodness-of-fit evaluation rather than being guided by constituent structure information such as the availability of an empty argument position. This is what might be expected if learners' ability to use

structural cues to interpretation is compromised in L2 compared to L1 processing, and relative to their ability to use semantic or pragmatic cues to interpretation (Clahsen & Felser, 2006a, 2006b). The current results refine Clahsen and Felser's original hypothesis by incorporating additional assumptions about the relative timing of different information sources. Clahsen and Felser (2006b) hypothesized that learners' reduced sensitivity to morphosyntactic and phrase structure information during L2 processing might be due to their L2 grammatical knowledge being either "incomplete, divergent, or of a form that makes it unsuitable for parsing" (p. 117). The observation that, unlike their sensitivity to semantic cues, learners' sensitivity to structural cues in object gap filling was delayed more specifically suggests that (certain parts of the) L2 grammar knowledge may be represented in such a way so as to make it inaccessible to first-pass parsing routines, and available only during later stages of processing. As a consequence, in terms of the dual pathways processing architecture assumed by Clahsen and Felser (2006a, 2006b), the shallow parsing route may dominate in L2 processing because it can operate faster than the full parsing route. This hypothesis clearly requires further testing, however.

In view of these current findings, it seems likely that the delayed filled-gap effects observed by Williams et al. (2001) and Williams (2006) also reflect a delay in L2 learners' use of structural information, rather than a delayed use of plausibility information, as hypothesized by Williams (2006). This would explain why plausibility had an immediate effect on learners' stop-making-sense decisions at the verb but delayed effects on their reading times of the postverbal region that contained a filled gap in Williams and colleagues' studies. Even if this alternative interpretation of their results is along the right lines, however, the question of the extent to which learners' sensitivity to different information sources may be subject to task effects (see Williams, 2006) is clearly worthy of further investigation.

CONCLUSION

The current results corroborate earlier findings showing that, like L1 speakers, L2 speakers are sensitive to island constraints in L2 processing tasks. At the same time, the observed L1-L2 differences in the timing of plausibility versus filled-gap effects point to differences in the nature of wh-dependency formation in L1 compared to L2 processing. Although the results from the L1 speakers support previous findings suggesting that filler integration in L1 processing is guided by subcategorization information and mediated by constituent structure, the results from the L2 speakers suggest that the initial stage of filler integration is semantically driven. This means that these findings provide evidence for incremental interpretation in L2 sentence processing that

is not contingent on, or even necessarily concurrent with, the computation of detailed hierarchical constituent structure representations. At the methodological level, the current study highlights the importance of using a variety of experimental methods and diagnostics for gaining a better understanding of the nature and time course of L2 processing.

(*Received 20 October 2010*)

NOTES

1. Stowe (1986), for example, found longer reading times at the direct object position of the verb *bring* in sentences such as *My brother wanted to know who Ruth will bring us home to at Christmas* compared to sentences that did not contain a filler-gap dependency. The measurable processing difficulty that Stowe and others have observed at filled object gaps indicates that the parser postulates direct object gaps immediately on encountering transitive verbs such as *bring*.

2. By taking into account responses to both ungrammatical and grammatical items, A′ scores provide a unified sensitivity index to a given stimulus property that corrects for potential response biases in binary forced-choice tasks. An A′ score of .50 indicates chance performance, and a score of 1.00 indicates perfect discrimination (see Grier, 1971).

3. To control for possible effects of individual differences in participants' working memory capacity on processing, all participants additionally underwent a reading span test (L1: Daneman & Carpenter, 1980; L2: Harrington & Sawyer, 1992). However, because the factor reading span was not found to affect the presence or timing of island effects in either the L1 or the L2 groups in either Experiment 1 or 2, full details of the working memory results are not reported here.

4. Complete lists of the experimental items used in Experiments 1 and 2 are available on request from the first author.

5. As the current study focuses on readers' initial sensitivity to extraction islands during processing, for which reading times at the ultimate gap site are not directly relevant, the results from the disambiguating region have not been included here. However, a second plausibility norming pretest was carried out to ensure that the experimental sentences did not differ in their degree of global plausibility. To this end, 10 L1 English speakers rated the plausibility of the untransformed sentences (e.g., *The hairdresser read about the magazine/shampoo*) on a 5-point scale. The results confirmed that the NPs used in the plausible (mean rating: 2.2) and implausible (mean rating: 2.1) conditions were considered to be equally plausible as complements of their lexical licenser, $t_1(9) = 0.524$, $p = .613$; $t_2(27) = 0.724$, $p = .475$.

6. As the extent to which learners show nativelike processing patterns may be affected by their L2 proficiency (Hopp, 2006), additional analyses with proficiency (as measured by the Quick Placement Test) as a covariate were carried out for the critical region. These yielded no significant three-way interactions with the factor proficiency for any of the three eye-movement measures (all Fs < 1.6), indicating that the timing of island effects was not affected by individual differences in the learners' general L2 proficiency.

7. Because there was no access to the full set of materials used by Traxler and Pickering (1996), it was not possible to directly compare the materials used here and theirs with regard to the exact kind of violations used or the degree of their severity.

8. To examine possible effects of individual differences in the learners' L2 proficiency on their reading time patterns at the critical region, additional analyses with proficiency as a covariate were carried out again. As in Experiment 1, these analyses showed no evidence that the presence or timing of the Constraint × Gap interaction was modulated by proficiency (all Fs < 1).

REFERENCES

Alexopoulou, T., & Keller, F. (2007). Locality, cyclicity and resumption: At the interface between the grammar and the human sentence processor. *Language*, *83*, 110–160.

Baayen, H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database* [CD ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Balota, D., Yap, M., Cortese, M., Hutchison, K., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. *Behavior Research Methods*, *39*, 445–459.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A festschrift for Morris Halle* (pp. 232–286). New York: Holt, Rinehart, & Winston.

Clahsen, H., & Featherston, S. (1999). Antecedent priming at trace positions: Evidence from German scrambling. *Journal of Psycholinguistic Research*, *28*, 415–437.

Clahsen, H., & Felser, C. (2006a). Continuity and shallow structures in language processing. *Applied Psycholinguistics*, *27*, 107–126.

Clahsen, H., & Felser, C. (2006b). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*, 3–42.

Clifton, C., Staub, A., & Rayner, K. (2007). Eye movements in reading words and sentences. In R. van Gompel (Ed.), *Eye movements: A window on mind and brain* (pp. 341–372). Amsterdam: Elsevier.

Cunnings, I., Batterham, C., Felser, C., & Clahsen, H. (2010). Constraints on L2 learners' processing of *wh*-dependencies: Evidence from eye movements. In B. VanPatten & J. Jegerski (Eds.), *Research in second language processing and parsing: Issues in theory and research* (pp. 87–110). Amsterdam: Benjamins.

Dallas, A., & Kaan, E. (2008). Second language processing of filler-gap dependencies by late learners. *Language and Linguistics Compass*, *2*, 372–388.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Felser, C., & Roberts, L. (2007). Processing *wh*-dependencies in a second language: A cross-modal priming study. *Second Language Research*, *23*, 9–36.

Frazier, L., & Clifton, C. (1989). Successive cyclicity in the grammar and the parser. *Language and Cognitive Processes*, *4*, 93–126.

Frenck-Mestre, C. (2005). Eye-movement recording as a tool for studying syntactic processing in a second language: A review of methodologies and experimental findings. *Second Language Research*, *21*, 175–198.

Garnsey, S., Tanenhaus, M., & Chapman, R. (1989). Evoked potentials and the study of sentence comprehension. *Journal of Psycholinguistic Research*, *18*, 51–60.

Gibson, E. (1998). Syntactic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1–75.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–125). Cambridge, MA: MIT Press.

Gibson, E., & Warren, T. (2004). Reading-time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax*, *7*, 55–78.

Grier, J. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin*, *75*, 424–429.

Harrington, M., & Sawyer, M. (1992). Working memory capacity and L2 reading skill. *Studies in Second Language Acquisition*, *14*, 25–38.

Hestvik, A., Maxfield, N., Schwartz, R., & Shafer, V. (2007). Brain responses to filled gaps. *Brain and Language*, *100*, 301–316.

Hofmeister, P., and Sag, I. (2010). Cognitive constraints and island effects. *Language*, *86*, 366–415.

Hopp, H. (2006). Syntactic features and reanalysis in near-native processing. *Second Language Research*, *22*, 369–397.

Juffs, A. (2005). The influence of first language on the processing of wh-movement in English as a second language. *Second Language Research*, *21*, 121–151.

Juffs, A., & Harrington, M. (1995). Parsing effects in second language sentence processing: Subject and object asymmetries in *wh*-extraction. *Studies in Second Language Acquisition*, *17*, 483–516.

Kluender, R. (2004). Are subject islands subject to a processing account? In A. Rodríguez, V. Chand, A. Kelleher, & B. Scheiser (Eds.), *Proceedings of West Coast Conference on Formal Linguistics* (Vol. 23; pp. 475–499). Somerville, MA: Cascadilla Press.

Kluender, R., & Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and Cognitive Processes*, *8*, 573–633.
Lee, M.-W. (2004). Another look at the role of empty categories in sentence processing (and grammar). *Journal of Psycholinguistic Research*, *33*, 51–73.
MacWhinney, B. (2005). A unified model of language acquisition. In J. Kroll & A. De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). Oxford: Oxford University Press.
Marinis, T., Roberts, L., Felser, C., & Clahsen, H. (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition*, *27*, 53–78.
Murray, W., & Rowan, S. (1998). Early, mandatory, pragmatic processing. *Journal of Psycholinguistic Research*, *27*, 1–22.
Nakano, Y., Felser, C., & Clahsen, H. (2002). Antecedent priming at trace positions in Japanese long-distance scrambling. *Journal of Psycholinguistic Research*, *31*, 531–571.
Nicol, J. (1993). Reconsidering reactivation. In R. Shillcock (Ed.), *Cognitive models of speech processing: The second Sperlonga meeting* (pp. 321–347). Mahwah, NJ: Erlbaum.
Phillips, C. (2006). The real-time status of island phenomena. *Language*, *82*, 795–823.
Pickering, M., & Barry, G. (1991). Sentence processing without empty categories. *Language and Cognitive Processes*, *6*, 229–259.
Pollard, C., & Sag, I. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.
Rayner, K., Warren, T., Juhasz, B., & Liversedge, S. (2004). The effect of plausibility on eye movements in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 1290–1301.
Roberts, L., & Felser, C. (2011). Plausibility and recovery from garden-paths in second-language sentence processing. *Applied Psycholinguistics*, *32*, 299–331.
Roberts, L., Marinis, T., Felser, C., & Clahsen, H. (2007). Antecedent priming at gap positions in children's sentence processing. *Journal of Psycholinguistic Research*, *36*, 175–188.
Ross, J. (1967). *Constraints on variables in syntax*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.
Saah, K., & Goodluck, H. (1995). Island effects in parsing and grammar: Evidence from Akan. *The Linguistic Review*, *12*, 381–409.
Sato, M. (2007). *Sensitivity to syntactic and semantic information in second language sentence processing*. Unpublished doctoral dissertation, University of Essex, United Kingdom.
Staub, A., & Rayner, K. (2007). Eye movements and online comprehension processes. In M. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 327–342). Oxford: Oxford University Press.
Stowe, L. (1986). Parsing *wh*-constructions: Evidence for on-line gap location. *Language and Cognitive Processes*, *1*, 227–245.
Traxler, M., & Pickering, M. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, *35*, 542–562.
Trueswell, J., Tanenhaus, M., & Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 528–553.
Wagers, M., & Phillips, C. (2009). Multiple dependencies and the role of the grammar in real-time comprehension. *Journal of Linguistics*, *45*, 395–433.
Williams, J. (2006). Incremental interpretation in second language sentence processing. *Bilingualism: Language and Cognition*, *9*, 71–88.
Williams, J., Möbius, P., & Kim, C. (2001). Native and non-native processing of English *wh*-questions: Parsing strategies and plausibility constraints. *Applied Psycholinguistics*, *22*, 509–540.