The effects of teaching parsing strategies on the processing and learning of L2 English relative clauses

Xiaoran Niu

PhD

University of York

Education

August 2021

Abstract

The effectiveness of explicit training on second language (L2) offline comprehension and production has been demonstrated by numerous studies. Yet only a few studies have investigated training effects on online processing, and the findings to date are inconclusive. The current study investigated the effects of teaching explicit knowledge with input-based practice on L2 English relative clauses learning. In addition, how native English speakers process and produce relative clauses was also investigated.

79 Chinese-speaking L2 learners and 21 native speakers were involved in the study. The L2 learners were randomly assigned to either a parsing (n=27), input flood (n=26), or test-only (n=26) group. The tests measured offline (aural sentence-picture matching) and online (self-paced reading and visual-world eye-tracking) comprehension, oral production (picture description), and metalinguistic knowledge (written sentence-picture matching with explanations), administrated in a pre-, one-week post- and three-week delayed post-test design. The L2 learners attended all test phases, and the native speakers only attended the pre-tests. Between the pre- and post-test, the parsing (parsing strategies with explicit information and practice) and the input flood (exposure to the target structures) group received two 30-minute training sessions.

The results from the native speakers demonstrated the expected asymmetry between subject relative clauses (SRC) and object relative clauses (ORC). The results from the L2 learners suggested that the explicit training developed offline comprehension, production and metalinguistic knowledge. Very small improvements were also found in one online measure, but only for SRCs. The input flood group did not have significant gains in any measures, relative to the test-only group. The findings contribute to our understanding about the effects of explicit training on online and offline processing, production and knowledge of syntax. Following small amount of practice, effects of explicit training seemed to be reliably observable on offline measures, potentially observable on online measures for a structure that is easier to process (SRCs), and not likely to be observed on online measures for a structure that is more difficult to process (ORCs).

Table of contents

List of tables

Chapter 3

Chapter 4

Appendices

List of figures

Chapter 2

Chapter 3

Chpater 4

Acknowledgements

First and foremost, I would like to express my heartfelt thanks to my supervisor Professor Emma Marsden. Thank you so much for your continuous academic support and encouragement. Every time when I met tough problems and felt frustrated, you always provided me with insightful advice and helped me to overcome the difficulties.

I would like to thank my TAP member Dr Heather Marsden. Thank you for your advice and suggestions on my research. My thanks also go to Professor Leah Robert and Dr Giulia Bovolenta. Thanks for your advice on my data analysis. I also would like to thank Dr Sophie Thompson-Lee. Thanks very much for your help in the design of the study, data analysis and recording of the training items. Thank you also to Nick Avery for proofreading my test items, to Katrina d'Apice and Catherine Bakewell for recording the test or training items of my study.

I am deeply grateful to all my participants who took part in the study. I know my study took a very long time. Thanks so much for your patience in coming to the lab on time for several sessions.

I would like to express my love and thanks to my parents Xin Qin and Niu Baiqi. Thanks very much for your endless love, understanding, encouragement and everything throughout my life. Love you!

Finally, I would like to thank my family and my friends for your emotional support. In particular, I would like to thank my friend Dr Yang Ruotao and Dr Yang Ting for helping me with mathematical problems and talking to me when I felt alone or upset. Thanks to my friend Hao Meng for your friendship lasting more than ten years and for helping me when I drew the pictures for the study. Thanks to my niece Li Lingxiao, my cousin Xin Ying, my friend Shao Man for helping me in creating pictures. My thanks also go to my cousin Niu Ping for helping me to tackle technical problems of software when I analysed the data. Thanks to all my friends that I met in the last four years. Thanks for your friendship.

Author’s declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Chapter 1 Introduction

1.1 The research context

Research into whether and how explicit training can affect L2 grammar learning has been widely conducted for several decades. Numerous studies have demonstrated that explicit training could facilitate L2 comprehension and production (see research into processing instruction, e.g., Marsden, 2006; Marsden & Chen, 2011; VanPatten & Cadierno, 1993a, 1993b; VanPatten & Wong, 2004; for reviews, see Long, 1983; Norris & Ortega, 2000). Nevertheless, the extent to which the effects of explicit training could be observed in online measures is still debatable.

Commonly, it is regarded that online measures detect the way of real-time processing (Keating & Jegerski, 2015). VanPatten and Cadierno (1993a) proposed that second language (L2) learners could be trained to process differently by providing them with strategies related to form-meaning connections. However, studies that looked into the training effects in online processing generated inconsistent results. Some studies showed evidence supporting the idea that training effects could be found in online measures (e.g., Hopp, 2016; McManus & Marsden, 2017), while the others indicated that L2 learners did not improve (increase the speed or accuracy of) their online processing (e.g. Andringa & Curcic, 2015; Dracos & Henry, 2021). In addition, the previous studies that have examined online processing over time focused on training morphosyntax structures that have referential meaning (such as number, tense, animacy) in the real world. To the best of our knowledge, no published study has yet investigated whether the processing of syntax (such as, in the case of the current study, the word order and role assignment in English relative clause) could be altered through explicit training. In addition, to the best of our knowledge, no study has yet used 'instruction and practice *during* parsing' to investigate whether such training can affect offline and/or online processing. It is important to have a better understanding of these issues for the purposes of both learning theory and pedagogy.

The English relative clause is a complex and important structure which is

frequently used in everyday communication and academic language. A great number of studies have been conducted to explore how native speakers and L2 learners comprehend and produce relative clauses (e.g., Diessel & Tomasello, 2005; Keenan & Comrie, 1977; Kim & O'Grady, 2016; Traxler, Morris & Seely, 2002). Those studies have demonstrated that the SRC is easier to comprehend and produce than the ORC, and the animacy of the head noun might affect the difficulty of the comprehension and production of ORC. However, to the best of our knowledge, no study has explored whether teaching grammatical role assignment of the two nouns (the one in the main sentence and one in the relative clause) in relative clauses could facilitate the learning of relative clauses and whether the training effects would be affected by the type of relative clause. Moreover, despite the existing body of research, it remains worth investigating how native English speakers assign grammatical roles to the two nouns during processing relative clauses in real-time and whether different patterns could be observed in different types of relative clauses.

The current study attempted to fill these knowledge gaps by investigating the extent to which teaching parsing strategies (i.e., providing information and practice during parsing) of relative clauses can benefit processing and learning of L2 English relative clauses. The performance of the native English speakers in comprehension and production was also documented as a baseline, in part to check the validity of the measures used. The study aimed to address the following research questions (RQs).

1) Which type of relative clause (SRC vs. ORC) is more difficult in online and offline comprehension, production, metalinguistic knowledge, and what is the role of animacy of the main clause noun?

2) To what extent can teaching parsing strategies, with explicit information and practice, exposure alone or no exposure (tests alone), develop the learning of relative clauses?

1.2 Outline of the thesis

Chapter 1 illustrates the context of the current study. Chapter 2 presents the literature review in three sections: 1) relative clauses, 2) instruction about helping the learning of

relative clauses and 3) rationale for the current study. In the first section, the studies related to processing SRC and ORC and the influence of the head animacy and the first language (L1) are discussed. In addition, this section also reviews studies about the relationship between metalinguistic knowledge and language ability in the relative clause. In the second section, the literature related to whether explicit training could facilitate online and offline processing and also production and contribute to L2 learning is discussed initially. Then, the research into processing instruction, sentence processing and parsing strategies, and input flood is reviewed critically. The third section provides the rationale and puts forward the research questions of the current study.

Chapter 3 presents the methodological considerations and methods used in the current study. The ethical considerations and the information about the participants are provided initially. Then, the design of the outcome measures and the training sessions are described in detail.

Chapter 4 presents the results related to each research question respectively, and the results are detailed separately for each outcome measure. The findings in relation to each research question are critically discussed in Chapter 5.

Chapter 6 is the conclusion. A summary of the study and the findings are provided. Some of the key limitations of the study are identified, and suggestions about further studies are put forward. Finally, the contribution of the study to the field of second language acquisition research is discussed.

Chapter 2 Literature review

This chapter reviews the previous literature from three aspects. First, the literature related to the target structures (relative clauses) in the current study will be discussed in section 2.1. Second, the research into explicit training of L2 learners will be analysed in section 2.2. Finally, section 2.3 will present the rationale of the study and put forward the research questions with hypotheses.

2.1 Relative clauses

Relative clause belongs to a type of subordinate clause which is used to modify a noun in the matrix (or main) clause (Tallerman, 2014). Sentences that contain relative clauses are complex sentences and are frequently used in both everyday and academic language. In a relative clause (e.g., *The cat that chases the dog is big*), the first noun of the sentence (*cat*) is called the 'head noun' and 'that' is the relative pronoun. A relative pronoun is used to introduce a clause. If the pronoun is the subject of the relative clause, the relative clause is called the subject relative clause (SRC), whereas if the pronoun is the object of the clause, the relative clause is called the object relative clause (ORC). There are two types of ORCs, direct ORC (e.g., *The cat that the dog chases is big*) and indirect ORC (e.g., *"The man whom I give the book to is my colleague"*, Izumi, 2003, p.288). The current study will use SRC and direct ORC as the target structures, and 'ORC' in this thesis refers to direct ORCs only.

2.1.1 Processing subject and object relative clauses

A variety of studies have demonstrated that SRCs are easier to comprehend and produce than ORCs (e.g., Diessel & Tomasello, 2005; Keenan & Comrie, 1977; Kim & O'Grady, 2016; Traxler, Morris & Seely, 2002). For instance, Diessel and Tomasello (2005) investigated how monolingual native-English children used relative clauses in sentence repetition tasks. In the study, the children were engaged in a game-based experiment, and they were required to repeat sentences after the researcher. The result suggested a subject-object asymmetry of relative clauses in that the children made fewer mistakes in repeating SRCs compared to the ORCs. In addition, similar findings were

found by Traxler et al. (2002), in a study conducted with adults that used reading-based eye-tracking tests. In the study, the participants spent more time reading ORCs than SRCs, which suggested that ORCs tended to be more difficult than SRCs.

In general, three potential reasons, "working memory limitations, syntactic factors, and perspective shifting" were usually used to explain this subject-object asymmetry (Traxler et al., 2002, p.70). These three reasons will be briefly described respectively as follows.

Working memory limitations

Working memory plays a role in language processing, and processing ORCs usually requires higher working memory demands than SRCs (e.g. Gibson, 1998; Traxler et al., 2002; Traxler, Williams, Blozis & Morris, 2005). The relative clause is a type of structure that contains a dependency between a 'filler' (the entity that will be needed later in the parsing) and a 'gap' (the place where the 'filler' will need to be activated in the parse), and filler-gap dependencies lead to the processing difficulty (Hawkins, 1999). The difficulty level of processing each type of relative clauses depends on the distance between filler and gap. Longer distances make relative clauses harder to process because the filler and the following information need to be kept in working memory until the gap appears (Hawkins, 1999; O'Grady, Lee, & Choo, 2003). In SRCs, the distance between the filler and the gap is one word, while there can be a four-word distance between the filler and the gap in ORCs (see example 1).

Example 1:

a. Subject relative

*the **man** that* [___*likes the woman*]

linear distance between the gap (shown by __) and the head (man) = 1 word (i.e., 'that')

b. Direct object relative

*the **man** that* [*the woman likes*___]

linear distance between the gap and the head = 4 words (i.e., that the woman likes)

(Example 1 is cited from O'Grady et al., 2003, p.434)

Note: *O'Grady et al. (2003) named filler as 'head'. The fillers in the example were marked in bold.*

To clarify, in the SRC, the processor only needs to keep the relative pronoun (in example 1, 'that') in working memory. However, the head noun 'man' has to be retained in working memory until the end of the clause when processing an ORC. Thus, SRCs are easier to process compared to ORCs according to an account based on limitations in working memory.

Syntactic factors

A syntactic factor could also be used to explain why SRC is easier than ORC. Sheldon (1974) put forward a parallel-function hypothesis. It states that the relative clauses in which the subject of the relative clause is the same as the sentential subject are easier to process than the relative clauses in which the subjects of the relative clause and that of the sentence are different (Diessel, 2004). In the SRC (e.g., *The cat that chases the dog is big*), 'cat' is the sentential subject as well as the subject of the clause; however, in the ORC (e.g., *The cat that the dog chases is big*), the sentential subject is 'cat', while the subject of the clause is 'dog'. Thus, when processing relative clauses, the parser would first attempt to regard the sentential subject as the subject of the clause (otherwise a possible gap would appear immediately after the subject, Traxler et al., 2002). This initial parsing attempt would cause the misanalysis of ORC, but not for SRCs, which leads to ORCs be more difficult to process compared to SRCs.

Perspective shifting

The perspective shifting account states that when the sentential subject and the subject of the clause are inconsistent, perspective shifting needs to happen (Traxler et al., 2002). In processing sentences containing ORCs, the processors would shift their perspective when they notice the subject of the clause is different from that of the matrix. In addition, they would need to shift the perspective back to the processing of the matrix when the relative clause finishes. However, because the subjects of the matrix and the clause are the same in SRCs, perspective shifting would not be required during the processing. As shifting perspectives is usually considered as "costly and

time-consuming" (Traxler et al., 2002, p.71), from this aspect, SRCs are easier to process than ORCs.

2.1.2 Influence of animacy in subject and object relative clause processing

In processing relative clauses, many studies have found that the animacy of the head noun can influence the difficulty of ORCs (e.g. Kidd, Brandt, Lieven & Tomasello, 2007; Macdonald, Brandt, Theakston, Lieven & Serratrice, 2020; Traxler et al., 2005). Corpus studies have shown that in natural English conversation, the ORCs with inanimate head nouns and animate nouns in the clauses are more frequently used than the ORCs with two animate nouns (Fox & Thompson, 1990; Kidd et al., 2007; Reali & Christiansen, 2007). They suggested that the structure that has a higher frequency of occurrence is easier to process. Thus, the ORCs with inanimate heads are predicted to be easier to process than those with animate heads.

In addition to corpus studies, some empirical studies have also found evidence that the animacy of the head noun could affect the difficulty of ORC processing, and ORCs with inanimate heads could be easier than those with animate heads under some circumstances. For example, MacDonald et al. (2020) investigated the influence of animacy in online and offline comprehension of subject and object relative clauses with children and adults. Visual-world eye-tracking tests with comprehension questions were used, and the animacy of the head noun (the first noun) and the noun of the clause (the second noun) was manipulated. For each item of the test, the participants heard one sentence and saw a pair of reversible pictures. The participants were asked to listen to the sentence and decide which picture matched the sentence once they knew the answer. Two experiments were conducted in the study. Experiment 1 manipulated the animacy of the *first* noun of the sentence, and the second noun was always animate (e.g. In "*Where is the tractor that the cow is chasing*" (p.10), the first noun is 'tractor' and the second noun is 'cow'). The results showed that the children had higher accuracy scores and faster reaction time of deciding the matched picture in SRCs relative to ORCs regardless of the animacy of the first noun, and ORCs with inanimate first noun generated higher accuracy scores than those with the animate

first noun. Meanwhile, the adults scored at ceiling for both types of relative clauses, and they always responded to SRCs faster than ORCs in both animate and inanimate conditions. In addition, it was observed that the inanimate first noun did not help children or adults to expect an ORC, and they were likely to initially misinterpret the ORCs as SRCs when the first nouns were inanimate. When the first nouns were animate, both children and adults parsed the SRCs and ORCs in a similar pattern.

Experiment 2 fixed the *second* noun as inanimate while the first noun either could be animate or inanimate. There were two differences between the results of experiment 1 and 2. In experiment 2, the inanimate first noun did not increase the accuracy or shorten the response time (relative to an animate first noun) in the comprehension of ORCs for children (whereas it did in experiment 1). Moreover, the adults did not show a preference for SRCs or ORCs when the first nouns were inanimate (whereas in experiment 1, they showed preferences for SRCs). Overall, MacDonald et al. (2020) indicated that an inanimate first noun might reduce the difficulty of ORC in children's offline comprehension when the second noun in the sentence was animate. However, children would expect an SRC when the head noun was inanimate regardless of the animacy of the second noun (adults showed only this tendency when the second noun was animate).

To summarise, the corpus analyses and empirical studies indicate that the ORCs with inanimate heads are likely to be easier than those with animate heads in offline comprehension. However, the inanimate heads are unlikely to alter the subject-object asymmetry and they are unlikely to make the ORCs easier than SRCs.

In the current study, SRCs and ORCs were used as target structures. Both online and offline comprehension and production of the target structures were examined. In previous studies, the influence of the head noun animacy on the difficulty of processing relative clauses has been considered, so in the current study, the animacy of the head noun was manipulated. The SRCs and ORCs with animate and inanimate heads (SRC-A, SRC-I, ORC-A, ORC-I) were addressed separately as four independent structures, and all the head nouns (the first noun in the sentence) of the relative

clauses served as the subject of the sentences. Each type of relative clause (SRC and ORC) had both animate and inanimate nouns of the clauses (the second noun), and the number of these items was balanced (there were equal numbers of relative clauses with animate and inanimate nouns). However, the research question in relation to animacy focused only on the animacy of the noun of the matrix clause (the first noun), not that of the noun in the relative clause (the second noun). This is because, for time and participant fatigue reasons, there were insufficient items to manipulate the animacy of the nouns in the relative clauses experimentally.

2.1.3 Metalinguistic knowledge and language ability of comprehending and producing relative clause

Metalinguistic knowledge "is the knowledge about knowledge" (Alderson, Clapham & Steel, 1997, p.95) which refers to "explicit knowledge about categories (e.g. 'noun'; 'verb'; 'adjective') as well as explicit knowledge about the relationship between categories (e.g. 'subject of the main clause')" (Roehr, 2006, p.183; Roehr, 2008).

Some researchers, working from a skill acquisition (or information processing) theoretical perspective, have suggested that metalinguistic knowledge is associated with language proficiency (e.g., Roehr, 2008). Roehr (2008) tested the L2 proficiency and the metalinguistic knowledge with English-speaking L2 learners of German. The participants included 34 first-year and 26 fourth-year undergraduate students who studied Advanced German at a University in the UK. It was found that the grammar and vocabulary knowledge of the advanced L2 learners (the fourth-year students) was strongly positively correlated with their metalinguistic knowledge. However, this correlation for the lower proficiency L2 learners was weaker compared to that of the higher proficiency L2 learners. The results revealed that the L2 learners who have higher language competence might have more and better metalinguistic knowledge. This finding was against Alderson et al. (1997), which found that metalinguistic knowledge was not correlated to L2 proficiency. The inconsistent results of the two studies might because the participants in Roehr (2008) received form-focused L2 treatment at the university and may have specific characteristics as learners, such as

high ability to understand grammatical rules. As noted by Roehr (2008), the findings in the study might not be generalised to other groups of L2 learners. Thus, the relationship between L2 proficiency and metalinguistic knowledge competence is inconclusive.

Native speakers who are not given (or retain) explicit information about their first language might not have much metalinguistic knowledge. Green and Hecht (1992) measured the metalinguistic knowledge of pupils about their L1. The pupils were asked to correct the errors and explain the roles. It was found that the native speakers could successfully correct the errors but lacked the ability to explain the nature of the errors (the underlying rules).

Some studies included metalinguistic knowledge test as a measure of training effects for L2 learners (e.g., Kasprowicz & Marsden, 2018). Kasprowicz and Marsden (2018) provided explicit information with input-based practice about German case-marking to L1 English speakers. It was found that the participants did not have metalinguistic knowledge about the target structures before the training, and they showed improvement after the training. This indicated that the explicit training could facilitate the gains of metalinguistic knowledge on target structures.

One concern of the current study is to investigate the extent to which the L2 learners could gain metalinguistic knowledge about relative clauses from training. A metalinguistic knowledge test was used to explore whether the native speakers and L2 learners have the metalinguistic knowledge about relative clauses, and whether the L2 learners could gain the knowledge from the training (with explicit knowledge given, versus exposure to the structure alone) or even from taking the tests alone.

2.1.4 Influence of L1 in relative clause processing

It is generally believed that L1 would influence L2 learning at least in some aspects of the learning process, almost regardless of the broad theoretical framework being adopted (e.g. N.C. Ellis, 2006; Ionin & Montrul, 2010; Schwartz, 1998). L1 can facilitate L2 learning and processing when the two languages have the same grammar feature; in contrast, if the grammar feature of L2 does not exist in L1 or if it is different in some

way, the L1 might impede L2 learning and processing (Hopp & Lemmerth, 2016). Tolentino and Tokowicz (2014) found that the effectiveness of L2 grammar instruction was modulated by the relationship between the L1 and the L2. They found that when a structure had existed in both the L1 and the L2 but it was different in some way, then instruction that simply highlighted the difference was useful (and giving a rule did not provide additional benefits). However, for a completely different L2 feature (that was unique to the L2), instruction that also provided a rule was beneficial.

Of relevance to the current study is that both English and Chinese languages have relative clauses, but the constructions are different. The differences in the construction in the two languages might influence how Chinese-speaking L2 learners of English process English relative clauses. As described in 2.1.1, it has been found that SRCs are easier than ORCs in English. Some researchers state that Chinese SRCs and ORCs have a similar asymmetry pattern to English (e.g. Lin & Bever, 2006; Pu, 2007). For example, Lin and Bever (2006) conducted a SPR test, and they found that the participants read the relative pronoun (*De* in Chinese) and the head noun faster in SRCs compared to ORCs. Pu (2007) stated that in the discourse level, Chinese SRCs are more frequently used than ORCs. These two studies indicate a subject preference in processing and producing Chinese relative clauses

However, many researchers argue that the asymmetry between Chinese SRCs and ORCs could be the opposite of the asymmetry found in English relative clauses. In other words, Chinese SRCs might be more difficult than ORCs (see Chen, Ning, Bi & Dunlap, 2008; He, Xu & Ji, 2017; Hsiao & Gibson, 2003; Su, Lee & Chung, 2007; etc.). One possible reason which could be used to explain the different asymmetry between English and Chinese relative clauses is the different word orders of the relative clauses in the two languages. English relative clauses are head-initial sentences in which the fillers (heads) appear *before the gaps*, while Chinese relative clauses have the fillers after the gaps (He et al., 2017; see example 2). From the perspective of working memory limitations, Chinese SRCs have more words between the filler and the gap (one-word distance) compared to ORCs (five-word distance), which means that

processing Chinese SRCs could take more memory load than ORCs. Thus, in Chinese, SRCs are likely to be more complex than ORCs.

Example 2:

a. SRC

[*gap* 攻击　议员　的] **律师** *filler* 不喜欢　那位　　政客。

[gapGongJi　YiYuan De] **LvShi***filler* BuXiHuan NaWei　Zheng Ke

[*gap* Attacked senator de] **lawyer** *filler* doesn't like that politician.

The **lawyer** *filler* [that _*gap*__ attacked the senator] doesn't like the politician.

b. ORC

[议员　攻击 *gap* 的]**律师** *filler* 不喜欢　　那位　政客。

[YiYuan　GongJi*gap* De] **LvShi***filler* BuXiHuan　NaWei　ZhengKe]

[Senator attacked *gap* de] **lawyer** *filler* doesn't like that politician.

The **lawyer** *filler* [that the senator attacked *gap*] doesn't like the politician.

(Example 2 adopted from He et al., 2017, p.1069)

Indeed, several empirical studies provide evidence that supports the hypothesis that the Chinese ORC is easier than SRC. Chen et al. (2008) investigated how low and high working memory span readers process Chinese SRC and ORC in self-paced reading tests. The results indicated that low memory span readers had slower reading time in the SRCs compared to the ORCs. Similar findings were observed by He et al. (2007). The experiment was conducted using sentence-picture selection and self-paced reading tests. It was found that when the head noun was the sentential subject, the ORCs had higher accuracy scores and could be read faster compared to SRCs.

In sum, Chinese relative clauses are different from those of English. Because the word orders of the relative clauses in the two languages are different, the asymmetry in terms of processing difficulty also might be different. Although the current study did not examine the existence or nature of this cross-linguistic difference (as examining relative clauses in Chinese L1 was beyond the scope of the study), the cross-linguistic difference nevertheless served as a rationale for selecting this linguistic feature as our target for a training experiment. It is possible that the cross-linguistic difference could

lead to difficulty for Chinese-speaking English learners in comprehending and producing English SRC and ORC, and therefore, they may perform in a different way relative to native English speakers.

(Note, in the current study, Chinese-speaking L2 learners of English and native English speakers will be involved, but the performance of L2 learners and native speakers will be analysed separately, as it was not appropriate to compare the two groups as they were different in many ways that could affect processing and producing: age, educational and language learning background).

2.2 What does the research into instruction tell us about helping the learning of relative clauses?

2.2.1 Can explicit training affect offline and online processing and help learning?

Explicit training is where the instructor or materials "provide learners with information about L2 grammar rules or direct them to search for rules" (Morgan-Short, Steinhauer, Sanz & Ullman, 2012, p.933). There are many types of explicit training, including instruction which calls learners' attention to meaning, forms, or the integration of forms and meaning (Norris & Ortega, 2000). So far, the effects of explicit training in L2 learning has been widely investigated (e.g. Andringa & Curcic, 2015; Hopp, 2016; Marsden & Chen, 2011; McManus and Marsden, 2017; Norris & Ortega, 2000; Long, 1983; Wong & Ito, 2018). It is generally agreed that explicit training can facilitate L2 learning at least in offline processing (i.e., comprehension demonstrated after the relevant input sentence is complete) and production. Some studies have also suggested that explicit information and some kinds of practice might also be able to promote online processing.

2.2.1.1 Effects of explicit training on offline processing

In theory, the reason that explicit training can benefit offline processing is that it can promote noticing. Krashen (1982) put forward the input hypothesis, which claims comprehensible input is necessary but not sufficient for language learning. In this view, the input plays an essential role in helping learners to construct mental representation

about grammar (VanPatten, 1996), and the initial stage of L2 learning is converting input to intake (VanPatten & Cadierno, 1993a). In the process of the conversion, noticing is regarded as a "necessary" and "sufficient" condition (Schmidt, 1990, p.129). Explicit input-based training consists of instructional interventions that aim to promote learning by changing learners' focal attention in processing target languages, and therefore increase the likelihood of the learners noticing the target features (Norris & Ortega, 2000). Based on these theories, explicit training could facilitate learners to pay attention to the targets and promote learning.

So far, there are an overwhelming number of studies that have demonstrated the effectiveness of explicit training in promoting offline comprehension and production (see for a small selection: Andringa, de Glopper & Hacquebord, 2011; Long, 1983; Marsden, 2006; Marsden & Chen, 2011; Norris & Ortega, 2000; Roehr-Brackin, 2014; VanPatten & Cadierno, 1993a; VanPatten & Oikkenon, 1996), and in metalinguistic knowledge (see Kasprowicz & Marsden, 2018; Tellier& Roehr-Brackin, 2013). Long (1983) investigated the role of instruction in L2 learning. It was found that compared to naturalistic exposure, explicit training is beneficial for both children and adults and for learners from beginning to advanced proficiency. In addition, Norris and Ortega (2002) reviewed 49 experimental studies related to L2 instruction and found that explicit instruction could lead to substantial effects which seemed durable. They also pointed out that different types of explicit training could have different effects in promoting learning. The instruction that focused on meanings tended to be more beneficial to learning than focus on forms instruction.

Some empirical studies have specifically compared the effectiveness of different type of instruction (e.g. Benati, 2005; Marsden, 2006; VanPatten & Cadierno, 1993a; VanPatten & Wong, 2004). For example, VanPatten and Cadierno (1993a) investigated the learning effects of a type of focus on form instruction which calls the participants' attention to the integration of form and meaning when listening and reading (input-based; this type of instruction will be introduced in detail in Section 2.2.2), and the effects of a type of focus on forms instruction which forced the participants to

produce sentences of target structures. The study was a pre- to post-test design, and the outcomes were measured by interpretation and written production tasks. It was argued that the input-based instruction was superior to the production-based mechanical practice-based instruction because the input-based instruction promoted gains on both interpretation and production measures while the production-based instruction was only beneficial to production (similar findings could be found in Benati, 2005; VanPatten & Wong, 2004; etc.).

In those studies, production-based instruction involved *output* practice and they found that, perhaps unsurprisingly, that the learners' ability in production following production practice was more likely to be promoted than that of interpretation. When the comparison instruction was also input-based, the results also suggested the effectiveness of explicit information and input-based training and gave additional insight to the role of *different types* of input-based training. Marsden (2006) carried out a classroom-based study to compare the effects of the focus on form instruction used in VanPatten and Cadierno (1993) (trained the integration of form and meaning) and another type of input-based training which was named enriched input. The enriched input training provided the participants with grammar explanation about target structures followed by activities which involved the targets but did not force them to understand the meaning of the inflections for person, number and tense. In a pre-, post- and delayed post-test design, the participants' listening, reading, speaking and writing competencies were measured. The results revealed that the learners who received form-meaning connection training had more learning gains than those who received enriched input, in both comprehension and production. The enriched input group did not show significant improvement across the time, especially the lower proficiency learners.

On the other hand, slightly different results were found by Kasprowicz and Marsden (2018), with younger learners on a different linguistic feature (case marking on German articles as cues for role assignment). They found that when the comparison group were told the rule and had spot the form each time by clicking on it (though not

connect it to a meaning each time), they performed as equally as well as those who had to connect the form to its meaning (role assignment function) each time, on all measures which were offline comprehension and production.

McManus and Marsden (2017, 2018, 2019) also conducted a series of empirical studies (which are summarised in more detail later because they also took online measures) that provide additional empirical studies which demonstrated that explicit information and input-based training can be beneficial for offline comprehension and production.

In sum, the role of input-based instruction on *offline* comprehension and production seems to indicate clear benefits of providing explicit information and practice some kind (whether it is spotting the form or connecting to its meaning or function).

The current study will involve the measures of offline comprehension, oral production and metalinguistic knowledge. Although the offline effects of explicit training have already been widely investigated, the instruction used in this thesis is innovative as it introduces a new type of practice which aims to intervene *during* a parse (and is explained in more detail in section 2.2.3). That is, given this new type of intervention, and the nature of the linguistic feature (syntax) the offline measure are examined and could provide additional evidence about the effectiveness of explicit information with input-based practice. In sum, the current study, therefore, investigate how effective this (new) kind of instruction will be in teaching relative clauses (relative to exposure alone and tests only) on offline measures, and whether the effects will be affected by the type of relative clauses.

2.2.1.2 Can explicit training affect online processing?

It remains unclear the extent to which the knowledge gained via explicit instruction, that is accessible via an offline test, could affect online processing. VanPatten and Cadierno (1993a, 1993b) put forward a basic idea (a simple model of learning) that second language acquisition includes three processes. The first one is converting input to intake; then, the intake would be internalised into learners' developing system ("the

mental representation of the second language the learner is constructing over time", VanPatten, 1996, p.5), and it is seen to be a type of internal grammar (VanPatten, 2020). Learners use the knowledge from this system to make output. However, not all input can convert to intake because learners 'filter' input during processing (VanPatten, 1996). In this simple conceptualisation of learning, this part of the input is called intake (VanPatten, 1996). The initial stage of language learning, the conversion of input to intake, involves making grammatical forms connect to their referential real-world meaning during real time comprehension (VanPatten, 1996, 2004; VanPatten & Cadierno, 1993b). This process is called input processing (VanPatten, 2004). VanPatten & Cadierno (1993a) argued that the traditional form-focused instruction that manipulates the output might not be able to change the developing system because such instruction does not require learners to pay attention to meaning during the processing of input. In order to effectively help learners to convert the input to intake, the process of input could be manipulated. They suggested that "instruction as direct intervention on learners' strategies in input processing should have a significant effect on the learners' developing system" (VanPatten & Cadierno, 1993a, p. 240). In other words, if providing learners with the strategies about how to make form-meaning connections, the way that the learners process the grammar could, in theory, be changed. Such changes in processing should be detectable through online measures like self-paced reading (SPR), eye-tracking and event-related brain potentials (ERPs) (see Keating & Jegerski, 2015 for a methodological discussion of these methods).

However, whether learners can be 'trained' to process the input differently, still needs more solid evidence. So far, some studies have explored the effects of explicit training on online processing (Andringa & Curcic, 2015; Dracos & Henry, 2021; Hopp, 2016; Issa & Morgan-Short, 2019; McManus & Marsden, 2017; VanPatten & Smith, 2019; Wong & Ito, 2018), yet the findings are inconclusive. Andringa and Curcic (2015) did not find that explicit knowledge facilitated online processing of direct object assignment in an artificial language. They provided Dutch L1 speakers with a 35-minute auditory session to train them to use a preposition to predict the animacy of direct

objects. The ability to use the target structure was measured by oral grammatical judgment test (GJT) and visual world eye-tracking tests. The effects of teaching explicit knowledge could only be observed in offline oral GJT but not in (online) eye-tracking tests. Similar findings were observed by Dracos & Henry (2021). They found that task-essential training (training to make the meaning of specific forms in the input essential to comprehension) could facilitate aural offline interpretation of Spanish verbal inflections but could not affect online processing examined via SPR. They proposed that it might because the training was not "strong enough" (p. 23) to help the learners to overcome their inefficient processing strategy.

In contrast, Hopp (2016) found that the intermediate L1 English learners of German showed predictive gender processing (measured by visual world eye-tracking tests) after receiving explicit training on lexical gender agreement (for similar findings see Wong & Ito, 2018). In addition, McManus and Marsden (2017) found that the learners who only received the explicit information and practice about L2 French *Imparfait* did not show significant online improvement in the SPR tests. The learners who received additional information and training in their L1 (English) showed gains in online comprehension, and the effects were durable in the delayed post-test. One possible reason to explain the inconsistency of the findings might be the L2 proficiency of the learners. Hopp (2016) pointed out that the previous knowledge about the target structure is a prerequisite for successfully using the language cues to process the structure during online processing. Thus, it might be more difficult to show online effects of training with learners who do not have previous knowledge about the target language (e.g., the artificial language in Andringa and Curcic, 2015) than the learners that have it (Hopp, 2016; McManus and Marsden, 2017; Wong & Ito, 2018). Critically, however, the participants recruited in Dracos and Henry (2021) also had existing knowledge about target structures, yet no significant improvement could be observed in online processing, thus suggesting that having prior (existing) knowledge of the target structure cannot fully explain the contradictory findings to data. In addition, the difference findings from Hopp (2016) and Dracos and Henry (2021) also indicated that

the online effects of explicit training might also vary as a function of the target structure being trained. Finally, the number of studies related to the training effects on online processing is somewhat limited, so this topic still merits further exploration.

The current study will investigate the extent to which providing explicit training benefits online and offline comprehension as well as the production of English relative clauses. A critical novel feature of the current study is that the previous studies that have examined the online effects of training all focused on morphology (or morphosyntax), while the current study will explore the effects of syntax (word order and role assignment in relative clauses). Processing syntax may differ from processing morphology as arguably it may have a more indirect relationship with the kind of real-world meanings (or functions) (such as person, number, tense, case, animacy gender of the noun) that have previously investigated; thus, different results with the previous findings might be found in the current study.

So far, as we have seen, plenty of studies have investigated the relations between L2 instruction and learning. In the following sections (sections 2.2.2, 2.2.3, 2.2.4), the research into the instruction that helps learners to process language will be reviewed in more detail. The studies related to processing instruction (PI), sentence processing, and input flood training will be reviewed, respectively.

2.2.2 Processing instruction

Processing instruction (PI) was put forward based on input processing theory (see 2.2.1), which, as noted above, aims to alter the way that learners perceive and process the input by training them to establish form-meaning connections, therefore, to change the learners developing system (VanPatten, 2005; VanPatten, 2015; VanPatten & Cadierno, 1993a). It has two fundamental characteristics (as conceived by VanPatten, 2002). First, PI could be regarded as a type of focus on form or input enhancement, which “uses a type of input to push learners away from the nonoptimal processing strategies” (VanPatten, 2002, p.764). Although PI is placed under the border of focus-on-form instruction, the activities involved in PI are different to drill-like activities

that focus-on-form instruction usually includes. To be specific, PI first help learners to identify the problematic processing strategies and then provide activities that force them to process sentences in a correct way. Second, during the instruction, the learners are expected to process and comprehend sentences in the activities while learning to connect meaning to form. They are not required to produce the target form, yet it is thought to be possible to generate when the developing system is shaped by the input-based training (VanPatten & Cadierno, 1993a & 1993b).

2.2.2.1 Components of PI

Typically, PI consists of three components: explicit information (EI) prior to the activities and two types of structured input activities, which are referential activities and affective activities (VanPatten, 2002, 2005).

EI

EI provided before the practice refers to the information about a target structure that guides the learners to establish the form-meaning connection and correctly process the structures (VanPatten, 2002). Example 3 shows the EI used to guide the learners to process French causative construction (Wong & Ito, 2018). Explanations about the target structure and the situations that the structure that could be used were provided.

Example 3:

We often ask or get people to do things for us by telling them to do something.

Paul says, " John, would you mind doing the dishes?"

If you and I were to describe what is happening we might say:

We say, "Paul gets John to do the dishes."

or

"Paul makes John do the dishes."

This is called a causative construction (because someone is causing a behavior in someone else.)
French has a similar structure using the verb *faire*. Let's repeat our examples from above.

Paul says, "Jean, pourrais-tu faire la vaisselle?"

We say, "Paul fait faire la vaisselle à Jean."

How would we describe the following scenario?

Wynne says, "Sara, pourrais-tu promener le chien?"

We would describe Wynne getting Sara to do it like this.

We say, "Wynne fait promener le chien à Sara."

(Wong & Ito, 2018, p.256)

Theoretically, EI would be beneficial in L2 learning because it helps learners to notice the form and therefore facilitates learning (Alanen, 1995; DeKeyser, 1997; Doughty & Williams, 1998). There are numerous empirical studies that investigated the effectiveness of EI, but the results were inconsistent.

Some researchers compared the effects of PI with or without EI, and argued that EI might not contribute towards the effectiveness of PI (e.g. Sanz & Morgan-Short, 2004; VanPatten & Oikkenon, 1996; Wong & Ito, 2018). For instance, VanPatten and Oikkenon (1996) argued the effects of PI should be attributed to the input practice instead of the EI itself. The effects of the EI and the activities on the interpretation and production of Spanish preverbal direct-object pronouns were examined. In the study, VanPatten and Oikkenon divided the participants into three groups: a regular PI group (receiving EI and input structured activities), EI only group and structured input activities only group. It was observed that the regular PI group and the structured input activities only group outperformed the EI only group in both interpretation and the production after training. The EI only group showed no improvement in the interpretation tests and very limited gains in production. This study indicated that EI *alone* might not benefit L2 learning or provided a very limited role at most. The findings of this study were in line with Sanz and Morgan-Short (2004). They investigated the role of EI (either provided before or after task-essential practice) in learning Spanish word order. They found that the participants who received explanation before practice (with or without corrective feedback after practice) did not outperform as well as the participants who received practice alone (without corrective feedback) in comprehending and producing the target form. In addition, the role of EI in PI also had been investigated with online measures. Wong and Ito (2018) compared the effects of regular PI and PI without EI on learning French causative visual-world eye-tracking tests. The results showed that after training, the participants significantly gained accuracy in selecting the correct picture, and their eye-movement pattern has been changed regardless of whether they received EI, so long as they received input-based practice in connecting form to meaning. In sum, the above studies indicate

that EI might not be necessary for PI because the effectiveness of PI seemed to be mainly due to the structured input activities.

Nevertheless, some researchers argue that although EI *alone* might not play a critical role in PI, it can *speed up* input processing to some extent, and the effects are mediated by the type of structure (e.g. Culman, Henry & VanPatten, 2009; Fernández, 2008; Henry, Jackson & Dimidio, 2017). Fernández (2008) investigated whether EI in PI could assist learners to process the linguistic targets sooner and faster. Two groups of participants were involved in the study and were taught both the Spanish object-verb-subject (OVS) word order and subjunctive (the design was within-subject for the linguistic feature). One group received the regular PI that included EI (structured input + EI) while the other group received the PI without EI (structured input). The effects of learning were examined during processing by a measure called '*trials to criterion'* (p.285) which set a criterion by which the researcher would deem a participant was correctly processing the input. It was assumed that if the learners began and continued to respond to target items at least three times correctly, they were able to process the input correctly. Thus, the number of practice items up until the criterion point were counted. In addition, the response time and accuracy were also analysed. The results revealed that the two groups had no difference in trials to criterion, speed or accuracy in processing one of the target features: the OVS sentences. However, it was observed that the group that received regular PI (with EI) outperformed the group that received structured input only (no EI) in processing subjunctive sentences. The PI group needed fewer items to reach the criterion and had faster reaction time as well as higher accuracy than the structured input group. The results indicated that the role of EI in PI might depend on the nature of processing the linguistic targets.

Culman et al. (2009) replicated the study of Fernández (2008) by investigating the role of EI in PI in processing German accusative case makers and OVS sentence structure. They used the same measure as Fernández (trials to criterion), and they found that the participants who received the PI with EI tended to reach the point of

criterion sooner than those who did not receive EI. Thus, they claimed that EI was beneficial in promoting faster acquisition. In addition, Henry et al. (2017) conducted a study to investigate the role of EI in processing German accusative case makers. The outcomes were measured by a comprehension and a production task through pre-test, immediate post-test and four-week delayed post-test. They found that the participants who received EI showed improvement in both comprehension and production at the immediate post-test, and the gains in comprehension were maintained on the delayed post-test. On the contrary, the participants who were not provided with EI only showed improvement in comprehension at the immediate post-test, and the improvement disappeared at the delayed post-test.

In sum, EI might not be a necessary component in PI; however, it has the possibility to assist learners in correctly processing the input sooner (that is, with less practice than when they do not receive EI). In addition, the previous studies provided evidence that indicates PI (structured input with EI) is beneficial to L2 learning, and the learning effects might be more durable than following solely structured input, though it seems that the effects of EI could be different depends on target linguistic structures.

Referential activities

Referential activities force the learners to pay attention to target grammatical form to get the meaning by asking them to choose an answer from two options (Culman et al., 2009; Marsden, 2006; Marsden and Chen, 2011; VanPatten, 2002). After each item, feedback indicating whether the response is correct or not is provided, but there is no further explanation about the target form (Benati, 2005; Marsden & Chen, 2011; Wong & Ito, 2018). For example, Marsden and Chen (2011) trained learners on English past tense verb inflection *-ed*. In referential activities, the participants were asked to respond to statements in a way which meant they had to notice the target and attend to its meaning (see Example 4). The statement sentences contained either a target form (*-ed* verb inflection) or a contrasting form (present tense form), and the participants were required to decide whether the action described in the sentences referring to something that happened before or happened regularly.

Example 4:

Some of Delia's diary entries have got smudged. Decide whether Delia has written about an event that happened in her *previous summer holidays* or if she is referring to something she *usually does in the summer holidays*.

1. I learn Spanish.

   a. last summer b. usually does

2. My family visited Paris.

   a. last summer b. usually does

(Marsden & Chen, 2011, p. 1067)

Referential activities aim to alter the way that the learners process linguistic structures. For instance, VanPatten's lexical preference principle predicates that "if grammatical forms express a meaning that can also be encoded lexically, then learners will not initially process those grammatical forms until they have lexical forms to which they can match them" (VanPatten, 2015, p.95). It means that if there are lexical items like 'yesterday' in the sentence, the learners would be likely to use those items instead of using the *–ed* verb inflection to get the meaning of past tense. (This phenomenon could, in fact, be for a variety of reasons, including the physical salience (e.g., length, prosody) of grammatical versus lexical forms; the influence of the fact that learners have an existing representation for the lexical items in their L1 and L2 (e.g., temporal adverbs or subject pronouns); the order of encountering these features in the sentences used in the studies. VanPatten does not distinguish between theses explanations). The referential activities in Marsden and Chen (2011) did not provide lexical items that indicated the tense, so the participants had to use the verb inflection cue to process the sentences. Therefore, in this way, referential activities might be able to push the learners away from inefficient processing strategies by training them on form-meaning connection.

Affective activities

Affective activities also belong to structured input. They are provided after referential activities and are used to reinforce form-meaning connection (Culman et al., 2009;

Marsden, 2006; Marsden and Chen, 2011; VanPatten, 2002). The idea behind them is that they provide additional exposure to the target form, and, because they follow the training provided by referential activities, the learners will be using their 'new' processing strategies (presumably embedded during training) by the time they encounter these affective activities. Different from referential activities, items in affective activities do not have right or wrong answers, and the participants are involved in processing sentences about the expressing some opinions or responding in some way to whole sentential meaning of sentences in which the target form is embedded (Culman et al. 2009; Marsden, 2006; Marsden and Chen, 2011; VanPatten, 2002). In addition, the sentences in affective activities only include the target feature, while those in referential activities contain another feature besides the target in order to juxtapose different pairs of form-meaning connections (as noted by Marsden and Chen, 2011). Looking at example 5, the affective activities in Marsden and Chen (2011) asked the participants to express their opinions on some events. The sentences that described the events included the target structure *–ed* past tense verb inflection, which helped the learners to reinforce the structure they learned from referential activities.

Example 5:

Delia has written a diary entry about her family's *last summer holidays*.

What do you think about her activities?

1. My family visited London.

   a. interesting b. boring

2. I learned Japanese.

   a. interesting b. boring

(Marsden & Chen, 2011, p. 1068)

Only very few studies have compared the role of referential and affective (or affective-style) activities in PI, but the results were inconsistent. Marsden (2006) compared PI ('EI + referential + affective') activities with 'EI + affective' only (enriched input in which the form was embedded but learners did not have to detect a form meaning mapping). Each group had had equivalent exposure to the same number and

type of target forms. Marsden found that the group *with* referential activities made reliable gains on all measures (comprehension and production) that were superior to those who only have affective activities. This suggested that the affective-style activities did not really add benefits to the learning. Marsden and Chen (2011) also suggested that the referential activities might play a more critical role than affective activities in PI. They investigated the effects of referential and affective activities in *isolation*. They assigned the participants into four groups: referential and affective group, referential group, affective group, and test-only group. They found that the gains (in comprehension and production) of the referential + affective group were similar to those of the referential only group. In addition, the affective group and the test-only group did not show any improvement across time. Thus, the study suggested that the observed gains of training should be attributed to referential activities, and the affective activities were not beneficial in learning English *-ed* past tense verb inflection. Both these studies emphasised the importance of drawing the learners' attention to the target structures.

However, Henshaw (2012) conducted a similar study to that of Marsden and Chen (2011) but had different results. She suggested that referential activities and affective activities, either be provided isolated or combined, may have similar effects in L2 learning. Henshaw (2012) investigated the role of referential activities and affective activities in combination and in isolation on the learning of Spanish subjunctive. Three groups of participants were involved in the study. All of them received EI first, and then they were treated in three training condition: 1) referential activities, 2) affective activities, or 3) the combination of the two types of activities. The recognition and interpretation of the target structure were tested in pre-, post- and two-week delayed post-test. The results showed that all three groups significantly improved at the post-test in both recognition and interpretation tasks. In addition, at the delayed post-test, it was found that the groups that received affective activities, regardless of whether or not they had received referential activities, were better at maintaining the gains in interpretation task compared to the group who received referential activities

only. One reason might be able to explain partially why the results of HenShaw (2012) and Marsden and Chen (2011) were inconsistent. Henshaw provided EI about the target structure to all the participants before they engaged in the activities, while Marsden and Chen did not. EI prior to the activities might be sufficient to raise the participants' attention to the target form, though affective activities did not force them to make the form-meaning connection. However, Marsden (2006) did provide the learners with EI, and yet the learners who received affective activities alone did not make substantial or reliable gains in all measures. Thus, the provision of EI cannot solely explain why Henshaw's study found gains in the + affective groups that were superior to those in the referential only group. See also Kasprowicz and Marsden (2018) who found that when learners where given EI followed by enriched kind of input (in which they had to spot that target form and also proved some opinion about the meaning of the sentence), they performed equally as well as learners who had EI followed by referential activities alone (in which they had to connect case marking on the article to its function). That study also suggested that input-flood style activities which provide lots of exemplars, can be effective for learning if learners are also given some EI and/or have to spot the target form during the affective-style activities.

In sum, three components (EI, referential activities, and affective activities) play different roles in PI. EI, shown before the structured input activities, provides learners with grammar rules of target structures. It might not be necessary for language learning, but some studies indicated it could speed up processing input (and some argue that the nature of the referential activities actually engenders explicit knowledge – that is, the corrective feedback leads to learners establishing some explicit knowledge about the target rule, even if they are not told it – see DeKeyser and Prieto Batano (2015). Referential activities force the learners to use the form-meaning connection to process sentences, while affective activities are designed to reinforce the form-meaning connection (on the assumption that the referential activities have helped learners to establish a new processing routine). It is not completely clear whether both referential and affective activities are needed to help learning. However,

if no EI is provided or the practice does not make noticing the target form in some way obligatory, it may be that affective activities seem to be less effective than referential activities and may not contribute to learning in isolation. The number of studies investigating this issue is low.

2.2.2.2 The first noun principle

As discussed previously, the aim of PI is to push the learners away from wrong processing strategies by helping them to make the correct form-meaning connection (e.g. VanPatten & Cadierno, 1993a, 1993b; VanPatten, 2002; VanPatten, 2020). VanPatten (1996, 2004, 2015) put forward that there are some universal strategies that all learners are likely to use when they process languages. Sometimes the universal strategies may cause wrong form-meaning connection and, therefore, result in the failure of converting input to intake and being acquired into the developing system (VanPatten, 2004). One is named as the First Noun Principle.

Principle 2: The First Noun Principle.

Learners tend to process the first noun or pronoun they encounter in a sentence as the subject/agent.

(VanPatten, 2004, p.14, p.18)

The design of PI should base on those input processing principles and steer learners away from inappropriate strategies by providing them with the correct ones (VanPatten, 2015). So far, plenty of empirical studies that investigated the effectiveness of PI have been carried out.

Evidence supporting Principle 2: The First Noun Principle

This principle addresses the grammatical role assignment of the two nouns of a sentence (VanPatten, 2004). Listeners (and readers) are likely to assign the role of agent or subject to the first noun phrase they encounter. This is an efficient strategy when processing sentences like English Subject-Verb-Object (SVO) (*e.g., The cat chases the dog*). However, this strategy can be problematic when the agent or the subject of the sentence is not the first noun phrase (e.g. English passive voice, Object-Verb-Subject (OVS) structures, case marking). For instance, L2 learners tend to

interpret sentences like *The cow was kicked by the horse* as *The cow kicked the horse* (VanPatten, 2004, p.15). This misinterpretation might be because the first noun of the sentence was the *cow*, the processors might assign the role of the subject to it and regard the *cow* as the subject.

Several studies provide the evidence to support the existence of the First Noun Principle in L2 studies (e.g. Allen, 2000; LoCoco, 1987; Henry, Jackson & Hopp, 2020; Hopp, 2015; Issa & Morgan-Short, 2019; Jackson, 2007; Kempe & MacWhinney, 1998; MacWhinney, Bates & Kliegl, 1984; Meyer, Mack & Thompson, 2012; VanPatten & Smith, 2019; Wong & Ito, 2018). For example, Jackson (2007) investigated the strategies that intermediate English-speaking L2 learners of German used in processing German sentences. In English sentences, the grammatical roles are assigned by word orders, while, in German, role assignment is most reliably expressed by case marking. The case marking system exists in German but not in English, so English L2 learners of German may ignore case markings when processing the sentences. The study examined L2 German learners (with L1 English) ability to interpret German SVO and OVS sentences by timed comprehension task. They were tested three times in an 8-month time span. It was found that the participants had significantly lower accuracy scores in OVSs compared to SVOs across time. The findings indicated that the L2 learners were using word order instead of the case-marking cue during processing, which demonstrated the First Noun Principle.

Similar findings indicating that L2 learners tend to ignore the case markings can be found in learning Latin (VanPatten & Smith, 2019). They investigated the strategy that native English speakers use in Latin comprehension. Latin is a language that has flexible word orders. It allows SVO, subject-object-verb (SOV), and object-subject-verb (OSV) structures to express the same meaning. For example, the three sentences in example 6 all mean '*the tiger loves the bear*'. The cue that decides who does the action is the nominative and accusative case markings of nouns. If they use the word orders rather than the case markings to assign the grammatical roles of nouns, they are likely to misinterpret the OSV sentences. For example, the sentence *Ursum tigris amat* might be

regarded as ‘the bear loves the tiger’.

Example 6:

SOV: Tigris ursum amat.

SVO: Tigris amat ursum.

OSV: Ursum tigris amat.

(VanPatten & Smith, 2019, p.409)

**Note: In the original paper, the example they gave for OSV structure is ‘Ursum amat tigris’. It is likely to be a typographical error. It has been corrected to ‘Ursum tigris amat’ here.*

All the participants had no previous exposure to Latin, and they received 100 training items of either SOV or SVO structure. After the training, the offline comprehension of SOV, SVO and OSV sentences was examined. The results indicated that the participants could comprehend SOV and SVO structures much better than OSVs. They had very low accuracy scores in interpreting OSV sentences.

The above two studies demonstrated that L2 learners’ interpret the first noun as the subject in processing a language containing case markings, based on the accuracy scores of offline comprehension tests. Henry et al. (2020) confirmed this preference in German with an online measure, a visual-world eye-tracking test. For each item of the test, the participants would see a picture depicting four nouns, including a potential agent and a potential patient. The participants heard a sentence either in SVO or OVS structure (see example 7), and their fixation proportions were analysed. The grammatical roles of nouns were assigned by the nominative (*der*) and accusative (*den*) case. If the learners used the case-marking cue in processing, they would look at the targets after hearing the verb because the action and the doer/receiver were clear by then. Example 7:

SVO: Der Wolf tötet den Hirsch der Hirsch der Jäger

The wolf kills the dear the dear the hunter

OVS: Den Wolf tötet der Jäger der Jäger der Hirsch

The wolf kills the hunter the hunter the deer

(Henry et al., 2020, p.12)

The results showed that the participants fixated on the agent after hearing the first noun and the verb regardless of the structure. It means that they were trying to integrate the noun and the verb to predict the meaning of the sentence, and they anticipated the second noun to be the patient. This pattern suggested that the learners did not use the case marking to process the sentences, and they preferred to regard the first noun they heard as the subject (agent).

In addition, some researchers suggested that L2 learners of French had difficulty in interpreting causative construction (Allen, 2000; VanPatten & Wong, 2004; Wong & Ito, 2018). In French, the verb *faire* that means *to do* or *to make* can be used as a common verb, and it also can be used as the main verb in causative construction. In Example 8, the first noun of the sentence is *Jean*, which might be regarded as the subject of the action *'buy milk'* when the processors attempt to use the First Noun strategy to process the sentence. However, in causative construction, the cue used to assign the grammatical roles to the nouns is '*fait* + infinitive verb' structure instead of the word orders.

Example 8:

Jean fait acheter du lait à Paul.

Jean makes to buy milk to Paul.

"Jean makes Pierre buy milk."

(Wong & Ito, 2018, p.243)

Wong and Ito (2018) investigated the strategy that intermediate L2 learners used in processing French causative through the visual-world eye-tracking test. For each item of the test, the participants saw two pictures and heard a sentence. They were required to choose the picture that matched the sentence after hearing the complete sentence. The results showed that the participants increasingly fixated on the incorrect picture when they heard the first noun of the sentence, and the accurate rate of choosing the picture was less than 5%. The findings are in line with those of Allen (2000) and VanPatten and Wong (2004). The two studies provide evidence that

supports the First Noun Principle during online comprehension.

From what has been discussed above, the First Noun Principle had been considered as a universal default processing strategy that applies to many languages, but even though it can lead to misinterpretation of some structures, like case marking, French causative construction and English passive voice.

So far, many researchers have attempted to push learners away from using the first noun strategy in processing by providing them with PI or other types of instruction. The previous PI studies that related to the role assignment will be reviewed in the following section.

PI studies related to teaching role assignment

So far, a substantial number of studies that investigated the effectiveness of PI in teaching role assignment have been carried out (e.g. Allen, 2000; Andringa & Curcic, 2015; Culman et al., 2009; DeKeyser & Sokalski, 1996; Kasprowicz & Marsden, 2018; VanPatten & Cardierno, 1993a; VanPatten & Smith, 2019; VanPatten & Wong, 2004; Wong & Ito, 2018). In general, the studies have demonstrated that PI is beneficial to learning role assignment of various languages.

The first PI study was carried out by VanPatten and Cardierno (1993a). They compared the effectiveness of PI and instruction that included EI and output practice in learning the Spanish case marker. The EI in PI not only introduced the target structure but also emphasised that Spanish allowed OVS structure where objects might be placed before the verb, and the objects were identified by prepositions. Aural and reading referential and affective activities were provided after the EI. Compared with PI, the output group received the EI that only included the explanation about the target without illustrating the processing problem. The oral and written production activities which involved form-oriented practice, meaningful practice, and open-ended communicative practice, were provided after EI. Besides the two training groups, there was a control group that did not receive any instruction. All three groups took part in the pre-test, one-day post-test, one-week post-test and one-month post-test, and each test phase included an aural comprehension and a written production test. The results

showed that the PI group had significant improvement in both comprehension and production, while the output group only gained in production. VanPatten and Cardierno (1993a) explained that the improvement of output group was due to the participants had received production training, and what they gained was a learned competence instead of acquisition (Krashen, 1982). Nevertheless, the PI group merely received interpretation tasks, but they still gained in production, which suggested that the participants had the ability to draw upon the knowledge from intake to make sentences. Thus, they claimed that PI was able to change the way that learners process the language while the output-based training could not.

After VanPatten and Cardierno (1993a), a series of replication studies have been carried out, and some of them are in line with the findings of VanPatten and Cardierno. For instance, VanPatten and Wong (2004) strictly created teaching and accessing materials following VanPatten and Cardierno (1993a). They found that PI could promote the interpretation and production of French causative structure, and PI was superior to output based training in facilitating interpretation. However, some studies showed different results. Allen (2000) conducted a conceptual replication of VanPatten and Cardierno (1993a) to investigate the effectiveness of PI and output-based training in French causative structure. It was observed that the two types of instruction were equally effective in promoting the interpretation of the target structure, and moreover, the output-based group performed better than the PI group in the production task. Hence, Allen put forward that both PI and output-based training could alter the way of processing French causative structure, and the superiority of PI that was found in VanPatten and Cardierno (1993a) might not be generalised to all the structures. VanPatten and Wong (2004) argued that Allen's results were different from theirs was mainly due to the materials that Allen used did not strictly follow the requirement of structured input activities.

It could be noted that these early studies used offline tests to measure comprehension. However, to further evaluate whether the processing mechanism itself has been changed by instruction, online measures were needed. Wong and Ito (2018)

evaluated the effects of PI and output instruction in altering how learners process French causative construction through visual-world eye-tracking tests. In the test, the participants would hear a sentence and see two pictures, and they were asked to choose the picture that matched the sentence. The accuracy scores and eye movements were analysed. They have carried out two experiments, and both of them used pre-, post-test design. In the first experiment, the PI and TI group did not receive EI, while the second experiment included EI prior to the activities for both groups. Before the intervention, all the participants tended to look at the incorrect picture throughout the sentence and had very low accuracy rates (less than 5%). After the intervention, the PI group, regardless of whether they have received EI or not, showed significantly higher accuracy scores and their eye-movement pattern was changed. Compared to the pre-test, the participants started to fixate on the target picture at the onset of the last prepositional phrase (e.g. *à Pierre* in *Marie fait faire la vaisselle à Pierre*), which indicated that the First Noun Principle tendency had been reduced. On the other hand, the output group also showed some improvement in accuracy, especially when the activities followed EI, but the gains were very limited and were much less than those of the PI group. For the eye-movement pattern, the participants who received output-based training without EI did not change over time. However, the output with EI group could fixate on the targets at the same point as the participants who received PI. This study provided evidence from online tests to support that PI, even without EI, might be able to change the processing mechanism as well as improve the accuracy. The output-based training, including EI, was also likely to alter the way of processing, but it was less effective in promoting accuracy. Another interesting finding of the study is that although the aim of the PI is to help the learners to stop using the word order to process sentences and to train them to use *fait* + infinitive verb cue, the after-training eye-movement pattern indicated that the learners prefer to use of *à*+ noun to process French causative constructions. Thus, Wong and Ito claimed that PI was superior to output-based training because it promotes more gains in accuracy scores. Although this study suggested that training could affect online processing and

reduce the First Noun Principle bias, the issue would still benefit from further investigation to examine, for example, how much training could enable L2 learners to process in a target way, and whether this finding might be generalizable to syntax (as their study focused learners' attention on lexical items, *faire + infinitive* and/or the presence of *à* + noun).

In the comparison between the effects of PI and output-based instruction, DeKeyser and Prieto Botana (2015) reviewed previous empirical studies and suggested that PI might be better than output-based instruction in promoting comprehension, though this could only be observed in half of the relevant studies. When output-based instruction contained more communicative practice than drill-like practice, the instruction could be more effective than PI in helping learner' production. Nevertheless, in spite of the controversial results of the studies into whether PI is superior to output-based instruction in L2 grammar acquisition, plenty of studies have demonstrated the benefits of PI itself. From what has been discussed above, PI can effectively promote L2 grammar offline comprehension and production and might be able to reduce the bias of the First Noun Principle in online processing.

Besides the role assignment problem, PI has been used to address other language processing problems, and its effectiveness has been demonstrated. For example, Marsden and Chen (2011) found that PI could promote comprehension and production of English *–ed* past tense inflection; Marsden (2006) suggested that PI was beneficial in interpreting and producing French verb inflections for tense, number and person; Cadierno (1995) showed that the participants who received PI significantly gained in Spanish past tense comprehension and production; Benati (2004) confirmed the effectiveness of PI in Italian gender agreement comprehension and production. However, all the linguistic features discussed above are related to morphology and/or the lexicon (e.g. case marking, '*fai*t + infinitive verb' in French causative construction, verb inflections).

In contrast, the current study will teach learners to use the cue of distribution in sentence to assign the grammatical roles in relative clauses. In addition, as the core

principle of PI is to help learners to connect form to its referential meaning, the linguistic features used in PI studies all have, arguably, direct referential meaning. Arguably this is (less) the case for relative clauses. Indeed, VanPatten (2002) suggested that abstract syntax (which was not defined very fully) may not be amendable to instruction. Thus PI might not be the most appropriate instruction to use for amending the processing strategies used when encountering relative clauses in the input.

The instruction used in the current study is based on the PI but has a key difference in that it aims to teach the learners in one of the conditions 'parsing strategies' while they are reading and hearing each sentence with relative clauses (instructing them at the point a critical feature is heard or read), rather than exposing them to whole sentences and asking the learners to make a decision about appropriate meaning. The instruction will be described in section 2.2.3.

Methodological considerations of previous PI studies

Although the effects of PI have been widely investigated, it could be noticed that most of them only included offline comprehension (visual or aural) tests like sentence-picture matching (e.g. Allen, 2000; Kasprowicz & Marsden, 2018; VanPatten & Cadierno, 1993a; VanPatten & Oikkenon, 1996), GJT (e.g. Marsden & Chen, 2011; Robinson, 1995) and production (written or oral) tests like gap-fill (e.g. Kasprowicz & Marsden, 2018; Marsden & Chen, 2011; VanPatten & Cadierno, 1993a), picture description (Benati, 2004; Benati, 2005). However, since PI is based on input processing theory which refers to "the linking of form and meaning during real-time comprehension" (VanPatten, 2015, p.93), online processing assessments would be more convincing in measuring the learning effects of PI. To the best of our knowledge, only very few PI studies have involved online measurements, including trials to criterion (e.g. Fernández, 2008, Henry, Culman & VanPatten, 2009; VanPatten & Borst, 2012), self-paced reading (Dracos & Henry, 2021; McManus & Marsden, 2017; VanPatten & Smith, 2019), and visual world eye-tracking (Wong & Ito, 2018).

The measure, trials to criterion, is used to assess online comprehension during the treatment. As mentioned in section 2.2.2.1, this measure sets a criterion for reaching

correct processing. Fernández (2008) used this measure to evaluate the role of EI in PI. However, this measure had some limitations. As Fernández (2008) admitted, this criterion lacked external validation and the cut off of three in a row was established arbitrarily. In addition, this measure could only provide information about how many items that the participants need to process the targets correctly, but it could not show how the participants process sentences at the time of processing them. Thus, compared to the measure of trials to criterion, other online measures like eye-tracking and self-paced reading could provide moment-by-moment data to reveal processing patterns.

However, at the point of writing we are aware of only four published PI-informed studies that utilised eye-tracking or self-paced reading as outcome measures. Wong and Ito (2018) used the eye-tracking test to evaluate the effects of PI in processing French causative construction. They found that PI did change the way of learners processing, but the learners did not parse the sentences in a way as the researchers expected. For self-paced reading studies, McManus and Marsden (2017) showed that the learners who received L1+L2 EI and practice became sensitive to the target linguistic structure (French *Imparfait* structure) after training, while the L2 EI + practice group did not show robust online improvement. VanPatten and Smith (2019) indicated that whether the L2 learners could show online learning effects depended on the type of target structure; in addition, Dracos and Henry (2021) found that task-essential training did not benefit the online processing of Spanish verbal inflections.

It could be observed that the number of studies that used online outcome measures is very limited, and the results are inconsistent. Thus, the online effects of PI still need further exploration. Moreover, the previous three studies all focussed learners' attention on morphology or lexical items (not syntax itself) and only contained one type of online test, which measured either listening or reading online comprehension. The current study will provide a more comprehensive view of online effects by using both self-paced reading and visual-world eye-tracking tests.

2.2.3 Sentence processing and parsing strategies

The terms processing and parsing are widely used in language research, and both of them "involve moment-by-moment computations of language during real-time comprehension" (VanPatten & Jegerski, 2010, p.4). These two terms used here are based on the sentence level. Sentence processing is broadly defined as "a process whereby the meaning of a sentence is understood", and parsing is a component of processing, which refers to "the process whereby a syntactic structure is built based on the activated lexical information, morphosyntactic cues (such as word order and case marking), and an individual's syntactic knowledge"(Jiang, 2018, p.244). In this section, the differences between the sentence processing of native speakers and L2 learners will be discussed initially; then, the parsing strategies used in sentence processing will be illustrated.

2.2.3.1 Differences between L1 and L2 sentence processing

A significant number of studies have investigated the sentence processing of L1 and L2 from the aspects of sentence ambiguity (e.g. Cunnings, Fotiadou & Tsimpli, 2017; Fujita & Cunnings, 2021; Juffs, 1998; Witzel, Witzel & Nicol, 2012) and syntactic dependencies (e.g. Dallas & Kaan, 2008; Juffs & Harrington, 1995; Marinis, Roberts, Felser & Clahsen, 2005; Omaki & Schulz, 2011). Some researchers state that the sentence processing of L2 learners is fundamentally different from that of native speakers. One influential hypothesis related to this statement is the Shallow Structure Hypothesis (SSH) (Clahsen & Felser, 2006a, 2006b), which suggests that the L2 learners construct shallower syntactic representations compared to native speakers during online processing. To be specific, native speakers are able to use both lexical-semantic cues and syntactic cues during parsing sentence, while L2 learners might mainly rely on the lexical-semantic cues. This hypothesis has some support from some empirical studies. For example, Juffs and Harrington (1995) investigated the way that advanced Chinese-speaking L2 learners of English process subject (e.g. "*Who did Ann say__ likes her friend*") and object (e.g. "*Which man did Jane say her friend likes __*") extraction (p.487). In the study, the native English speakers and Chinese-speaking L2 learners of

English were examined the comprehension of subject and object extraction through self-paced moving window GJT and full-sentence GJT. It was found that relative to the native English speakers, the L2 learners had similar accuracy scores in object extractions but had less accuracy scores in subject extraction. Meanwhile, the reading time data also indicated that the L2 learners showed a greater extent of slowdown at the region after the first verb in subject extraction than object extraction sentences. Thus, Juffs and Harrington claimed that the difficulty of parsing subject extraction of L2 learners might be due to the reanalysis problem instead of competence difference.

Clahsen and Felser (2006a) suggested that the L2 learners in Juffs and Harrington (1995) tended to use a "lexically driven strategy" during processing syntactic dependencies (p. 25) because the main verb of the sentence was misinterpreted initially, which was consistent with SSH. In addition, the results of Marinis et al. (2005) are consistent with Juffs and Harrington (1995). Marinis et al. (2005) compared the processing of long-distance filler-gap dependencies between native English speakers and L2 learners of different language backgrounds (Chinese, Japanese, German, and Greek). The participants were tested through the SPR test and offline comprehension questionnaire. The native speakers and the L2 learners performed almost equally well in offline comprehension; however, the native speakers were able to use the intermediate syntactic gap in online processing (e.g. "*The nurse who the doctor argued___ that the rude patient had angered___ is refusing to work late*" p.74), while the L2 learners did not show the evidence of using the filler-driven strategy during processing. The L2 learners tended to use lexical cues to process long-distance *wh*-dependencies as they attempted to link the previous *wh*-phrase to its lexical subcategorizer. This study also supports SSH and provides evidence that the L2 learners mainly rely on lexically driven strategy rather than syntactic cues during real-time processing (see also Felser & Roberts, 2007; Grüter, Lau & Ling, 2020; Papadopoulou & Clahsen, 2003).

Nevertheless, some researchers challenged the hypothesis that L2 learners process sentences fundamentally different from that of native speakers, instead, they

argued that L2 learners are able to process sentences like native speakers (e.g. Fujita & Cunnings, 2021; Kim, Montrul, & Yoon, 2015; Omaki & Schulz, 2011; Williams, Mobius & Kim, 2001; Witzel, Witzel & Nicol, 2012) at least for some structures. For instance, Omaki and Schulz (2011) carried out a study to examine whether advanced Spanish-speaking L2 learners of English could use the relative clause island constraints to construct filler-gap dependencies. The offline and online comprehension was tested through the accessibility judgment task and SPR task, respectively. Both offline and online results indicated that the native speakers as well as the L2 learners demonstrated the ability of using relative clause island constraint during processing. In addition, some studies suggested that L2 learners could use syntactic cues to process as the native speakers for some language features, but they are less sensitive to grammatical information for the others. For instance, Williams et al. (2001) found that both native speakers and L2 learners could use filler-driven strategy in processing *wh*-dependencies; however, the L2 learners had difficulty in recovering from the initial misanalysis even in offline tests, while the native speakers did not encounter this difficulty. Similar findings could be observed in Fujita and Cunnings (2021), which compared the reanalysis of native speakers and L2 learners in processing temporarily ambiguous sentences (e.g. "*While Mary dressed the baby laughed happily*" p.1). Their findings were partially in line with Williams et al. (2001). They also found that L2 learners were more persistent with initial misinterpretation compared to native speakers. However, they suggested that the L2 learners, as well as native speakers, experienced garden-path effects during processing, and they reanalysed the sentences in a similar way (see also Kim et al., 2015).

In light of the findings that the L2 learners might be able to generate native-like sentence processing pattern, Clahsen and Felser (2018) modified the SSH and claimed that native speakers and L2 learners were not restricted to use only one pattern to process sentences. Both L1 and L2 speakers could have *shallow* structure processing (i.e., depend more on non-grammatical information relative to the grammar information in sentence processing), but it occurs more often in L2 learners. In addition,

the L2 learners have reduced ability in using syntactic cues instead of totally could not use the cues.

The above studies present a debate of whether L2 learners can process sentences based on syntactic cues in a similar way as native speakers, and their findings are generally consistent with the modified SSH (Clahsen & Felser, 2018). For example, some studies indicated that L2 learners could only use non-grammatical information like semantic to parse the sentences, while others argued that L2 learners were also sensitive to grammar features during processing. However, even though some evidence supports that advanced L2 learners can use syntactic cues to parse the sentences (see Fujita & Cunnings, 2021; Williams et al. 2001), the L2 learners still could not use the cues as efficient as native speakers.

The current study will directly train L2 learners to use syntactic based parsing strategies and investigate whether they will be able to use them to process relative clauses in real-time comprehension. Nevertheless, the current study will not focus on whether the L2 learners can be trained to process like native speakers; instead, the changes of L2 processing after training will be examined. In addition, many previous studies about processing relative clauses relate to garden-path effects of reduced relative clause and how temporarily ambiguous sentences can be solved. The current study will investigate the processing of relative clauses from a new angle, that is, with a view to avoiding ambiguity occurring during processing. In relative clauses, the position that decides whether the sentence contains an SRC or an ORC is the first word after the relative pronoun (see example 9). After the relative pronoun (*that*), if it is a verb (*chases*), the clause is an SRC, and the cat does the action of chasing; if it is a noun phrase (*the dog*), the clause is an ORC, and the dog does the action of chasing.

Example 9:

SRC: The cat that chases the dog is big.

ORC: The cat that the dog chases is big.

Thus, the word following the relative pronoun can be used as a syntactic cue to assign the grammatical roles of the two nouns in the relative clause. The L2 learners of the

current study will receive training about this syntactic cue, and the training effects will be measured in a pre-, post-, and delayed post-test design.

2.2.3.2 Predictive processing

The linguistic cue that will be used in the current study is the word after a pronoun in English relative clauses. The class of the word following the relative pronoun assigns the grammatical roles of the two nouns. If the word is a verb, the noun before the pronoun does the action, and the noun after the verb receives the action (e.g. *SRC: The cat that chases the dog is big*); a noun phrase showing after the pronoun indicates that that noun does the action while the noun before the pronoun receives the action (e.g. *ORC: The cat that the dog chases is big*). This syntactic cue could perhaps be used to predictively and retrospectively interpret the meaning of the sentence, as learners will be looking at a visual display of two potential meanings. The training materials will involve incomplete sentences, while the learner has to choose the meaning by selecting one of the two pictures (described in section 2.2.3.1). The information after the syntactic cue will not be provided (e.g., *SRC: The cat that chases…; ORC: The cat that the dog…*), so the participants will be forced to used the cue to parse the sentences.

Thus, in the following section, a short overview of the research agenda on L1 and L2 predictive processing will be given. However, it is emphasised that this body of literature is not core to the current study because when learners encounter the disambiguating word (noun or verb after *'that'*) they can *retrospectively* integrate this information in order to parse the matrix clause they just heard as either the subject or object of the sentence (they do not need to anticipate the rest of the sentence). Thus in the current study, the disambiguating word allows the language user to assign the grammatical role of the previous noun *and* the upcoming noun.

In order to parse sentences accurately, "comprehenders need to assign the intended structure and meaning to sentence incrementally, taking advantage of the grammatical constraints of the language" (Phillips & Ehrenhofer, 2015, p. 412). The words or phrases of a sentence are related to each other, and when the parsers

encounter one word or phrase, they would associate the word or phrase with previous information or wait for further information to integrate (Phillips & Ehrenhofer, 2015). In other words, language processing includes information integration and prediction (Grüter & Rohde, 2013). During sentence processing, listeners and readers might be able to use language cues such as syntactic structure, word category and lexical items to anticipate upcoming information (Kaan, 2014). This process is called predictive processing. Philips and Ehrenhofer (2015) emphasised the important role of predictive processing in language learning. They argued that the learners who have the ability to predict what will come next are likely to gain more information from input compared to those who passively parse the given information. If learners can make predictions about the upcoming materials, they would compare the words that actually occur to what they have expected. When the upcoming words mismatch the expectation, the parsers could learn from it and adjust future prediction to minimize the opportunity of making the same errors (Coumel, Ushioda & Messenger, 2020; Jaeger & Snider, 2013; Johnson, Turk-Browne & Goldberg, 2013; Kaan, 2014). Based on this theory, Grüter, Zhu and Jackson (2021) examined the effects of forcing prediction on learning English double object structure (DO, *"The girl fed the squirrel some nuts"*) (p.15). In the study, the experimental group was asked to describe, in writing, a picture using the given information (see figure 2.1), and then they would be presented with a written sentence of the target structure on screen. The participants were forced to compare their own production to the target sentence; thus, they were expected to learn the structure from the prediction error (that is, the 'prediction' being operationalised as 'what they had produced', and the 'error' as the difference between their production and the input they heard). In addition, the control group would see the target structure directly, and they were asked to copy down the sentence and then produce in writing their own sentence. The control group were not asked to compare between the target sentence and their own sentence. It was found that both groups showed improvement in producing the target structure at the post-test compared to the baseline test, but the gains of the experimental group were significantly greater than those of the control

group. The findings support the statement that prediction error can facilitate L2 learning.

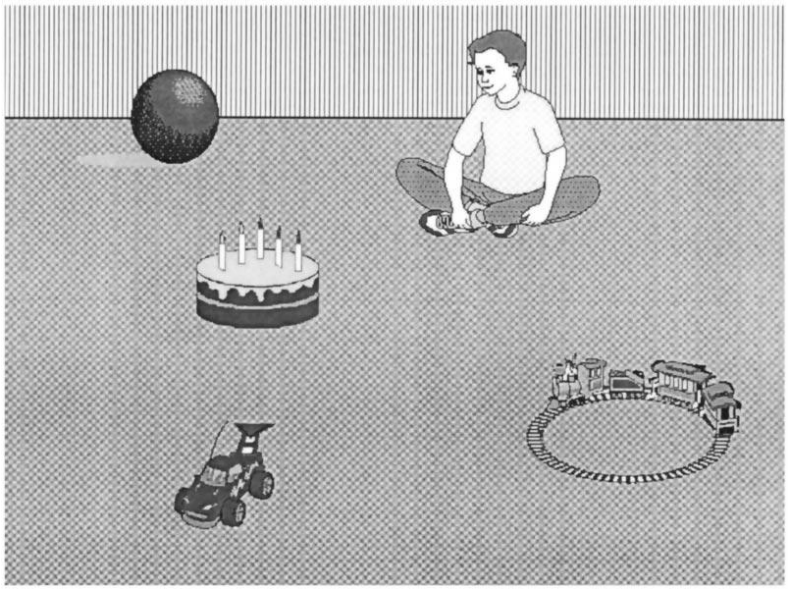
Figure 2. 1 Example training item for the experimental group

(Grüter et al., 2021, p.54)

To date, a large quantity of studies has been conducted to investigate predictive processing of native speakers and L2 learners (e.g. Altmann & Kamide, 1999; Brouwer, Sprenger & Unsworth, 2017; Chambers, Tanenhaus, Eberhard, Filip & Carlson, 2002; Chen, Bowerman, Huettig & Majid, 2010; Grüter et al., 2020; Grüter, Lew-Williams, & Fernald, 2012; Kamide, Scheepers & Altmann, 2003; Hopp, 2015; Hopp & Lemmerth, 2018; Lew-Williams & Fernald, 2007, 2010; Mitsugi & Macwhinney, 2016; Staub & Clifton, 2006; Van Berkum, Brown, Zwitserlood, Kooijman & Hagoort, 2005). There seems to be a consistent finding that native speakers show anticipatory behaviour during sentence processing. For instance, Altmann and Kamide (1999) found that native speakers could use real-world knowledge to make the prediction. Twenty-four native English speakers were involved in a visual-world eye-tracking test. For each set of items, the participants would see a semi-realistic visual scene that included a person and several objects and heard a sentence that contained one object of the scene. The eye movements towards the target object were recorded and analysed. For example, for the scene shown in figure 2.2, the participants were either heard "*the boy will move the cake*" or "*the boy will eat the cake*" (p.250). When the participants heard *eat*, they could predict the following item should be something that could be eaten. However, the anticipatory behaviour would not be observed when the participants heard *move*,

as everything in the scene could be moved. In other words, *eat* could be used as a semantic cue to make the prediction, but not for *make*. Thus, the participants were expected to saccade to the *cake* after hearing *eat* because the *cake* was the only edible object of the scene.

Figure 2. 2: Example visual scene (Altmann & Kamide, 1999, p. 250)



The results revealed that the participants started to fixate on the *cake* earlier in the *eat* condition than in the *move* condition. The first fixation towards the target object (*eat*) happened before presenting the actual word in *eat* condition but not in the *move* condition. It was demonstrated that the participants were sensitive to the language cue and could use it to make predictions with their real-world knowledge.

In addition, anticipatory behaviour also could be observed in processing other structures such as grammatical gender (e.g. Brouwer et al., 2017; Hopp & Lemmerth, 2018; Lew-Williams & Fernald, 2010), case-marking (e.g. Hopp, 2015; Kamide et al., 2003; Mitsugi & Macwhinney, 2016), syntactic structures (e.g. Chen, 2010; Grüter et al., 2020; Staub & Clifton, 2006), discourse context (e.g. Van Berkum et al., 2005) with native speakers. In a word, L1 predictive processing has been demonstrated by substantial literature.

In contrast, whether L2 learners can use predictive parsing strategies during sentence processing was unclear (e.g. Dijkgraaf, Hartsuiker & Duyck, 2017; Grüter et al., 2012; Hopp, 2015; Lew-Williams & Fernald, 2010). Many researchers stated that L2

learners had difficulty with predictive processing. Grüter and Rohde (2013) proposed a Reduced Ability to Generate Expectations (RAGE) hypothesis, which claimed that L2 learners, especially the low proficient L2 learners, had less ability in predicting upcoming information during processing. Many studies provided evidence for this hypothesis. Lew-Williams and Fernald (2010) found that intermediate English-speaking L2 learners could not use the gender of articles in Spanish as a predictive cue to predict upcoming nouns as well as native Spanish speaker, measured by visual world eye-tracking tests. In addition, Grüter et al. (2012) confirmed the findings in Lew-Williams and Fernald (2010). It was found that even highly proficient English-speaking L2 learners of Spanish who scored at ceiling in offline sentence comprehension task could not use gender cues as efficiently as native Spanish speakers. Nevertheless, some studies also indicated that advanced learners could have a native-like predictive processing pattern when the two languages have similar grammar property (see Foucart & Frenck-Mestre, 2011; Hopp & Lemmerth, 2016; Weber & Paris, 2004). This finding suggested that the predictive processing of L2 learners might be associated with learners' native language and their L2 proficiency.

To the best of our knowledge, no study to date has investigated this syntactic cue in processing English relative clauses for both native speakers and L2 learners. The current study will explore how native English speakers and Chinese-speaking L2 learners use this syntactic cue in aural and written sentence processing, respectively, and whether L2 learners can be trained to use it more (or differently) and whether this impacts performance in offline measures too. The small body of literature that address this later question will be reviewed in the following section.

2.2.3.3. Research into teaching predictive parsing strategies

So far, to the best of our knowledge, only two studies have investigated the effects of teaching predictive parsing strategies on L2 sentence processing (Andringa & Curcic, 2015; Hopp, 2016). As reviewed above in 2.2.1.2, the results of the two studies were inconsistent. Andringa and Curcic (2015) trained the learners to use a preposition as the cue to anticipate the animacy of direct objects through a brief EI with practice

which only lasted for 35 minutes. This linguistic feature did not exist in the L1 of the learners. After training, the learners could comprehend the target structure examined by GJT, but no evidence showed they were using the cue in online comprehension measured by the visual-world eye-tracking test. Nevertheless, Hopp (2016) argued that English-speaking L2 learners of German were able to use the determine-noun sequences to predict the grammatical gender of nouns after exposure to training activities (no EI was provided). Each participant attended two sessions. The first session was a pre-test, and the second session that involved training and immediate post-test took place one week later. The participants made significant gains in production and online comprehension and showed predictive processing at the post-test. The L2 learners in the two studies had different L2 language proficiency. The participants in Hopp (2016) had existing knowledge of the target language, while the participants in Andringa and Curcic (2015) did not have. However, previous studies indicated that L2 proficiency was important in predictive processing, and only advanced L2 learners might be able to use the predictive cues in processing (Hopp & Lemmerth, 2016). The lack of knowledge of the target language might be the reason to explain why the participants in Andringa and Curcic (2015) did not generate anticipatory behaviour after the training.

The participants who will be involved in the current study are supposed to have existing knowledge about English relative clauses, and their English proficiency is upper-intermediate. Thus, according to the arguments above, training in parsing strategies may help them to use a syntactic cue to predictively and retrospectively interpret the meaning of the sentence in order to decide which picture matches the sentence they hear or read (with the pictures depicting an SRC or ORC). An additional novelty of the current study is that the two previous studies focused learners' attention on morphology, and the current study orients attention to the part of speech.

2.2.4 Input flood instruction (Incidental learning)

The current study will compared the training in parsing strategies with a condition in which a similar amount of exposure to the target form will be given but in a condition

that does not orient attention to the disambiguating role of the word after 'that'. This condition is akin to a kind of input flood, a treatment that has been investigated for its role in promoting incidental learning. Thus, it is now appropriate to provide a short review of research into input flood and incidental learning.

Incidental learning is defined as "the acquisition of a word or expression without the conscious intention to commit the element to memory" (Hulstijn, 2012, p.1). The term normally applies to learners picking up language features from reading or hearing some materials (though could in theory also refer to learning from producing language). Input flood instruction could be considered as a kind of instruction that might promote incidental learning because it enables learners to be exposed to a large number of examples of target form without providing any explicit instruction or corrective feedback (Hernández, 2018). Input flood conditions ask learners to do something with the input that that does not require that they pay attention to the target feature. Indeed, a PI component, i.e., affective activities (introduced in section 2.2.2.1), can be described as a type of input-based activity which provide learners with multiple exemplars of target form but does not force learners to process the meaning of the targets (Marsden, 2006; VanPatten, 2002) and as such is a type of input-flood.

Hernández (2018) suggested that one disadvantage of input flood instruction was that it might not make specific feature salient enough to induce noticing of the target forms. This may not therefore be a very beneficial condition for learning according to proponents of a necessary of facilitatory role for noticing or awareness (such as Schmidt, 1990). To address this concern, some researchers employed some visual or aural enhancement techniques (e.g. **bolding**, <u>underlining</u>, *italicisation*, speak it more loudly) in the input to increase the likelihood of the target forms being noticed (e.g. Izumi, 2002; Lee & Huang; 2008), often termed 'input enhancement'. Although input enhancement highlighted the target forms, it is similar to input flood (Hernández, 2011) in that it does not directly provide EI to learners. In sum, input flood, input enhancement, and affective activities of PI share this same characteristic. All three types of input-based activities allow learners to attend to the target forms if they are

driven to do so by their own internal mechanisms (e.g., if they perceive it or analyse the input consciously) but the activities do not *require* learners to notice or comprehend the feature. Thus, if they contribute to learning, it could be evidence of incidental learning. However, incidental learning is normally used to describe learning conditions where the participants are not aware an upcoming test. Thus, it must be acknowledge that in a research study where the learners aware that there is going to be a test, it is possible that the learners themselves strive to become aware of a target feature and learn it.

A few studies examined whether input flood itself could contribute to L2 learning, but the results were inconclusive. Hernández (2008, 2011) investigated the effects of input flood with and without explicit instruction on learning Spanish discourse markers, but the results of the two studies were inconsistent. Both studies involved two groups; one group received input flood with explicit instruction while another group received input flood treatment only. Hernández (2008) found that only the group that received input flood with explicit instruction significantly improved at post-test (no delayed post-test was included in this study). However, Hernández (2011) observed that both groups gained in using Spanish discourse markers at the immediate and delayed post-test, which could indicate that explicit instruction had not enhanced the effects of the input flood. Two differences between the two studies might account for the inconsistency between the two sets of results. The most important one was Hernández (2011) had 49 training items, but only 15 training items were included in Hernández (2008). This increase in exposure was likely to increase the chance of noticing the forms during input flood training in Hernández (2011). Indeed, Uchihara, Webb & Yanagisawa (2019) confirmed the important role of repetition in L2 incidental learning. By analysing 45 effect sizes from 26 incidental vocabulary learning studies, Uchihara et al. (2019) claimed that the positive association between frequency of encounters and incidental vocabulary learning was of medium strength. Hence, the effectiveness of input flood might depend on whether sufficient training items are being provided. The second difference between Hernández (2008) and Hernández (2011) was the

proficiency of the participants. The participants in Hernández (2011) had higher L2 proficiency than those in Hernández (2008), which could indicate that the learners with higher proficiency might find it easier to notice the form from exposure to the targets.

In addition, some researchers compared the effectiveness of incidental learning instruction (e.g. input flood, input enhancement, affective activities) to that of explicit training (e.g., PI) in L2 learning (e.g. Henshaw, 2012; Issa & Morgan-Short, 2019; Marsden, 2006; Marsden & Chen, 2011), and inconsistent results were observed. As discussed in section 2.2.2.1, both Marsden and Chen (2011) and Henshaw (2012) investigated the extent to which referential activities and affective activities can contribute to learning L2 morphosyntax. Marsden and Chen (2011) found that the affective activities alone could not promote learning English *–ed* past tense verb inflection, while Henshaw (2012) claimed that affective activities had effects equivalent to those of referential activities in learning Spanish subjunctive. One notable difference between the two studies was that Henshaw (2012) provided the learners with EI before affective activities while Marsden and Chen (2011) did not. Issa and Morgan-Short (2019) found like Marsden (2006), that input enhancement activities were inferior to the structured input activities (no EI was provided for either group) in learning Spanish direct object pronoun. These studies were in line with Hernández (2008), which found that input flood training was effective in L2 learning only when combined with EI.

Some benefits of incidental learning instruction with EI also could be observed in Marsden (2006). Marsden (2006) also investigated the impact of PI on enriched input, but on comprehension and production of L2 French verb inflections. The enriched input that the study used was similar to input flood. The study included two experiments, one experiment in each of two secondary schools. The participants in experiment 1 had lower L2 proficiency relative to the participants in experiment 2. In each experiment, the participants were assigned to either PI or enriched input groups, and both groups received EI prior to the training activities. The results of experiment 1 confirmed the superiority of PI because the PI group showed significant improvement in both comprehension and production tests while the enriched input group did not

show any gains. On the other hand, in experiment 2, the participants who received enriched input made identical gains with the PI group in oral and written production tests, but they did not gain in comprehension. Marsden (2006) suggested that enriched input was generally less effective than PI in learning French verb inflections, and learners who had higher L2 proficiency were more likely to gain from enriched input than those with lower proficiency.

In addition, the influence of *enhancement* in improving the effects of input flood seems to be unclear. Lee and Huang (2008)'s meta-analysis of input enhancement on L2 learning found that overall the learners who received enhanced input outperformed those who received input flood only with a very small effect size (*d*=.22), a very small to neligible effect (as between .40 and .70 in L2 research is deemed to be small by Plonsky & Oswald's 2014 meta-analysis of effect sizes in the field). In addition, Issa and Morgan-Short (2019) investigated the effects of input enhancement on attentional allocation and L2 learning. It was found that although input enhancement triggered the learners' attention to the target form, it did not facilitate learning.

We have seen that the extent to which input flood (and similar input-based instruction like input enhancement) can benefit learning is far from conclusive. Some studies showed that learners could learn from input flood alone (Hernández, 2011); some showed that input flood could be effective when it follows EI (Henshaw, 2012; Hernández, 2008) or when it is combined with a 'spot the form'/noticing condition as in Kasprowicz and Marsden (2018); and some revealed that input flood could not benefit learning regardless whether or not it is combined with EI (Marsden, 2006; Marsden & Chen, 2011). One possible reason that could account for the inconsistency might be the different target forms. Previous studies that compared the magnitude of effects of referential style input activities with those of an input flood style of activities mainly focused on learning morphosyntax; none of them have examined syntax. To address this gap, the current study will use an input flood style of activity as comparative instruction in which one group of participants will receive input flood training on relative clauses. It is emphasised that the main purpose of involving input

flood group is to evaluate whether learners can gain from simple exposure to the target forms. Thus, no explicit information will be given to the participants who will receive input flood training, and no enhancement techniques will be used to emphasize the target structure.

2.3 Rationale and research questions

2.3.1 Rationale

The current study will address the extent to which teaching parsing strategies can benefit the L2 learning of English relative clauses. This research aims to address gaps in research to date as follows.

First, the previous studies have demonstrated that English SRC is easier than ORC, and the animacy of the head noun might affect the difficulty of ORC in offline comprehension but not in online processing. However, to the best of our knowledge, no study has investigated whether native English speakers and L2 learners could use the part of speech of the disambiguating linguistic cue, the word after the relative pronoun, to process relative clauses (if it is a verb straight after relative pronoun, the sentence will contain an SRC, e.g. *The cat that chases the dog is big*; if it is a noun straight after relative pronoun, the sentence will contain an ORC, e.g. *The cat that the dog chases is big*). The current study attempts to investigate whether native speakers and L2 learners are sensitive to this cue during online aural and reading comprehension of relative clauses and whether it will be affected by the animacy of the head noun.

Second, it is debatable whether explicit training can affect online processing. So far, few published studies have investigated the effect of explicit training on online processing (Andringa & Curcic, 2015; Dracos & Henry, 2021; Hopp, 2016; McManus & Marsden, 2017; VanPatten & Smith, 2019; Wong & Ito, 2018), and the studies generated inconsistent results. Some studies suggested that the nature of online processing was not changed by explicit training (Andringa & Curcic, 2015; Dracos & Henry, 2021), while the others showed evidence that explicit training could alter online

processing to some extent (Hopp, 2016; McManus & Marsden, 2017; VanPatten & Smith, 2019; Wong & Ito, 2018). For example, Wong and Ito (2018) produced nuanced findings suggesting that although the training did change the pattern of online processing, the learners could not use the cues that they learned in training to process sentences predictively. In addition, the previous studies related to this issue focused on teaching morphology processing, and no study has investigated the effects of explicit training on the learning and processing of syntax. This study will address the gap by investigating the extent to which explicit training could affect L2 relative clauses online processing.

Third, the instruction that will be used in the current study originates from the referential activity component within PI, but it will be slightly different from PI. PI aims to push the learners away from inefficient processing strategies by training them to make the form-meaning connection, so grammatical features used in PI studies to date always have referential meaning. However, the language cue that the current study will include does not have referential meaning, and it belongs to syntax. In addition, the previous research into teaching predictive processing (Andringa & Curcic, 2015, Hopp, 2016) addressed the issue of whether the learners could use a morphological cue to predict the upcoming language after training (e.g. gender agreement). No study has investigated whether learners can be trained to use part of speech (noun versus verb) to predict the meaning of the sentence. The current study will test whether the learners can be trained to use these syntactic cues to interpret the sentence meaning (by assigning roles to the first and second noun) during real-time processing.

Fourth, the previous studies that examined the effects of explicit teaching on processing itself have some methodological limitations. Some researchers used trials to criterion (e.g. Fernández, 2008; Henry, Culman & VanPatten, 2009; VanPatten & Borst, 2012) to measure what they referred to as online comprehension during training. However, this measure lacked external validation and the criterion that determined whether learners can have correct processing was arbitrary (Fernández, 2008). Keating and Jegerski (2015) pointed out three online measures: SPR, eye-tracking and ERPs. To

date, the previous studies that addressed this issue only included *either* SPR *or* eye-tracking test. Hence, within one study, only online aural processing *or* reading processing has been examined. The current study will involve both SPR and visual-world eye-tracking tests. Eye-tracking and SPR tests will tap into a different type of knowledge: the eye-tracking test will be based on auditory input, while the SPR tests measure online reading comprehension. Participants may show some differences in these two types of knowledge. Also, the eye-tracking tests will measure whether participants can process relative clauses using the syntactic cue, while the SPR tests focus on sensitivity to the cue when the sentence and the picture mismatch. It is possible that the participants may have some kind of sensitivity towards the cue (observable in eye-movements) but lack the ability to use the cue to detect anomalies with a picture when processing sentences, or vice versa. Therefore, involving both online measures in a study might be able to provide a more comprehensive picture in investigating the training effects of a teaching method.

Finally, the instruction that will be employed in the current study is different from those used in previous PI studies. The previous studies utilised morphological cues to train learners to understand their referential meanings/functions, while the current study will teach a syntactic cue which does not have referential meaning. In addition, the training items provided to the participants were stopped after the syntactic cue; thus, the participants were forced to use the syntactic cue to assign roles in the sentence. Because of this innovative teaching method, the training effects of offline comprehension and production will also be evaluated.

The current study will be in a pre-, post-, and delayed post-test design, and all the L2 learners will take part in the three test phases with same time intervals. Participants will be randomly assigned into either parsing group, input flood group or test-only group. The parsing group will be provided with EI and practice about parsing strategies of relative clauses. The input flood group will receive an equal number of practice items as the parsing group, but they will not receive EI, and the training will expose them the target structures but not orient their attention to it. The test-only group will

not receive any training, and they will only take the tests.

2.3.2 Research questions

The current study seeks to address the following research questions:

Main RQ1: Which type of relative clause (SRC vs. ORC) is more difficult, and what is the role of animacy of the noun in the main clause?

1) in offline comprehension?

2) in online comprehension measured by self-paced reading?

3) in online comprehension measured by eye-tracking?

4) in oral production?

5) in a metalinguistic knowledge test?

Predictions: Based on previous studies into processing and producing relative clauses, in general, the ORC is expected to be more difficult to process and produce than SRC for both native speakers and L2 learners.

ORCs with the inanimate head are predicted to be easier than ORCs with animate heads.

The performance of NSs and L2 learners may vary among the five outcome measures. The NSs are expected to score at ceiling in offline comprehension. However, they might have difficulty in producing ORCs and a lack of metalinguistic knowledge for both SRC and ORC. In online measures, the NSs are expected to be more sensitive to the language cues and be more likely to use the cue to predict the upcoming information in SRCs compared to ORCs.

Before receiving the training, the L2 learners are predicted to have higher accuracy scores for SRC than ORC in offline comprehension and production, and might not have metalinguistic knowledge for either type. Because L2 learners are likely to have a reduced ability to generate expectations relative to native speakers and/or a reduced sensitivity to processing syntax, the L2 learners might not be sensitive to the syntactic cues as measured by SPR and might not be able to use the part of speech cue to parse SRC and ORC at pre-test.

Main RQ2: To what extent can teaching parsing strategies, with explicit information and practice, exposure alone or no exposure (tests alone), develop the learning of relative clauses?

To what extent are effects observable?

1) in offline comprehension?

2) in online processing during self-paced reading?

3) in online comprehension as measured by predictive eye movements, and if so at what point in the sentence are effects observable?

4) in oral production?

5) in the metalinguistic knowledge test?

***Prediction*:** Based on the findings of PI studies, the participants who will receive training in parsing strategies are expected to gain in offline comprehension, oral production and metalinguistic knowledge. The participants in the current study are upper-intermediate L2 learners who have existing knowledge about the structures. Based on the findings of Hopp (2016), the parsing group is predicted to improve in online measures in terms of being more able than the other groups to show sensitivity to a mismatch between a relative clause and a picture stimulus, and more likely to use the part of speech cue predictively to make anticipatory looks in a visual world paradigm. The participants of the input flood group are expected to have a limited improvement in offline comprehension and production, perhaps due to awareness raising during the tests and/or training, but not in online comprehension and metalinguistic knowledge. The test-only group is expected to have no gains over time or some negligible to small gains due to taking the test batteries multiple times.

Chapter 3 Methodology and methods

This chapter illustrates the methods used in the current study and discusses the methodological considerations. Ethical considerations, participants, design of the study, outcome measures, training methods and the findings of the pilot study are discussed in sequence.

3.1 Ethical considerations

Several ethical issues were considered before data collection. This was considered to be a low-risk project. All the participants involved in the study were over 16 years old, and they did not belong to vulnerable or high-risk groups. They were not asked for opinions about sensitive or potentially distressing topics. The following steps were taken to ensure adherence to ethical considerations.

First, the participants were invited to voluntarily take part in the study. The participants were recruited from the University of York, and they were invited to the research through an email sent by their department administrators or through an advertisement that the researcher posted on the social media software, Wechat. The participants were informed that the research was entirely separate from the University courses, and their participation was optional. The participants were free to withdraw their participation during the experiment and within six weeks after the final session. After that time, they were not allowed to withdraw the data because their data would have been anonymised for analysis.

Second, anonymity and confidentiality of the participants' information were protected. The participants were assigned numerical identifiers that were used in the experiment to ensure anonymity. The links of the names with the numerical identifiers were kept in a sheet and stored separately and securely. The personal information of the participants will not be shared openly.

Third, the digital-based data were stored on a password-protected computer, and the paper-based data were stored in a locked drawer. Only the researcher could access the raw data, and the researcher's supervisor had the right to view them. The data

were anonymised, and only the final data sets will be shared openly after thesis submission.

Fourth, there were two potential benefits to participants in the study: the L2 learners could learn English during the study, and all the participants were rewarded some money at the end of the final session. Native speakers who took part in one test session (which lasted for around one hour) received £6; L2 learners who took part in three test sessions (which lasted for around three hours) received £15 and the training materials after the final session; L2 learners who took part in three test sessions and two training sessions (which lasted for four hours) received £20 after the final session.

The participants were required to a) read an information page that included the ethical considerations listed above, the purpose of the study, and the requirement for participation (i.e. first language should be English or Chinese; Chinese-speaking participants should have IELTS score between 6.0 and 7.0), and b) sign a consent form (see Appendix 1) to confirm that they had read and understand the information. The consent form had been approved by the Ethics Committee of the Department of Education at the University of York.

3.2 Participants

79 Chinese-speaking L2 learners of English and 21 native English speakers were involved in the study. All the participants reported that they had the good hearing, normal or corrected-to-normal vision, and did not have dyslexia. Each participant filled in a background information questionnaire (see Appendix 2) at the beginning of the first session. The information collected from the questionnaire about native speakers and L2 learners are provided here.

3.2.1 Native English speakers

The 21 native English speakers were 17 female and 4 male. Their age ranged from 18 to 43 with a mean of 25.77 (SD = 7.03). All of them were students of the University of York. Six were undergraduate students, 11 of them studied at master level, and 4 PhD students.

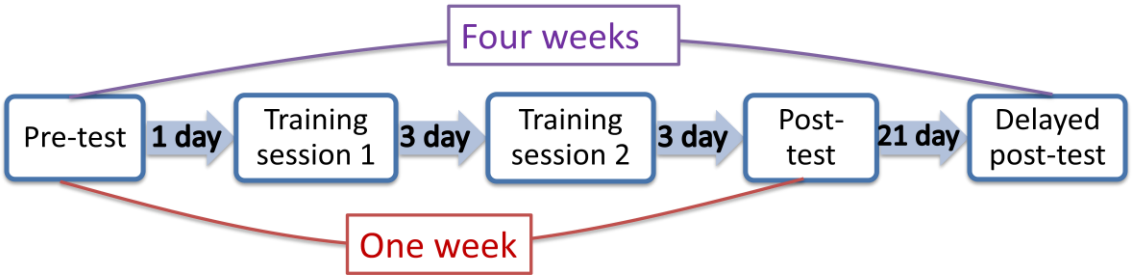
3.2.2. Chinese-speaking L2 learners

79 L2 learners were involved in the test, and they were randomly assigned into two training groups - the parsing group (n = 27) and the input flood group (n = 26) - and the test-only group (n = 26). Among 79 participants, 68 were females, 10 were males, and one did not provide gender information. Their average age was 23.95 (SD = 3.08, Min = 17, Max = 35). 4 of them were undergraduate students, 70 were master students, and four were doing their PhD. They had stayed in the UK for an average of 6.40 months (SD = 7.35) when they took part in the study. 14 of them reported they had been to an English-speaking country for one month or more before coming to the UK (mean = 6.38 months, SD = 5.73, Min = 1, Max = 20). All participants reported their highest overall IELTS scores were between 6.0 and 7.0, with an average of 6.46 (SD = 0.42). In addition, the participants were asked to rate their confidence in English grammar from 1 to 5. "1" indicates "not at all proficient" while "5" indicates "very proficient", and the mean of self-rated proficiency was 3.09 (SD = 0.69, Min = 2.00, Max = 4.00).

3.3 Design of the study

The current study included three groups of L2 participants (parsing group, n = 27; input flood group, n = 26; test-only group, n = 26) and a group of native English speakers (n = 20). The native English speakers only took part in one test phase, and the three L2 groups attended three test phases (pre-test, post-test, three-week delayed post-test). The two training groups, parsing and input flood groups, received two approximate 35 or 30-minute training sessions between the pre-test and the post-test. The first training session took place on the day following the pre-test, and the interval between the first and the second training session was three days for all the participants. The post-test was held three days after the second training session, and the delayed post-test was administrated 21 days later. The test-only group only took part in the three test phases, and intervals between the pre-, post-, and delayed post-test were the same as the training. They took the post-test seven days after the pre-test and took the delayed post-test 21 days after the post-test (see figure 3.1).

Figure 3. 1 Experimental design of the study



3.3.1 Stimuli

The training activities and tests included a great number of pictures and aural stimuli. Every picture in the study had a size of 400*400 pixels. Each picture included an agent and a patient involved in an event or an action. The agent and the patient either could be two animated or two inanimate nouns, or one animate and one inanimate noun. The pictures were created by the researcher or edited from free online images (some online images came from a picture database MultiPic (www.bcbl.eu/databases/multipic). The aural stimuli were recorded by three female British accented native English speakers. All the tests stimuli were recorded by one speaker, and the stimuli for training activities were recorded by the other two speakers.

The relative clauses used in the study were all in present simple tense and active voice, and the clauses were centre-embedded in the sentence (e.g. The cat that chases the dog is big). All the relative clauses were introduced by the relative pronoun "*that*". This level of homogeneity across the stimuli was necessary so that to avoid introducing confounding variables (such as additional syntactic complexity) for which we did not have sufficient time in the study to manipulate or control.

3.4 Outcome measures

The current study used a battery of measures to tap into the knowledge and the use of relative clauses. Eye-tracking and self-paced reading tests were used to measure online comprehension, while offline comprehension was measured by the aural sentence-picture matching test. In addition, oral production and metalinguistic knowledge were measured.

Each outcome measure was designed in four versions because, as Conklin,

Pellicer-Sánchez, and Carrol (2018) state, the number of versions of tests (often called 'lists' in psychology) should equal the number of different 'linguistic conditions'. In the current study, there were four linguistic conditions (subject relative clause with animate head (SRC-A), subject relative clause with inanimate head (SRC-I), direct object relative clause with animate head (ORC-A), and direct object relative clause with inanimate head (ORC-I), as explained in the Literature Review); and therefore, four versions of outcome measures were developed. In the outcome measures, each pair of pictures could create a group of four different stimuli, which are two SRCs, and two ORCs. Within one group of stimuli, the verb type and the number of words were kept constant. In order to balance the degree of difficulty of each version of the test, each of the four stimuli based on one pair of pictures were allocated to one of the four different versions.

In this study, there were three test phases (pre-, post-, and delayed post-test). The test versions were counter-balanced across each test phase, within condition. That is, in each test phase, all four outcome measure versions were administered to the participants. The participants in each Treatment Condition (parsing strategy group n=27, input flood group n=26, and test only group n=26) were randomly divided into 4 sub-groups, each of 6 or 7 participants; each sub-group completed a different version at each test phase (see Table 3-1).

Table 3. 1 Illustration of the allocation of test versions in pre-, post-, and delayed post-test, in each treatment Condition.

Treatment Condition A	Pre-test	Post-test	Delayed post-test
Sub-group 1 (*n=6 or 7)*	Version 1	Version 2	Version 3
Sub-group 2 (*n=6 or 7)*	Version 2	Version 3	Version 4
Sub-group 3 (*n=6 or 7)*	Version 3	Version 4	Version 1
Sub-group 4 (*n=6 or 7)*	Version 4	Version 1	Version 2

For each measure, the methodological considerations will be discussed initially, with a view to justifying the choice of measures used in the current study. Then, the

detailed design of each measure will be laid out. The two online measures, visual-world eye-tracking tests and SPR, will be introduced first, followed by offline comprehension production and metalinguistic knowledge measures

3.4.1 Online measures

Online measures that provide information about how participants process language in real time are favoured by an increasing numbers of second language (L2) researchers. The most commonly used measures in recently published studies are SPR, eye-tracking and event-related brain potentials (ERP) tests (Keating & Jegerski, 2015). In this section, eye-tracking and SPR measures will be discussed in detail.

3.4.1.1 Methodological considerations

Eye-tracking test

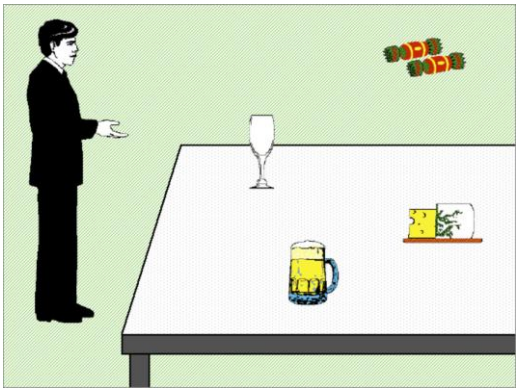
"Eye-tracking is a technology that measures fixations, saccades and regressions in response to visual input, while an eye-tracker is the device that does this" (Conklin, Pellicer-Sánchez, & Carrol, 2018, p.2). Eye-tracking technology has been widely used in L2 studies, and it has two major advantages. First, eye-tracking experiments reduce the influence of working memory since it does not require participants to recall something. Second, this technique allows participants to read or listen to relatively natural stimuli (without the need to introduce ungrammaticalities, though sometimes this can be done), and it can measure learners' processing of sentences without imposing an additional task (Conklin, Pellicer-Sánchez, & Carrol, 2018). Therefore, eye-tracking tests have been adopted by many researchers in recent years (Huettig, Rommers & Meyer, 2011; Roberts & Siyanova-Chanturia, 2013; Winke, Godfroid & Gass, 2013).

In general, there are two types of eye-tracking experiments, text-based where participants read the written input, and the visual-world paradigm where participants listen to input and are presented with one or more pictures or videos related to the aural input. Text-based eye-tracking tests measure online reading comprehension, and there are plenty of researchers using this type of test to examine anomaly detection (Keating, 2009), ambiguity resolution (Witzel, Witzel, & Nicol, 2012) and syntactic dependency formation (Cummings, Batterham, Felser & Clahsen, 2010). However, the

visual-world paradigm which measures spoken language comprehension attracts researchers' attention in the recent two decades, and the current section will focus on this type of eye-tracking test.

The visual-world paradigm eye-tracking tests are based on the hypothesis that there is a referential relation between aural and visual stimuli (Cooper, 1974). In other words, people tend to look at the elements that they are hearing; thus, the eye movements can reflect the "linguistic representations" that the people use while hearing the aural stimuli (Godfroid, 2019, p.89). Based on the theory that there is a link between language processing and eye movement, researchers have used the visual-world paradigm to investigate word recognition (Marian & Spivey, 2003), anticipatory eye movements in sentence processing (Altmann & Kamide, 1999, 2007; Andringa & Curcic, 2015; Curcic, Andringa & Kuiken, 2019; Dijkgraaf, Hartsuiker, & Duyck, 2017; Hopp & Lemmerth, 2018; Mitsugi & MacWhinney, 2016; etc.), and referential processing which refers to the ability of establishing reference using the information that the processors hear (Cunnings et al., 2017; Kim et al., 2015; Sekerina & Sauermann, 2015; Sekerina & Trueswell, 2011). Among those empirical studies, Altmann and Kamide (2007) demonstrated that native speakers could use morphosyntactic cues of tense to predict the meaning of upcoming information. In the study, the participants heard sentences like "*The man will drink the beer*" or "*The man has drunk the wine*" (p. 505) while looking at a scene containing a man, a full glass of beer, an empty wine glass and some distractors (see Figure 3.2).

Figure 3. 2 Scene used in Altmann & Kamide, 2007, p.505



It was found that the participants had more looks towards the empty wine glass than

the full glass of beer when they heard *has drunk* and vice versa. This study suggested that the anticipatory eye movements could reveal the linkage between the language input and objects that had (or had not yet) been named.

The current study investigated using syntactic cue to interpret the meaning of the sentence, to be specific, the grammatical roles assignment of nouns in relative clauses. The word order (part of speech of the word) after the relative pronoun could be used as a syntactic cue to determine the role of nouns in a relative clause. Participants were expected to distinguish the agent and patient in a relative clause, and it was expected that sensitivity to this cue would be observable through visual-world eye-tracking tests. In the tests, for each item, participants saw two reversible pictures depicting an action while hearing a sentence about one of the pictures. The proportions of their eye fixations on the target picture versus the distractor at various points after the syntactic cues were recorded and analysed.

Self-paced reading test (SPR)

Self-paced reading (SPR) is another widely used online comprehension measure, which requires participants to press a key at their own pace to read sentences broken into words or segments (for a review of the method, see Marsden, Thompson & Plonsky, 2018). It can be used to examine similar types of questions as eye-tracking, for example, ambiguity resolution (Dussias & Cramer Scaltz, 2008; Roberts & Felser, 2011), anomaly detection (McManus & Marsden, 2017, 2018; Roberts & Liszka, 2013), syntactic dependency formation (Williams, 2006; William, Mobus, & Kim, 2001). In SPR studies, the reaction times (RTs) of each keypress are recorded and analysed. The speed of the reactions is interpreted according to the specific question being examined. For example, in some studies, a faster RT for whole sentences might indicate higher proficiency if sentential comprehension is the purpose of the study.

However, participants can slow down for some reasons, revealing sensitivity to ungrammaticality, to unexpected words, or to ambiguity and/or its resolution. Native speakers and advanced proficient L2 users tend to spend more time processing ungrammatical or unexpected information than low proficiency language learners

(Roberts, 2016). Therefore, in SPR studies, the speed taken to process critical regions is also analysed. For example, McManus and Marsden (2017) conducted an SPR experiments by using sentence-picture match or mismatch task. They defined it as a type of anomaly detection as it contained context-stimulus anomalies. It is true that the term 'anomaly detection' is not being used in its normal sense (detecting a linguistic anomaly in verbal input); but the task is sensitive to whether or not the participant 'detects an anomaly' between the verbal stimuli and the image.For each item of the test, the participants saw a picture and a sentence, and the sentence either matched or mismatched the sentence. The RTs on critical words in matched and mismatched items were analysed. The analysis was used to investigate sensitivity to 'violations' (mismatches with the meaning depicted by the pictures) in a type of French morphosyntax. If participants were sensitive to the violation of the morphosyntax, slower RTs would be found on critical words of mismatched items relative to matched items.

In the current study, the aim of using the SPR test was to examine whether the participants were sensitive to the syntactic cue (i.e., the word order after the relative pronoun) instead of merely investigating which type of relative clauses was more difficult to process. Thus, anomalies were considered to be included in the critical regions. However, since the only difference between SRC and ORC was the word order after the relative pronoun (i.e., in SRC *The cat that* ***chases the dog*** *is big*, after '*that'* is a verb '*chases';* ORC *The cat that* ***the dog chases*** *is big,* after '*that'* is a noun '*the dog'*), the ungrammaticality was not applicable under this condition. The test was designed following McManus and Marsden (2017), and sentence-picture anomalies were included.

For each item in the tests, participants saw one picture and read a sentence at their own pace. The sentence might match or mismatch the picture. If the participants could process the linguistic cue appropriately, they would slow down when the words they read mismatch the meaning conveyed by the visual stimuli. L2 learners might also change their processing behaviour after the training sessions (as in McManus &

Marsden, 2017), for example, showing increased sensitivity to the morphosyntax they had been trained on.

3.4.1.2 Design of the tests

Both the eye-tracking and the SPR test each had 120 items. They were 40 critical items (10 SRC-A, 10 SRC-I, 10 ORC-A, 10 ORC-I) and 80 non-critical items (32 distractors and 48 fillers). 30 of these 80 non-critical items were followed by comprehension questions. In both tests, the items were displayed randomly – in a different order for different participants

The sentences used as visual or aural stimuli contained either eight or nine words, but the number of words up to and including the critical region (the noun and the verb following 'that') were six for all the critical items. For critical items, the sentences created by one (in SPR test) or a pair of (in eye-tracking test) had exactly the same words, except the word orders of the critical region were different.

Balancing critical and non-critical items

In an eye-tracking or SPR test, in case of data loss, the ratio for critical items and non-critical items is recommended to be more than 1:1 (Keating & Jegerski, 2015). In other words, in an eye-tracking or an SPR test, it was desirable to have at least 50% of items as noun critical items, and ideally, this proportion can be up to 75%.

In addition, because psycholinguistic measures like eye-tracking and SPR tests usually lose 8-15% of data during collection, in order to collect sufficient data to conduct analysis, 8-12 items were recommended for each linguistic condition (Keating & Jegerski, 2015). In this study, there were four linguistic conditions:

1) subject relative clauses with animate heads (SRC-A),

2) subject relative clauses with inanimate heads (SRC-I),

3) direct object relative clauses with animate heads (ORC-A),

4) direct object relative clauses with inanimate heads (ORC-I).

Each condition had ten items, so there were 40 critical items in total. Ideally, the number for non-critical items would be 120. However, in the pilot study, we noticed participants easily got fatigued, so the number of the non-critical items was reduced to

80. Therefore, the ratio for critical and non-critical items was 1:2. Within the 80 non-critical items, four of them were used as practice items which were displayed before the real test.

Comprehension questions following some items

In order to keep participants engaged in reading or hearing the stimuli, comprehension questions (CQs) followed some of the items. However, repetitive CQs on specific regions would cause longer reaction times (RTs) on those regions (as suggested by Marsden, Thompson, Plonsky, 2018), so in the current study, CQs only appeared after *non-critical* items to reduce the chance of CQs affecting RTs. With a view to reducing the chances that participants would become aware of the specific target features or the purpose of the critical stimuli, the comprehension questions asked participants to respond to their comprehension of beginning, middle and ending parts of the distractor and the filler sentences. The frequency of comprehension questions has been recommended to be one for every three or four stimuli (Keating and Jegerski, 2015). In the current study, 30 non-critical items, constituting 25% of the total number of items, was followed by comprehension questions; thus, the ratio for comprehension questions after distractors and fillers was 2:3. All the CQs were *yes/no* questions, and the answers to 50% of questions were *yes*, while the others were *no*. In the eye-tracking tests, the participants were instructed to answer the questions according to what they heard and saw, while in SPR tests, the participants were required to respond to the questions based on what they had read. The participants would not receive any feedback to tell them whether the response was correct or not.

Design and administration of the eye-tracking test

Design of eye-tracking test

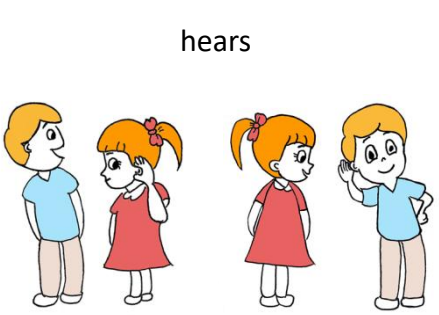
For each item of the eye-tracking test, the participants were asked to look at two pictures while hearing one sentence which matched only one of the pictures. There were 40 critical items and 80 non-critical items. The regions of the two pictures in each visual scene were defined as the regions of interest (ROI).

Critical items

Based on one pair of pictures, two SRCs and two ORCs were created and randomly distributed in four versions of tests. The pictures used in critical items were reversible. The participants' attention was expected to focus on the syntax in order to interpret which picture was being described. The critical regions were the three words after the relative pronoun "that". The participants were expected to have a fixation on the target picture when they heard the onset of the first critical word.

The examples for critical items are shown in figure 3.3.

Figure 3. 3 An example of a critical item in the eye-tracking test



Aural stimulus:

1. The boy that hears the girl has blond hair. (SRC-A)
2. The boy that the girl hears has blond hair. (ORC-A)
3. The girl that hears the boy has blond hair. (SRC-A)
4. The girl that the boy hears has blond hair. (ORC-A)

(As described above, each of these stimuli were distributed across different versions of the test, so that one individual participant would see this picture with only one of these aural stimuli in any one test phase).

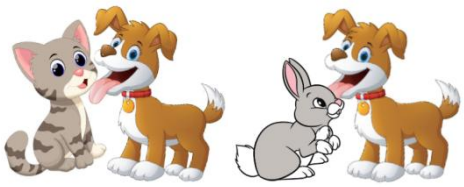
Non-critical items

Non-critical items could be divided into two types, distractors, which had similar characteristics to critical items, and fillers, which were not related to critical items (Keating & Jegerski, 2015). The pictures used in non-critical items were non-reversible. This study had 32 distractors (approximately 27% of the total items) and 48 fillers (40% of the total items). For distractors, SRCs and ORCs were used as distractors as well as

serving as critical items. That is, when the relative clauses were used as distractors, participants were able to choose the matched pictures based on *lexical* cues (i.e., the semantics of the nouns), whereas, for critical items, they would have to use syntactic cues to make the decision. The four types of relative clauses each served as eight distractor items. The examples of the distractors are shown in figure 3.4 (more examples see Appendix7, and the list of critical items see Appendix 12).

Figure 3. 4 An example of a distractor in the eye-tracking test

licks



Aural stimulus, one of either:

1. The dog that licks the cat is brown. (SRC-A)

or 2. The cat that the dog licks is grey. (ORC-A)

CQ: Does the dog kiss the cat? *No*

or 3. The dog that licks the rabbit is brown. (SRC-A)

or 4. The rabbit that the dog licks is grey. (ORC-A)

CQ: Is the rabbit grey? *Yes*

For fillers, there were 4 types of sentences: present simple, present simple + prepositions, present simple + passive voice and present progressive. Each type had 12 items. See figure 3.5.

Figure 3. 5 The examples of eye-tracking fillers

washing

Aural stimulus, one of either:

1. The boy is washing his hands in cold water. (Present progressive)

or 2. The boy is washing his hands very carefully. (Present progressive)

CQ: Is the boy washing the clothes? *No*

or 3. The girl is washing a lot of dishes. (Present progressive)

or 4. The girl is washing dishes in warm water. (Present progressive)

Administration of eye-tracking test

The tests were administrated with an Eyelink 1000 plus eye-tracker (SR). The stimuli of the test were displayed on a desktop computer, and the size of the screen of the monitor was 48*27 cm. In the test, each participant sat on a chair and fitted themselves on a chin and forehead rest. The distance from the person's eye to the screen was fixed at around 80 cm. The positions of target and distractor pictures were balanced. To be specific, in the 120 items, 60 items had the target pictures presented on the left of the screen, and 60 items had the target pictures on the right.

For each item in the eye-tracking test, the participants initially saw a fixation dot, and they saw two pictures and they saw a written verb which described the action of both the two pictures (e.g., 'wash'). The pictures were presented at the right centre and the left centre of the screen. 2000 ms were allowed for participants to observe the pictures before the aural stimulus start. After previewing the pictures, an audio recorded sentence which matched one of the pictures was played. Participants were

not given any explicit instruction to look at anything in particular while listening (i.e., they were left to listen and look as they wished, without direction to orient their attention to the picture they were hearing about). There were comprehension questions after some items. The test proceeded directly to the next item if there was no comprehension question for the item. One eye-tracking test, including the time for calibration, took around 25 minutes.

Design and administration of SPR test

Design of SPR test

The SPR test was an anomaly detection test, and half of the items had an anomaly (a mismatch) between the meaning of the sentence and the meaning of the picture. For each item, there was one picture and one sentence, and the sentence might match or mismatch the picture. The verb that described the action in the picture was provided alongside the picture, in writing. The ratio of the matched items and mismatched items was 1:1. One picture generated four sentences, and they were randomly distributed in the four versions of the tests. The participants were asked to view the picture and read the sentence word by word at their own pace controlled by pressing a key. They were instructed (in English) to try to understand the meaning of the sentences as fully as possible instead of reading as fast as possible.

Critical items

In each version of SPR tests, 40 critical items were designed. One picture generated two SRCs and two ORCs, and one SRC/ORC matched the picture while another one mismatched the picture. For critical items, the anomaly occurred when the sentence and the picture described the same action and yet, the agent and the patient were opposite. The three words after the relative pronoun "*that*" were regarded as critical region 1, 2 and 3. If participants were sensitive to the linguistic cue, they were expected to have slower RTs when they read the critical regions in mismatched items compared to reading those regions in matched items.

The examples for critical items are shown in figure 3.6 (more examples see appendix 8, and the list of critical items see Appendix 13).

Figure 3. 6 An example of a critical item in the SPR test

follows

Sentences for critical items, one of:

1) The tiger that follows the lion is relaxed. (Match) (SRC-A)

Critical region: 1 2 3

or 2) The tiger that the lion follows is relaxed. (Mismatch) (ORC-A)

Critical region: 1 2 3

or 3) The lion that follows the tiger is yellow. (Mismatch) (SRC-A)

Critical region: 1 2 3

or 4) The lion that the tiger follows is yellow. (Match) (ORC-A)

Critical region: 1 2 3

Non-critical items

The design of non-critical items for SPRs was similar to that of the eye-tracking test. The distractors (k = 32) and the fillers (k = 48) had the same features as those of the eye-tracking test. SRCs and ORCs were used as distractors as well as critical items. However, the relative clauses as distractors mismatched the pictures based on lexical differences (see figure 3.7, in the picture, the girl draws the picture, but the mismatched sentence expresses the boy draws the picture).

Figure 3. 7 An example of a distractor in the SPR test

draws

Example distractor sentences, one of:

1) The girl that draws the picture is cute. (Match) (SRC-A)

or 2) The picture that the girl draws is colourful. (Match) (ORC-I)

Comprehension question: Does a boy draw the picture? *Answer: No*

or 3) The boy that draws the picture is cute. (Mismatch) (SRC-A)

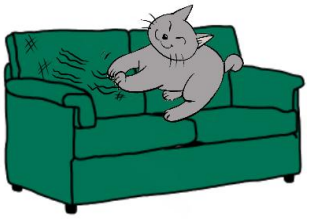
Comprehension question: Is there a picture? *Answer: Yes*

or 4) The picture that the boy draws is colourful. (Mismatch) (ORC-I)

For the filler items, 4 types of sentences (present simple, present simple + prepositions, present simple + passive voice and present progressive) used in the eye-tracking tests were also used in SPR tests. 12 items were created for each type of sentence (the example of fillers see figure 3.8).

Figure 3. 8 An example of a filler in the SPR test

scratches / scratched



Example filler sentences, one of:

1) The grey cat scratches a nice green sofa. (Match) (Present simple)

or 2) The green sofa is scratched by a cat. (Match) (Passive voice)

or 3) The brown dog scratches a nice green sofa. (Mismatch) (Present simple)

Comprehension question: Is the sofa green? *Answer: Yes*

or 4) The green sofa is scratched by a dog. (Mismatch) (Passive voice)

Administration of SPR test

The SPR tests were administrated on a laptop through the software OpenSesame (Mathôt, Schreij & Theeuwes, 2012). For each item, the participants first saw a fixation dot lasting for 745ms, and then a picture and a verb expressing the action in the picture were presented automatically at the centre of the screen. Before showing the first word, the participants were able to look at the picture and the verb for 1000 ms. When the last word of a sentence had been read, the fixation dot for the next item would show directly, if a comprehension question did not follow that item. For the items with comprehension questions, the test would proceed to the next item when the question had been responded to. One SPR test normally lasted for 20 to 25 minutes, varying by individuals' reading speeds.

3.4.1.3 Data cleaning and analysis

Data cleaning

After data collection, the data of the eye-tracking and SPR tests were cleaned and prepared for analysis. The data cleaning process was conducted in Microsoft Excel. First, the participants who might not have engaged in the test were removed from the analysis. The CQs after non-critical items were used to examine the engagement of the participants; the inclusion threshold is typically set between 70% and 80% (Godfroid, 2019). In the current study, the participants whose correct rates were below 75% were excluded from the data pool. Second, outliers were removed based on Standard Deviation (SD)-based boundaries and lower and upper time-based cut-off ranges. The means and SDs by participants and test-phase were calculated. Then, the fixation durations that exceeded or were below mean ± 2.5 SD were removed. This will be explained in detail in the following section.

Cleaning eye-tracking data

For the eye-tracking tests, one native speaker and one L2 learner scored lower than 75%

for the eye-tracking tests they attended to. In addition, the CQs data for one L2 learner's delayed post-test were missing. Thus, the whole dataset for two people and the data of one person's delayed post-test were excluded from the analysis.

In dealing with outliers, most visual-world eye-tracking studies did not report their standards of defining outliers. Altmann and Kamide (2007) reported that they removed the fixations of durations below 100ms. The way of dealing with outliers in the current study was partially in line with Altmann and Kamide. First, the fixations outside the ROI were removed. Then, the fixation durations that did not belong to the range of the mean ± 2.5 SD were removed. The aim of using the upper boundary was to get rid of the overly long fixations due to simply staring at the screen. In addition, it was noticed that mean-2.5 SDs were usually below zero. However, Inhoff and Radach (1998) suggested that fixations less than 50ms might not be underpinned by (or reflective of) cognitive processes. Thus, 50 ms was set as the lower-time boundary, and the fixation durations below 50 ms were also removed. In total, 6% of data were regarded as outliers and have been removed from the analysis.

Cleaning SPR data

For the SPR tests, the correct rates of comprehension questions of five participants in the parsing group (1 out of 27 at the pre-test, 1 out of 27 at the pre-test, 3 out of 27 at the delayed post-test), ten participants in the input flood group (2 out of 26 at the pre-test, 4 out of 26 at the post-test, 4 out of 26 at the delayed post-test), six participants of the test-only group (2 out of 26 at each of the pre-, post-, and delayed post-test), and one participant of the native group (1 out of 21) were lower than 75%. Those affected data were removed from the analysis.

In dealing with outliers, Nicklin and Plonsky (2020) conducted a meta-analysis and suggested that the lower time-based boundary was usually set around 150 ms in word-level self-paced reading, and the upper boundary could be set at up to 10000ms. Normally, the cutoff range of the upper boundary has been set between 2000 and 6000ms (Keating & Jegerski, 2015). In the current study, the RTs less than 150ms or greater than 3000ms and above or below 2.5 *SD*s of an individual's mean per word per

test phase were removed. 4.91% of the data in total were influenced, and 5.36% of the words in the critical items were removed.

Data analysis

For the eye-tracking tests, the fixation proportions were analysed through visualisation and inferential statistics. For the SPR tests, the RTs at the first, second, third critical words and the whole sentence was analysed through descriptive statistics, visualisation and inferential statistics.

For both tests, mixed-effects regression models were used to conduct inferential statistical analysis. They were carried out through the "lme4" package (Bates, Mächler, Bolker & Walker, 2015) in R (R Core Team, 2018). In recent years, this statistical tool has been widely used in second language research, and it was considered to be superior to the analysis of variance (ANOVA) (Godfroid, 2019). ANOVA analyses the data either by item or by subject, while mixed-effects analyses allow researchers to include all the variances, due to individual participants and items, within a model (Godfroid, 2019). Mixed-effects models analyse dependent variables by fixed effects and random effects (Linck & Cunnings, 2015). The fixed effects are the independent variables controlled by researchers, like training conditions; the random effects refer to the "independent variables that result from random sampling from a population", for example, the individual differences can be included as random effects (Godfroid, 2019, p. 277). The fixed effects usually include the interactions between all the controlled independent variables, and the random effects include random intercepts and may have random slopes. A random intercept explains the variance among subjects or items within one condition, while a random slope accounts for the variance among subjects or items in repeated measures (Godfroid, 2019). When an experiment has an entire within-group design, only random intercepts are needed. If the experiment has between-group variables, the random slopes are recommended to include as well as random intercepts (Linck & Cunnings, 2015). It is worth mentioning that the independent variable "*group"* normally cannot be included as a by-subject random slope. It is because one subject cannot be involved in different groups, so the performance of a subject cannot vary

among groups (Linck & Cunnings, 2015).

In selecting random effects, Barr, Levy, Scheepers & Tily (2013) suggested that in testing a confirmatory hypothesis, the maximal random effects structure should be adopted in order to take every possible variance into account. However, in practice, maximal random effects are usually too complex to converge. In this situation, the random effects are suggested to be trimmed down, starting by removing the interactions with random slopes. If the model still fails to converge, some random slopes could be removed step by step (Godfroid, 2019). In the current study, for the confirmatory hypothesis, the model started with the maximal random effects. In order to avoid overfitting of the model, the model was trimmed down until it only included random intercepts. During this process, the likelihood ratio tests (LRT) (conducted through "lmtest" package, Zeileis & Hothorn, 2002) and Akaike information criterion (AIC) (conducted through "AICcmodavg" package, Mazerolle, 2020) were used to select the best fitting model. For AIC, the model that had a smaller AIC was likely to fit the data better than the one with a larger AIC (Godfroid, 2019). For LRT, following the suggestion of Godfroid (2019), the significant level α of .20 was adopted. In the comparison of the two models, if the *p*-value was more than .20, the simpler model (i.e., the model with fewer random slopes) would be adopted; on the contrary, when the *p*-value was smaller than .20, the more complex model was likely to fit the data better. When the results of AIC and LRT conflicted, the model that was consistent with the results of effect sizes and descriptive results would be adopted. For exploratory hypothesis test, Linck and Cunnings (2015) put forward that the model could only include random intercepts, unless the random slopes could make the model fit the data significantly better than without them. Thus, in the current study, for the exploratory hypothesis, only the random intercepts were included.

However, in running mixed-effects models, dummy coding method was adopted, and R took the first alphabetical group to compare to as the baseline. This coding method followed Kim, Skalicky and Jung (2020) and Thompson-Lee (2021). It should be acknowledged that sum coding and Helmert coding methods would be more efficient

than the dummy coding as they would allow multiple comparisons to be investigated. However, due to capacity of the study, these two coding methods were not adopted in this thesis. In the current study, in order to compare the effects of each pair of the three groups – as that was of theoretical interest (were there differences between the two training conditions? Did each of the different training conditions differ from the test-only group?) –the models had to be run with more than one baseline. In analysing the difficulty differences of each type of relative clause (RQ1), models with the ORC-I, ORC-A, and SRC-I as the baselines were run respectively. For the analysis of teaching effect (RQ2), the test-only and the input flood group were used as the baseline groups, respectively in separate analyses.

In addition, to evaluate how much the variances have been explained, marginal R^2 is used to refer to the proportion of variance explained by fixed effects, and conditional R^2 to refer to the proportion of variance explained by fixed and random effects were provided. Plonsky and Ghanbar (2018) suggested that the $R^2 \leq .20$ and $R^2 \geq .50$ could be generally regarded as small and large respectively in the per cent of explained variance. R^2 was calculated through "MuMIn" package (Bartoń, 2020).

In reporting the inferential statistical results, the estimate *b* with 95% confidence interval (CI), standard error (SE), *t*-value and *p*-value of the fixed effects, as well as marginal and conditional R^2 of the model were reported.

Data analysis of eye-tracking test

In the eye-tracking tests, the fixation proportions from the onset of the first critical word were analysed through line plots and inferential statistical analysis. The plots were created using "ggplot2" package (Wickham, 2016), and the statistical analysis was conducted with "lme4" package (Bates, Mächler, Bolker & Walker, 2015) in R (R Core Team, 2018).

The line charts, each covering 3000ms depicting the fixation proportions of looking at target and distractor from the onset of the first critical word were created for each type of relative clauses. Then, a mixed-effects growth curve analysis was used to conduct the inferential statistical analysis. This method was designed to analyse the

time course data, and it includes 'time' as a predictor (Barr, 2008; Mirman, Dixon & Magnuson, 2008). Changes over time are not always linear, and they might be quadratic and cubic. Hence, choosing an appropriate time vector was important in using growth curve analysis. Barr (2008) recommended that researchers do not go beyond the third-order time term in choosing the time vector.

Because each datapoint of looking at either the target or the distractor within one time bin is provided as binary data, it was necessary to collapse several times bins to a larger time bin and calculate the empirical log-odds before running growth curve analysis (Godfroid, 2019). Following the steps of Barr (2008), five 10ms time bins were collected into a 50ms time bin. The empirical log-odds were calculated using the formula below via Microsoft excel.

$$\text{elog(looks)} = \ln\left(\frac{\text{looks} + 0.5}{\text{non-looks} + 0.5}\right)$$ (Godfroid, 2019, p.299)

In sum, in inferential statistical analysis of the eye-tracking test, the empirical log odds of fixation proportion were analysed as a dependent variable.

In answering the first research question, 'which type of relative clauses is more difficult in eye-tracking test', the fixed effects included the interaction between the relative clause type and time order predictor. The maximal random effects included by-subject and by-item intercepts and slopes of the interaction between type and time order. The models with maximal random effects were trimmed down step by step until they only included by-subject and by-item intercepts. The data of native speakers and L2 learners were analysed separately. During this process, LRT and AIC were used to select the best-fitting model.

In answering the second research question, 'to what extent are training effects observable in online comprehension measured by eye-tracking test', the fixed effects included the interaction between test phase, group and time order. As this study is the first one (to the best of our knowledge) to investigate the online effects of parsing strategies training on the learning of a syntactic structure. The approach was broadly

exploratory in that we did not have very specific directional predictions that were informed by previous research. Linck and Cunnings (2015) suggested that for the exploratory research, which had a multitude of fixed effects, it might be unpractical to include the maximal random effects. Thus, they suggested that under this condition, only the random slope which could significantly improve the model to fit the data is needed. In the current study, the fixed effects were rather complex (includes the interaction between group, test phase and time order), and the model with maximal random effects failed to converge. Triming down the model step by step to only include random intercepts would generate too many possible models (could up hundreds of models) to compare with each other. It was too difficult to decide which random slope was potentially necessary. Thus, the random effects only included by-subject and by-item random intercepts. In selecting time order vector, following Andringa and Curcic (2015), the primary model started with the first-order time vector. The second-order and the third-order time vectors were added step by step. The best-fitting model was selected using LRT and AIC.

Data analysis of SPR tests

In SPR tests, the RTs of the first, second, third critical words and the whole sentences were analysed. Before data analysis, it was found that many RTs of the first word of the sentences had extremely low (<150ms) or high (>2000ms) values (occupied 20% of total datapoint of the first word). Two possible reasons might explain it: a) the fixation dot (lasting for 745 ms) and the picture page (lasting for 1000 ms) were presented initially, and the sentence would be shown automatically. Some participants continuously pressed the keyboard during the showing of the fixation dots, which leads to the extremely fast RTs, and b) the participants might have noticed that the occurrence of the fixation dots would introduce a new sentence. They might prefer to take a very short break before start reading the new sentence, which leads to extremely slow RTs. Because the first word of every sentence was ‘the’, removing the RT data from this first word was unlikely to influence the results of the whole sentence analysis. Thus, RTs of the first word ‘the’ were excluded in analysing the RTs of the

whole sentence.

In order to control the differences in word length, the residual RTs were used in the analysis. The way of calculating the residuals followed Lee, Lu, and Garnsey (2013) using R studio (R Core Team, 2018). Before computing the residuals, the cleaned RTs were log-transformed and multiplied by 100 to avoid extremely small parameter estimates. Then, the residuals were calculated based on a regression model predicting the log-transformed RTs from word length based on every word of both the critical and non-critical sentences per person per test phase. The results of the visuals, effect sizes and inferential statistics were calculated based on the residual RTs. However, since the residual RTs were less intuitive than the raw RTs, the raw RTs are used to calculate the means and the SDs for each condition in the descriptive analysis.

For descriptive analyses, the means and SDs of raw RTs for matched and mismatched items, and the mean RTs for the mismatched items that had been subtracted from those of the matched items of each critical words and whole sentences were provided. Then, line charts were created to depict the residual RT differences between mismatched and matched items. In addition, within-group effect sizes (Cohen's *d* with 95% CI) were calculated through the "effectsize" package in R (Ben-Shachar, Lüdecke, Makowski, 2020). The benchmarks of reliable sensitivity to anomaly detection were adopted from Avery and Marsden (2019) (L2 learners: *d* = .19 [.09, .29]; Native speakers: *d* = .41 [.29, .54]).

For inferential statistical analysis, the residual RTs of the first, second, third and whole sentence were analysed separately.

For the first research question, 'which type of relative clauses is more difficult in the SPR test', each model included the fixed effects of the interaction between the type of relative clause and whether the sentence matched the picture or not. Following the recommendation of using the maximal random-effects structures, the models with the by-subject and by-item random slopes of the interaction the type and the match or mismatch were run initially. The models for native speakers and L2 learners were run separately, and the best-fitting model was selected by LRT and AIC tests.

For the second research question, 'to what extent are training effects observable in online comprehension measured by SPR test', the fixed effects included the interaction between the group, test phase, and whether the sentence-picture matches or not. The random effects only included by-subject and by-item intercepts (same reason with that of the eye-tracking test).

3.4.2 Offline measures

The three offline measures that tapped into offline comprehension, oral production and metalinguistic knowledge were used in the current study. In this section, the methodological considerations, test design and the administration, and data analysis of three offline measures will be discussed.

3.4.2.1 Methodological considerations

Offline comprehension: Aural sentence-picture matching test

Sentence-picture matching test has been used in previous studies, and sometimes it is also called picture selection test (Friedmann, Belletti, & Rizzi, 2009; Friedmann & Novogrodsky, 2004; Frizelle, Thompson, Duta & Bishop, 2019; Izumi, 2003; O'Grady, Lee, & Choo, 2003; Sanz, & Morgan-Short, 2004, etc.). In sentence-picture matching tests, participants are provided with a sentence (usually provided as aural input), and then they will be asked to choose the matched picture from a set of pictures. In some of the studies, two reversible pictures which contain the same two figures performing the same action are presented as visual stimuli. The only difference between the two pictures is the elements (e.g. people, animals, things) in pictures playing reversed roles in the action (Friedmann, Belletti, & Rizzi, 2009; Friedmann & Novogrodsky, 2004; Sanz, & Morgan-Short, 2004). For example, in Friedmann and Novogrodsky (2004), for each item, participants saw a pair of pictures (see figure 3.9) and heard a sentence in Hebrew whose meaning was "*This is the grandmother that the girl is kissing*" (p.670). The participants could only correctly choose the matched picture when they understood the meaning of the syntax. It was because except the roles of nouns were different, the other elements were exactly the same in the two pictures. Thus, the sentence-picture matching test can be regarded as an effective method in measuring

participants' offline comprehension. This measure had been adopted in the current study.

Figure 3. 9 A picture pair used in Friedmann & Novogrodsky, 2004

Friedmann & Novogrodsky, 2004 (p.671)

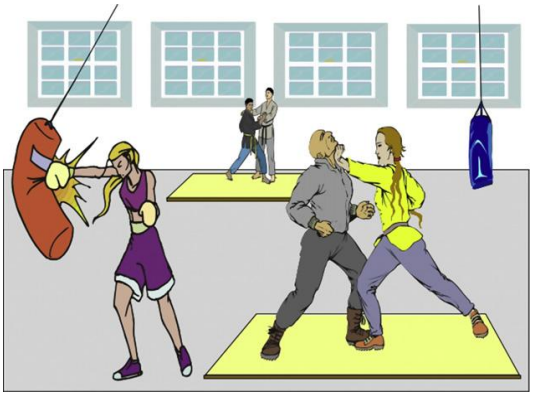
Oral production: Sentence description test

In previous studies, a wide range of techniques has been used to elicit the production of relative clauses. Gass (1982), Doughty (1991), and Izumi (2003) conducted written sentence combination tests with L2 learners to measure their production competence. In the three studies, participants were required to combine two simple sentences into a complex sentence containing a relative clause. In order to prompt participants to produce relative clauses instead of coordinate sentences or other type of sentences, the participants were instructed not to use any coordinating conjoiners like *and, but, because,* and so on (Izumi, 2003). However, although, to some extent, this type of test can successfully elicit relative clauses production, the way of eliciting tends to be less natural compared to presenting participants scenarios to describe.

In recent studies, pictures or videos are used in relative clauses production tests (Friedmann, Belletti & Rizzi, 2009; Gennari, Mirkovic, & MacDonald, 2012; Kim & O'Grady, 2016; Zukowski, 2009). Hamburger and Crain (1982) described the method of designing the production tests. To elicit relative clauses, two identical animals which had unique features were used, and the participants were asked to describe one of them. As simple sentences could not fully express the feature of the animal, relative clauses were likely to be produced. Following this idea, Gennari, Mirkovic & MacDonald (2012) conducted scene depicting tests to elicit relative clauses production. In each

scene of the test, there were two people who performed the same action. One actor did the action to another person, while the other actor did it to an object. Meanwhile, additional people and objects which served as distractors were also contained in the scene. Each person or the object in the scene had a unique feature, and the participants were required to answer a question about the agent or the patient of an action. For example, in figure 3.10, there is a red bag and a blue bag, and the question is *Which bag is red?* This prompted the production of relative clauses because the participants needed to specify the information about the object by referring to the action that it received. This design has effectively elicited relative clauses in both active and passive voices. However, in the current study, the aim was to train and measure participants' production of subject and object relative clauses in active voice, so it was necessary to elicit object relative clauses rather than the passive subject relative clauses.

Figure 3. 10 An example picture stimulus by Gennari et al., 2012



Gennari et al., 2012 (p.147)

To prompt the participants to produce ORCs, some salient cues could be provided. Kim and O'Grady (2016) attempted to elicit direct object and oblique relative clauses with pictures. In their study, participants listened to the descriptions of the pictures in simple sentences with active voice before they were asked to describe one of the elements in the pictures. The results showed that more than 70% of adults produced direct object and oblique relative clauses as expected. The study indicated that although the pre-provided simple sentences could not completely prevent participants from producing some passive subject relative clauses, this technique might be

beneficial for eliciting ORCs. Therefore, this technique was adapted for use in the current study in picture description tasks.

In the current study, each item in the oral production test involved two pictures depicting the same action. The participants were asked to describe either the agent or the patient of one of the pictures based on the given information. The details of the test are described in section 3.4.2.2.

Metalinguistic knowledge test

Metalinguistic knowledge tests typically contained three sections, 1) grammatical judgments of sentences, 2) error identifications and corrections, and 3) explanations about the violated grammar rules (Renou, 2000; Roehr, 2006, 2008). Roehr (2008) conducted a metalinguistic knowledge test with L1 English – L2 German speakers. The metalinguistic knowledge test measured L2 learners' ability in language correction, description and explanation. The test involved 12 sentences and three short passages. For the sentences, each of them contained a highlighted error that needed to be corrected, described and explained. The three short passages were written in an inappropriate manner, and the participants were required to describe and explain why they were inappropriate.

The metalinguistic knowledge test in the current study was informed by Roehr (2008). However, because the current study focused on the difference of the grammatical roles of two nouns in SRCs and ORCs, it was not feasible to measure learners' knowledge of this difference by including grammatical *errors* in the target structures – that is, an 'error' in the target feature (e.g., replacing a noun by a verb, after 'that') could render a correct sentence (i.e., change an ORC to an SRC). Hence, a sentence-picture anomaly detection test was used in the study. For each item in the test, one picture and one sentence were involved. The sentences used in the metalinguistic knowledge test were all grammatically correct but half mismatched the picture. Participants were required to decide whether the sentence matched the picture or not. For mismatched items, participants needed to explain why it mismatched and correct the sentence to match the picture by moving one word in the

sentence, as described in section 3.4.2.2.

3.4.2.2 Design of the test

Offline comprehension: Aural sentence-picture matching test

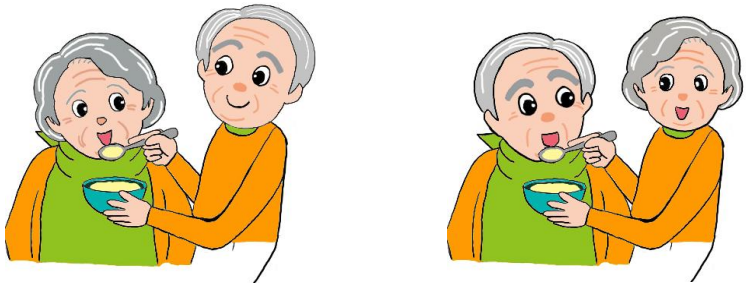
Aural sentence-picture matching test was used to measure participants' offline comprehension of the target forms. In total, there were 24 sets of items. For each set of items, four sentences including two SRCs and two ORCs (the animacy of the head nouns could be both animate or both inanimate; or one animate and one inanimate) were created based on one picture (see figure 3.11). The four sentences were intended to be randomly distributed in the four versions of test, and leaded to each version of tests include six SRC-A, six SRC-I, six ORC-A, and six ORC-I. Among the test items, 12 of them were exclusively created for offline comprehension test, and 12 were taken from eye-tracking tests. The adoption of items from the eye-tracking test aimed to compare the score differences between the existing and the new created items (though this analysis was not done because the limitation of the capacity). The numbers of newly created items and existing items were balanced across the four types of relative clauses. In addition, the offline comprehension test did not have fillers, considering about the time and energy of the participants. As noted at the beginning of section 3.4, the whole study had five measures. The two online measures took the participants around 40 to 50 minutes to finish, so the offline and production measures did not involve fillers to avoid fatigue.

However, it should be acknowledged that in the administration of the test, due to the researcher's mistake, the number of items for each type of relative clause was not equal in version 1, 2, 3, and 4. Within one version of the four versions of the test, the number of each type of relative clauses could be 4 or 8, and the total number of items was 24. The versions were counterbalanced across pre-, post-, and delayed post-test within groups.

For each item of the test, participants first saw a fixation dot lasting for 745ms. Then they saw a set of two reversible pictures and heard an audio recorded sentence (see figure 3.11). The aural stimuli and the pictures were played simultaneously. The

participants were asked to choose which picture matched the sentence by pressing the keys on the keyboard as soon as possible. More examples see Appendix 9, and the list of all the items see Appendix 14.

Figure 3. 11 An example item of offline comprehension test



Participants would hear one of the stimuli (depending on the version of the test):

1) The man that feeds the woman has grey hair. (left) (SRC-A)

or 2) The man that the woman feeds has grey hair. (right) (ORC-A)

or 3) The woman that feeds the man has grey hair. (right) (SRC-A)

or 4) The woman that the man feeds has grey hair. (left) (ORC-A)

Oral production: Picture description test

The sentence description test was used to measure learners' oral production. Each version of the test included 20 test items. In this test, each three pictures was used to create four sentences (each pair of pictures generate two SRCs or two ORCs, and the head none could either be animate or inanimate, see figure 3.12). The four sentences were assigned into the four versions of test. Each version of the test was intended to have five items for each type of the target structures (SRC-A, SRC-I, ORC-A and ORC-I). However, in the administration, each version of the test included four SRC-A and six SRC-I or six SRC-A and four SRC-I; versions were counterbalanced across pre-, post-, and delayed post-test, within groups. No filler was involved in this test (same reason as that of the offline comprehension).

For each item, the participants saw two pictures described an action. As shown in figure 3.12, each pair of pictures had two similar elements that were different by one

specific feature (e.g., *a brown door and a blue door*) and two totally different elements (e.g. *a boy and a girl*). The similar elements in the two pictures could either be the agent or the patient of the action. While looking at the pictures, the participants heard and saw two simple sentences in active voice describing the two pictures respectively. Then, a question about the feature of one of the similar elements was provided aurally and visually, and the participants were guided to start the sentence with the given words. They were also informed that the conjunctions and conjunctive adverbs like 'and', 'but', 'because', 'when' were not allowed to use, and they were not allowed to describe the position of the element (e.g. *the door on the left/second picture*).

More examples see Appendix 10, and the full list of items see Appendix 15.

Figure 3. 12 Example items for oral production test

Aural and visual stimuli:

In the first picture, the boy opens the door. The door is brown. In the second picture, the girl opens the door, the door is blue.

1) Question: Which door is blue?

Start with: The door…

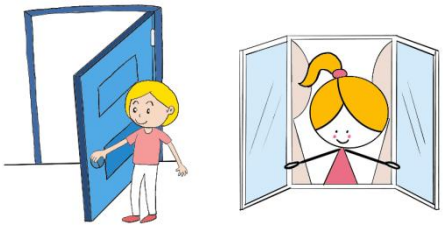
Expected answer: The door that the girl opens is blue. (ORC-I)

Or

2) Question: Which door is brown?

Start with: The door…

Expected answer: The door that the boy opens is brown. (ORC-I)



Aural and visual stimuli:

In the first picture, the girl has short hair. The girl opens the door. In the second picture, the girl has long hair. The girl opens the window.

3) Question: Which girl has short hair?

Start with: The girl…

Expected answer: The girl that opens the door has short hair (SRC-A).

Or

4) Question: Which girl has long hair?

Start with: The girl…

Expected answer: The girl that opens the window has long hair (SRC-A).

Metalinguistic knowledge test

Each version of the metalinguistic knowledge test included 16 items in total. Each picture generated four sentences (two SRCs and two ORCs, the head noun could be two animate nouns, two inanimate nouns, or one animate noun and one inanimate noun; one of the SRCs and ORCs matched the picture, and another one mismatched, see figure 3.13). The sentences were distributed in the four versions of tests. Within one version of the test, each type of relative clauses had four items; half of the sentences matched (k = 2) the picture and half mismatched (k = 2). Eight of the 16 items were adopted from the SPR test, and eight were newly created. The new and adopted items numbers were balanced across the relative clause type and match/mismatch. No filler item was involved in the test.

For each item, participants saw one picture and one sentence. Participants were required to decide whether the sentence matched the picture or not. For mismatched items, the participants were asked to correct the sentence by moving only one word and explain the reason why there was a mismatch as fully as possible (see figure 3.13).

More examples see Appendix 11, and the full list of test items see Appendix 16.

Figure 3. 12 An example item for the metalinguistic knowledge test

The woman that the man calls has long hair. (ORC-A)

Match☑ Mismatch☐

Or
The man that the woman calls has short hair. (ORC-A)
Or
The woman that calls the man has long hair. (SRC-A)
Or
The man that calls the woman has short hair. (SRC-A)

3.4.2.3 Administration of the tests

The offline comprehension and oral production tests were delivered through the software OpenSesame (Mathôt et al., 2012), and all the items were displayed randomly. The metalinguistic knowledge test was paper-and-pen-based, and the items were presented in a fixed order. All three offline tests were untimed, and no practice items were provided before the tests.

For the offline comprehension test, the test would not proceed to the next item until the response had been made. In order to encourage the participants to try their best to concentrate on the test, the participants were informed that they would be provided with the overall accuracy scores at the end of the test. The whole test lasted for approximately 5 minutes.

For the oral production test, the participants were asked to speak out the answers aloud into a microphone, and the responses were recorded using the software Audacity (Audacity Team, 2018). After finishing one item, a key pressing was needed in order to proceed to the next item. If the participants made more than one response for

an item, only the last one would be transcribed and analysed. The test lasted for about 8 minutes.

For the metalinguistic knowledge test, multiple tasks were involved within one item. In case of the participants forgetting the test requirements, the instructions printed on a separate page was provided during the test. The duration of the test varied among participants over a wide range (2 to 45 minutes). Most of them finished the test in around 10 to 15 minutes, but some participants who knew the grammar rules very well or clearly did not have any explicit knowledge about the target structures finished the tests very fast (within 5 minutes). Meanwhile, some participants who could not correct the sentence initially were willing to spend time addressing the problems. They could take up to 45 minutes to complete the metalinguistic knowledge test.

3.4.2.4 Data scoring and coding

Data Scoring

For the offline comprehension and the oral production tests, each correct response was rewarded one point, and incorrect responses were scored zero.

For the metalinguistic knowledge test, one point was awarded for correctly choosing match or mismatch, one point for correctly making the sentence match the picture, and one point for clearly explaining the reason for the mismatch. For matched items, only the first task, deciding match or mismatch, was analysed. For mismatched items, the three tasks were analysed separately.

Data coding

The oral production and the metalinguistic knowledge tests involved qualitative data; thus, the data needed to be coded before scoring.

Oral production tests

Defining 'correctness'

A correct answer included: the first noun, the second noun and the verb in the correct order, and a relative pronoun. Errors involving pronunciation, articles, and verb inflections for tense and aspect were ignored during the scoring, for all relative clause

types.

For ORCs, responses could omit the relative pronoun 'that' and still be deemed correct (*e.g., The woman the boy helps has blond hair*, is an adequate answer and a correct relative clause). In contrast, for SRCs, correct responses always needed 'that', as the relative pronoun is obligatory. Reduced SRCs (without a relative pronoun and BE) were regarded as correct if the order of the noun phrases and the verb were the same as the relative clauses would have been had a relative pronoun and BE been included (*e.g., The box keeping the orange is brown*). However, when the target structures were ORCs, productions of passive SRCs and reduced passive SRCs were regarded as incorrect. For both SRCs and ORCs, answers that did not provide the full verbal complement were counted as correct as long as they included correct relative clauses (*e.g. The man that lifts the boy [has blond hair])*. The codes for incorrect responses are shown in table 3.2.

Table 3. 2 Codes for incorrect responses

Codes	Examples from the dataset
No relative clauses	The table is brown with a girl hided in it.
Omission of the relative pronoun in SRCs	The television presents the man is yellow.
Passive SRCs	The man that is carried by the balloon wears green t-shirt.
Non-adjacency	The door is brown opened by the boy.

Excluded data

In addition, the irrelevant responses were excluded from the analysis. For example, in the item, *in the first picture, the rabbit is grey, the snake licks the rabbit. In the second picture, the rabbit is white, the dog licks the rabbit. Which rabbit is grey?* The expected response was *The rabbit that the snake licks is grey.* However, if the response from the participant was *The snake that licks the rabbit is green.* Since the specific question was not answered (the question was about the rabbit, not the snake), such responses were taken out from the data pool, even though they did produce a relative clause. That is, it was not regarded as 'incorrect' (or indeed 'correct') even though it was an SRC rather

than the expected ORC. This decision was taken because it was unknown whether they could not produce the target ORC or whether they just misunderstood the purpose of the question. This way, these items did not serve as a denominator in the proportion of correct answers given.

Metalinguistic knowledge test

Defining 'providing' of metalinguistic knowledge

For the reason explanation task, the data were coded as providing and non-providing of metalinguistic knowledge. The responses that included syntax related keywords like *subject, object, word order, agent, patient,* or expressed the same meanings (*e.g., the woman is the one who acts)* were regarded as providing metalinguistic knowledge. Grammatical errors were ignored, as long as the explanations were understandable (*e.g., The boy is the one who do the action of finding the girl*). The semantic explanations that described the picture (*e.g. because the boat is chasing ship*) or simply repeated the target sentences (*e.g., The shirt that the woman wets is blue*) were deemed non-providing of metalinguistic knowledge because such explanations could not show evidence of knowing the target structures. In addition, no explanation was regarded as non-providing.

Excluded data

For sentence correction, successfully making the sentence match the picture by moving one word was awarded one point. No answer or the sentence still mismatching the picture after moving the word was given 0. However, some responses containing the exchange of the two words were taken out from the data pool. For instance (see figure 3.14), in the picture, the man is calling the woman, but the sentence means that the woman is calling the man. The participants were expected to move *calls* after *the man* to make the sentence match the picture. Answers that swapped around *the woman* and *the man* sometimes occurred. Such responses could not be counted as either correct or incorrect because it was unknown whether the occurrence of the responses was due to the lack of knowledge or simply due to forgetting or not following the task requirement.

Figure 3. 13 An example sentence of metalinguistic knowledge test

Sentence: The woman that calls the man has long hair.

3.4.2.5 Data analysis

The data from offline measures were analysed in four steps. First, descriptive analysis, including the means and SDs of each group in each condition, was conducted. Second, violin plots with included box plots were used to visualise the distribution of data.

Third, effect sizes (Cohen's *d* with 95% CI) that measured the mean differences were calculated. For the first research question (which type of relative clause is more difficult), within-group effect sizes reflecting the differences between each pair of relative clauses were calculated with the group of native speakers and the group of L2 learners separately. For the second research question (the extent to which the training effects are observable), within-group effect sizes, reflecting changes over time (pre-, post, and delayed post-test) and between-group effect sizes, reflecting differences between groups within a test phase were calculated. In addition, for between-group effect sizes, the adjusted effect sizes, referring to the *changes* of effect sizes from pre- to post-, delayed post-test were provided (i.e., accounting for any baseline differences between groups). This method was in line with McManus and Marsden (2017). The benchmarks of Cohen's *d* used for interpreting the results followed the suggestion of Plonsky and Oswald (2014): within-group *d* = .60 (small), 1.00 (medium), 1.40 (large); between-group *d* = .40 (small), .70 (medium), 1.00 (large).

Last, logistic regression models run via the "lme4 package" (Bates et al., 2015) were used to conduct inferential statistical analysis. The logistic regression model is defined as a type of a generalised linear mixed-effects model, which is used for binary data (Linck & Cunnings, 2015). Unlike normal linear mixed-effects models that treat the

data as continuous variables, logistic regression estimates the effects based on log-odds (Linck & Cunnings, 2015). In the current study, the scores of each measure were used as the dependent variable.

In analysing the first research question (which type of relative clause is more difficult), the fixed effects only included the type of relative clause. The primary random effects included the by-subject and by-item random intercepts, while the maximal random effects included the by-subject and by-item random intercepts and random slopes of the relative type.

For the second research question (the extent to which the training effects are observable), the interaction between group and test phase was used as fixed effects. The maximal random effects included by-subject random effects with the intercepts and the slope of test phase, and by-item random effects with intercepts and the slope of the interaction between test phase and group; the primary random effects contained the by-subject and by-item random intercepts.

For both research questions, the model with maximal random effects was run initially, and then the model was trimmed down step by step to the primary model. The best-fitting model was selected by LRT and AIC. The estimate *b* with 95% CI, SE, *z*-value, *p*-value and odds ratio about the estimate *b* with 95% CI of the fixed effects, marginal and condition R^2 of the selected model were reported.

3.4.2.6 Instrument Reliability

Kuder-Richardson 20 (KR-20) is a type of Cronbach's alpha, which is designed to test instrument reliability for binary data (Raykov & Marcoulides, 2019). In the current study, the internal reliability of the offline measures was tested by KR-20. For each measure, the reliability for the native speakers and L2 learners were tested separately. In addition, for L2 learners, the reliability of the pre-, post- and delayed post-test were calculated separately (see table 3.3).

In terms of interpreting the instrument reliability, Plonsky and Derrick (2016) found that in published L2 research, the median of instrument reliability of L2 learners was .81, and that of native speakers was .87. In the current study, table 3.3 suggested

that only the oral production test and the reason explanation task of the metalinguistic knowledge test reached this median across the time for both L2 learners and native speakers. However, they also admitted that, it was possible that the reliability reported in the published studies might be higher than the studies that was not be reported, as the unreported reliability might because they were calculated but too low (Plonsky & Derrick, 2016). In addition, Brown (2014) suggested the benchmarks of reliability in L2 research: .00-.30 = virtually none; .31-.50 = slight; .51-.70 = fair; .71-.89 = moderate; .90-1.00 = substantial.

For L2 learners, according to the benchmarks of Brown (2014), the oral production test had moderate (i.e., the pre-test) to substantial reliability (i.e., the post- and the delayed post-test). The reliability of the offline comprehension test could be regarded as slight (i.e., the pre- and the delayed post-test) to fair (i.e., the post-test). For the metalinguistic knowledge test, at the pre-test, the reliability was slight (i.e., deciding match or mismatch), fair (i.e., sentence correction task) or moderate (i.e., reason explanation task), and reached the benchmark of moderate at the post- and delayed post-test. The rather low reliability of the offline comprehension and the metalinguistic knowledge tests might because the participants had ceiling or ceiling scores for SRCs but had comparatively lower score for ORCs (relative to SRCs), which leads to the inconsistency of the items.

For native speakers, the oral production test and the reason explanation task of the metalinguistic knowledge test could be considered as moderate, and the offline comprehension test had the slight reliability. In addition, it could be noticed that the alpha value for native speakers in metalinguistic knowledge (deciding match or mismatch task and sentence correction task) was negative. The occurrence of the abnormal values might because 1) the native speakers scored at or near ceiling for deciding match or mismatch task; 2) in sentence correction task, some responses were regarded as invalid and were removed from the data pool (criterion see 3.4.2.4), which might lead to the data points of the native speakers being too few. These factors might negatively influence the KR-20 results of the measure.

Table 3. 3 KR-20 for the offline comprehension, the oral production and the metalinguistic knowledge test

Measures	NSs	L2 learners		
		pre-	post-	delayed
offline comprehension	.48	.50	.67	.41
oral production	.87	.81	.90	.92
metalinguistic knowledge test (deciding match or mismatch)	-1.26	.38	.71	.73
metalinguistic knowledge test (sentence correction)	-21.00	.67	.78	.74
metalinguistic knowledge test (reason explanation)	.81	.89	.85	.83

Note: NSs = native speakers

3.5 The design of the training sessions

The parsing and the input flood group received two training sessions delivered one-to-one via OpenSesame (Mathôt, et al., 2012) between the pre- and post-test. Two versions of training materials were designed. Half of the participants used version 1 for the first session and used version 2 for the second session, while half of the participants did the two versions in the opposite order.

The sessions for the parsing group included being provided with explicit information (EI) about the target forms and two listening and two reading activities to train the participants how to use the linguistic cues (i.e. grammatical role assignment) to parse the relative clauses. Each training session for the parsing group lasted for approximately 35 minutes.

The input flood group proceeded to the training activities without receiving any EI, and the activities did not call attention to the linguistic cues. Each input flood training session lasted around 30 minutes.

3.5.1 The parsing strategies training

EI for the parsing group

The EI about relative clauses was delivered at the beginning of each training session. First, the function of the relative clause (i.e. Relative clauses give extra information about nouns) was introduced. The participants were allowed to read this information for at least 4000ms, and after that, the training would automatically proceed to the

next step. Then, the participants saw two examples of SRC and ORC respectively. In each example, a picture, a relative clause and explanations of the relative clause were involved. The participants first saw a picture that contained two elements performing an action, and they were guided to press 'SPACE BAR' to see the relative clause and the explanations. The relative clause was shown word by word at a pace controlled by participants' key pressing. After each word, an explanation about the function of the word in the sentence was shown in red. When the syntactic cue had been explained, an interactive question '*can you predict what comes next*?' was presented in green. The participants did not need to answer the question aloud or type into the answer, and the purpose of providing the interactive question was to encourage the participants to actively think of the cue. After the full relative clause being provided, the answer to the interactive question was provided. Then, the rest of the sentences showed together, and the explanation about the meaning of the sentence would be shown (see figure 3.15).

Figure 3. 14 An example of EI for parsing strategy group

Subject relative clause

The

The woman

'woman' is a noun.

The woman that

When you see 'that' straight after a noun, you know a relative clause is coming.

The woman that washes

‘washes’ is a verb. When you see the verb straight after ‘that’, you know the first noun ‘the woman’ does the action.

Can you predict what comes next?

The woman that washes the

The woman that washes the dog

It is ‘the dog’! Did you predict it correctly?

Now you know the action is ‘the woman washes the dog’.

The woman that washes the dog has blond hair.

Now, you know, ‘the woman washes the dog’, and ‘the woman has blond hair’.

Training activities for the parsing group

After EI, two reading activities and two listening activities were delivered. All four activities involved sentence-picture matching. Each picture used in the activities depicting an action, and the agent and the patient were reversible. Two SRCs and two ORCs were created based on a pair of reversible pictures. In order to force the learners to use the syntactic cues to parse the sentence, the stimuli were not presented in complete sentences. The words after the syntactic cues were omitted (*e.g.,* for SRCs: *The cat that chases…;* of for ORCs: *The cat that the dog…*), and the participants were required to decide sentence-picture matching based on sentence segments.

Each activity contained 24 training items (6 SRC-A, 6 SRC-I, 6 ORC-A, 6 ORC-I) and four practice items (one item for each type of relative clause). The practice items were shown in sequence, and the training items were displayed randomly. After each item, the participants received feedback informing them whether they had correctly responded to the question. A congratulations page was presented for 1000 ms after each correct response, and then the activity proceeded to the next item. For incorrect responses, the participants were provided with corrective feedback which included the complete sentence and the explanation about how to use the syntactic cue to parse the sentence. For the four practice items, the feedback emphasising the syntactic cue was provided for both correct and incorrect responses, and the correct responses would also receive the congratulation page. In the reading activities, the participants

needed to press a key after reading the corrective feedback to proceed to the next item, while in the listening activities, the activities would automatically start the next item after the end of the corrective feedback.

The design of the four activities will be illustrated below.

Training activity 1 & 2: Decide which sentence segment matches the picture

In training activity 1 and 2, the participants were shown one picture and two sentence segments. The sentence segments and the corrective feedback were provided visually in activity 1, while in activity 2, the stimuli were played aurally.

For each item, the participants first saw a picture and a verb which described the action of the picture. 1000ms later, the sentence segments were displayed visually or aurally. The participants were required to decide which sentence segments matched the picture by pressing the LEFT or RIGHT key on the keyboard. The feedback was provided after each item. It was noticed that, in the listening activity, the participants were able to make the decision after hearing the first sentence segments. However, in order to ensure every participant could receive exactly the same amount of input, the participants were not allowed to respond to the question before both of the sentence segments stimuli being displayed. The examples of SRC and ORC items are shown in figure 3.16 (more examples see Appendix 3).

Figure 3. 15 Example items for the parsing strategies training activity 1 & 2

Example item (SRC):

carries

left: The dog that carries… right: The dog that the bag…

Matched sentence segment: The dog that carries… (left)

Feedback for correct response:

Congratulations! You are right!

Feedback for incorrect response:

Full sentence: The dog that carries the bag is brown.

When you see "carries" straight after "that", you know "the dog" does the action. You can predict the action is "the dog carries the bag".

Example item (ORC):

carries

left: The dog that carries… right: The dog that the bag…

Matched sentence segment: The dog that the bag… (right)

Feedback for correct response:

Congratulations! You are right!

Feedback for incorrect response:

Full sentence: The dog that the bag carries is brown.

When you see "the bag" straight after "that", you know "the bag" does the action. You can predict the action is "the bag carries the dog".

Training activity 3 & 4: Decide which picture matches the sentence segments

Activity 3 and 4 were designed in the same pattern with activity 3 based on reading and activity 4 based on listening.

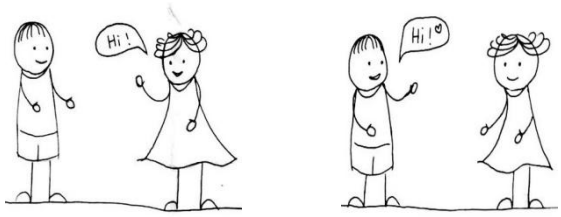
For each item of the activities, two reversible pictures and the verb related to the pictures were presented initially. The participants were allowed to observe the pictures and the verb for 2000ms. Then, they saw or heard a sentence segment describing one of the pictures, and they needed to decide which picture matched the sentence segments by pressing the LEFT or the RIGHT key. After each response, the feedback would be given. The examples of SRC and ORC items are shown in figure 3.17 (more

examples see Appendix 4).

Figure 3. 16 Example items for the parsing strategies training activity 3 & 4

Example item (SRC):

greets



The girl that greets…

Matched picture: left

Feedback for correct response:

Congratulations! You are right!

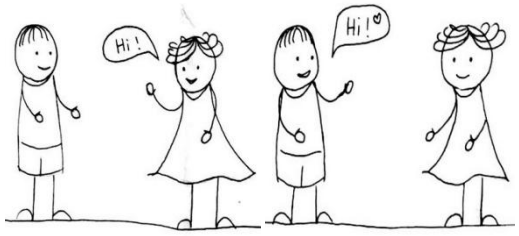
Feedback for incorrect response:

Full sentence: The girl that greets the boy wears a dress.

When you see "greets" straight after "that", you know "the girl" does the action. You can predict the action is "the girl greets the boy".

Example item (ORC):

greets



The girl that the boy…

Matched picture: right

Feedback for correct response:

Congratulations! You are right!

Feedback for incorrect response:

Full sentence: The girl that the boy greets wears a dress.

When you see "the boy" straight after "that", you know "the boy" does the action. You can predict the action is "the boy greets the girl".

3.5.2 The input flood training

The input flood group took part in the four training activities, without being provided with EI. They received the same number of training items with the parsing group, but the attention required them to respond to the items was focused on the meaning of the nouns instead of having to understand the syntax. There were two major differences between the input flood training activities and parsing strategies training activities. First, the stimuli were presented in complete sentences in the input flood training, while the parsing strategies training utilised sentence segments. Second, the pictures used in the input flood training were non-reversible, and the sentences mismatched the pictures due to differences in the nouns.

The feedback was provided for each item, and a congratulations page lasting 1000ms was shown for a correct response. The corrective feedback for incorrect responses emphasised the noun differences between sentence and picture. The activities proceeded to the next item after the congratulation page or after the key-pressing for corrective feedback. All the feedback was shown visually for both reading and listening activities.

Training activity 1 & 2: Decide which sentence matches the picture

Training activity 1 and 2 were designed in the same pattern as each other, which asked the participants to choose the sentence that matched the picture from two complete sentences. The sentences in activity 1 were provided visually, while those in activity 2 were displayed aurally. The pictures used in the two activities were the same as the ones used in the parsing strategies training activity 1 and 2.

For each item of the activities, participants first saw a picture, and 1000ms later, they would see or hear two sentences. The sentences were in the same syntactic structure, which means that both sentences were SRCs or ORCs. The first nouns or the second nouns of the two sentences were different (the items were balanced). The participants were asked to decide which sentence matched the picture and then saw

the feedback. The examples of items are shown in figure 3.18 (more examples see Appendix 5).

Figure 3. 17 Example items for the input flood training activity 1 & 2

Example item (SRC):

left: The man that follows the dog wears a hat. (matched)

right: The man that follows the cat wears a hat. (mismatched)

Matched sentence: left

Feedback for correct response:

Congratulations! You are right!

Feedback for incorrect response: No! There is no cat in the picture!

Example item (ORC):

left: The dog that the man follows is brown. (matched)

right: The dog that the woman follows is brown. (mismatched)

Matched sentence: left

Feedback for correct response:

Congratulations! You are right!

Feedback for incorrect response: In the picture, there is a man. So “the man” follows the dog.

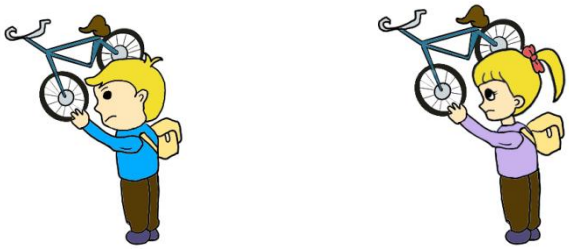
Training activity 3 & 4: Decide which picture matches the sentence

Activity 3 (based on reading) and 4 (based on listening) required participants to choose the matching picture from two options for a sentence. In these two activities, the sentences were the same as the ones in the parsing strategy group. In a pair of pictures, one is adopted from the activities used in the parsing strategy group, and another one depicting the same action but with a different agent or patient (e.g., *one picture depicted a dog chasing a cat, and one picture showed a dog chasing a rabbit*).

For each item in activity 3 and 4, before the sentence was shown or heard, the participants were allowed to observe the two pictures for 2000ms. The participants were required to decide which picture matched the sentence. Feedback was provided after each response. The examples of items are shown in figure 3.19 (more examples see Appendix 6).

Figure 3. 18 Example items for the input flood training activity 3 & 4

Example item (SRC):



The boy that carries the bike is strong.

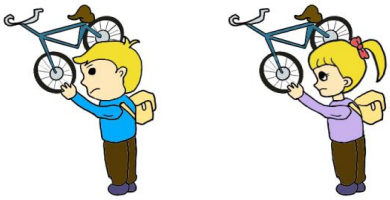
Matched picture: left

Feedback for correct response:

Congratulations! You are right!

Feedback for incorrect response: No! There is no girl in the sentence!

Example item (ORC):



The bike that the boy carries is blue.

Matched picture: left

Feedback for correct response:

Congratulations! You are right!

Feedback for incorrect response: No! There is no girl in the sentence!

3.6 The pilot study

There were two purposes of carrying out the pilot study. The first one was to check whether the equipment and the software could work smoothly. The second one was to find out whether the materials were feasible and detect any problems with the materials.

The pilot study was conducted with three Chinese-speaking L2 learners of English and four native English speakers. Each native speaker piloted one version of the outcome measures, and the L2 learners piloted the measures by attending pre- and post-test. The two versions of parsing strategies training materials were piloted by two learners, while one L2 learner piloted the two versions of input flood training materials. After the pilot study, several amendments had been made.

First, in the pilot study, the eye-tracking and the SPR tests had 160 items (40 critical items and 120 non-critical items) in total, and each test took around 30 minutes. It was found that the participants would get fatigued after 20 minutes. Thus, the non-critical items in each of the two measures were reduced to 80 items, which was in line with the recommendation that the non-critical items should constitute 50% to 75% of the total items in online measures (Keating & Jegerski, 2015).

Second, in the self-paced reading tests, for each item, the participants first saw a picture and then read the sentence controlled by their own pace. By analysing the pilot data of native speakers, it was found that they did not slow down when they read the mismatched items during the critical regions. It was realised that one possible reason might be that more than one verb could be used to describe the action of the picture, and if they had expected one verb to be used and then another was used, that might be a source of cognitive load for them. During reading the sentences, the participants might expect the following information could make the sentence match the picture. In

order to reduce the effects of such a confound we reduced the influence of the uncertainty of the verbs in the main study by presenting the verb used in the relative clause alongside the picture *before* the occurrence of the sentence.

Third, in piloting the metalinguistic knowledge test, the pictures were printed in black and white. However, some sentences of the test included the expression of colour (*e.g., The basket that the dog carries is brown*). Some participants wrote that the sentence mismatched the picture because the basket in the picture was not brown. Thus, to avoid such misunderstandings, the pictures were printed in colour in the main study. In addition, for mismatched items, the participants were instructed to move one word to make the sentence match the picture. However, many of them tended to swap around the two nouns of the sentence. In the main study, a requirement was added to the instruction to explain that exchanging the position of *two* nouns was not allowed.

Fourth, some problems like spelling mistakes, incorrect size of few pictures, few inappropriate expressions (*e.g., 'the soup is sweet' has been changed to 'the soup is hot'*) were solved.

Chapter 4 Results

In this chapter, the results in relation to the first research question will be presented in Section 4.1, and Section 4.2 will show the results for the second research question.

4.1 Which type of relative clause (SRC vs. ORC) is more difficult in online, offline comprehension, production, metalinguistic knowledge and the role of animacy of main clause noun?

This section will analyse the pre-test performance of L2 learners and the data of Native English speakers (NSs) on each outcome measure. For offline comprehension, SPR, oral production and metalinguistic knowledge tests, the descriptive results including means and standard deviations (SDs), plots, examination of effect sizes with 95% CIs, and the inferential statistical results will be reported respectively. In the interpretation of the inferential statistics, the effects were regarded as statistically significant when $p \leq .05$. In addition, when the *p* value was more than .05 but smaller than .10, and the 95% CI around the estimate *b* did not pass through zero (followed the interpretation of 95% CI in McManus and Marsden, 2017), the effects were regarded as reliable. For the eye-tracking data, the visualisation and inferential statistics will be reported.

Section 4.1.1 to 4.1.5 presents the results in offline comprehension, self-paced reading, eye-tracking, oral production and metalinguistic knowledge tests respectively.

4.1.1 Which type of relative clause (SRC vs. ORC) is more difficult in offline comprehension: aural sentence-picture matching test?

The aural sentence-picture matching test was used to measure offline comprehension. Following previous research findings, the accuracy scores for SRCs were expected to be higher than ORCs for both native speakers and L2 learners. In addition, the ORC-I might be easier than ORC-A for both groups.

4.1.1.1 Descriptive analysis

The descriptive statistics (see in table 4.1.1) showed that the NSs scored at ceiling for SRCs and ORC-A, but had slightly lower scores in ORC-I. For L2 learners, they also had higher accuracy scores in SRCs and ORC-A compared to ORC-I.

Table 4. 1. 1 Mean (*SDs*) accuracy scores in offline aural sentence-picture matching test of native English Speakers and L2 learners

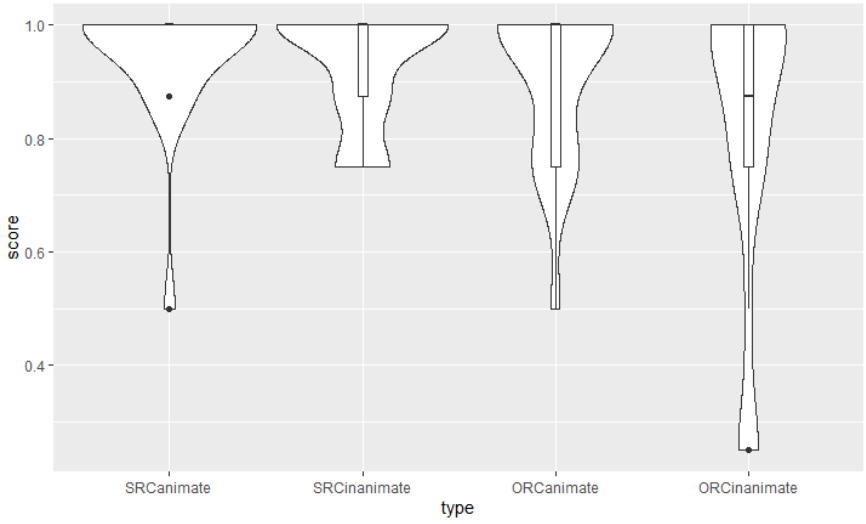
Structure	NSs	L2 learners
SRC-A (k=4 or 8 [a])	.94 (.25)	.86 (.35)
SRC-I (k=4 or 8 [a])	.93 (.26)	.86 (.34)
ORC-A (k=4 or 8 [a])	.91 (.29)	.84 (.37)
ORC-I (k=4 or 8 [a])	.84 (.37)	.77 (.42)

***Note**:* SRC-A = subject relative clause – animate head; SRC-I = subject relative clause – inanimate head; ORC-A = object relative clause – animate head; ORC-A = object relative clause – inanimate head; [a] Due to an oversight, numbers of items differed between version 1 and version 2, and between version 3 and version 4 of the tests, respectively; versions were counterbalanced across pre-, post-, and delayed post-test, within groups.

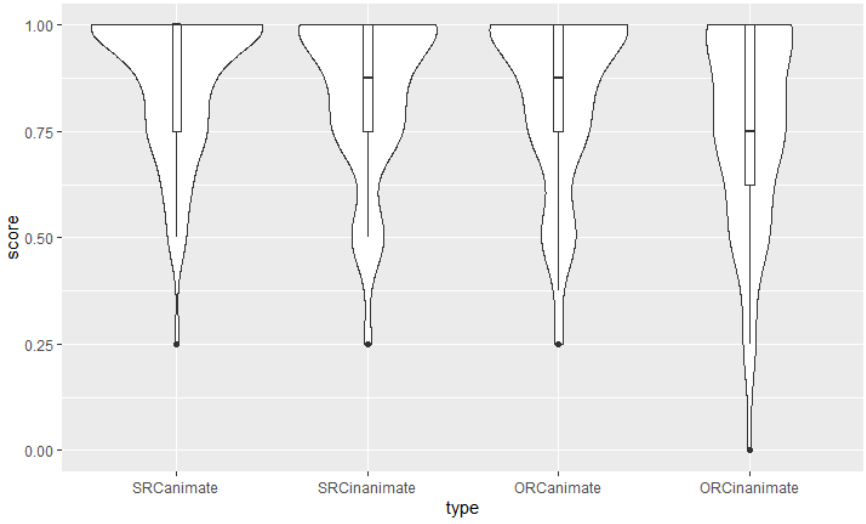
4.1.1.2 Plots

Figure 4.1.1 presents violin plots with included box plots of the accuracy scores of SRC-A, SRC-I, ORC-A and ORC-I for the native English speakers and L2 learners respectively. The plots suggested that more NSs scored at ceiling for SRCs relative to ORCs. In addition, the number of NSs who scored at ceiling was higher in ORC-A compared to ORC-I. Moreover, the plots of ORC-I might be able to explain why this structure had lower mean accuracy scores than the other structures. It might due to a few participants scored at zero. For L2 learners, the proportions of the participants that scored at ceiling were similar across the SRC-A, SRC-I and ORC-A structures, and the ORC-I structure had lower accuracy scores than those other three structures. Moreover, as the inserted box plots showed, for L2 learners, the median of the SRC-A was higher than SRC-I and ORC-A structures, which indicated that the SRC-A was the easiest and the ORC-I was the most difficult across the four structures.

Figure 4. 1. 1 comparisons of mean accuracy scores of each type of relative clauses for native English speakers and L2 learners in offline aural sentence-picture matching test



Native Speakers



L2 learners

4.1.1.3 Examination of effect sizes

Table 4.1.2 presents the within-group effect sizes, reflecting differences between each pair of relative clauses of the NSs and L2 learners separately. All the effects were found to be negligible or very small. For the NSs group, the score differences between SRCs (including both SRC-A and SRC-I combined) and the ORC-I were reliable, because the 95% CI did not pass through zero. For L2 learners, reliable effects could be observed in the comparisons between the ORC-I and the other three structures. Thus, generally, the ORC-I structure was likely to the most difficult one across the four types of relative

clauses for both NSs and L2 learners.

Table 4. 1. 2 Within-group effect sizes for (Cohen's *d*) [95% CI] for offline aural sentence-picture matching test

	NSs	L2 learners
SRC-A vs. SRC-I	.00 [-.18, .18]	-.01 [-.10, .08]
SRC-A vs. ORC-A	.08 [-.10, .26]	.02 [-.07, .12]
SRC-A vs. ORC-I	**.22 [.05, .40]**	**.17 [.08, .27]**
SRC-I vs. ORC-A	.06 [-.11, .23]	.05 [-.04, .14]
SRC-I vs. ORC-I	**.21 [.03, .39]**	**.18 [.08, .27]**
ORC-A vs. ORC-I	.14 [-.04, .31]	**.14 [.05, .23]**

Note: Bold typeface indicates that the CIs did not pass through zero

4.1.1.4 Inferential statistical analysis

For NSs, the best-fitting models suggested by the AIC and LRT were inconsistent (AIC and LRT results see Appendix 17). The AIC suggested that the primary model which only included by-subject and by-item random intercepts was the best-fitting one, while the LRT suggested that the model with the maximal random structure (including by-subject and by-object random slopes of the type of relative clauses) was preferred ($AIC_{primary}$ = 321.37, $AIC_{maximal}$ = 324.16; $LRT_{primary}$: $LRT_{maximal}$: $\chi^2(18) = 33.21$, $p = .016$). However, the results of the maximal model were inconsistent with the descriptive and the effect sizes results, and the results of the odds ratio were extreme values (e.g. ORC-I vs. ORC-A: *b* [CI] = 31.12 [19.02, 43.22], SE = 7.36, *z* = 4.23, *p* < .001, OR [CI] = 3.28e+13 [1.82e+08, 5.89e+18]). Thus, the basic model might be more suitable than the maximal model to fit the data of the NSs and these results are presented in table 4.1.3. The statistically significant effect was found with the comparison between the ORC-I and SRC-A. The odds ratio predicted that the NSs were 2.94 more likely to provide correct answers in the SRC-A compared to the ORC-I. In addition, the comparison between ORC-I and SRC-I was reliable, because the 95% CI of the estimate *b* did not pass through zero. The odds ratio for this indicated that the NSs were 2.76 times more likely to correctly answer the questions about SRC-I compared to the ORC-I. No statistically significant effect could be observed for other comparisons.

For L2 learners, the selected best-fitting model (see 4.1.4) included the by-subject

random slope of the type of relative clauses, and the by-item intercepts (AIC and LRT results see Appendix 17). The results showed that the effect between the ORC-I and the SRC-A was reliable, because the 95% CI of the estimate *b* did not pass through the zero. The odds ratio predicted L2 learners to be 2.01 times more likely to provide correct answers in the SRC-A compared to the ORC-I.

Table 4. 1. 3 The fixed effects of the model analysis of accuracy scores for native English Speakers in offline comprehension: aural sentence-picture matching test

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
ORC-I vs. ORC-A	.73 [-.11, 1.57]	0.51	1.42	.155	2.07 [.89, 4.78]
ORC-I vs. SRC-A	**1.08 [.19, 1.97]**	**0.54**	**2.00**	**.046***	**2.94 [1.21, 7.16]**
ORC-I vs. SRC-I	**1.02 [.14, 1.89]**	**0.53**	**1.91**	**.056**	**2.76 [1.15, 6.64]**
ORC-A vs. SRC-A	.35 [-.59, 1.30]	0.58	0.62	.539	1.42 [.55, 3.67]
ORC-A vs. SRC-I	.29 [-.63, 1.22]	0.56	0.52	.605	1.34 [.53, 3.38]
SRC-I vs. SRC-A	.06 [-.92, 1.04]	0.60	0.11	.917	1.06 [.40, 2.84]

Note: Model formula: model4=glmer(score ~ type + (1|subject) + (1|item), data=offline_comprehension, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.04, Conditional R^2=.25; bold typeface indicates a reliable effect; *significantly differently from zero when α≤ .05

Table 4. 1. 4 The fixed effects of the model analysis of accuracy scores for L2 learners in offline comprehension: aural sentence-picture matching test

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
ORC-I vs. ORC-A	.45 [-.19, 1.10]	.39	1.16	.246	1.57 [.83, 2.99]
ORC-I vs. SRC-A	**.70 [.02, 1.37]**	**.41**	**1.69**	**.091**	**2.01 [1.02, 3.95]**
ORC-I vs. SRC-I	.58 [-.05, 1.22]	.39	1.51	.132	1.79 [.95, 3.38]
ORC-A vs. SRC-A	.24 [-.46, .94]	.43	.57	.568	1.27 [.63, 2.56]
ORC-A vs. SRC-I	.13 [-.52, .78]	.40	.33	.744	1.14 [.59, 2.18]
SRC-I vs. SRC-A	.11 [-.58, .79]	.42	.26	.793	1.12 [.56, 2.21]

Note: Model formula: model3=glmer(score ~ type + (type|subject) + (1|item), data=offline_comprehension, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.01, Conditional R^2=.36; bold typeface indicates a reliable effect

4.1.1.5 Summary of the results in the offline comprehension test

In summary, the results suggested that in offline aural comprehension, ORC-I was the most difficult structure and the SRC-A was the easiest one among the four types of relative clauses (SRC-A, SRC-I, ORC-A and ORC-I) for both NSs and L2 learners. The other three types of relative clauses did not have statistically significant differences in

terms of difficulty in the offline comprehension test.

4.1.2 Which type of relative clause (SRC vs. ORC) is more difficult in online comprehension measured by self-paced reading?

The self-paced reading test was adopted to measure the online comprehension, and the reaction times (RTs) on the first, second, third critical words and the average RTs of the whole sentence were analysed. In the self-paced reading, if the participants were sensitive to the violation of the picture and the sentence, they would have the slower RTs in mismatched items compared to the matched items. For NSs and L2 learners, they were expected to be more sensitive to the violation in the use of the SRCs compared to the ORCs.

In the analysis of the self-paced reading test, except the descriptive analysis which was based on the raw RTs, the other analyses (including visuals, effect sizes and inferential statistical analysis) were based on the residual RTs.

4.1.2.1 Descriptive analysis

The table 4.1.5 and table 4.1.6 present the mean (SDs) RTs of the matched items, mismatched items and the differences between matched and mismatched items (mean RTs of the mismatched items subtract from matched items) of each type of relative clause for NSs and L2 learners respectively. The results indicated that NSs and L2 learners had slower RTs in mismatched items compared to the matched items at the third critical word (SRCs: noun, ORCs: verb) for SRC-A, SRC-I and ORC-A. For ORC-I, the NSs were observed to have slower RTs for mismatch items across the three critical words, though the RT difference between mismatched and matched items was small, which indicated that sensitivity might just be due to chance. In addition, the L2 learners did not show sensitivity to the mismatched items for ORC-I structure, though, at the third critical word, the RT difference between mismatched and matched items was slightly over zero.

Table 4. 1. 5 Mean (SDs) reaction times of native English speakers in self-paced reading

Type	word	matched	mismatched	mismatched-matched
SRC-A (k=5)	1st critical word	377.97 (186.56)	358.14 (197.66)	-19.82
	2nd critical word	382.64 (201.63)	361.67 (189.90)	-20.97
	3rd critical word	436.76 (295.94)	478.43 (435.62)	41.68
	Whole sentence	481.44 (274.72)	432.71 (242.19)	-48.73
SRC-I (k=5)	1st critical word	384.63 (181.99)	370.12 (196.98)	-14.51
	2nd critical word	361.79 (158.18)	366.78 (172.37)	4.99
	3rd critical word	402.53 (211.65)	517.80 (423.66)	115.27
	Whole sentence	461.95 (255.86)	456.64 (247.63)	-5.31
ORC-A (k=5)	1st critical word	355.82 (156.79)	358.19 (164.46)	2.37
	2nd critical word	398.59 (245.54)	357.13 (167.35)	-41.46
	3rd critical word	455.11 (331.19)	492.53 (374.42)	37.43
	Whole sentence	429.00 (227.68)	451.10 (241.60)	22.10
ORC-I (k=5)	1st critical word	339.93 (143.06)	365.43 (156.12)	25.50
	2nd critical word	355.32 (166.49)	379.48 (188.73)	24.16
	3rd critical word	472.91 (390.28)	502.17 (379.25)	29.25
	Whole sentence	438.68 (232.63)	472.92 (249.47)	34.24

Table 4. 1. 6 Mean (SDs) reaction times of L2 learners in self-paced reading

Type	word	matched	mismatched	mismatched-matched
SRC-A (k=5)	1st critical word	423.00 (250.06)	435.66 (230.05)	12.66
	2nd critical word	396.88 (206.06)	410.46 (185.10)	13.58
	3rd critical word	450.54 (271.88)	509.52 (359.05)	58.99
	Whole sentence	451.70 (192.47)	501.72 (226.75)	50.02
SRC-I (k=5)	1st critical word	446.75 (272.02)	448.74 (235.31)	1.99
	2nd critical word	417.78 (207.32)	423.53 (203.57)	5.75
	3rd critical word	448.77 (252.35)	537.99 (435.63)	89.22
	Whole sentence	484.23 (219.76)	521.25 (251.38)	37.03
ORC-A (k=5)	1st critical word	379.77 (173.50)	369.51 (157.70)	-10.26
	2nd critical word	429.36 (213.18)	415.53 (205.59)	-13.84
	3rd critical word	484.61 (311.19)	523.79 (363.97)	39.18
	Whole sentence	477.60 (203.88)	503.55 (223.98)	25.95
ORC-I (k=5)	1st critical word	401.41 (184.23)	377.52 (149.92)	-23.89
	2nd critical word	461.55 (262.93)	416.10 (236.50)	-45.45
	3rd critical word	502.60 (315.63)	512.03 (339.65)	9.43
	Whole sentence	499.24 (228.30)	506.52 (241.84)	7.28

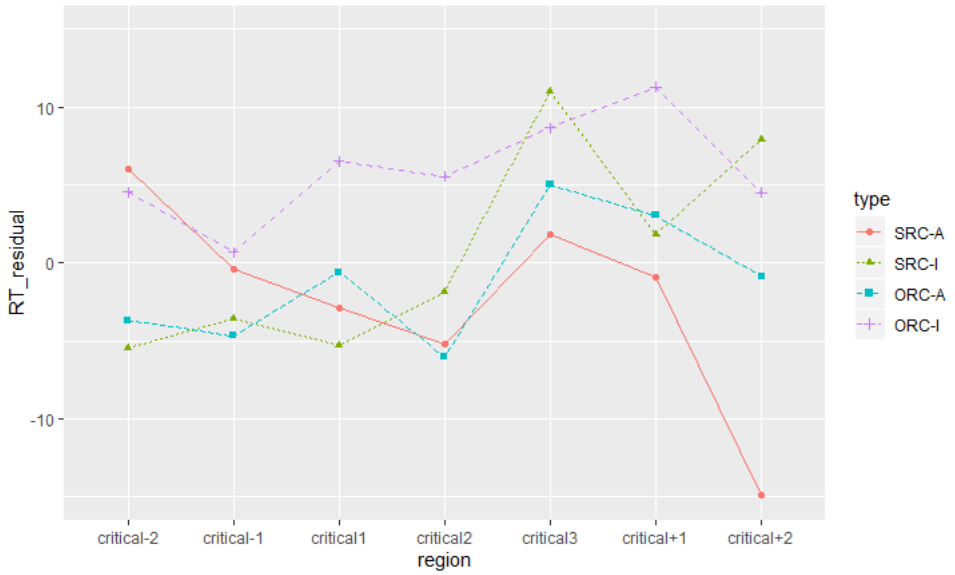
4.1.2.2 Plots

Figure 4.1.2 presents the line charts of the residual RT differences (subtract mean residual RTs of mismatched items from that of the matched items) in each type of relative clauses for NSs and the L2 learners respectively.

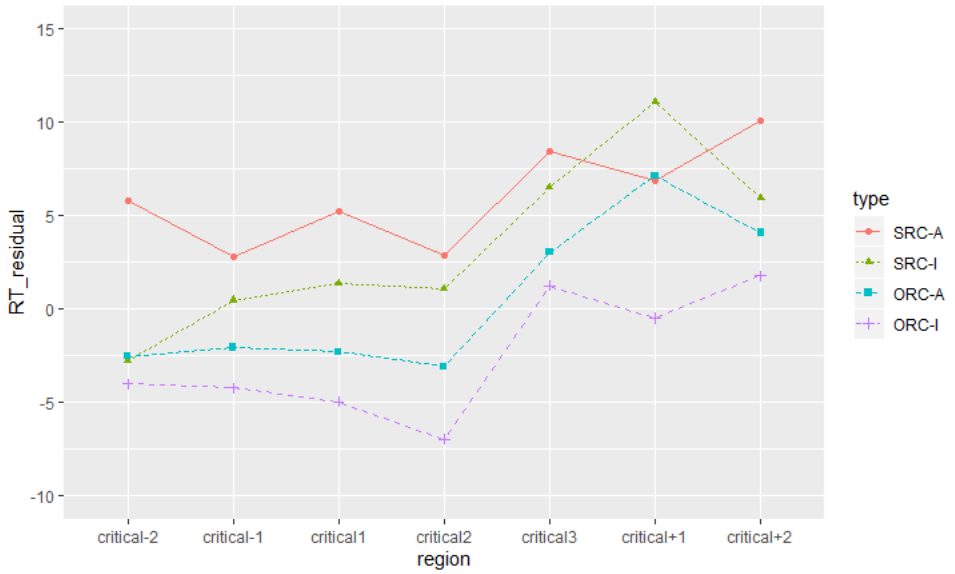
For the NSs, the line chart indicated that they became sensitive to the violation of sentence-picture matching in the SRC-A, SRC-I and ORC-A items at the third critical word. At the third critical word, residual RT differences between mismatched items and matched items were greater than zero, and the increase of the residual differences from the second critical word to the third critical word was salient. In addition, at the third critical word, the residual difference of the SRC-I was the biggest among the SRC-A, SRC-I, and ORC-A structures, and the difference in ORC-A structure was bigger than that of SRC-A. For ORC-I structure, the NSs showed slower residual RTs in mismatched items compared to the matched items in all the three critical words and even in reading the whole sentence. However, the residual RT differences did not have salient change across the critical words. The increase of RT difference for ORC-I could be observed at the next word after the critical regions.

For the L2 learners, the reading patterns of the four types of relative clauses were similar. The residual RT differences had salient increase at the third critical word compared to those of previous words. At the three critical words, all the type of relative clauses had slower residual RTs in mismatched items relative to matched items, and the residual RT differences were bigger for SRCs compared to the ORCs. However, for ORC-I, the residual RT difference was very small, and was just over zero. In addition, after the critical regions, for SRCs and ORC-A, the L2 learners still had slower residual RTs in mismatched compared to the matched items. For SRC-I and ORC-A structure, the maximal residual RT differences were found at the next word following the critical regions.

Figure 4. 1. 2 Comparisons of residual reaction time (RT) differences (mismatched RTs – matched RTs) in each type of relative clause for native English speakers and L2 learners in self-paced reading



Native Speakers



L2 learners

Note: Example of SRC-A: The cat (critical-2) that (critical-1) chases (critical1) the (critical2) dog (critical3) is (critical+1) big (critical+2);
Example of SRC-I: The cat (critical-2) that (critical-1) the (critical1) dog (critical2) chases (critical3) is (critical+1) big (critical+2);
Example of ORC-A: The car (critical-2) that (critical-1) chases (critical1) the (critical2) bike (critical3) is (critical+1) red (critical+2);
Example of ORC-I: The car (critical-2) that (critical-1) the (critical1) bike (critical2) chases (critical3) is (critical+1) red (critical+2).

4.1.2.3 Examination of effect sizes

The within-group effect sizes, reflecting differences between mismatched and matched items, are presented in table 4.1.7. All the effect sizes were found to be negligible or very small. For the NSs, reliable effects (95% CI did not pass through zero) could be observed at the third critical word of the SRC-I and at the whole sentence of ORC-I, though the effects did not reach the benchmark of the reliable sensitivity to the mismatch for native speakers (*d* = .41 [.29, .54]) found by Avery and Marsden's (2019) meta-analysis.

The L2 learners showed the reliable sensitivity to the sentence-picture anomaly for the SRC-A structure based on the whole sentence (benchmark of reliable sensitivity for L2 learners in anomaly detection *d*= .19 [.09, .19], Avery & Marsden, 2019). In addition, the comparisons between mismatched and matched items were reliable (because the 95% CI did not pass through zero) at the third critical word for SRC-A, and at the third critical word as well as across the whole sentence for SRC-I. For SRCs, the reliable effects observed for the whole sentence might be because the learners were still or becoming more sensitive to the sentence-picture anomaly after the critical regions.

Table 4. 1. 7 Within-group (mismatched vs. matched) effect sizes for (Cohen's *d*) [95% CI] for self-paced reading

		SRC-A (k=5)	SRC-I (k=5)	ORC-A (k=5)	ORC-I (k=5)
NSs	1st critical	-.06 [-.27, .14]	-.16 [-.37, .04]	-.02 [-.23, .18]	.16 [-.04, .37]
	2nd critical	-.12 [-.33, .08]	-.08 [-.29, .13]	-.15 [-.35, .06]	.15 [-.06, .36]
	3rd critical	.02 [-.18, .23]	**.24 [.04, .45]**	.10 [-.10, .31]	.18 [-.03, .39]
	whole	-.19 [-.39, .01]	.02 [-.18, .22]	.09 [-.12, .29]	**.24 [.03, .44]**
L2 learners	1st critical	.10 [-.01, .20]	.03 [-.08, .13]	-.06 [-.16, .04]	-.12 [-.22, .01]
	2nd critical	.07 [-.04, .17]	.03 [-.08, .13]	-.06 [-.16, .04]	-.13 [-.24, -.03]
	3rd critical	**.13 [.03, .24]**	**.12 [.01, .22]**	.05 [-.05, .16]	.02 [-.09, .13]
	whole	**.23 [.12, .33]**	**.13 [.02, .24]**	.06 [-.04, .17]	.02 [-.09, .12]

Note: Bold typeface indicates that the CIs did not pass through zero

4.1.2.4 Inferential statistical results

The inferential statistical tests were conducted to test the residual RT differences between each pair of relative clauses. The tests were run respectively for each critical word and the whole sentence. The AIC and LRT results of models for NSs and L2 learners at each critical words and whole sentence are shown in Appendix 18.

Native English Speakers

At the first critical word, the LRT and AIC tests suggested that the primary model (see table 4.1.8), which only included the by-subject and by-item random intercepts, was selected as the best-fitting model. A statistically significant effect was found with the interaction between type (ORC-I vs. SRC-I) and match or mismatch (match vs. mismatch), which indicated that compared to the ORC-I, the residual RTs for SRC-I were faster in mismatched items relative to matched items. In addition, the NSs were more sensitive to the violation in the use of ORC-I than the SRC-A, as the 95% CI of the estimate *b* did not pass through zero.

Table 4. 1. 8 The fixed effects (related to the interactions between type and match or mismatch) of the model analysis of residual reaction times (first critical word) for native English speakers in self-paced reading test

Fixed effects	Estimate [CI]	SE	*t*-value	*p*-value
ORC-I vs. ORC-A: match vs. mismatch	-7.15 [16.24, 1.95]	5.53	-1.29	.198
ORC-I vs. SRC-A: match vs. mismatch	**-9.38 [-18.48, -.29]**	**5.53**	**-1.70**	**.092**
ORC-I vs. SRC-I: match vs. mismatch	**-11.82 [-20.93, -2.70]**	**5.54**	**-2.13**	**.035***
ORC-A vs. SRC-A: match vs. mismatch	-2.23 [-11.38, 6.92]	5.56	-.40	.689
ORC-A vs. SRC-I: match vs. mismatch	-4.67 [-13.84, 4.50]	5.58	-.84	.404
SRC-I vs. SRC-A: match vs. mismatch	2.43 [-6.74, 11.61]	5.58	.44	.663

Note: Model formula: model1=lmer(resid ~ type*match_mismatch + (1|subject) + (1|item), data=SPR, control = lmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); marginal R^2 = .01, conditional R^2 = .03; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$

At the second critical word (see table 4.1.9), the primary model, which included the by-subject and by-item random intercepts, was adopted. The interactions between the type (ORC-I vs. ORC-A & SRC-A) and the match of mismatch (match vs. mismatch) were reliable (95% CIs of the estimate *b* did not pass through zero), though the *p*

values were more than .05, which indicated that the NSs were more sensitive to the violation in the use of the ORC-I compared to the ORC-A and SRC-A.

Table 4. 1. 9 The fixed effects (related to the interactions between type and match or mismatch) of the model analysis of residual reaction times (second critical word) for native English speakers in self-paced reading test

Fixed effects	Estimate [CI]	SE	*t*-value	*p*-value
ORC-I vs. ORC-A: match vs. mismatch	**-11.93 [-22.06, -1.79]**	**6.16**	**-1.94**	**.055**
ORC-I vs. SRC-A: match vs. mismatch	**-10.92 [-21.03, -.82]**	**6.14**	**-1.78**	**.078**
ORC-I vs. SRC-I: match vs. mismatch	-7.77 [-17.92, 2.39]	6.17	-1.26	.210
ORC-A vs. SRC-A: match vs. mismatch	1.00 [-9.07, 11.07]	6.12	.16	.870
ORC-A vs. SRC-I: match vs. mismatch	4.16 [-5.95, 14.28]	6.15	.68	.500
SRC-I vs. SRC-A: match vs. mismatch	-3.16 [-13.24, 6.93]	6.13	-0.52	.607

Note: Model formula: model1=lmer(resid ~ type*match_mismatch + (1|subject) + (1|item), data=SPR, control = lmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); marginal R^2=.01, conditional R^2=.06; bold typeface indicates a reliable effect

At the third critical word, the model (see table 4.1.10), which included the by-subject random slope of the match or mismatch and the by-item random intercept was the best-fitting one. No statistically significant effect was found by this model.

Table 4. 1. 10 The fixed effects (related to the interactions between type and match or mismatch) of the model analysis of residual reaction times (third critical word) for native English speakers in self-paced reading test

Fixed effects	Estimate [CI]	SE	*t*-value	*p*-value
ORC-I vs. ORC-A: match vs. mismatch	-4.13 [-16.93, 8.67]	7.78	-.53	.596
ORC-I vs. SRC-A: match vs. mismatch	-7.17 [-19.83, 5.50]	7.70	-.93	.354
ORC-I vs. SRC-I: match vs. mismatch	2.05 [-10.66, 14.76]	7.73	.27	.791
ORC-A vs. SRC-A: match vs. mismatch	-3.04 [-15.67, 9.60]	7.68	-.40	.693
ORC-A vs. SRC-I: match vs. mismatch	6.18 [-6.50, 18.85]	7.71	.80	.424
SRC-I vs. SRC-A: match vs. mismatch	-9.21 [-21.75, 3.32]	7.62	-1.21	.229

Note: Model formula: model5_3=lmer(resid ~ type*match_mismatch + (match_mismatch|subject) + (1|item), data=SPR, control = lmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); marginal R^2=.01, conditional R^2=.10

For the whole sentence (see table 4.1.11), the primary model including the by-subject and by-items random intercepts was selected. A statistically significant effect was found with the interaction between the type (ORC-I vs. SRC-A) and the

match or mismatch (match vs. mismatch). In addition, the interaction between the type (ORC-A vs. SRC-A) and the match or mismatch (match vs. mismatch) were found to be reliable, as the 95% CI of the estimate *b* did not pass through zero. The results indicated that the NSs were more sensitive to the violation in the ORC-I and ORC-A items relative to SRC-A.

Table 4. 1. 11 The fixed effects (related to the interactions between type and match or mismatch) of the model analysis of residual reaction times (whole sentence) for native English speakers in self-paced reading test

fixed effects	Estimate [CI]	SE	t-value	p-value
ORC-I vs. ORC-A: match vs. mismatch	-5.81 [-15.65, 4.02]	5.98	-.97	.332
ORC-I vs. SRC-A: match vs. mismatch	**-16.92 [-26.68, -7.16]**	**5.93**	**-2.85**	**.005****
ORC-I vs. SRC-I: match vs. mismatch	-8.22 [-17.98, 1.54]	5.93	-1.39	.168
ORC-A vs. SRC-A: match vs. mismatch	**-11.11 [-20.96, -1.26]**	**5.99**	**-1.86**	**.065**
ORC-A vs. SRC-I: match vs. mismatch	-2.41 [-12.26, 7.44]	5.99	-.40	.688
SRC-I vs. SRC-A: match vs. mismatch	-8.70 [-18.47, 1.07]	5.94	-1.46	.145

Note: Model formula: model1=lmer(resid ~ type*match_mismatch + (1|subject) + (1|item), data=SPR, control = lmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); marginal R^2=.02, conditional R^2=.06; bold typeface indicates a reliable effect; ** significantly differently from zero when α < .01

L2 learners

At the first critical word, the primary model that included the by-subject and by-item random intercepts was selected (see table 4.1.12). The results indicated that the L2 learners were significantly more sensitive to the sentence-picture anomaly in the SRC-A compared to the ORC-A and ORC-I. In addition, another reliable effect (the 95% CI of the estimate *b* did not pass through zero) was found with the interaction between the type (ORC-I vs. SRC-I) and the match or mismatch (match vs. mismatch), which suggested that the L2 learners were more sensitive to the anomaly in the SRC-I relative to ORC-I.

Table 4. 1. 12 The fixed effects (related to the interactions between type and match or mismatch) of the model analysis of residual reaction times (first critical word) for L2 learners in self-paced reading test

fixed effects	Estimate [CI]	SE	t-value	p-value
ORC-I vs. ORC-A: match vs. mismatch	2.55 [-3.57, 8.66]	3.72	.69	.494
ORC-I vs. SRC-A: match vs. mismatch	**10.16 [4.02, 16.30]**	**3.73**	**2.72**	**.007****
ORC-I vs. SRC-I: match vs. mismatch	**6.22 [.10, 12.34]**	**3.72**	**1.67**	**.096**
ORC-A vs. SRC-A: match vs. mismatch	**7.61 [1.49, 13.74]**	**3.73**	**2.04**	**.043***
ORC-A vs. SRC-I: match vs. mismatch	3.68 [-2.43, 9.78]	3.71	.99	.324
SRC-I vs. SRC-A: match vs. mismatch	3.94 [-2.20, 10.07]	3.73	1.06	.293

Note: Model formula: model1=lmer(resid ~ type*match_mismatch + (1|subject) + (1|item), data=SPR, control = lmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); marginal R^2=.01, conditional R^2=.02; bold typeface indicates a reliable effect; *significantly differently from zero when α≤ .05; ** significantly differently from zero when α < .01

At the second critical word, the model that included the by-subject random slope of the relative clause type and the by-item random intercept was selected as the best-fitting model. The L2 learners were found to have statistically slower residual RTs in the mismatched items than in the matched items for SRC-A and SRC-I compared to the ORC-I.

Table 4. 1. 13 The fixed effects (related to the interactions between type and match or mismatch) of the model analysis of residual reaction times (second critical word) for L2 learners in self-paced reading test

Fixed effects	Estimate [CI]	SE	*t*-value	*p*-value
ORC-I vs. ORC-A: match vs. mismatch	4.01 [-2.12, 10.13]	3.72	1.08	.283
ORC-I vs. SRC-A: match vs. mismatch	**9.76 [3.65, 15.88]**	**3.72**	**2.63**	**.009****
ORC-I vs. SRC-I: match vs. mismatch	**7.93 [1.81, 14.06]**	**3.73**	**2.13**	**.035****
ORC-A vs. SRC-A: match vs. mismatch	5.76 [-.32, 11.84]	3.70	1.56	.122
ORC-A vs. SRC-I: match vs. mismatch	3.93 [-2.16, 10.01]	3.70	1.06	.290
SRC-I vs. SRC-A: match vs. mismatch	1.83 [-4.25, 7.91]	3.70	.50	.621

Note: Model formula: model4_3=lmer(resid ~ type*match_mismatch + (type|subject) + (1|item), data=SPR, control = lmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); marginal R^2=.01, conditional R^2=.05; bold typeface indicates a reliable effect; *significantly differently from zero when α≤ .05; ** significantly differently from zero when α < .01

At the third critical word, the model that included the by-subject random slope of

match or mismatch and the by-item random intercepts was the best-fitting model. There was no statistically significant effect that could be found in this model.

Table 4. 1. 14 The fixed effects (related to the interactions between type and match or mismatch) of the model analysis of residual reaction times (third critical word) for L2 learners in self-paced reading test

Fixed effects	Estimate [CI]	SE	*t*-value	*p*-value
ORC-I vs. ORC-A: match vs. mismatch	1.63 [-7.30, 10.55]	5.43	.30	.765
ORC-I vs. SRC-A: match vs. mismatch	6.56 [-2.31, 15.43]	5.39	1.22	.225
ORC-I vs. SRC-I: match vs. mismatch	4.86 [-4.04, 14.76]	5.41	.90	.370
ORC-A vs. SRC-A: match vs. mismatch	4.94 [-3.92, 13.79]	5.38	.92	.361
ORC-A vs. SRC-I: match vs. mismatch	3.23 [-5.65, 12.11]	5.40	.60	.550
SRC-I vs. SRC-A: match vs. mismatch	1.70 [-7.12, 10.52]	5.36	.32	.751

Note: Model formula: model5_3=lmer(resid ~ type*match_mismatch + (match_mismatch|subject) + (1|item), data=SPR, control = lmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); marginal R^2=.01, conditional R^2=.06

For the whole sentence, the primary model including the by-subject and by-item random intercepts were suggested to the best-fitting model. The interaction between the type (ORC-I vs. SRC-A) and match or mismatch (match vs. mismatch) were found to be statistically significant, and another interaction between the type (ORC-A vs. SRC-A) and match or mismatch (match vs. mismatch) were reliable (95% CI of the estimate *b* did not pass through zero). The results suggested that the L2 learners were more sensitive to sentence-picture anomaly in the SRC-A compared to the ORC-A and ORC-I.

Table 4. 1. 15 The fixed effects (related to the interactions between type and match or mismatch) of the model analysis of residual reaction times (whole sentence) for L2 learners in self-paced reading test

Fixed effects	Estimate [CI]	SE	*t*-value	*p*-value
ORC-I vs. ORC-A: match vs. mismatch	1.67 [-12.11, 5.65]	4.14	.40	.688
ORC-I vs. SRC-A: match vs. mismatch	**8.76 [-13.76, 4.04]**	**4.14**	**2.12**	**.036***
ORC-I vs. SRC-I: match vs. mismatch	4.44 [-7.12, 10.52]	4.14	1.07	.286
ORC-A vs. SRC-A: match vs. mismatch	**7.10 [.30, 13.90]**	**4.13**	**1.72**	**.088**
ORC-A vs. SRC-I: match vs. mismatch	2.77 [-4.04, 9.58]	4.14	.67	.505
SRC-I vs. SRC-A: match vs. mismatch	4.33 [-2.48, 11.13]	4.14	1.05	.297

Note: Model formula: model1=lmer(resid ~ type*match_mismatch + (1|subject) + (1|item), data=SPR, control = lmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); marginal R^2=.01, conditional R^2=.05; bold typeface indicates a reliable effect; *significantly differently from zero when $\alpha \leq .05$

4.1.2.5 Summary of the results in the self-paced reading test

For NSs, the descriptive analysis of raw RTs and the visualisations of the residual RTs indicated that the NSs were more sensitive to the sentence-picture anomaly in the SRCs relative to the ORCs at the third critical word. On the other hand, the inferential statistical results suggested that at the first critical word and for the whole sentence, the NSs tended to be more sensitive to the anomaly in the ORC-I compared to the SRCs. However, these statistically significant effects might not have practical meaning, because before the critical regions, the NSs already had slower residual RTs in the mismatched items compared to the matched items, while this phenomenon did not show in other structures. It might just happen by chance, as the participants were not able to decide whether the sentence matched the picture or not before the coming of the first critical word. In addition, for ORC-I, at the first critical word, although the NSs had slower residual RTs, the effect size, reflecting residual RT differences between mismatched items and matched items, were negligible. Thus, the statistical results might not be sufficient evidence to support the advantage of ORCs over the SRCs at the first critical word and the whole sentence. Looking at the visuals (figure 4.1.2), all the four types of relative clauses had the salient increase of the residual RT differences (mismatched minus matched) at the third critical word, so at the third critical word, the inferential statistical results which related to the comparison between the types of relatives clauses were more likely to have explanatory power than at the other regions. Nevertheless, no statistically significant effects could be observed at the third critical word. In sum, the descriptive analysis and the line charts showed that the NSs became sensitive to the sentence-picture mismatch at the third critical word for SRCs and ORC-A, and the within-group effect size (comparing the differences between matched and mismatched residual RTs) was reliable for SRC-I at the third critical word. Thus, there was some tentative evidence of a tendency for SRCs especially SRC-I to be easier than ORCs for NSs in the SPR test, though it did not have statistically significant effect.

For L2 learners, similar to the NSs, the models that compared the four types of relative clauses at the third critical word were more meaningful to explain the difficulty

differences between the different types of relative clauses relative to the models for other regions. It is because the descriptive analysis and the line chats showed that the RT differences (RTs for mismatched minus matched) had notably increase at the third critical word for all the structures. At the third critical region, the effect sizes of the two SRCs were reliable, though no statistically significant effect could be found. However, the statistically significant advantages of the SRCs over ORCs were found at the first two critical regions, but it might just happen by chance. It was because the effect size results showed that the residual RT differences between mismatched and matched items were negligible for all the structures at the first and the second critical words. Thus, it could be inferred that there was some tentative evidence that L2 learners tended to show some increased sensitivity to the sentence-picture anomaly for SRCs relative to ORCs, though no statistically significant effects could be found.

4.1.3 Which type of relative clause (SRC vs. ORC) is more difficult in online comprehension measured by eye-tracking?

The visual world eye-tracking tests were used to measure the listening online comprehension. In each item of the eye-tracking test, the participant listened to a sentence and saw a target picture, a distractor, and a verb on the screen. In the eye-tracking test, the SRCs were predicted to be easier than the ORCs for both NSs and L2 learners. To be specific, the participants were expected to fixate on the target picture earlier in SRCs compared to ORCs. The data of the NSs and the L2 learners were analysed separately.

The fixation proportions of looking at the targets and distractors at the critical regions for each type of relative clause were presented through line charts and analysed by inferential statistics. Each line chart included a line for looking at the target and a line for looking at the distractor, describing the changes of the fixation proportion of a type of relative clause from 200ms prior to the onset of the first critical word and lasting for 3000ms. The offsets of the first, second and the third critical words were marked by vertical lines.

In addition, mixed-effects growth curve analysis was used to test whether the

fixations towards the targets over the duration of the three critical words between structures were statistically significant or not. The time window was defined from the onset of the first critical word to the offset of the third critical word. First, the models that had the fixed effects of the interaction between the relative clause type and the time vector and the by-subject and by-item random intercepts were run. In order to find the appropriate time vector, the first-, second-, and third-order time vectors were added step by step, and the models were compared using AIC and LRT. For NSs, the AIC and LRT suggested the first-order time vector fitted the data best. For L2 learners, the results of the AIC and LRT were conflicted. The AIC showed that the model with the first-time order vector was the best, while the LRT indicated the model with the combination of the first-, second- and third-order time vectors was the best-fitting model. Looking at the plots, the curves for the looking at the targets of the four structures had more than one bend, so the combination of the first-, second-, and the third-order time vector might be better than the first-order time vector in fitting the data of the L2 learners. Then, the random slopes were added to the selected models (the AIC and LRT results for selecting time vector and random effects for both NSs and L2 learners see Appendix 19).

The model for NSs, following the suggestion of maximal random effects structures, included the by-subject and by-item random intercepts and the slopes of the first-order time vector and the relative clause type. To avoid over fitting of the model, the random slopes were trimmed down step by step until the model only had random intercepts, and the models were examined by AIC and LRT.

For L2 learners, the model that had the maximal random effects was too complex to converge. The maximal random effects model that could be converged included the by-subject and by-item random intercepts, and the slopes of one time vector and the relative clauses type. The models that included each possible option of the slopes were run separately and tested by AIC and LRT. The AIC and LRT suggested that the model with the by-subject and by-object random intercepts was the best-fitting model.

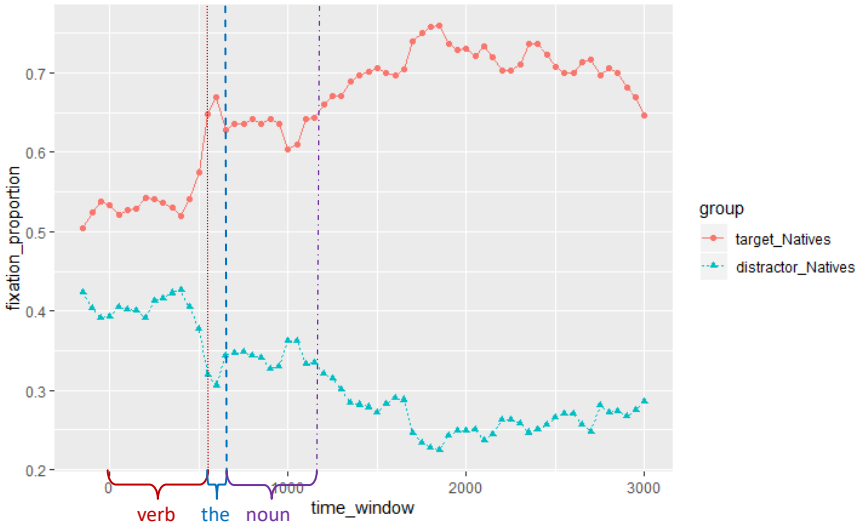
However, the R^2 of the models for NSs and L2 learners were extremely low

(Marginal $R^2_{model\ NSs}$ =.00; Conditional $R^2_{model\ NSs}$ = .01; Marginal $R^2_{model\ L2s}$ =.00; Conditional $R^2_{model\ L2s}$ = .01), thus the data might not be explained very well by the independent and random variables included in the model.

For the NSs, figure 4.1.3 presents the fixation proportions of looking at the targets and the distractors of the four types of the relative clauses. The plots indicated that for the SRCs, the NSs could fixate on the target pictures before the offset of the first critical word. For SRC-A, although the participants had higher proportion of looking at the targets compared to looking at the distractors before the onset of the first critical word, a steady increase in the line of the looking at the target could be observed at around 400ms. Compared to SRCs, the divergent point of looking at the target and looking at the distractor was later in ORCs. For ORC-A, the point that the proportion of looking at the target exceeded .50 was around the end of the second critical word (around 600ms after the onset of the first critical word). For ORC-I, the line of looking at the target diverged from looking at the distractor around 800ms, and this point was within the third critical word. In sum, the plots indicated that the time that the NSs became sensitive to the target structures was earlier in SRCs relative to ORCs. However, the inferential statistical results did not show any statistically significant difference between structures in the examination of the differences in looking at the targets (see table 4.1.16).

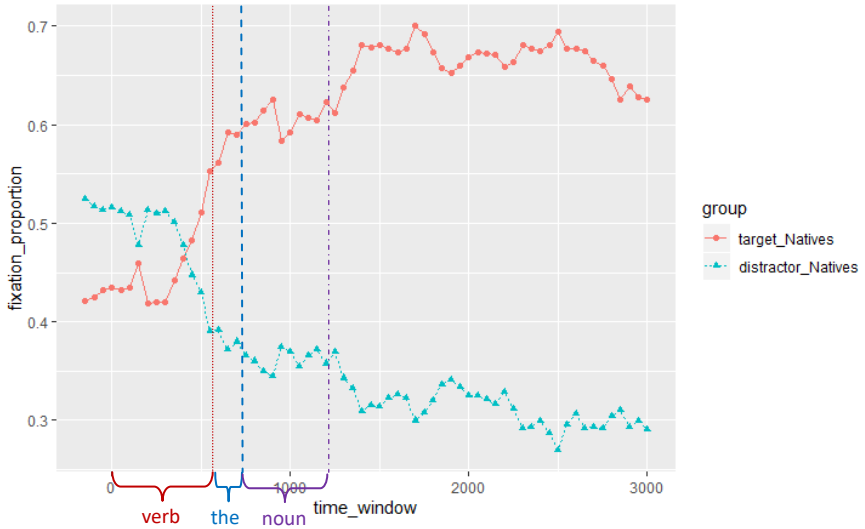
For L2 learners, the fixation proportions of looking at the targets and the distractors were shown in figure 4.1.4. For the SRCs, the L2 learners could fixate at the target pictures before the end of the first critical word (around 450ms after the onset of the first critical word) for SRC-A, but the divergent point occurred later in SRC-I, which showed at the third critical word (around 800ms after the first critical word). For ORCs, the plots showed that the L2 learners could only become sensitive to the language cues of ORCs at the third critical word (around 800ms after the first critical word) regardless of animacy of the first noun. Nevertheless, the inferential statistical results indicated that the differences in the proportion of looking at the targets in the critical words were not significant (see table 4.1.17).

Figure 4. 1. 3 Fixation proportions of looking at the targets and distractors for native English Speakers



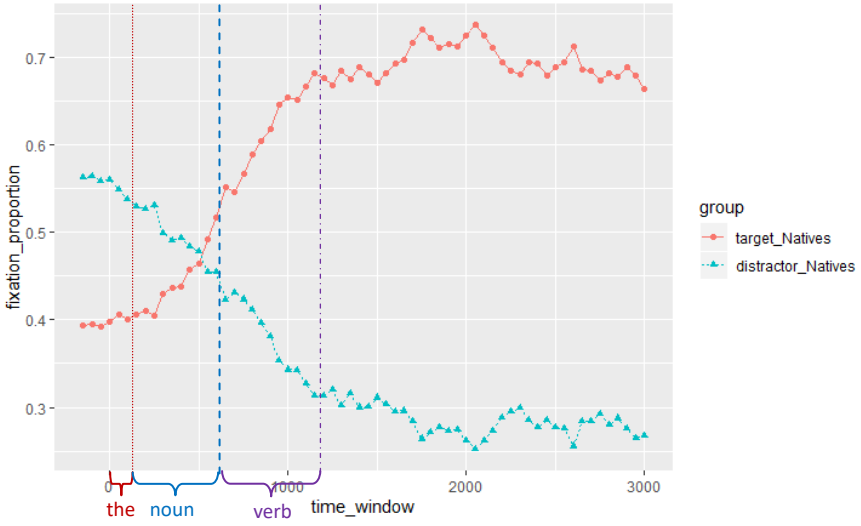
SRC-A

Example SRC-A: The cat that chases the dog is big.



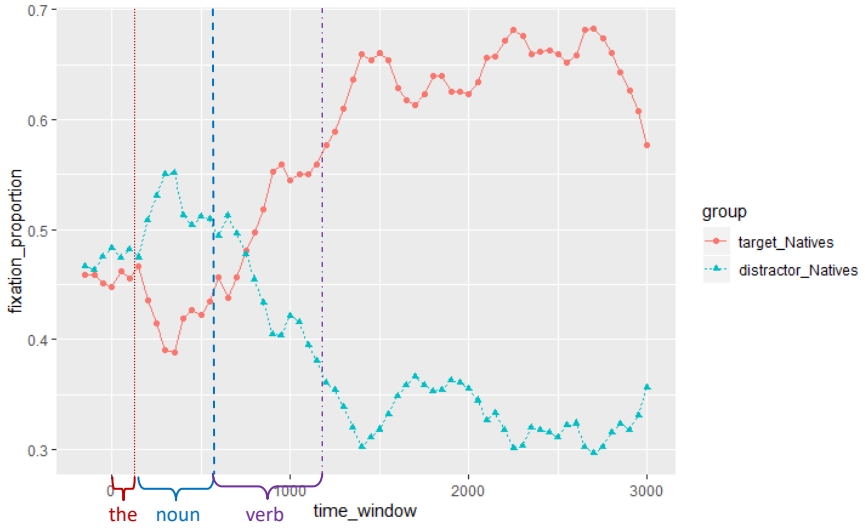
SRC-I

Example SRC-I: The car that chases the bike is red.



ORC-A

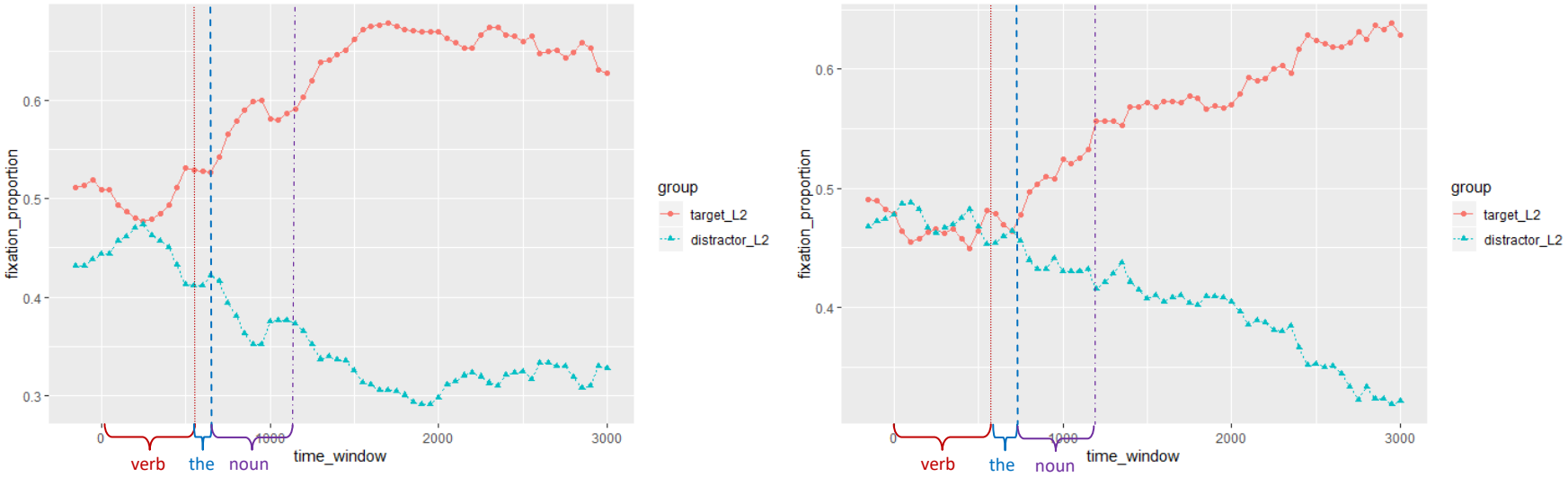
Example ORC-A: The cat that the dog chases is big.



ORC-I

Example ORC-I: The car that the bike chases is red.

Figure 4. 1. 4 Fixation proportions of looking at the targets and distractors for L2 learners

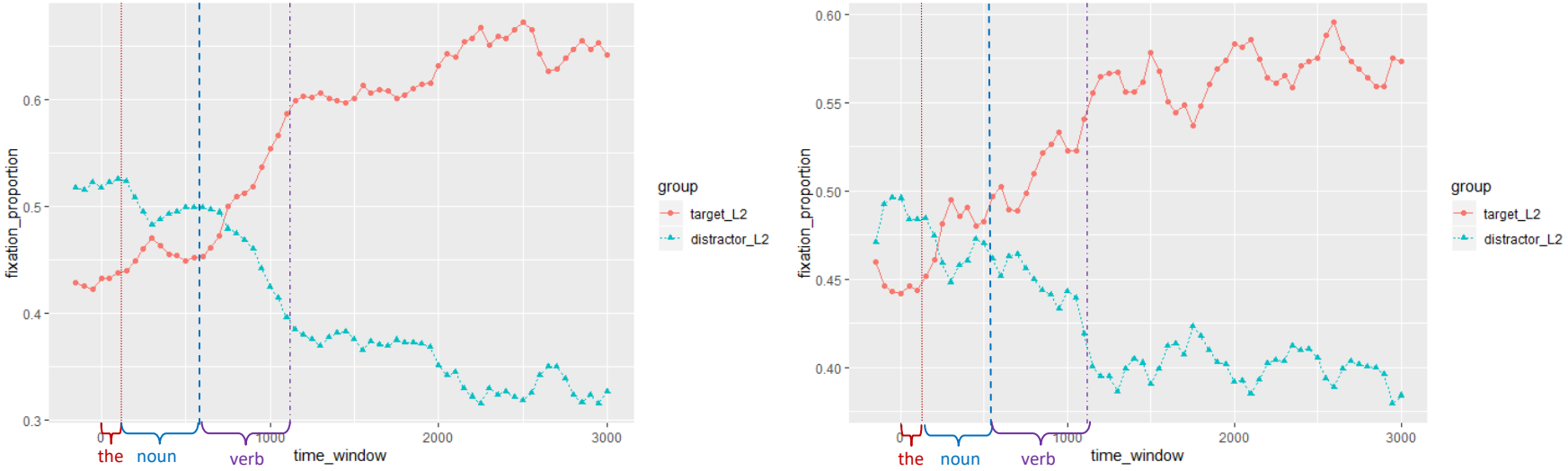


SRC-A

Example SRC-A: The cat that <u>chases the dog</u> is big.

SRC-I

Example SRC-I: The car that <u>chases the bike</u> is red.



ORC-A

Example ORC-A: The cat that <u>the dog chases</u> is big.

ORC-I

Example ORC-I: The car that <u>the bike chases</u> is red.

Table 4. 1. 16 The fixed effects of the model analysis of the empirical log odds of fixation proportion for native English speakers in eye-tracking test

Fixed effects	Estimate [CI]	SE	*t*-value	*p*-value
ORC-I vs. ORC-A	-.02 [-.08, .04]	.04	-.52	.605
ORC-I vs. SRC-A	-.01 [-.07, .05]	.03	-.30	.761
ORC-I vs. SRC-I	-.01 [-.06, .05]	.04	-.17	.862
Liner × ORC-I vs. ORC-A	.15 [-.14, .43]	.17	.86	.388
Liner × ORC-I vs. SRC-A	.17 [-.11, .45]	.17	.99	.322
Liner × ORC-I vs. SRC-I	.21 [-.08, .49]	.17	1.20	.229
ORC-A vs. SRC-A	.01 [-.05, .06]	.03	.23	.819
ORC-A vs. SRC-I	.01 [-.04, .07]	.03	.36	.720
Liner × ORC-A vs. SRC-A	.02 [-.25, .30]	.17	.13	.897
Liner × ORC-A vs. SRC-I	.06 [-.22, .34]	.17	.36	.723
SRC-I vs. SRC-A	.00 [-.06, .05]	.03	-.14	.893
Liner × SRC-I vs. SRC-A	-.04 [-.31, .24]	.17	-.23	.820

Note: Model formula: gca_model1 = lmer(e_log ~ (ot1)*type +
+ (1 | subject) +
+ (1 | item),
+ control = lmerControl(optimizer="bobyqa"),
+ data=eye_data);
Marginal R^2=.00, conditional R^2=.01

Table 4. 1. 17 The fixed effects of the model analysis of the empirical log odds of fixation proportion for L2 learners in eye-tracking test

Fixed effects	Estimate [CI]	SE	*t*-value	*p*-value
ORC-I vs. ORC-A	-.01 [-.04, .02]	.02	-.51	.611
ORC-I vs. SRC-A	-.02 [-.05, .01]	.02	-.92	.358
ORC-I vs. SRC-I	-.01 [-.04, .02]	.02	-.59	.555
Liner × ORC-I vs. ORC-A	-.04 [-.19, .11]	.09	-.42	.676
Liner × ORC-I vs. SRC-A	-.07 [-.22, .09]	.09	-.71	.481
Liner × ORC-I vs. SRC-I	-.04 [-.19, .11]	.09	-.43	.669
Quadratic × ORC-I vs. ORC-A	.13 [-.02, .28]	.09	1.40	.162
Quadratic × ORC-I vs. SRC-A	.11 [-.04, .27]	.09	1.25	.211
Quadratic × ORC-I vs. SRC-I	.02 [-.13, .17]	.09	.22	.830
Cubic × ORC-I vs. ORC-A	-.08 [-.23, .07]	.09	-.90	.367
Cubic × ORC-I vs. SRC-A	-.10 [-.25, .05]	.09	-1.13	.257
Cubic × ORC-I vs. SRC-I	-.02 [-.18, .13]	.09	-.26	.795
ORC-A vs. SRC-A	-.01 [-.04, .02]	.02	-.42	.677
ORC-A vs. SRC-I	.00 [-.03, .03]	.02	-.09	.928
Liner × ORC-A vs. SRC-A	-.03 [-.17, .12]	.09	-.30	.767
Liner × ORC-A vs. SRC-I	.00 [-.14, .15]	.09	-.02	.987

Quadratic × ORC-A vs. SRC-A	-.01 [-.16, .13]	.09	-.16	.872
Quadratic × ORC-A vs. SRC-I	-.11 [-.26, .04]	.09	-1.19	.233
Cubic × ORC-A vs. SRC-A	-.02 [-.17, .13]	.09	-.24	.814
Cubic × ORC-A vs. SRC-I	.06 [-.09, .21]	.09	.65	.518
SRC-I vs. SRC-A	-.01 [-.04, .02]	.02	-.32	.750
Liner × SRC-I vs. SRC-A	-.03 [-.17, .12]	.09	-.28	.780
Quadratic × SRC-I vs. SRC-A	.09 [-.05, .24]	.09	1.04	.300
Cubic × SRC-I vs. SRC-A	-.08 [-.23, .07]	.09	-.88	.380

Note: Model formula: gca_model3 = lmer(e_log ~ (ot1+ot2+ot3)*type +
+ (1 | subject) +
+ (1 | item),
+ control = lmerControl(optimizer="bobyqa"),
+ data=eye_data);
Marginal R^2=.00, conditional R^2=.01

In summary, for both NSs and L2 learners, the inferential statistics showed that the proportion of looking at the targets over the three critical words did not have statistically significant difference between the types of relative clauses. However, differences between structures could be observed in the plots. The line charts showed that the NSs looked at the target picture earlier in SRCs relative to ORCs, and the L2 learners fixated on the targets earlier in the SRC-A compared to the three other conditions: SRC-I, ORC-A, and ORC-I. To sum up, the ORCs seemed to be more difficult to process relative to SRCs for NSs, descriptively but not statistically. For the L2 learners, the SRC-A was easier to interpret than the other three types of relative clauses, and those three types of relative clauses had similar degree of difficulty according to the eye-tracking data.

4.1.4 Which type of relative clause (SRC vs. ORC) is more difficult in oral production?

In the oral production test, the participants were asked to describe a picture based on the given information, and the accuracy scores were used in the analysis. The NSs and L2 learners were expected to have higher accuracy scores for SRCs compared to the ORCs.

4.1.4.1 Descriptive analysis

Table 4.1.18 presents the mean (*SD*s) accuracy scores for the NSs and L2 learners in the four types of relative clauses. It could be observed that for both NSs and L2 learners,

the accuracy scores of the ORCs were lower than those of the SRCs. The NSs scored at ceiling for SRCs, but they could only correctly produce around half of the ORCs. For NSs, the ORCs with inanimate heads had higher accuracy scores compared to those with animate heads. For L2 learners, the animacy of the head noun did not influence the difficulty in producing ORCs.

Table 4. 1. 18 Mean (*SDs*) accuracy scores in oral production of native English Speakers and L2 learners

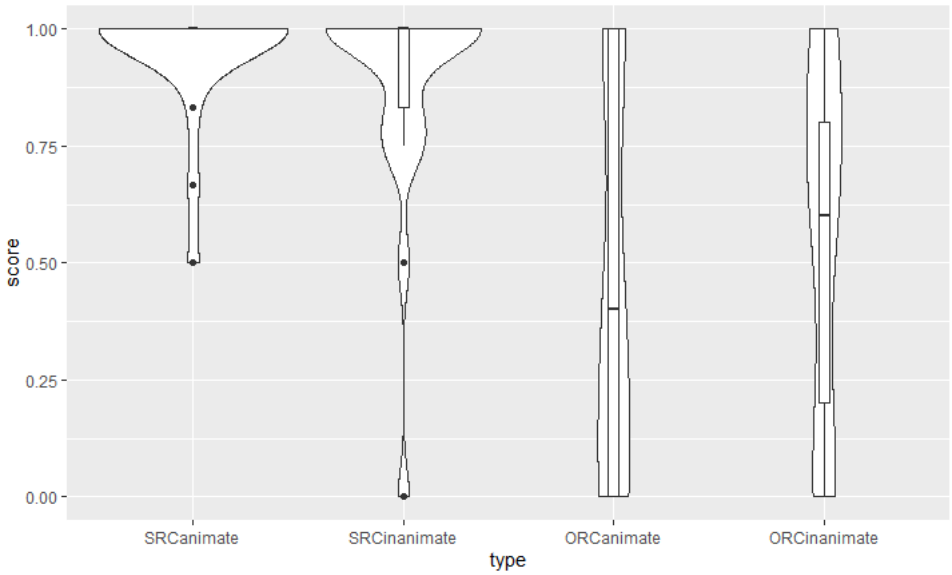
Structure	Natives	L2 learners
SRC-A (k=4 or 6 [a])	.94 (.23)	.51 (.50)
SRC-I (k=4 or 6 [a])	.98 (.31)	.53 (.50)
ORC-A (k=5)	.44 (.50)	.33 (.47)
ORC-I (k=5)	.55 (.50)	.33 (.47)

Note: [a] Numbers of items differed between version 1 and version 2, version 3 and version 4 of the tests, respectively; versions were counterbalanced across pre-, post-, and delayed post-test, within groups.

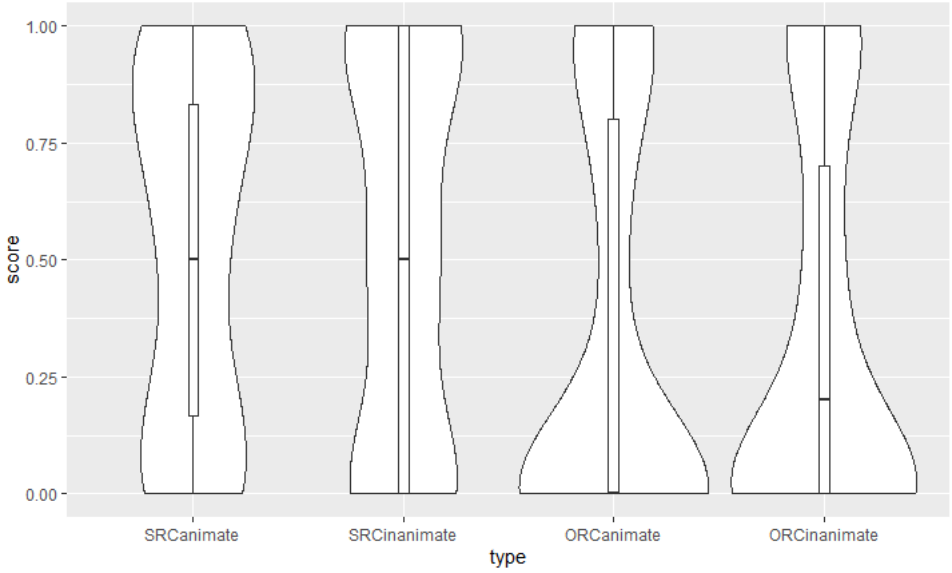
4.1.4.2 Plots

The violin plots with inserted Boxplots presented in the figure 4.1.5 show the comparisons of mean accuracy scores of the four types of relative clauses for NSs and L2 learners respectively. Looking at the plots, the NSs had much larger proportion of participants who scored at ceiling in the SRCs compared to the ORCs. For L2 learners, the proportions of the participants who scored at ceiling were similar across the four types of relative clauses, but the more participants scored at bottom for ORCs compared to the SRCs.

Figure 4. 1. 5 comparisons of mean accuracy scores of each type of relative clauses for native English speakers and L2 learners in oral production



Native English Speakers



L2 learners

4.1.4.3 Examination of effect sizes

Table 4.1.19 presents the within-group effects, reflecting differences between each pair of relative clauses for NSs and L2 learners respectively. For NSs, the higher scores for SRCs over ORCs had small effects, and difference between ORC-A and ORC-I was reliable, because the 95% CI did not pass through zero (though did not reach benchmark of small effects). For L2 learners, the differences in the comparisons between SRCs and ORCs were reliable (95% CI did not pass through zero), but did not

reach Plonsky and Oswald's (2014) benchmark of small effects.

Table 4. 1. 19 Within-group effect sizes for (Cohen's *d*) [95% CI] for oral production

	NSs	L2 learners
SRC-A vs. SRC-I	.13 [-.07, .32]	-.04 [-.14, .06]
SRC-A vs. ORC-A	**.97 [.74, 1.20]**	**.25 [.15, .36]**
SRC-A vs. ORC-I	**.80 [.58, 1.02]**	**.26 [.16, .36]**
SRC-I vs. ORC-A	**.90 [.67, 1.13]**	**.30 [.20, .40]**
SRC-I vs. ORC-I	**.61 [.40, .82]**	**.31 [.21, .41]**
ORC-A vs. ORC-I	**-.23 [-.43, -.04]**	.00 [-.10, .10]

***Note**:* Bold typeface indicates CIs that did not pass through zero

4.1.4.4 Inferential statistical analysis

The inferential statistical analysis was conducted using mixed-effects logistic regression models. The AIC and LRT results for model selection see Appendix 20.

For NSs (see table 4.1.20), the primary model that included the by-subject and by-item random intercepts was selected as the best-fitting model. The results indicated that every comparison between the two types of relative clauses was statistically significant. The odds ratio of the model suggested that the NSs were significantly more likely to produce correct SRC-A compared to ORC-I (168.27 times greater likelihood), to ORC-A (470.87 times greater likelihood) and to SRC-I (5.27 times greater likelihood). In addition, the odds ratio predicted the NSs to be 31.93 and 89.35 more likely to correctly produce SRC-I compared to ORC-I and ORC-A respectively. Moreover, it could be found that the NSs were 2.80 times more likely to correctly produce ORC-I compared to ORC-A. However, the higher likelihood of producing SRC-A relative to SRC-I might not have practical meaning, because the mean accuracy scores of the two structures were at ceiling.

Table 4. 1. 20 The fixed effects of the model analysis of accuracy scores for native English Speakers in oral production

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
ORC-I vs. SRC-A	5.13 [3.62, 6.63]	.91	5.61	<.001***	168.27 [37.45, 756.04]
ORC-I vs. SRC-I	3.46 [2.34, 4.59]	.69	5.05	<.001***	31.93 [10.34, 98.59]
ORC-A vs. ORC-I	1.03 [.21, 1.85]	.50	2.07	.038*	2.80 [1.24, 6.33]
ORC-A vs. SRC-A	6.15 [4.53, 7.78]	.99	6.24	<.001***	470.87 [93.01, 2383.83]
ORC-A vs. SRC-I	4.49 [3.25, 5.73]	.75	5.96	<.001***	89.35 [25.86, 308.77]
SRC-I vs. SRC-A	1.66 [.35, 2.98]	.80	2.08	.038*	5.27 [1.42, 19.62]

Note: Model formula: model4=glmer(score ~ type + (1|subject) + (1|item), data=oral_production, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.35, Conditional R^2=.82; *significantly differently from zero when α≤ .05; ***significantly differently from zero when α<.001

Table 4.1.21 presents the inferential statistical results for L2 learners. The model that included the by-subject random slope of the type of relative clauses and the by-items random intercept was suggested as the best-fitting model. The results indicated that the L2 learners had significantly higher probability in correctly producing SRCs relative to ORCs. The odds ratio predicted the L2 learners to be 8.96 times and 21.14 more likely to correctly produce SRC-A compared to ORC-I and ORC-A respectively; to be 9.13 times and 22.21 more likely to correctly produce SRC-I compared to ORC-I and ORC-A.

Table 4. 1. 21 The fixed effects of the model analysis of accuracy scores for L2 learners in oral production

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
ORC-I vs. ORC-A	-0.89 [-1.90, .12]	0.61	-1.45	.146	.41 [.15, 1.13]
ORC-I vs. SRC-A	2.16 [1.10, 3.22]	0.64	3.36	<.001***	8.69 [3.02, 25.04]
ORC-I vs. SRC-I	2.21 [1.17, 3.25]	0.63	3.49	<.001***	9.13 [3.22, 25.91]
ORC-A vs. SRC-A	3.05 [1.53, 4.58]	0.927	3.291	<.001***	21.14 [4.60, 97.13]
ORC-A vs. SRC-I	3.10 [1.61, 4.60]	0.91	3.41	<.001***	22.21 [4.98, 99.04]
SRC-I vs. SRC-A	-0.05 [-.47, .37]	0.26	-0.19	.848	.95 [.62, 1.45]

Note: Model formula: model3=glmer(score ~ type + (type|subject) + (1|item), data=oral_production, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.09, Conditional R^2=.81; ***significantly differently from zero when α<.001

4.1.4.5 Summary of the results in the oral production test

In summary, the descriptive and inferential statistics showed that SRCs were significantly easier than ORCs to produce for both NSs and L2 learners. In addition, for NSs, the ORC-I was easier than the ORC-A.

4.1.5 Which type of relative clause (SRC vs. ORC) is more difficult in the metalinguistic knowledge test?

The metalinguistic knowledge test consisted of three sections: decide whether the sentence matches the picture or not; if there was a mismatch, correct the sentence to match the picture by moving a word of the sentence; explain the reason why the sentence did not match the picture. The scores of the three sections were analysed separately. The NSs and L2 learners were expected to have higher accuracy scores in SRCs relative to ORCs.

4.1.5.1 Analysis of accuracy scores in deciding whether the sentence match the picture

The first task of the metalinguistic knowledge test was to decide whether the sentence matched the picture. The data of matched items and mismatched items were analysed separately.

a) Descriptive analysis

Table 4.1.22 presents the means (*SD*s) of the NSs and L2 learners in matched and mismatched items. For the matched items, both natives and the L2 learners scored at ceiling for all the four types of relative clauses. For the mismatched items, the NSs scored at ceiling for SRCs and ORC-I, and had slightly lower accuracy scores in ORC-A compared to the other three structures. In addition, the L2 learners scored higher in SRC-A and ORC-A compared to the SRC-I and ORC-I.

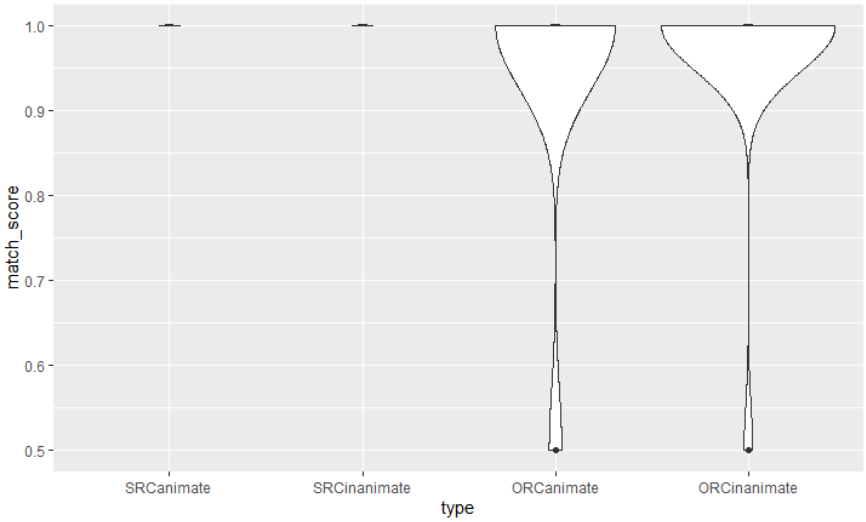
Table4. 1. 22 Mean (*SDs*) accuracy scores in metalinguistic knowledge test (decide whether the sentence matches the picture) of native English Speakers and L2 learners

Structure	Matched items		Mismatched items	
	Natives	L2 learners	Natives	L2 learners
SRC-A (k=2)	1.00 (.00)	.96 (.19)	.93 (.26)	.84 (.37)
SRC-I (k=2)	1.00 (.00)	.97 (.18)	.98 (.15)	.74 (.44)
ORC-A (k=2)	.95 (.22)	.95 (.22)	.81 (.40)	.80 (.40)
ORC-I (k=2)	.98 (.15)	.96 (.21)	.93 (.26)	.61 (.49)

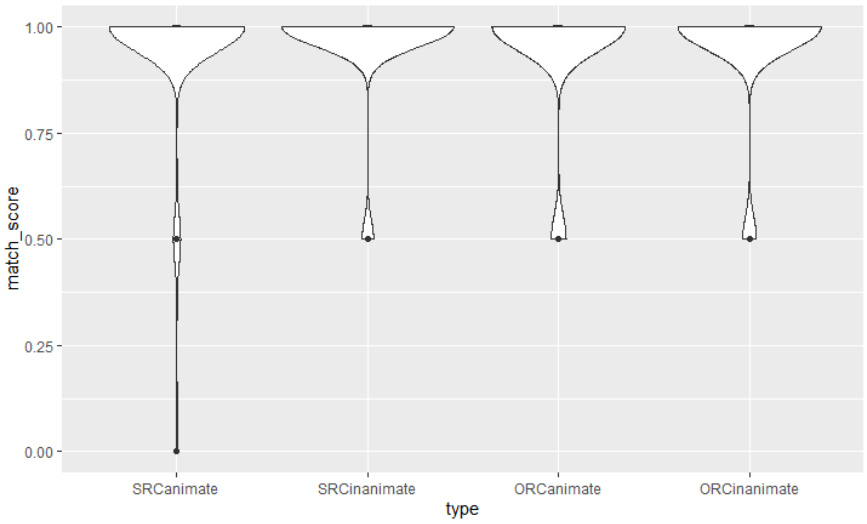
b) Plots

The mean accuracy scores of each type of relative clauses for NSs and L2 learners were presented through violin plots with inserted Boxplots. The plots for matched items are shown in the figure 4.1.6. Looking at the plots, both NSs and the L2 learners scored at ceiling regardless of the relative clauses type, and especially for the NSs who had correctly responded all the SRC-items.

Figure 4. 1. 6 comparisons of mean accuracy scores of each type of relative clauses for native English speakers and L2 learners in matched items of the metalinguistic knowledge test (decide whether the sentence matches the picture)



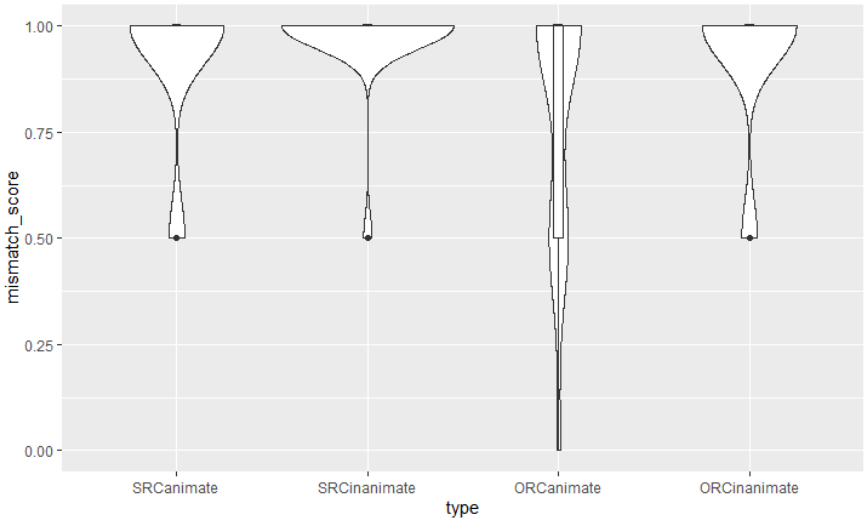
Native Speakers



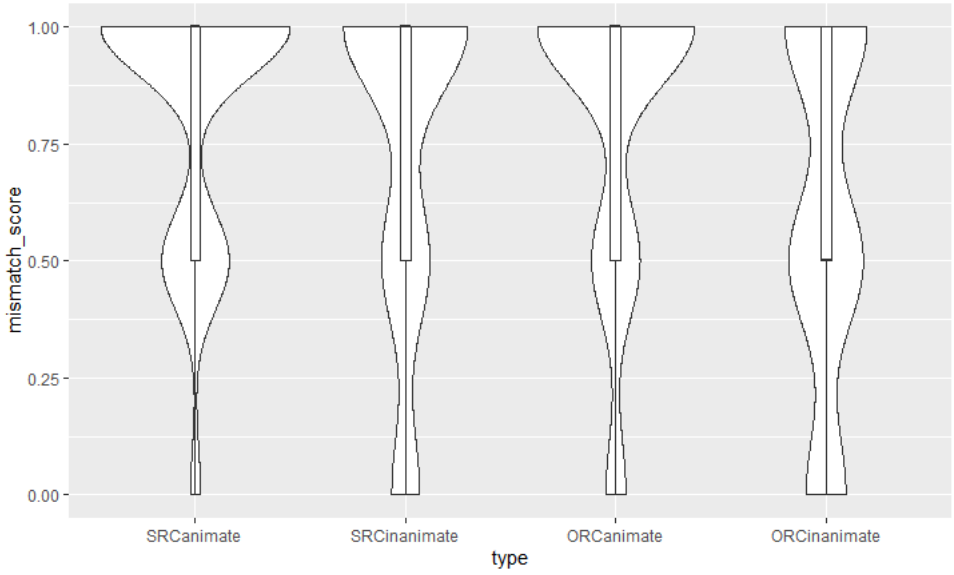
L2 learners

The plots in the figure 4.1.7 present the mean accuracy scores of the four types of relative clauses in the mismatched items. For NSs, the median of the mean accuracy scores for all the four structures was 1.00, however, the proportion of the participants who scored at ceiling was much lower in ORC-A compared to the other three structures. For L2 learners, there were more participants who scored at ceiling for SRC-A relative to the other three structures. Moreover, the L2 learners had the lowest median (.50) of the mean accuracy scores for ORC-I among the four structures, and in ORC-I structure, more participants averagely scored .50 than the participants who averagely scored 1.00.

Figure 4. 1. 7 comparisons of mean accuracy scores of each type of relative clauses for native English speakers and L2 learners in mismatched items of the metalinguistic knowledge test (decide whether the sentence matches the picture)



Native English Speakers



L2 learners

c) Examination of effect sizes

Table 4.1.23 presents the within-group effect sizes, reflecting the differences between structures. It could be found that all the effects in the comparisons between two types of relative clauses in the matched items were negligible, because the 95% CI passed through zero. For the mismatched items, the effects in the NSs and L2 learners could be described as extremely small to negligible. Reliable effects (95% CI did not pass

through zero) could be found with the NSs in the comparison between SRC-I and ORC-A. In addition, in the group of the L2 learners, the comparisons between ORC-I and each of the other three structures (SRC-A, SRC-I and ORC-A) were reliable, though none of them reached the benchmark of small effects.

Table 4. 1. 23 Within-group effect sizes for (Cohen's *d*) [95% CI] in metalinguistic knowledge test (decide whether the sentence matches the picture or not)

	Matched items		Mismatched items	
	Native Speakers	L2 learners	Native Speakers	L2 learners
SRC-A vs. SRC-I	N/A	-.03 [-.19, .13]	-.15 [-.46, .15]	.19 [.04, .35]
SRC-A vs. ORC-A	.22 [-.09, .53]	.04 [-.11, .20]	.26 [-.05, .58]	.07 [-.09, .22]
SRC-A vs. ORC-I	.15 [-.15, .46]	.03 [-.13, .18]	.00 [-.31, .31]	**.41 [.25, .58]**
SRC-I vs. ORC-A	.22 [-.09, .53]	.07 [-.09, .22]	**.38 [.07, .70]**	-.13 [-.28, .03]
SRC-I vs. ORC-I	.15 [-.15, .46]	.06 [-.10, .21]	.15 [-.15, .46]	**.24 [.08, .40]**
ORC-A vs. ORC-I	-.09 [-.40, .22]	-.02 [-.18, .14]	-.24 [-.55, .07]	**.36 [.20, .52]**

Note: Bold typeface indicates that the CIs did not pass through zero; N/A refers to the effect of the comparison could not be estimated, because the *SD* of one or both of the groups was zero

d) Inferential statistical analysis

The mixed-effects logistic regression models were used to examine whether the scores of correctly deciding the match or mismatch of the sentence and the picture were different between the four types of relative clauses. AIC and LRT results for model selection see Appendix 21.

For the matched items, none of the models could be converged in the NSs group, because the variance of the independent variable, the scores, was extremely small (see table 4.1.22). Under this circumstance, it was very unlikely to have the statistical difference between different types of relative clause, thus, the statistical results would not be provided. In addition, for the L2 learners, only the base model which included the by-subject and by-item random intercepts could be converged. No statistically significant effect could be found (see table 4.1.24).

Table 4. 1. 24 The fixed effects of the model analysis accuracy scores of the matched items for L2 learners in metalinguistic knowledge test (decide whether the sentence matches the picture or not)

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
ORC-I vs. ORC-A	-.21 [-1.55, 1.12]	.81	-.26	.792	.81 [.21, 3.07]
ORC-I vs. SRC-A	.20 [-1.20, 1.59]	.85	.23	.817	1.22 [.30, 4.91]
ORC-I vs. SRC-I	.38 [-1.04, 1.81]	.87	.44	.657	1.47 [.35, 6.11]
ORC-A vs. SRC-A	.41 [-.96, 1.78]	.83	.49	.622	1.51 [.38, 5.92]
ORC-A vs. SRC-I	.60 [-.80, 2.00]	.85	.70	.482	1.82 [.45, 7.37]
SRC-I vs. SRC-A	-.19 [-1.64, 1.27]	.88	-.21	.832	.83 [.19, 3.55]

Note: Model formula: model4=glmer(match_score ~ type + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.01, Conditional R^2=.48

For mismatched items, the basic models that included the by-subject and by-item random intercepts were selected at the best-fitting model for the data of both the NSs and L2 learners. For NSs, no statistically significant effect could be found (see table 4.2.25). However, the comparison between ORC-A and SRC-I was found to be reliable (the 95% CI of the estimate *b* did not pass through zero). The odds ratio predicted the NSs to be 13.11 more likely to correctly decide whether the sentence matches the picture in the SRC-I compared to the ORC-A. For L2 learners, the statistically significant effects could be observed in the comparisons between ORC-I and ORC-A and between ORC-I and SRC-I (see table 4.2.26). The odds ratio of the model suggested that the L2 learners were significantly more likely to correctly decide whether the sentence matches the picture in ORC-A (4.61 times greater likelihood) and SRC-A (6.37 times greater likelihood) compared to ORC-I. In addition, the difference between the scores in ORC-I and SRC-I was found to be reliable (the 95% CI of the estimate *b* did not pass through zero). The odds ratio predicted the L2 learners to be 2.93 times more likely to respond correctly in SRC-I compared to the ORC-I.

Table 4. 1. 25 The fixed effects of the model analysis accuracy scores of the mismatched items for native English speakers in metalinguistic knowledge test (decide whether the sentence matches the picture or not)

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
ORC-I vs. ORC-A	-1.40 [-3.34, .54]	1.18	-1.19	.235	.25 [.04, 1.71]
ORC-I vs. SRC-A	-.02 [-2.10, 2.06]	1.26	-.01	.989	.98 [.12, 7.87]
ORC-I vs. SRC-I	1.17 [-1.31, 3.66]	1.51	.78	.437	3.23 [.27, 38.80]
ORC-A vs. SRC-A	1.38 [-.55, 3.32]	1.18	1.18	.240	3.98 [.57, 27.62]
ORC-A vs. SRC-I	**2.57 [.19, 4.95]**	**1.45**	**1.78**	**.075**	**13.11 [1.21, 141.70]**
SRC-I vs. SRC-A	-1.19 [-3.67, 1.29]	1.51	-.79	.430	.30 [.03, 3.64]

Note: Model formula: model4=glmer(mismatch_score ~ type + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.13, Conditional R^2=.49; bold typeface indicates a reliable effect

Table 4. 1. 26 The fixed effects of the model analysis accuracy scores of the mismatched items for L2 learners in metalinguistic knowledge test (decide whether the sentence matches the picture or not)

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
ORC-I vs. ORC-A	**1.53 [.52, 2.54]**	**.61**	**2.49**	**.013***	**4.61 [1.68, 12.62]**
ORC-I vs. SRC-A	**1.85 [.82, 2.88]**	**.63**	**2.96**	**.003****	**6.37 [2.28, 17.84]**
ORC-I vs. SRC-I	**1.08 [.08, 2.07]**	**.60**	**1.78**	**.075**	**2.93 [1.08, 7.92]**
ORC-A vs. SRC-A	.32 [-.72, 1.37]	.64	.51	.610	1.38 [.49, 3.94]
ORC-A vs. SRC-I	-.45 [-1.47, .57]	.62	-.73	.467	.64 [.23, 1.77]
SRC-I vs. SRC-A	.78 [-.26, 1.82]	.63	1.23	.219	2.17 [.77, 6.15]

Note: Model formula: model4=glmer(mismatch_score ~ type + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.05, Conditional R^2=.44; bold typeface indicates a reliable effect; **significantly differently from zero when α<.01; *significantly differently from zero when α<.05

4.1.5.2. Analysis of accuracy scores in correcting the sentence to match the picture

The second task of the metalinguistic knowledge test was to move one word within a mismatched sentence to make the sentence match the picture. Thus, only the data of the mismatched items were analysed in this section.

a) Descriptive analysis

The mean (*SDs*) accuracy scores of the NSs and L2 learners in the sentence correct task are presented in the table 4.1.27. The table showed that the NSs scored higher in SRCs

relative to ORCs, regardless of the animacy of the head noun. For L2 learners, the mean score for ORC-A was higher than the other three structures and the ORC-I had the lowest accuracy among the four types of relative clauses.

Table 4. 1. 27 Mean (*SDs*) accuracy scores of native English Speakers and L2 learners in the sentence correct task of the metalinguistic knowledge test

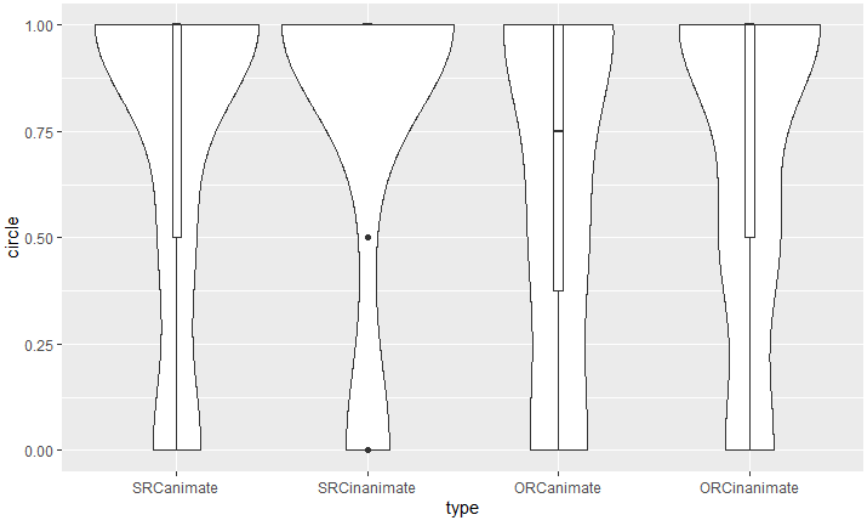
Structure	NSs	L2 learners
SRC-A (k=2)	.75 (.44)	.64 (.48)
SRC-I (k=2)	.78 (.42)	.59 (.49)
ORC-A (k=2)	.64 (.49)	.72 (.45)
ORC-I (k=2)	.68 (.47)	.54 (.50)

b) Plots

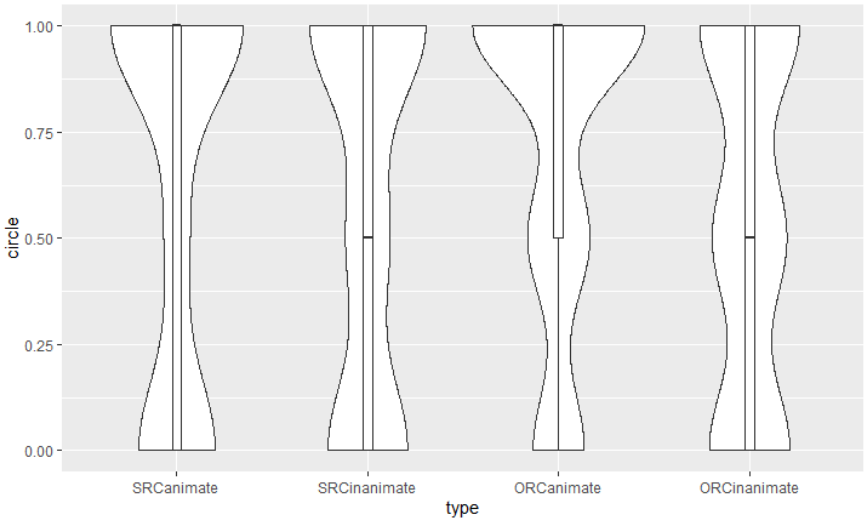
Figure 4.1.8 shows the plots of the accuracy scores of NSs and L2 learners of the sentence correction task through the violin plots with the inside Boxplots. For the NSs, the plots indicated that more participants scored at ceiling in SRCs relative to ORCs. Moreover, the median of the mean accuracy score of ORC-A was the lowest (.75) among the four types of relative clauses, while that of the other three structures was 1.00. It indicated that for NSs, the ORC-A was the most difficult structure.

For the L2 learners, they had more people scored at ceiling in ORC-A compared to the SRC-A, SRC-I and ORC-I. In addition, the median of the SRC-I and ORC-I (.50) was lower than that of the SRC-A and ORC-A (1.00), which suggested that the SRC-I and ORC-I were more difficult to L2 learners compared to the other two structures. Moreover, compared to SRC-I, there were more L2 learners who averagely scored at .50 and fewer individuals who scored at 1.00 in the ORC-I structure. Thus, the plots indicated that for the L2 learners, the ORC-I was the most difficult structure and the ORC-A was the easiest one among the four types of relative clauses.

Figure 4. 1. 8 comparisons of mean accuracy scores of each type of relative clauses for native English speakers and L2 learners in sentence correction task of the metalinguistic knowledge test



Native English Speakers



L2 learners

c) Examination of effect size

The table 4.1.26 shows the within-group effect sizes, reflecting the score differences between different structures. Except one comparison, all the effects were negligible, because the 95% CI passed through zero. Reliable effects were with the L2 learners in the comparisons between the SRC-A and ORC-I, and between ORC-A and ORC-I, but they did not reach Plonsky and Oswald's (2014) benchmark of small effect.

Table 4. 1. 28 Within-group effect sizes for (Cohen's *d*) [95% CI] in the sentence correction task of metalinguistic knowledge test

	NSs	L2 learners
SRC-A vs. SRC-I	-.07 [-.38, .24]	.08 [-.08, .24]
SRC-A vs. ORC-A	.15 [-.17, .47]	-.13 [-.29, .04]
SRC-A vs. ORC-I	.09 [-.22, .41]	**.17 [.01, .34]**
SRC-I vs. ORC-A	.22 [-.10, .55]	-.21 [-.38, -.05]
SRC-I vs. ORC-I	.20 [-.11, .51]	.09 [-.07, .25]
ORC-A vs. ORC-I	-.13 [-.45, .19]	**.28 [.12, .45]**

Note: Bold typeface indicates 95% CIs that did not pass through zero

d) Inferential statistical analysis

The inferential statistical analysis was conducted with the generalised mixed-effects model. AIC and LRT results for model selection see Appendix 21.

For NSs, the best-fitting model included the by-subject and by-item random intercepts. No statistically significant effect that was observed.

For L2 learners, the best-fitting model included the by-subject random slope of relative clause type and the by-subject and by-item random intercepts. The comparison between the ORC-I and ORC-A was found to be statistically significant. The odds ratio predicted the L2 learners to be 3.97 times more likely to successfully correct the sentence to match the picture in the ORC-A items relative to that in the ORC-I items. In addition, the comparison between the ORC-I and SRC-A was found to be reliable, because the 95% CI of the estimate *b* (*b*= 2.91 [.14, 5.68]) did not pass through zero. The odds ratio suggested that the L2 learners were more likely to provide a correct response in SRC-A compared to ORC-I (with a 18.33 times greater likelihood).

Table 4. 1. 29 The fixed effects of the model analysis in the accuracy scores of the correct sentence task for native English Speakers in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
ORC-I vs. ORC-A	-.29 [-1.75, 1.18]	.89	-.32	.748	.75 [.17, 3.25]
ORC-I vs. SRC-A	.64 [-.87, 2.14]	.91	.69	.488	1.89 [.42, 8.50]
ORC-I vs. SRC-I	.95 [-.58, 2.48]	.93	1.02	.307	2.58 [.56, 11.91]
ORC-A vs. SRC-A	.92 [-.61, 2.45]	.93	.99	.323	2.51 [.54, 11.61]
ORC-A vs. SRC-I	1.23 [-.32, 2.79]	.95	1.31	.192	3.44 [.73, 16.27]
SRC-I vs. SRC-A	-.31 [-1.88, 1.25]	.95	-.33	.741	.73 [.15, 3.49]

Note: Model formula: model4=glmer(circle ~ type + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.02, Conditional R^2=.66

Table 4. 1. 30 The fixed effects of the model analysis in the accuracy scores of the correct sentence task for L2 learners in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
ORC-I vs. ORC-A	1.38 [.41, 2.34]	.59	2.35	.019*	3.97 [1.51, 10.43]
ORC-I vs. SRC-A	**2.91 [.14, 5.68]**	**1.68**	**1.73**	**.084**	**18.33 [1.15, 292.14]**
ORC-I vs. SRC-I	.49 [-.46, 1.44]	.58	0.85	.395	1.63 [0.63, 4.21]
ORC-A vs. SRC-A	1.53 [-1.26, 4.31]	1.69	0.90	.367	4.61 [.28, 74.77]
ORC-A vs. SRC-I	-.89 [-.1.94, .16]	.64	-1.39	.165	.41 [.14, 1.18]
SRC-I vs. SRC-A	2.42 [-.19, 5.02]	1.58	1.53	.127	11.22 [.83, 152.14]

Note: Model formula: model3=glmer(circle ~ type + (type|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.03, Conditional R^2=.92; bold typeface indicates a reliable effect

4.1.5.3 Analysis of providing metalinguistic knowledge in the 'reason explanation' task

The third task of the metalinguistic knowledge test was to explain the reason for why the sentence does not match the picture in mismatched items. Only the data for mismatched items were analysed.

a) Descriptive analysis

The table 4.1.31 presents the mean (*SD*s) scores of providing metalinguistic knowledge. The results indicated that both NSs and L2 learners were almost unable to provide the metalinguistic knowledge for all the structures.

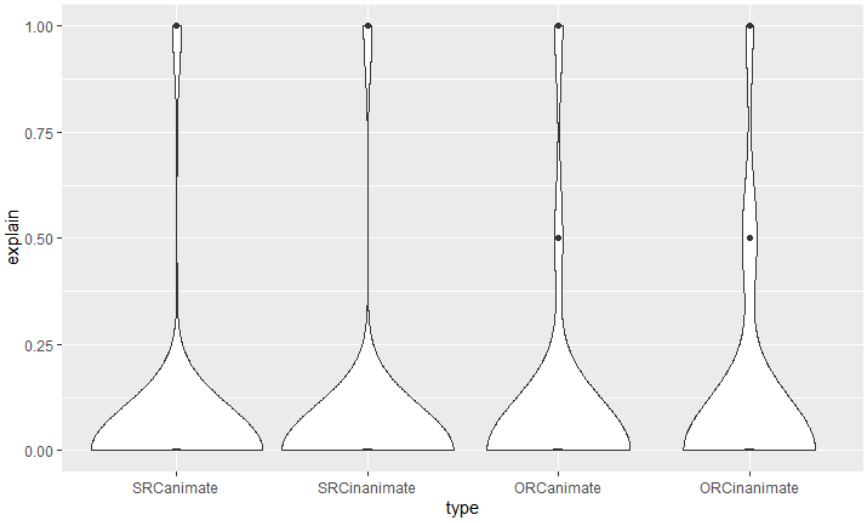
Table 4. 1. 31 Mean (*SDs*) scores of providing metalinguistic knowledge of native English Speakers and L2 learners in the reason explanation task

Structure	Natives	L2 learners
SRC-A (k=2)	.07 (.26)	.08 (.27)
SRC-I (k=2)	.05 (.22)	.06 (.24)
ORC-A (k=2)	.07 (.26)	.08 (.27)
ORC-I (k=2)	.10 (.30)	.06 (.24)

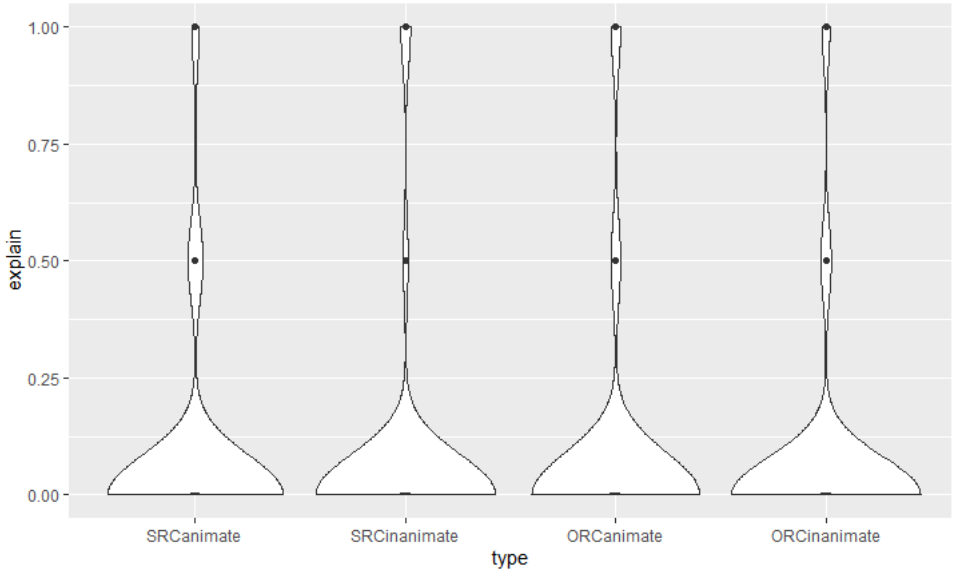
b) Plots

The figure 4.1.9 presents the violin plots with inserted Boxplots of the scores of providing metalinguistic knowledge for NSs and L2 learners respectively. The plots showed that for both NSs and L2 learners, the majority of participants failed to provide the metalinguistic knowledge across all the four structures, and no substantial difference could be observed in the different structures.

Figure 4. 1. 9 comparisons of mean scores of each type of relative clauses for native English speakers and L2 learners in providing the metalinguistic knowledge



Native English Speakers



L2 learners

c) Examination of effect size

The within-group effect sizes shown in the table 4.1.32 reflect the score differences between different types of relative clauses for NSs and L2 learners respectively. All the effects were negligible, because the 95% CI passed through zero.

Table 4. 1. 32 Within-group effect sizes for (Cohen's *d*) [95% CI] in providing metalinguistic knowledge

	Native Speakers	L2 learners
SRC-A vs. SRC-I	.15 [-.15, .46]	.06 [-.09, .22]
SRC-A vs. ORC-A	N/A	.00 [-.16, .16]
SRC-A vs. ORC-I	-.09 [-.40, .22]	.05 [-.11, .21]
SRC-I vs. ORC-A	-.15 [-.46, .15]	-.06 [-.21, .10]
SRC-I vs. ORC-I	-.15 [-.46, .15]	.00 [-.16, .16]
ORC-A vs. ORC-I	-.09 [-.40, .22]	.06 [-.10, .21]

Note: N/A refers to the effect of the comparison could not be estimated

d) Inferential statistical analysis

For the sake of completeness, the inferential statistical analysis was conducted with generalised mixed-effects models. However, arguably this was not necessary given the exceptionally low scores across both groups of participants in all the relative clause conditions. AIC and LRT results for model selection see Appendix 21.

For NSs (see table 4.1.33), the best-fitting model included the by-subject and by-item random intercepts. The statistically significant effect and the reliable effects

(95% CI about the estimate *b* did not pass through zero) could be found with the comparisons between the ORC-I and each of the other three structures. However, the odds ratios of the significant and reliable effects were extremely large (e.g. 5.12e+08) and had 95% CI around them were also enormous (e.g. 2.22e+16). The abnormal results might be attributed to two potential reasons. First, the variances between the scores of each type of relative clause were very small, and the sample size of the NSs data was rather small. In addition, numerically, the NSs scored higher in the ORC-I (mean= .10) relative to the SRC-A (mean= .07), SRC-I (mean= .05) and ORC-A (mean= .07). The mean score of the ORC-I was two times of the SRC-I, which might lead to the statistical comparison between the two groups being significant. Therefore, although the conditional R^2 of the model was very high, the results might not be robust, and might not be able to explain the independent variable. The statistically significant effects might not have practical meaning.

For L2 learners, the random effects of the best-fitting model included the by-subject and by-item intercepts (see table 4.1.34). In this model, no statistically significant effect was observed.

Table 4. 1. 33 The fixed effects of the model analysis in providing metalinguistic knowledge for native English Speakers

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
ORC-A vs. ORC-I	**20.05 [2.47, 37.64]**	**10.69**	**1.88**	**.061**	**5.12e+08 [11.8, 2.22e+16]**
ORC-A vs. SRC-A	12.93 [-4.44, 30.31]	10.56	1.22	.221	4.133+05 [.01, 1.45e+13]
ORC-A vs. SRC-I	-1.13 [-25.82, 23.57]	15.01	-.08	.940	.32 [6.11e-12, 1.72e+10]
SRC-I vs. ORC-I	**21.18 [3.82, 38.54]**	**10.55**	**2.01**	**.045***	**1.58e+09 [45.81, 5.46e+16]**
SRC-I vs. SRC-A	**14.06 [.04, 28.08]**	**8.52**	**1.65**	**.099**	**1.28e+06 [1.04, 1.56e+12]**
SRC-A vs. ORC-I	**20.05 [2.31, 37.80]**	**10.79**	**1.86**	**.063**	**5.12e+08 [10.03, 2.62e+16]**

Note: Model formula: model4=glmer(explain ~ type + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.01, Conditional R^2=.99; bold typeface indicates a reliable effect; *significantly differently from zero when α<.05

Table 4. 1. 34 The fixed effects of the model analysis in providing metalinguistic knowledge for L2 learners

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
ORC-I vs. ORC-A	.55 [-.76, 1.86]	.80	.69	.492	1.73 [.47, 6.43]
ORC-I vs. SRC-A	.55 [-.76, 1.86]	.80	.69	.489	1.74 [.47, 6.47]
ORC-I vs. SRC-I	.04 [-1.31, 1.40]	.82	.05	.961	1.04 [.27, 4.04]
ORC-A vs. SRC-A	.00 [-1.26, 1.27]	.77	.01	.995	1.00 [.28, 3.56]
ORC-A vs. SRC-I	-.51 [-1.82, .80]	.80	-.64	.524	.60 [.16, 2.23]
SRC-I vs. SRC-A	.51 [-.80, 1.82]	.80	.64	.520	1.67 [.45, 6.20]

Note: model4=glmer(explain ~ type + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.00, Conditional R^2=.96

4.1.5.4. Summary of the results in the metalinguistic knowledge test

In the first task of the metalinguistic knowledge test, deciding whether the sentence matched the picture, the ORC-A was more difficult for NSs relative to the other three structures (SRC-A, SRC-I and ORC-I), and for the L2 learners, ORC-I was more difficult than the other structures (SRC-A, SRC-I and ORC-A). The evidence could be observed in the descriptive, plots, effect sizes and inferential statistical results for the mismatched items.

In the second task, correct the mismatched sentence to match the picture, the NSs had higher accuracy in SRCs relative to ORCs, but the advantage of the SRCs over ORCs was not statistically significant. For L2 learners, the score in the ORC-I was significantly lower than SRC-A and ORC-A. Thus, ORC-I seemed to be the most difficult structure among the four types of relative clauses for L2 learners.

In the third task, explain why the sentence mismatched the picture, neither NSs nor L2 learners had the metalinguistic knowledge about the relative clauses. Although for NSs, the inferential statistics showed the score in ORC-I was significantly higher than the other structures, it was unlikely to have practical meaning. The statistically significant effects were highly likely due to the small number of participants who took part in the test and the almost negligible variance between the scores of each type of relative clause and within the scores.

4.2 To what extent can teaching parsing strategies (with explicit information and practice), exposure alone, or no exposure (tests alone), develop the learning of relative clauses?

This section will present the comparison of L2 learners' performance on each outcome measures from pre-, post-, and delayed post-tests. For each outcome measure, except the eye-tracking test where only plots and statistical analysis will be provided, the results will be reported in the following order:

a) Descriptive analysis including the mean score (M) and the standard deviations (SD)

b) Plots about the descriptive results

c) Examination of effect size (Cohen's *d*) with 95% confidence interval (CI) around *d*

d) Inferential statistical analysis (mixed-effects regression models)

e) Summary

Sections 4.2.1, 4.2.2, 4.2.3, 4.2.4, 4.2.5 will provide the results of offline comprehension, self-paced reading, eye-tracking, oral production, and metalinguistic knowledge respectively.

4.2.1 To what extent are effects observable in offline comprehension: the aural sentence-picture matching test?

The aural sentence-picture matching test was utilised to measure the offline comprehension. The participants of parsing group were expected to have more gains in accuracy compared to the input flood and the test-only groups at the post- and the delayed post-test, and the input flood group was expected to outperform the test-only group.

4.2.1.1 Descriptive analysis

For SRCs, the descriptive statistics (see Table 4.2.1) indicated that the parsing group scored higher at the post- and delayed post-test compare to the pre-test, and the input flood group gained at the delayed post-test for SRC-A. However, the test-only group did not show substantial improvement in SRCs. For ORCs, all the three groups showed

gains at the post- and delayed post-test.

Table 4. 2. 1 Mean (*SDs*) accuracy scores for offline aural sentence-picture matching test

Structure	test phase	parsing	input flood	test-only
SRC-A (k=4 or 8 [a])	pre-test	.84 (.36)	.88 (.32)	.85 (.36)
	post-test	.94 (.24)	.87 (.34)	.89 (.32)
	delayed post-test	.96 (.19)	.96 (.19)	.90 (.30)
SRC-I (k=4 or 8 [a])	pre-test	.87 (.34)	.86 (.34)	.87 (.34)
	post-test	.90 (.30)	.87 (.35)	.84 (.37)
	delayed post-test	.94 (.23)	.89 (.32)	.88 (.33)
ORC-A (k=4 or 8 [a])	pre-test	.84 (.37)	.86 (.35)	.83 (.37)
	post-test	.93 (.25)	.94 (.25)	.90 (.30)
	delayed post-test	.96 (.19)	.93 (.26)	.91 (.29)
ORC-I (k=4 or 8 [a])	pre-test	.77 (.42)	.75 (.43)	.78 (.41)
	post-test	.88 (.32)	.83 (.37)	.86 (.35)
	delayed post-test	.88 (.33)	.84 (.37)	.89 (.31)

Note: SRC-A = subject relative clause – animate head; SRC-I = subject relative clause – inanimate head; ORC-A = object relative clause – animate head; ORC-A = object relative clause – inanimate head; [a] Numbers of items differed between version 1 and version 2, version 3 and version 4 of the tests, respectively; versions were counterbalanced across pre-, post-, and delayed post-test, within groups.

4.2.1.2 Plots

The violin plots with included boxplots (see Figure 4.2.1, 4.2.2, 4.2.3 and 4.2.4) were created based on the mean scores of all the items of a target structure from each participant. Looking at the score distribution of the three groups, the parsing group had more individuals scoring at ceiling at the post- and delayed post-test phases compared to the other two groups for SRC-A, SRC-I, and ORC-A, but not for ORC-I. One notable exception to this pattern is that for the input flood group, at delayed post-test, for SRC-A only, there was a high proportion of individuals who scored at or near ceiling, with only a few who got lower scores.

Figure 4. 2. 1 Comparison of mean accuracy scores of three learner groups in different test phase for SRC-A structure in offline aural sentence-picture matching test

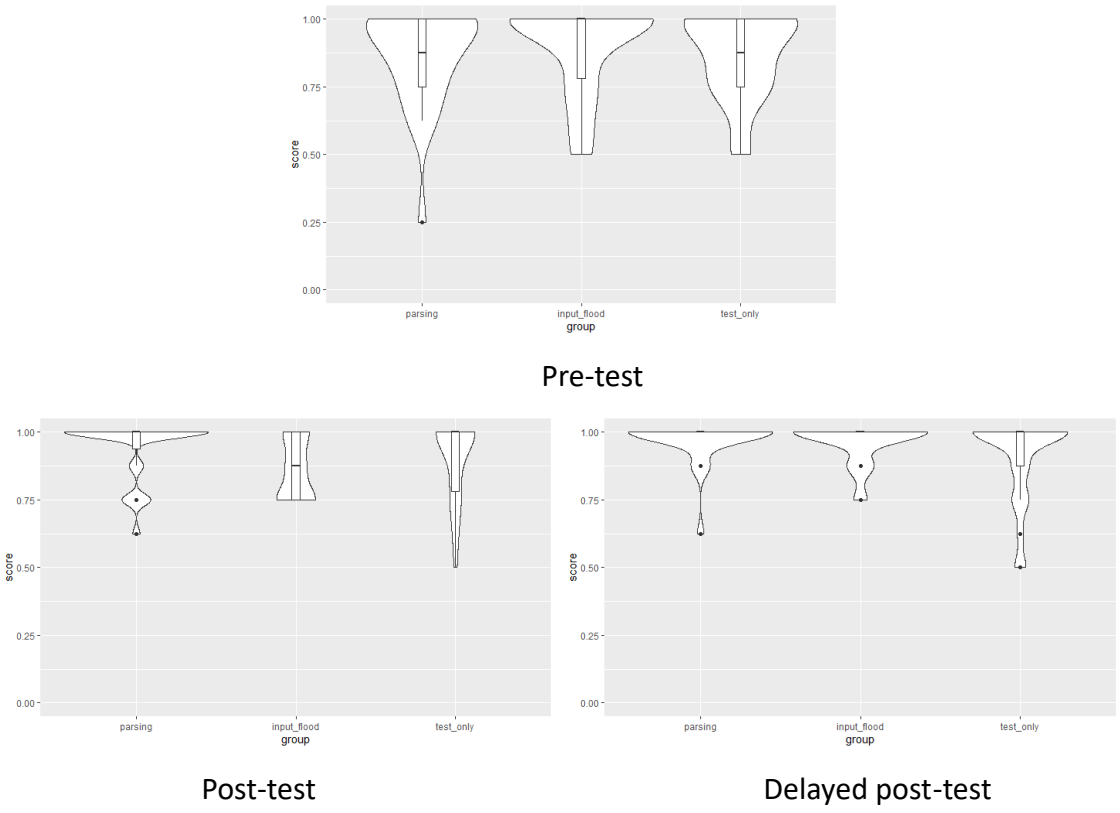


Figure 4. 2. 2 Comparison of mean accuracy scores of three learner groups in different test phase for SRC-I structure in offline aural sentence-picture matching test

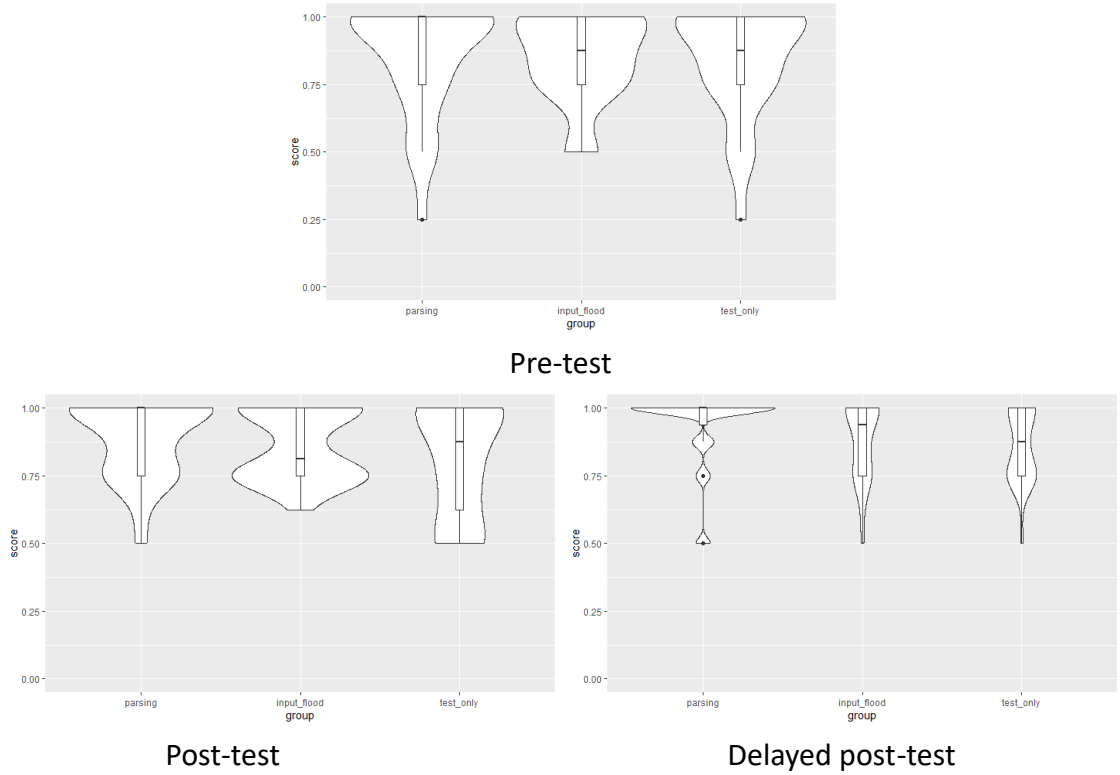


Figure 4. 2. 3 Comparison of mean accuracy scores of three learner groups in different test phase for ORC-A structure in offline aural sentence-picture matching test

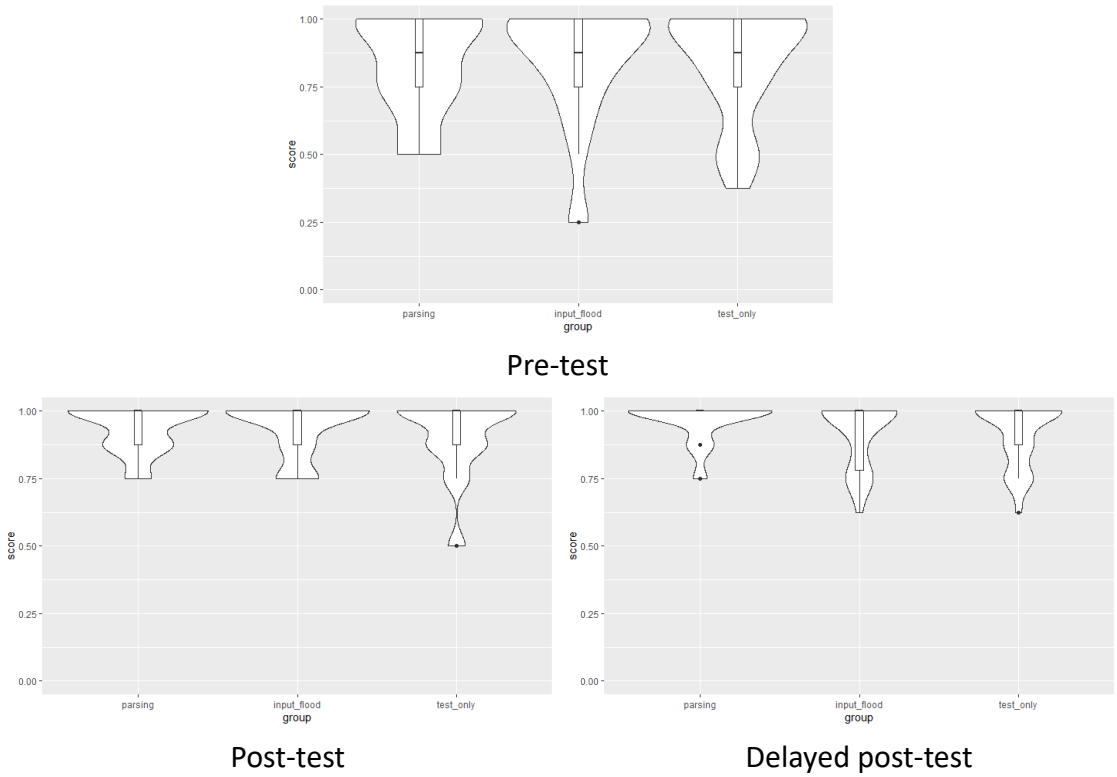
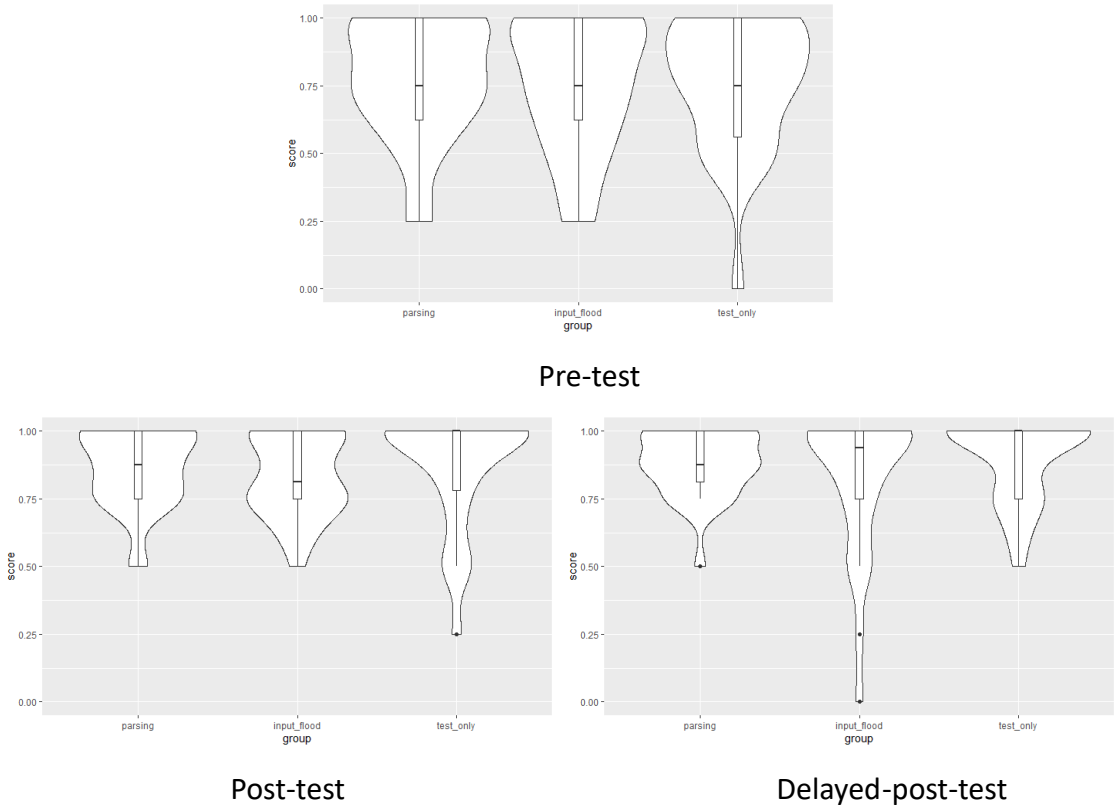


Figure 4. 2. 4 Comparison of mean accuracy scores of three learner groups in different test phase for ORC-I structure in offline aural sentence-picture matching test



4.2.1.3 Examination of effect size

For within-group contracts, reflecting change over time, (see Table 4.2), all effect sizes could be described as extremely small to negligible (i.e., with upper limits of their 95% CI smaller than the lowest point of the range (0.60) found by Plonsky & Oswald's 2014 field-general benchmarks, and many of the 95% CIs passed through zero). However, it is clear (see the bold typeface in Table 4.2.2), that only the parsing group made reliable, though extremely small, gains (where the CIs did not pass through zero) across all four target structures, between pre- and delayed post-tests. In addition, for the parsing group, extremely small gains were also observable in pre- to post-test in SRC-A, ORC-A and ORC-I. Two other noteworthy features of these results are: The input flood group made extremely small, though reliable, pre- to post-test gains for ORC-A structure, pre- to delayed post-test gains for SRC-A, ORC-A and ORC-I structures, post- to delayed post-test gains for SRC-A structure, and these gains were extremely small; the test-only group made reliable but extremely small gains in the two ORC structures at the post-test.

Table 4. 2. 2 Within-group effect size (Cohen's *d*) [95% CI] for offline aural sentence-picture matching test

structure	group	pre-post	pre-delayed	post-delayed
SRC-A (k=4 or 8)[a]	parsing	**.20 [.05, .36]**	**.28 [.12, .44]**	.08 [-.07, .24]
	input flood	-.04 [-.20, .12]	**.21 [.05, .37]**	**.24 [.08, .40]**
	test-only	.10 [-.06, .26]	.11 [-.05, .27]	.01 [-.14, .17]
SRC-I (k=4 or 8)	parsing	.08 [-.08, .23]	**.19 [.04, .35]**	.12 [-.04, .27]
	input flood	.01 [-.14, .17]	.05 [-.11, .21]	.03 [-.13, .19]
	test-only	-.03 [-.18, .13]	.03 [-.13, .19]	.05 [-.11, .21]
ORC-A (k=4 or 8)	parsing	**.21 [.06, .37]**	**.29 [.13, .45]**	.10 [-.06, .25]
	input flood	**.20 [.05, .36]**	**.17 [.01, .33]**	-.02 [-.18, .14]
	test-only	.16 [.00, .32]	**.17 [.01, .33]**	.02 [-.14, .18]
ORC-I (k=4 or 8)	parsing	**.21 [.06, .37]**	**.20 [.05, .36]**	0 [-.16, .16]
	input flood	.14 [-.02, .31]	**.19 [.03, .35]**	.01 [-.14, .17]
	test-only	.13 [-.03, .29]	**.22 [.06, .38]**	.13 [-.03, .29]

Note: Bold typeface indicates that the CI did not pass through zero; [a] Numbers of items

differed between version 1 and version 2, version 3 and version 4 of the tests, respectively; versions were counterbalanced across pre-, post-, and delayed post-test, within groups.

For between-group contrasts, reflecting differences between groups (see Table 4.2.3), the very small but reliable effects (CI did not pass through zero, and the upper limits of the CI reach the benchmark, .40, of small effect) were found in the comparison between the parsing and the input flood group at the post-test for SRC-A structure and between the parsing and the test-only group at the delayed post-test for SRC-I structure. However, the changes of Cohen's *d* from the pre- to the post-test and from the pre- to the delayed post-test did not reach the benchmark of the small effect (.40) across the structures.

Table 4. 2. 3 Between-group effect size (Cohen's d) [95% CI] for offline aural sentence-picture matching test

structure	test phase	parsing vs. input flood	parsing vs. test-only	input flood vs. test-only
SRC-A (k=4 or 8)[a]	pre-test	-.11 [-.33, .11]	-.01 [-.23, .21]	.10 [-.12, .33]
	post-test	**.24 [.02, .46]**	.18 [-.05, .40]	-.07 [-.29, .15]
	delayed post-test	.04 [-.18, .25]	.25 [-.03, .47]	.22 [.00, .44]
	pre-post d change	.35 [N/A]	.19 [N/A]	-.17 [N/A]
	pre-delayed d change	.15 [N/A]	.26 [N/A]	.12 [N/A]
SRC-I (k=4 or 8)	pre-test	.01 [-.21, .23]	0 [-.22, .22]	-.01 [-.23, .21]
	post-test	.10 [-.12, .32]	.19 [-.02, .41]	.10 [-.12, .32]
	delayed post-test	.20 [-.02, .42]	**.24 [.02, .46]**	.04 [-.18, .27]
	pre-post *d* change	.09 [N/A]	.19 [N/A]	.11 [N/A]
	pre-delayed *d* change	.19 [N/A]	.24 [N/A]	.05 [N/A]
ORC-A (k=4 or 8)	pre-test	-.06 [-.28, .16]	.01 [-.21, .22]	.06 [-.16, .28]
	post-test	-.01 [-.23, .21]	.12 [-.10, .34]	.13 [-.09, .35]
	delayed post-test	.15 [-.07, .38]	.22 [.00, .45]	.07 [-.15, .30]
	pre-post *d* change	.05 [N/A]	.11 [N/A]	.07 [N/A]
	pre-delayed *d* change	.21 [N/A]	.21 [N/A]	.01 [N/A]
ORC-I (k=4 or 8)	pre-test	.04 [-.18, .27]	-.03 [-.25, .19]	-.08 [-.30, .15]
	post-test	.14 [-.08, .36]	.08 [-.15, .30]	-.06 [-.28, 16]
	delayed post-test	.12[-.10, .33]	-.05 [-.27, .19]	-.16[-.38, .05]
	pre-post *d* change	.10 [N/A]	.11 [N/A]	.02 [N/A]
	pre-delayed *d* change	.08 [N/A]	-.02 [N/A]	-.08 [N/A]

Note: Bold typeface indicates that the CI did not pass through zero; [a] Numbers of items differed between version 1 and version 2, version 3 and version 4 of the tests, respectively; versions were counterbalanced across pre-, post-, and delayed post-test, within groups.

4.2.1.4 Inferential statistical analysis

Model selection

The mixed-effects logistic regression models were run separately for different language structures, SRC-A, SRC-I, ORC-A, and ORC-I. The LRT and AIC indicated that the primary model that included the fixed effects for the interaction between group and test phase, and the random intercepts for subjects and items fitted the data best for all four relative clause structures. Thus, the results of the primary model will be presented. The AIC and LRT results see Appendix 22.

In order to investigate the relationship between each two groups, the test-only group and the input flood group were used as the baseline group respectively (again, it should be acknowledged that other more efficient coding methods could have been used; however, in this thesis, the dummy coding method was adopted). The models with the baseline of the test-only group will be reported in detail in this Chapter. All the fixed effects regardless of whether they were statistically significant will be presented. In addition, the models that used the input flood group as the baseline are used as *supplementary* models, and the full results are given in Appendix 26. For the input flood baseline model, *only* the statistically significant effects that related to the comparison between the input flood and the parsing group will be reported in this Results Chapter, as that comparison is arguably the more pertinent theoretically (and for reasons of spaces).

Model analysis

Table 4.2.4 to Table 4.2.7 present the fixed effects for SRC-A, SRC-I, ORC-A, and ORC-I structures respectively. For SRC-A and SRC-I, the statistically significant effect was shown in the two-way interaction between test-phase (pre- vs. delayed post-test) and group (parsing vs. test-only group). It indicated that the parsing group significantly

outperformed the test-only group in the delayed post-test for SRC-A and SRC-I. In addition, in the models with the baseline of the input flood group, it was found that compared to the input flood group, the parsing group was 3.59 times more likely to correctly respond to the SRC-A items at the post-test relative to the pre-test.

For ORC-A and ORC-I, statistically significant effects were found for the interaction between the pre- and delayed post-test. It could be interpreted as the three groups' scores increased in the delayed post-test overall, but no group significantly better than the other. No statistically significant effect could be observed in the model with the baseline of the input flood group. According to the descriptive data, for ORC-A, all the three groups showed improvement. Thus, the parsing group and the test-only group contributed to the significant improvement from the pre- to the delayed post-test for ORC-I structure.

Table 4. 2. 4 The fixed effects of the model analysis of accuracy scores for SRC-A in offline comprehension: aural sentence-picture matching test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	2.18 [1.61, 2.76]	.35	6.25	<.001***	8.88 [5.00,15.78]
test vs. parsing	-.03 [-.68, .62]	.40	-.08	.936	.97 [.51, 1.86]
test vs. input	.29 [-.39, .97]	.42	.70	.487	1.34 [.67, 2.65]
pre- vs. post-	.48 [-.12, 1.09]	.37	1.32	.186	1.62 [.89, 2.97]
pre- vs. delayed-	.51 [-.10, 1.11]	.37	1.37	.171	1.66 [.90, 3.05]
test vs. parsing × pre- vs. post-	.54 [-.35, 1.44]	.55	1.00	.319	1.72 [.70, 4.23]
test vs. input × pre- vs. post-	-.73 [-1.58, .12]	.52	-1.42	.157	0.48 [.21, 1.13]
test vs. parsing × pre- vs. delayed-	**1.16 [.16, 2.15]**	**.60**	**1.91**	**.056**	**3.18 [1.17, 8.59]**
test vs. input × pre- vs. delayed-	.69 [-.29, 1.68]	.60	1.15	.249	2.00 [.74, 5.36]

***Note**:* Model formula: model1<-glmer(score ~ group*phase + (1|subject) + (1|item), data=new, family=binomial, control = glmerControl(optimizer = "bobyqa")); Marginal R^2=0.07, Conditional R^2=0.35; pasing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; *** significantly differently from zero when $\alpha < .001$

Table 4. 2. 5 The fixed effects of the model analysis of accuracy scores for SRC-I in offline comprehension: aural sentence-picture matching test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	2.33 [1.70, 2.96]	.38	6.10	<.001***	10.29 [5.49, 19.28]
test vs. parsing	.01 [-.65, .67]	.40	.02	.982	1.01 [.52, 1.94]
test vs. input	-.03 [-.69, .63]	.40	-.07	.948	.97 [.50, 1.88]
pre- vs. post-	-.32 [-.89, .26]	.35	-.91	.363	.73 [.41, 1.29]
pre- vs. delayed-	.21 [-.41, .83]	.38	.56	.573	1.24 [.66, 2.30]
test vs. parsing × pre- vs. post-	.84 [-.01, 1.70]	.52	1.62	.105	2.32 [.99, 5.45]
test vs. input × pre- vs. post-	.49 [-.34, 1.33]	.51	.97	.332	1.63 [.71, 3.76]
test vs. parsing × pre- vs. delayed-	**1.15 [.17, 2.13]**	**.60**	**1.93**	**.054**	**3.16 [1.19, 8.44]**
test vs. input × pre- vs. delayed-	.21 [-.68, 1.09]	.54	.39	.700	1.23[.51, 2.98]

***Note**:* Model formula: model1<-glmer(score ~ group*phase + (1|subject) + (1|item), data=new, family=binomial, control = glmerControl(optimizer = "bobyqa")); Marginal R^2=0.04, Conditional R^2=0.38; pasing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect ; *** significantly differently from zero when $\alpha < .001$

Table 4. 2. 6 The fixed effects of the model analysis of accuracy scores for ORC-A in offline comprehension: aural sentence-picture matching test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	1.95 [1.42, 2.48]	.32	6.05	<.001***	7.05 [4.14, 11.98]
test vs. parsing	-.01 [-.64, .62]	.38	-.02	.983	.99 [.53, 1.86]
test vs. test	.18 [-.47, .82]	.39	.45	.652	1.19 [.63, 2.28]
pre- vs. post-	.61 [.01, 1.21]	.36	1.67	.095	1.84 [1.01, 3.35]
pre- vs. delayed-	**.74 [.11, 1.37]**	**.38**	**1.95**	**.052**	**2.10 [1.12, 3.92]**
test vs. parsing × pre- vs. post-	.50 [-.39, 1.39]	.54	.92	.356	1.64 [.68, 4.00]
test vs. test × pre- vs. post-	.35 [-.57, 1.26]	.56	.62	.533	1.41 [.57, 3.53]
test vs. parsing × pre- vs. delayed-	.98 [-.03, 2.00]	.62	1.60	.110	2.68 [.97, 7.37]
test vs. test × pre- vs. delayed-	.18 [-.75, 1.10]	.56	.31	.755	1.19 [.47, 3.01]

***Note**:* Model formula: model1<-glmer(score ~ group*phase + (1|subject) + (1|item), data=new, family=binomial, control = glmerControl(optimizer = "bobyqa")); Marginal R^2=0.07, Conditional R^2=0.29; pasing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; *** significantly differently from zero when $\alpha < .001$

Table 4. 2. 7 The fixed effects of the model analysis of accuracy scores for ORC-I in offline comprehension: aural sentence-picture matching test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	1.70 [1.15, 2.26]	.34	5.05	<.001***	5.48 [3.15, 9.55]
test vs. parsing	-.13 [-.77, .52]	.39	-.33	.742	.88 [.46, 1.67]
test vs. input	-.21 [-.86, .43]	.39	-.54	.586	.81 [.42, 1.54]
pre- vs. post-	**.57 [.03, 1.12]**	**.33**	**1.74**	**.081**	**1.78 [1.03, 3.05]**
pre- vs. delayed-	.95 [.36, 1.53]	.35	2.67	.008**	2.58 [1.44, 4.61]
test vs. parsing × pre- vs. post-	.28 [-.49, 1.05]	.47	.59	.556	1.32 [.61, 2.85]
test vs. input × pre- vs. post-	.06 [-.69, .81]	.46	.14	.890	1.06 [.50, 2.25]
test vs. parsing × pre- vs. delayed-	-.14 [-.94, .65]	.48	-.30	.766	.87 [.39, 1.92]
test vs. input × pre- vs. delayed-	-.35 [-1.14, .44]	.48	-.74	.462	.70 [.32, 1.55]

***Note**:* Model formula: model1<-glmer(score ~ group*phase + (1|subject) + (1|item), data=new, family=binomial, control = glmerControl(optimizer = "bobyqa")); Marginal R^2=0.03, Conditional R^2=0.34; pasing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect ;** significantly differently from zero when α < 0.01; *** significantly differently from zero when α < .001

4.2.1.5 Summary of the results in the offline comprehension

In the offline comprehension test, all the three groups showed some gains at the post- and the delayed post-test. In addition, the parsing group statistically gained more than the input flood group and the test-only group for SRCs across the time.

4.2.2 To what extent are effects observable in online processing during self-paced reading?

The self-paced reading test (SPR) was used to measure the online reading comprehension, and the reaction times (RTs) of each word were analysed. Each item in the test contains a picture and a sentence which may match or mismatch the picture. Compared to the input flood and the test-only group, the parsing group was expected to be more sensitive to the sentence-picture anomaly at the post- and the delayed post-test, which would result in the slower RTs in reading the mismatched items than the matched items.

In this section, the RTs of the three critical words (e.g. in a SRC sentence, 'The cat that chases the dog is big', the three critical words are 'chases', 'the' and 'dog'

respectively; in an ORC sentence, 'The cat that the dog chases is big', the three critical words are 'the', 'dog' and 'chases' respectively) and the average RT of each whole sentence are analysed.

4.2.2.1 Descriptive analysis

Table 4.2.8 to 4.2.11 show the descriptive results of SRC-A, SRC-I, ORC-A and ORC-I. Each table includes the mean RTs (raw) with SDs for matched and mismatched items of the first, second, third words and the whole sentence, and mean RTs for the mismatched items that had been subtracted from those of the matched items.

SRC-A

The parsing group seemed to gain sensitivity to the sentence-picture anomaly at the post-test, but did not maintain this at the delayed post-test. At the pre-test, they had slower RTs for matched items compared to mismatched items at the third critical word and the whole sentence. At the post-test, slower RTs for the mismatched items than the matched items could be observed on all the three critical words and the whole sentence, and at the third critical word the RT difference was the biggest. At the delayed post-test, the parsing group showed sensitivity to the mismatch only at the second critical word, but the RT difference was rather small (15.86 ms).

The input flood group did not have salient change across the time. At the pre-test, they read mismatched items slower than the matched items at the first and the third critical words, and in the whole sentence, which indicated that they started to be sensitive to the mismatch at the first critical word. However, at the post-test, although they had slower RTs for the mismatched items relative to the matched items across the three critical words and the whole sentence, the RT difference was rather small, which might be a chance finding. At the delayed post-test, it seemed that they started to show sensitivity to the mismatch at the second critical word, so the time point was later than that of the pre-test.

The test only group was sensitive to the sentence picture mismatch at the second and the third critical words at the pre-test. However, they almost did not show sensitivity to the mismatch at the post- and the delayed post-test.

SRC-I

For SRC-I, all the three groups did not have clear gains across the time. *For the parsing group*, the RT differences indicated that they were sensitive to the sentence-picture mismatch at the third critical word (103.01 ms). At the post-test, they had slower RTs in the mismatched items compared to the matched items at all the three critical words and the whole sentence. However, the RT differences at the first (23.14 ms) and the second (23.03 ms) critical words were rather small, and sensitivity to the mismatch seems happened at the third critical word (49.06 ms). At the delayed post-test, it seems that they were not sensitive to the sentence-picture mismatch for all the three critical words and the whole sentence.

For the input flood group, they showed sensitivity to the sentence-picture mismatch at the third critical words across the time, but the RT differences were rather small. It seems that they did not gain in sensitivity after training.

For the test-only group, at the pre-test, they had read mismatched items slower than the matched items at the third critical word and the whole sentence. At the post-test, they were still sensitive to the mismatch at the third critical words and the whole sentence, but the RT differences were smaller than those at the pre-test. In addition, at the delayed post-test, they were insensitive to the mismatch at the critical regions and the whole sentence.

ORC-A

For ORC-A structure, no gains could be observed for all the three groups. The parsing group and the input flood group showed some sensitivity to the sentence-picture mismatch at the third critical word at the pre-test. However, it seems that they were no longer sensitive to the mismatch anymore at the post- and the delayed post-test across the three critical words.

ORC-I

The parsing group showed some gains at the delayed post-test compared to the pre-test at the second critical word and the whole sentence. However, even at the delayed post-test, the RT differences between mismatched and matched items were

very small, which might be chance finding.

T*he input flood group* showed improvement in sensitivity to the sentence-picture mismatch at the post- and the delayed post-test compared to the pre-test. At the post-test, they showed sensitivity at the third critical word and the whole sentence, and they read the mismatched items slower than the matched items across the three critical words and the whole sentence at the delayed post-test.

The test-only group did not become more sensitive to the mismatch at the post- and the delayed post-test compared to the pre-test. At the pre-test, they had slower RTs in the mismatched items than the matched items at the third critical word and the whole sentence. However, sensitivity to mismatch could not be found at the post-test and the delayed post-test.

Table 4. 2. 8 Means (SDs) of the reaction times for the SRC-A (k=10) structure for the whole sentence and the first, second, third critical words in self-paced reading tests

		Pre-test			Post-test			Delayed post-test		
Word		parsing	input flood	test-only	parsing	input flood	test-only	parsing	input flood	test-only
1st critical	matched	422.87	383.76	461.35	361.37	284.64	343.39	340.83	297.29	319.37
		(211.10)	(179.29)	(331.30)	(174.13)	(166.59)	(177.42)	(144.46)	(137.36)	(127.71)
	mismatched	411.51	418.04	481.23	370.78	309.90	341.60	330.63	292.88	329.27
		(214.39)	(222.18)	(249.72)	(181.72)	(141.94)	(168.28)	(159.60)	(130.90)	(144.63)
	mismatched-matched	-11.36	34.28	19.88	9.42	25.26	-1.79	-10.20	-4.40	9.90
2nd critical	matched	411.19	397.71	380.53	356.90	315.93	348.59	339.02	314.25	321.57
		(189.12)	(264.82)	(149.34)	(145.68)	(146.80)	(135.35)	(132.58)	(172.71)	(135.71)
	mismatched	412.19	386.88	432.72	375.56	319.01	337.89	354.88	358.97	338.25
		(189.64)	(139.98)	(216.63)	(158.77)	(166.72)	(125.97)	(154.40)	(308.43)	(139.61)
	mismatched-matched	1.00	-10.83	52.19	18.65	3.08	-10.71	15.86	44.72	16.69
3rd critical	matched	447.33	467.78	436.64	362.46	324.39	353.92	344.88	309.17	378.27
		(282.53)	(282.34)	(249.86)	(177.29)	(219.45)	(183.38)	(133.90)	(125.73)	(309.95)
	mismatched	470.52	512.01	550.54	410.04	335.17	362.84	339.91	332.99	376.85
		(303.93)	(376.33)	(394.74)	(235.40)	(195.39)	(153.11)	(139.53)	(183.20)	(243.81)
	mismatched-matched	23.19	44.23	113.91	47.58	10.78	8.92	-4.96	23.82	-1.43
whole	matched	452.98	432.27	468.97	397.24	331.07	367.75	378.68	349.00	349.19
		(195.38)	(163.33)	(213.89)	(179.64)	(136.11)	(136.51)	(141.85)	(137.41)	(139.03)
	mismatched	481.44	479.66	545.45	420.73	343.19	392.70	363.38	375.46	373.58
		(240.89)	(185.94)	(243.36)	(175.88)	(121.14)	(167.90)	(154.12)	(207.43)	(176.27)
	mismatched-matched	28.47	47.39	76.48	23.49	12.12	24.95	-15.30	26.46	24.40

Note: Example SRC-A: The cat that chases (first critical) the (second critical) dog (third critical) is big; the first word of the sentences (the) was excluded in the whole sentence analysis.

Table 4. 2. 9 Means (SDs) of the reaction times for the SRC-I (k=10) structure for the whole sentence and the first, second, third critical words in self-paced reading tests

		Pre-test			Post-test			Delayed post-test		
Word		parsing	input flood	test-only	parsing	input flood	test-only	parsing	input flood	test-only
1st critical	matched	417.17	427.41	499.85	363.63	307.55	365.68	342.23	314.90	311.65
		(212.04)	(252.92)	(337.71)	(165.02)	(146.90)	(240.43)	(148.90)	(155.06)	(128.26)
	mismatched	421.79	436.10	490.08	386.78	305.22	355.59	345.73	320.86	327.10
		(209.62)	(221.71)	(268.39)	(211.66)	(138.65)	(182.10)	(162.55)	(188.46)	(157.03)
	mismatched-matched	4.62	8.68	-9.77	23.14	-2.32	-10.09	3.50	5.96	15.45
2nd critical	matched	412.74	419.12	422.11	373.69	332.67	334.64	338.52	342.52	314.30
		(190.74)	(220.98)	(212.42)	(183.87)	(198.87)	(177.16)	(151.92)	(191.93)	(174.47)
	mismatched	403.07	424.02	445.65	396.72	298.71	350.82	352.05	336.72	323.19
		(215.80)	(181.81)	(210.48)	(188.33)	(174.18)	(177.54)	(150.15)	(212.06)	(125.28)
	mismatched-matched	-9.67	4.90	23.55	23.03	-33.96	16.19	13.53	-5.80	8.88
3rd critical	matched	418.91	459.08	471.91	369.67	324.45	344.86	356.70	344.68	362.97
		(209.77)	(255.49)	(289.52)	(190.94)	(165.09)	(226.89)	(186.20)	(198.21)	(246.61)
	mismatched	521.93	495.73	596.83	418.73	350.13	392.50	386.62	362.70	373.11
		(482.59)	(350.00)	(451.47)	(260.88)	(266.84)	(248.77)	(274.05)	(233.48)	(243.97)
	mismatched-matched	103.01	36.66	124.91	49.06	25.68	47.64	29.92	18.02	10.14
whole	matched	467.01	486.68	500.93	409.62	330.31	377.28	375.63	370.63	358.51
		(215.50)	(221.95)	(222.75)	(184.62)	(120.85)	(174.29)	(153.11)	(182.21)	(129.17)
	mismatched	488.71	512.99	567.76	447.06	351.91	398.80	397.93	372.02	379.56
		(239.69)	(233.82)	(276.00)	(215.92)	(159.78)	(169.06)	(173.21)	(166.54)	(176.03)
	mismatched-matched	21.70	26.31	66.83	37.45	21.61	21.52	22.30	1.39	21.05

Note: Example SRC-I: The car that chases (first critical) the (second critical) bike (third critical) is red; the first word of the sentences (the) was excluded in the whole sentence analysis.

Table 4. 2. 10 Means (SDs) of the reaction times for the ORC-A (k=10) structure for the whole sentence and the first, second, third critical words in self-paced reading tests

Word		Pre-test			Post-test			Delayed post-test		
		parsing	input flood	test-only	parsing	input flood	test-only	parsing	input flood	test-only
1st critical	matched	370.31	370.83	398.65	352.61	301.16	305.42	319.91	292.81	302.57
		(176.18)	(186.44)	(156.64)	(144.57)	(188.90)	(103.57)	(127.15)	(118.76)	(120.17)
	mismatched	374.65	351.26	382.34	342.47	299.10	323.50	305.05	301.77	309.89
		(131.99)	(137.54)	(197.81)	(133.97)	(148.57)	(136.22)	(91.26)	(139.15)	(146.97)
	mismatched-matched	4.34	-19.57	-16.31	-10.14	-2.06	18.08	-14.86	8.96	7.32
2nd critical	matched	411.27	420.00	459.45	337.28	303.37	347.43	324.03	313.25	312.27
		(198.21)	(207.13)	(233.38)	(138.29)	(210.74)	(155.44)	(143.05)	(166.45)	(126.18)
	mismatched	404.96	393.57	448.78	345.10	311.02	338.04	322.64	322.29	335.85
		(170.99)	(192.52)	(246.18)	(141.25)	(152.94)	(141.91)	(126.96)	(204.16)	(170.27)
	mismatched-matched	-6.31	-26.43	-10.67	7.82	7.65	-9.40	-1.39	9.04	23.58
3rd critical	matched	461.50	487.15	507.83	392.40	340.10	380.12	364.61	357.68	361.31
		(292.91)	(291.01)	(349.90)	(256.66)	(197.89)	(258.37)	(190.05)	(273.11)	(236.26)
	mismatched	515.24	528.77	527.97	406.49	353.53	394.14	355.75	332.76	379.64
		(370.69)	(368.77)	(354.71)	(229.30)	(229.09)	(275.88)	(152.50)	(192.02)	(308.10)
	mismatched-matched	53.74	41.62	20.14	14.08	13.43	14.01	-8.86	-24.92	18.34
whole	matched	461.81	471.70	500.80	390.88	329.13	375.12	380.44	350.65	358.33
		(217.52)	(179.71)	(211.43)	(155.45)	(120.41)	(150.99)	(168.52)	(190.07)	(168.23)
	mismatched	494.98	484.01	532.68	410.11	353.22	405.64	383.86	372.90	378.73
		(215.17)	(213.50)	(242.11)	(171.63)	(157.99)	(189.71)	(159.93)	(168.51)	(182.60)
	mismatched-matched	33.17	12.31	31.88	19.23	24.09	30.53	3.43	22.24	20.40

Note: Example ORC-A: The cat that the (first critical) dog (second critical) chases (third critical) is big; the first word of the sentences (the) was excluded in the whole sentence analysis.

Table 4. 2. 11 Means (SDs) of the reaction times for the ORC-I (k=10) structure for the whole sentence and the first, second, third critical words in self-paced reading tests

		Pre-test			Post-test			Delayed post-test		
Word		parsing	input flood	test-only	parsing	input flood	test-only	parsing	input flood	test-only
1st critical	match	401.25	384.16	419.46	356.28	288.08	320.78	317.68	301.43	309.84
		(188.51)	(154.26)	(206.51)	(167.43)	(99.56)	(130.68)	(100.20)	(114.88)	(116.39)
	mismatch	382.72	371.90	377.62	338.28	288.79	317.12	326.97	323.81	327.35
		(164.03)	(150.99)	(133.96)	(127.24)	(125.09)	(121.43)	(125.82)	(162.69)	(149.45)
	mismatch-match	-18.53	-12.26	-41.84	-18.01	0.72	-3.66	9.29	22.39	17.51
2nd critical	match	439.70	434.67	515.14	381.67	309.52	342.05	322.99	332.84	359.29
		(242.21)	(219.08)	(317.47)	(164.20)	(147.45)	(151.43)	(105.66)	(182.98)	(243.80)
	mismatch	408.45	406.22	433.80	357.02	318.11	342.14	341.36	369.78	344.89
		(204.50)	(279.47)	(223.87)	(143.03)	(198.85)	(149.84)	(131.64)	(281.23)	(162.10)
	mismatch-match	-31.25	-28.45	-81.34	-24.64	8.59	0.09	18.37	36.94	-14.40
3rd critical	match	481.84	500.77	527.26	413.41	312.18	403.49	376.90	354.29	366.30
		(312.41)	(277.80)	(353.86)	(249.56)	(136.26)	(268.31)	(213.16)	(205.85)	(271.89)
	mismatch	485.91	487.58	564.56	400.68	349.29	380.50	382.51	402.04	394.75
		(308.21)	(313.76)	(390.16)	(230.53)	(216.14)	(279.44)	(269.03)	(345.98)	(291.19)
	mismatch-match	4.07	-13.19	37.30	-12.73	37.11	-22.99	5.61	47.75	28.46
Whole	match	490.38	491.07	517.51	403.77	342.76	399.11	378.42	376.94	373.42
		(244.44)	(187.80)	(248.12)	(171.08)	(127.97)	(169.75)	(141.08)	(171.46)	(172.71)
	mismatch	495.59	489.36	535.52	425.94	372.69	398.38	412.59	395.42	393.90
		(253.54)	(207.72)	(259.02)	(185.60)	(170.70)	(186.12)	(177.23)	(216.78)	(204.60)
	mismatch-match	5.21	-1.71	18.01	22.17	29.93	-0.73	34.17	18.49	20.49

Note: Example ORC-I: The car that the (first critical) bike (second critical) chases (third critical) is red; the first word of the sentences (the) was excluded in the whole sentence analysis.

4.2.2.2 Plots

Figures 4.2.5 to 4.2.8 are the line charts of residual RT differences between matched and mismatched items of SRC-A, SRC-I, ORC-A and ORC-I structures respectively. Each line of a chart presents the residual RT difference of a group, calculated from the mean residual RTs of the mismatched items minus that of the matched items.

For SRC-A, at the pre-test, overall, the input flood group and the test-only group had bigger residual RT differences than the parsing group. At the first critical word 'chases', the input flood and the test-only groups showed sensitivity to the mismatch of sentence and the picture. At the third critical word, the test-only group had much bigger residual RT differences than the other two groups. At the post-test, the input flood group started to be sensitive to the sentence-picture mismatch at the first critical word, which was earlier than the other two groups. At the third critical word 'dog', the parsing group had bigger residual RT differences than the other two groups. At the delayed post-test, the RT differences of the three groups were similar across the three critical regions, and all the three groups were sensitive to the sentence-picture mismatch at the second critical word 'the'.

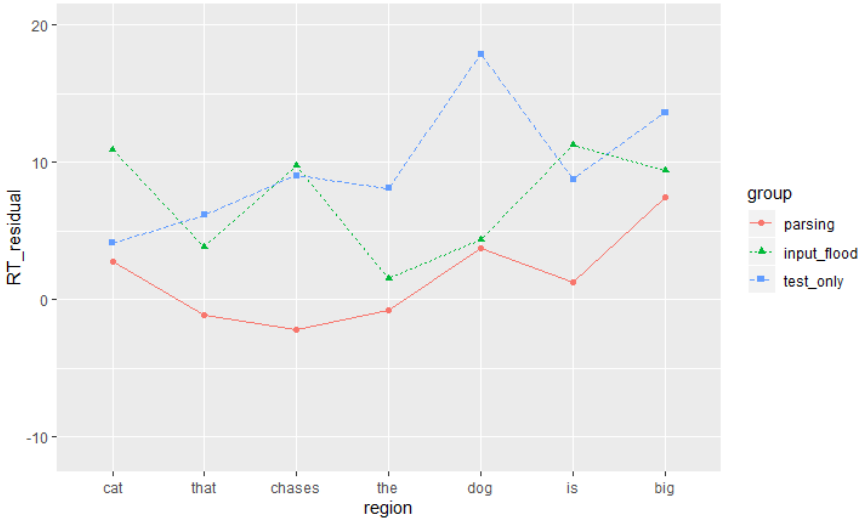
For SRC-I, at the pre-test, in general, the test-only group were more sensitive to the sentence-picture mismatch than the parsing and the input flood group. Although all the three groups showed some sensitivity to sentence-picture mismatch at the third critical word '*bike*', the residual RT difference for the test-only group was much bigger than the other two groups. At the post-test, the parsing group and the test-only group had similar sensitivity to the mismatch across the sentence, while the input flood group did not. At the delayed post-test, the three groups performed similar, and sensitivity to the mismatch could be observed at the word after the third critical word.

For ORC-A, at the pre-test, the three groups had similar residual RT differences across the sentence. They showed some sensitivity to the sentence-picture mismatch at the third critical word, and the word after it. At the post-test, the parsing group had slower residual RTs in reading mismatched items than the matched items at the third critical word. The input flood and the test-only group seemed showed sensitivity to the

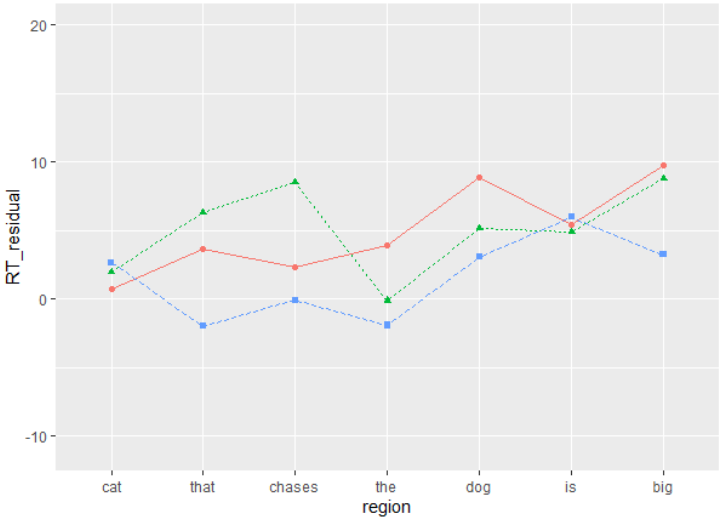
mismatch at the word after the third critical word. At the delayed post-test, the three groups showed similar reading patterns at the critical regions, and sensitivity to the mismatch was negligible (or not sensitive at all). The test-only group seemed to be sensitive to the mismatch at the word after the third critical word.

For ORC-I, at the pre-test, the parsing and the test-only group showed had slightly slower residual RTs in mismatched items than the matched items at the third critical word. At the post-test, only the input flood group was sensitive to the sentence-picture mismatch at the third critical word. The parsing and the test-only group showed sensitivity to the mismatch at the word after the third critical word. At the delayed post-test, all the three groups had slower residual RTs in mismatched items than the matched items across the three critical words. However, the changes of the residual RT differences across the critical regions were very small. Sensitivity to the mismatch could be observed at the word after the critical region for the parsing and the test-only group.

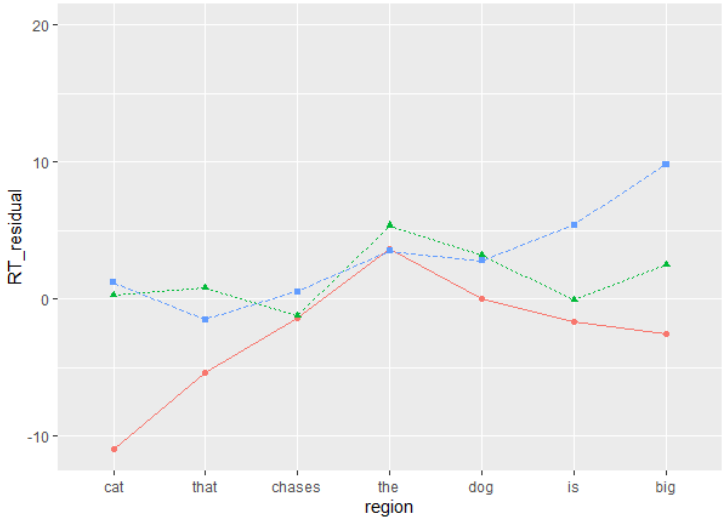
Figure 4. 2. 5 Comparison of residual RT differences (residual mean mismatched – residual mean matched RTs) of three groups for SRC-A structure in self-paced reading



Pre-test

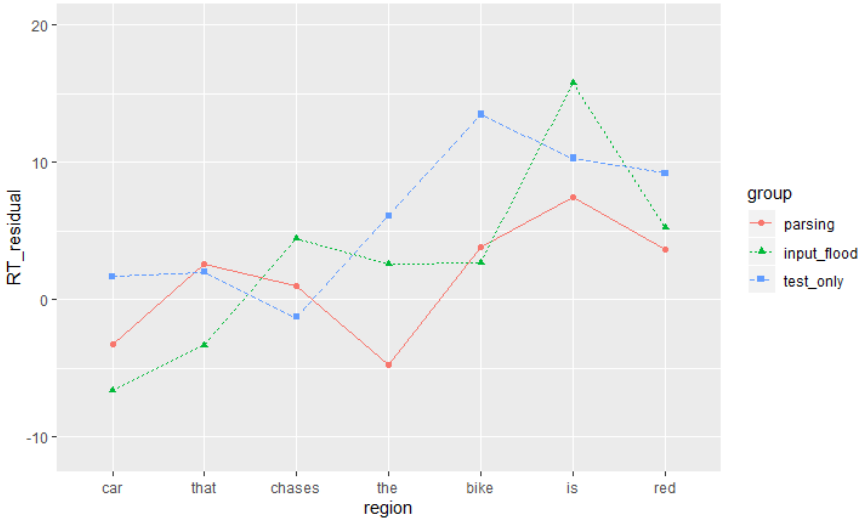


Post-test

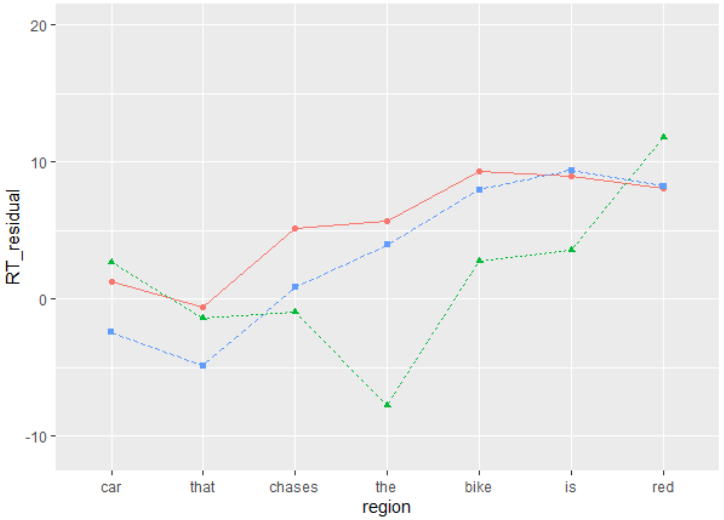


Delayed post-test

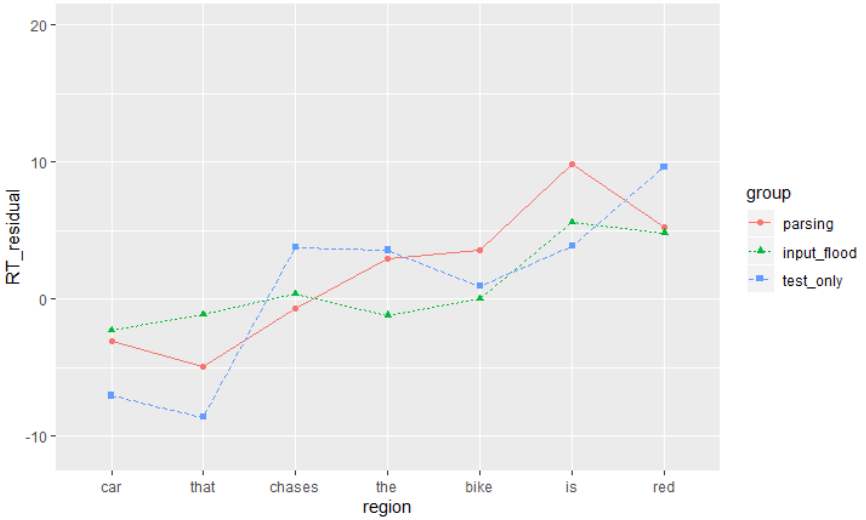
Figure 4. 2. 6 Comparison of residual RT differences (residual mean mismatched – residual mean matched RTs) of three groups for SRC-I structure in self-paced reading



Pre-test

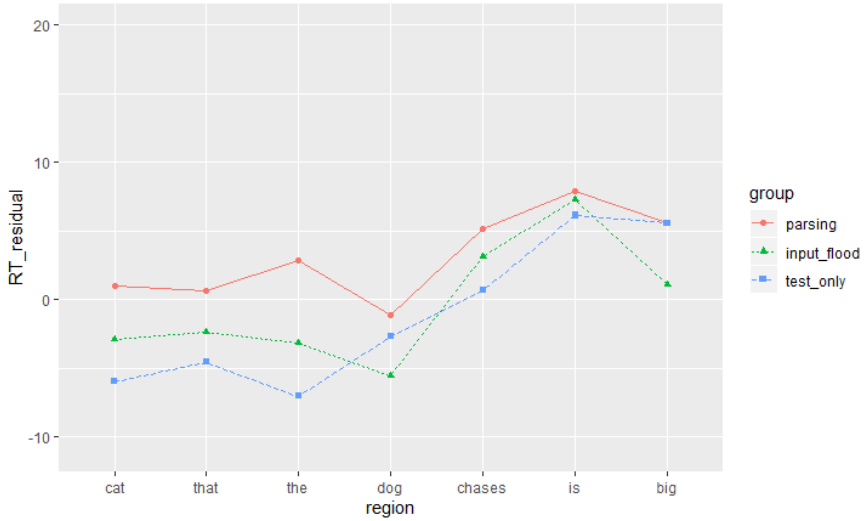


Post-test

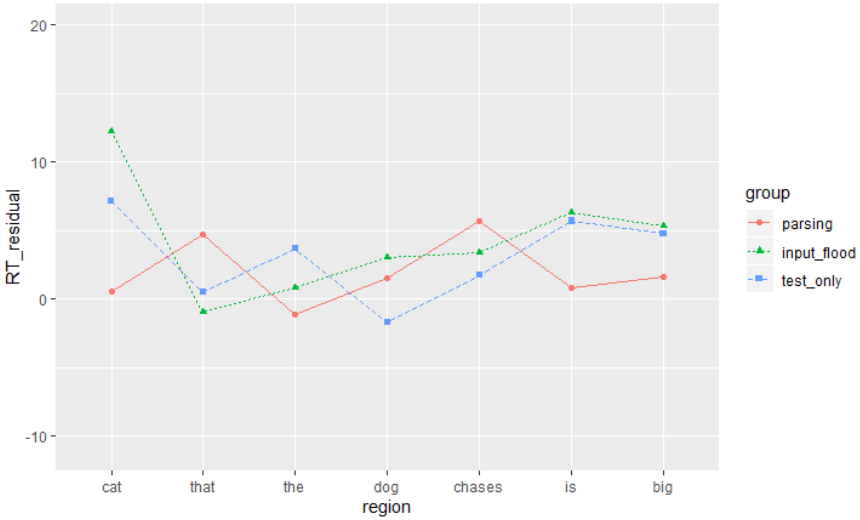


Delayed post-test

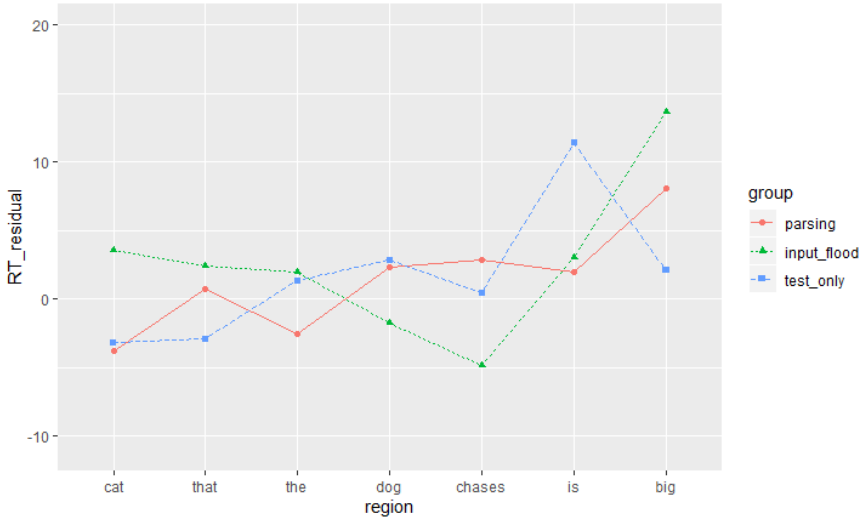
Figure 4. 2. 7 Comparison of residual RT differences (residual mean mismatched – residual mean matched RTs) of three groups for ORC-A structure in self-paced reading



Pre-test

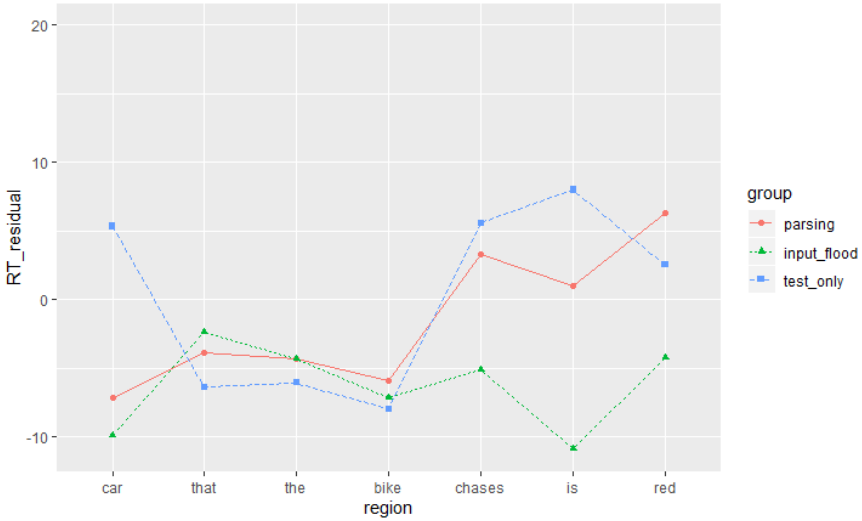


Post-test

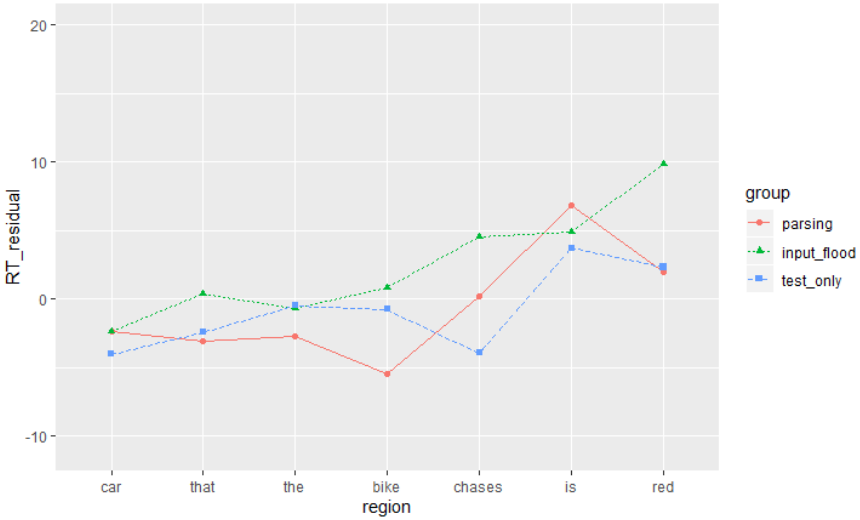


Delayed post-test

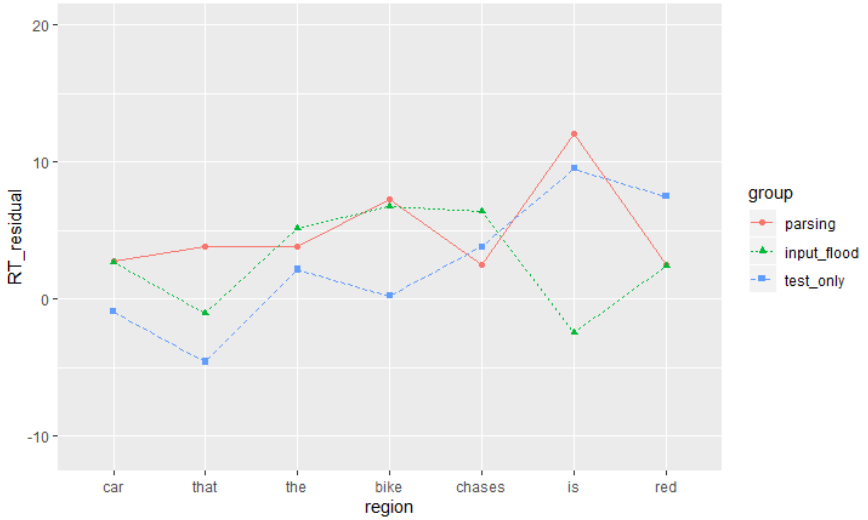
Figure 4. 2. 8 Comparison of residual RT differences (residual mean mismatched – residual mean matched RTs) of three groups for ORC-I structure in self-paced reading



Pre-test



Post-test



Delayed post-test

4.2.2.3 Examination of effect sizes

Table 4.2.12 presents the within-group effect sizes, reflecting differences between mismatched and matched items for the first, second, third critical words and the whole sentence at the pre-, post, and delayed post-test of each group.

Overall, most of the effects were negligible, because the 95% CI passed through zero. The reliable sensitivity to anomaly detection (i.e., following Avery & Marsden, 2019; *d* = .19 [.09, .29]) at the post-test or the delayed post-test could mainly be observed with the parsing group.

For the parsing group, the reliable sensitivity could be observed in the SRC-A (at the third critical word) and the SRC-I (at the third critical word and the whole sentence) items at the post-test phase. In addition, for ORC-I structure, sensitivity at the delayed post-test at the second critical word and the whole sentence were regarded as reliable.

For the input flood and the test-only group, reliable sensitivity only could be found at the pre-test phase for SRCs. This indicated that these two groups were unlikely to gain sensitivity across the time for these structures. Nevertheless, the input flood group had reliable sensitivity to the mismatch for the ORC-I structure based on the whole sentence reading time at the delayed post-test. It indicated that the input flood group might have some gains from the training, but sensitivity could not be observed at the critical words.

Table 4. 2. 12 Within-group effect sizes (Cohen's d) [95% CI] between mismatched and matched items for residual reaction times in the self-paced reading test

			parsing	input flood	test-only
SRC-A (k=10)	1st critical word	pre-test	-.06 [-.24, .12]	.17 [-.02, .36]	.17 [-.02, .36]
		post-test	.06 [-.12, .24]	.19 [-.01, .39]	.00 [-.19, .18]
		delayed post-test	-.04 [-.22, .15]	-.03 [-.23, .16]	.02 [-.17, .20]
	2nd critical word	pre-test	-.01 [-.19, .16]	.03 [-.15, .21]	**.20 [.01, .38]**
		post-test	.11 [-.07, .29]	-.03 [-.22, .17]	-.05 [-.23, .13]
		delayed post-test	.10 [-.09, .28]	.12 [-.07, .32]	.11 [-.08, .30]
	3rd critical word	pre-test	.06 [-.11, .24]	.08 [-.11, .26]	**.29 [.10, .48]**
		post-test	**.19 [.01, .38]**	.13 [-.06, .33]	.07 [-.11, .26]
		delayed post-test	-.01 [-.19, .18]	.09 [-.11, .28]	.04 [-.15, .22]
	Whole sentence	pre-test	.08 [-.10, .26]	**.25 [.06, .44]**	**.36 [.17, .55]**
		post-test	.15 [-.03, .33]	.12 [-.07, .32]	.12 [-.07, .30]
		delayed post-test	-.09 [-.28, .10]	.11 [-.08, .30]	.16 [-.03, .34]
SRC-I (k=10)	1st critical word	pre-test	.02 [-.16, .19]	.08 [-.11, .26]	-.04 [-.22, .15]
		post-test	.11 [-.07, .30]	-.02 [-.22, .18]	.03 [-.16, .22]
		delayed post-test	.00 [-.19, .19]	.00 [-.20, .19]	.12 [-.07, .13]
	2nd critical word	pre-test	-.14 [-.31, .04]	.05 [-.13, .23]	.13 [-.06, .32]
		post-test	.15 [-.03, .33]	-.19 [-.39, .00]	.09 [-.10, .28]
		delayed post-test	.10 [-.09, .29]	-.02 [-.22, .17]	.10 [-.09, .29]
	3rd critical word	pre-test	.06 [-.12, .24]	.02 [-.18, .21]	**.23 [.04, .42]**
		post-test	**.21 [.03, .39]**	.06 [-.14, .25]	.16 [-.03, .35]
		delayed post-test	.09 [-.10, .28]	.01 [-.19, .20]	.00 [-.19, .19]
	Whole sentence	pre-test	.04 [-.13, .21]	.12 [-.07, .30]	**.24 [.05, .43]**
		post-test	**.23 [.04, .41]**	.19 [.00, .39]	.17 [-.01, .35]
		delayed post-test	.15 [-.03, .34]	.00 [-.19, .19]	.14 [-.05, .32]
ORC-A (k=10)	1st critical word	pre-test	.08 [-.09, .26]	-.09 [-.27, .10]	-.19 [-.37,.00]
		post-test	-.03 [-.20, .15]	.04 [-.16, .24]	.10 [-.08, .29]
		delayed post-test	-.07 [-.26, .11]	.04 [-.16, 24]	.04 [-.15, .22]
	2nd critical word	pre-test	-.02 [-.20, .15]	-.10 [-.29, .08]	-.05 [-.23, .14]
		post-test	.04 [-.13, .22]	.08 [-.12, .27]	-.02 [-.21, .18]
		delayed post-test	.09 [-.09, .28]	-.04 [-.24, .15]	.07 [-.12, .26]

	3rd critical word	pre-test	.08 [-.10, .27]	.06 [-.13, .24]	.04 [-.15, .23]
		post-test	.11 [-.07, .30]	.07 [-.14, .27]	.04 [-.16, .23]
		delayed post-test	.06 [-.13, .25]	-.11 [-.31, .09]	.01 [-.18, .20]
	Whole sentence	pre-test	**.20 [.03, .38]**	-.02 [-.20, .17]	.03 [-.16, .21]
		post-test	.13 [-.05, .31]	.15 [-.05, .35]	.14 [-.05, .33]
		delayed post-test	.09 [-.10, .28]	.16 [-.03, .35]	.16 [-.03, .35]
ORC-I (k=10)	1st critical word	pre-test	-.12 [-.30, .06]	-.11 [-.30, .07]	-.14 [-.33, .05]
		post-test	-.07 [-.25, .11]	-.01 [-.20, .19]	-.01 [-.20, .17]
		delayed post-test	.09 [-.09, .28]	.13 [-.06, .33]	.06 [-.13, .25]
	2nd critical word	pre-test	-.16 [-.34, .02]	-.13 [-.32, .05]	-.13 [-.32, .06]
		post-test	-.15 [-.33, .03]	.03 [-.16, .23]	-.01 [-.20, .17]
		delayed post-test	**.22 [.03, .42]**	.12 [-.07, .32]	.00 [-.19, .19]
	3rd critical word	pre-test	.09 [-.10, .27]	-.08 [-.27, .11]	.07 [-.12, .26]
		post-test	.00 [-.17, .18]	.12 [-.08, .32]	-.08 [-.27, .11]
		delayed post-test	.05 [-.14, .24]	.12 [-.07, .32]	.07 [-.12, .26]
	Whole sentence	pre-test	.02 [-.16, .19]	-.01 [-.19, .18]	.06 [-.13, .24]
		post-test	.14 [-.04, .32]	**.22 [.02, .42]**	.00 [-.18, .18]
		delayed post-test	**.20 [.02, .39]**	.07 [-.12, .26]	.12 [-.07, .31]

Note: Bold typeface refers to the effects indicating the reliable sensitivity to the sentence-picture anomaly

4.2.2.4 Inferential statistical analysis

The mixed effects models were used to conduct the inferential statistical analysis of the residual RTs in self-paced reading test. As stated in the Methodology and Methods Chapter, the analyses of the online measures were more likely to be exploratory than confirmatory. Thus, the random effects only included the by-subject and the by-item intercepts (formula: model1=lmer(resid ~ group*stage*match_mismatch + (1|subject) + (1|item), data=SPR, control = lmerControl(optimizer = "bobyqa", optCtrl = list (maxfun=100000)))).

The residual RTs of the first, second, third critical words and the whole sentence of each type of relative clause were analysed separately. The models with the baseline of the test-only group and the input flood group will be reported separately. For the full model results see Appendix 27 (baseline of the test-only group) and Appendix 28 (baseline of the input flood group).

SRC-A

At the first critical word (verb) (marginal R^2 = .01, conditional R^2 = .01), the statistically significant effects were found with the comparison between match and mismatch (*b* [CI] = 9.02 [2.07, 15.97], SE = 4.23, *t* = 2.14, *p* = .033), which indicated that in general, all the group as a whole had slower residual RTs in mismatched items than the matched items regardless of the test phase. In addition, a three-way interaction between group (test-only vs. parsing), test phase (pre-test vs. post-test) and match or not (match vs. mismatch) were found to be reliable, because the 95% CI of the estimate *b* did not pass through zero (*b* [CI] = 13.59 [.04, 27.17], SE = 8.24, *t* = 1.65, *p* = .099), though the *p* value was more than .05. In the model with the baseline of the input flood group, no meaningful statistically significant effect could be found.

At the second critical word (*the*) (marginal R^2 = .01, conditional R^2 = .01), the participants as a whole had statistically significant slower residual RTs in reading mismatched items than the matched items, regardless of the test phase (*b* [CI] = 8.07 [1.55, 14.58], SE = 3.96, *t* = 2.04, *p* = .042). Compared to the test-only group, the parsing group had significantly slower residual RTs in the mismatched items than

matched items at the post-test relative to the pre-test (*b* [CI] = 14.64 [2.06, 27.23], SE = 7.65, *t* = 1.91, *p* = .056). No statistically significant effect could be observed in the model which had the input flood group as the baseline.

At the third critical word (*noun*) (marginal R^2 = .02 conditional R^2 = .04), the statistically significant effects were found with the comparison between matched and mismatched items, regardless group and test phase (*b* [CI] = 11.15 [2.06, 27.23], SE = 4.59, *t* = 2.43, *p* < .001). In addition, compared to the test-only group, overall, the parsing group (*b* [CI] = -14.40 [-24.90, -3.90], SE = 6.38, *t* = -2.26, *p* = .024) and the input flood group (*b* [CI] = -13.72 [-24.44, -2.99], SE = 6.52, *t* = -2.10, *p* = .035) had significantly faster residual RTs in mismatched items than the matched items. Compared to the pre-test, the participants as a whole had significantly faster residual RTs in mismatched items in matched items at the post-test (*b* [CI] = -15.17 [-25.97, -4.38], SE = 6.56, *t* = -2.31, *p* = .021) and the delayed post-test (*b* [CI] = -15.46 [-26.27, -4.65], SE = 6.57, *t* = -2.35, *p* = .019). Moreover, a three-way interaction between the group (test-only vs. parsing), test phase (pre- vs. post-) and match or mismatch (match vs. mismatch) (*b* [CI] = 20.43 [5.50, 35.36], SE = 9.08, *t* =2.25, *p* = .025) had statistically significant effect. No statistically significant effects could be found in the model with the input flood group as the baseline.

In the whole sentence (marginal R^2 = .02 conditional R^2 = .03), the comparison between matched and mismatched items were found statistically significant (*b* [CI] = 15.79 [9.82, 21.77], SE = 3.63, *t* = 4.35, *p* < .001). The statistically significant two-way interactions were: test-only vs. parsing: match vs. mismatch (*b* [CI] = -11.93 [-20.00, -3.87], SE = 4.90, *t* = -2.44, *p* = .015), pre- vs. post-: match vs. mismatch (*b* [CI] = -11.52 [-19.75, -3.28], SE = 5.01, *t* = -2.30, *p* = .022), and pre- vs. delayed post-: match vs. mismatch (*b* [CI] = -10.03 [-18.22, -1.84], SE = 4.98, *t* = -2.21, *p* = .044). In addition, compared to the test-only group, the parsing group had significantly slower residual RTs in mismatched items relative to matched items at the post-test than the pre-test (*b* [CI] = 13.92 [2.47, 25.37], SE = 6.96, *t* = 2.00, *p* = .046). In the model of the input flood group as the baseline, no statistically significant effect related to the comparisons

between the input flood and the parsing group could be observed.

SRC-I

At the first critical word (*verb*) (marginal R^2 = .01 conditional R^2 = .04), no interaction that related to sensitivity to sentence-picture mismatch had statistically significant effect in both test-only and the input flood baselines models.

At the second critical word (*the*) (marginal R^2 = .01 conditional R^2 = .03), a two-way interaction between group and match or mismatch was found statistically significant (test-only vs. parsing: match vs. mismatch: *b* [CI] = -10.78 [-19.69, -1.86], SE = 5.42, *t* = -1.99, *p* = .047). The model which adopted input flood group as the baseline showed that compared to the input flood group, the parsing group read significantly slower in mismatched items relative to matched items at the post-test than the pre-test (*b* [CI] = 20.17 [7.39, 32.96], SE = 7.77, *t* = 2.60, *p* = .009).

At the third critical word (*noun*), (marginal R^2 = .01 conditional R^2 = .06), the statistically significant effect was found with the comparison between match and mismatch (*b* [CI] = 13.71 [5.45, 21.97], SE = 5.02, *t* = 2.73, *p* = .007). In addition, all the groups as a whole had significantly shorter residual RTs in mismatched items relative to matched items at the delayed post-test compared to the pre-test (*b* [CI] = -12.77 [-23.50, -2.04], SE = 6.52, *t* = -1.96, *p* = .050). In the model with the input flood group as the baseline, no statistically significant effect could be observed.

In the whole sentence (marginal R^2 = .02 conditional R^2 = .08), the comparison between matched and mismatched items were found significantly significant (*b* [CI] = 10.60 [4.06, 17.14], SE = 3.98, *t* = 2.67, *p* = .008). In addition, a two-way interaction between the group and match or mismatch was found to be statistically significant: test-only vs. parsing: match vs. mismatch (*b* [CI] = -9.14 [0.03, 21.63], SE = 4.72, *t* = -1.94, *p* = .054). In the model with the baseline of the input flood group, there was no statistically significant effect that could be found.

ORC-A

At the first critical word (*the*) (marginal R^2 = .01 conditional R^2 = .02, there were two two-way interactions that were found statistically significant: test-only vs. parsing:

match vs. mismatch (*b* [CI] = 9.94 [1.69, 18.18], SE = 5.01, *t* = 1.98, *p* = .048) and pre-test vs. post-test: match vs. mismatch (*b* [CI] = 10.82 [2.39, 19.25], SE = 1.53, *t* = 2.11, *p* = .035). In addition, a three-way interaction between the group (test-only vs. parsing), test phase (pre-test vs. post-test) and match or mismatch (match vs. mismatch) had statistically significant effect (*b* [CI] = -14.90 [-26.57, -3.22], SE = 7.10, *t* = -2.10, *p* = .036). No statistically significant effect could be observed in the model which had the input flood group as the baseline.

At the second critical word (*noun*) (marginal R^2 = .01 conditional R^2 = .02), no statistically significant effect could be found with the model that either had the test-only or the input flood group as the baseline.

At the third critical word (*verb*) (marginal R^2 = .01 conditional R^2 = .02), neither the model that had the test-only group as the baseline nor had the input flood group as the baseline had the statistically significant effect.

In the whole sentence (marginal R^2 = .01 conditional R^2 = .03), no statistically significant effects could be observed with the model that had the test-only group as the baseline and the model that that had the input flood group as the baseline.

ORC-I

At the first critical word (*the*) (marginal R^2 = .01 conditional R^2 = .04), there was no statistically significant effect that could be found with the model that had the test-only group as the baseline and with the model whose baseline was the input flood group.

At the second critical word (*noun*) (marginal R^2 = .01 conditional R^2 = .02), the three groups as a whole had reliable faster residual RTs in reading mismatched items compared to the matched items (*b* [CI] = -8.18 [-15.31, -1.04], SE = 4.34, *t* = -1.89, *p* = .060). In the model with the baseline of the input flood group, no statistically significant effect related to the comparison between the input flood group and parsing group could be observed.

At the third critical word (*verb*) (marginal R^2 = .01 conditional R^2 = .05), one three-way interaction between group (test-only vs. input flood), test phase (pre-test vs. post-test) and match or mismatch (match vs. mismatch) was found significant (*b* [CI] =

19.20 [2.91, 35.49], SE = 9.90, *t* = 1.94, *p* = .053). No statistically significant effects could be found in the model with the baseline of the input flood group.

In the whole sentence (marginal R^2 = .01 conditional R^2 = .07), no statistical effect could be found in the model with the baseline of the test-only or the input flood group.

4.2.2.5 Summary of the results in the self-paced reading test

In summary, teaching parsing strategies seemed to facilitate sensitivity to the sentence-picture mismatch in SRCs, which was seen in the within-group effect sizes and the inferential statistical results. Yet, the teaching effects could only be observed at the post-test and did not last through to the delayed-post. The within-group effect sizes, reflecting differences between the residual RTs of mismatched and matched per group per test-phase, showed that the parsing group was not sensitive to the sentence-picture mismatch at any critical word or the whole sentence at the pre-test. However, they showed the reliable sensitivity to the mismatch at the post-test at the third critical word for SRC-A and SRC-I, and in the whole sentence for SRC-I. The other two groups might have been sensitive to the sentence-picture mismatch at the pre-test, but sensitivity was not observed at the post- and delayed post-test. In addition, the inferential statistics indicated that for SRC-A, the parsing group had more gains in sensitivity to the mismatch than the test-only group in the comparison between the pre- and the post-test at all the three critical regions and the whole sentence. However, it must be acknowledged that these significant effects might not only be because the parsing group changed over time, but might also due to the test-only group being sensitive to the mismatch at pre-test but did not show sensitivity at the post-test (see 4.2.12). For SRC-I, the statistical results indicated an advantage of the parsing group over the input flood group in terms of sensitivity to the mismatch at the second critical word at the post-test.

For ORCs, the descriptive results and the effect sizes between the mismatched and matched items showed that the parsing group did not gain at the post- and delayed post-test. In addition, the effect sizes showed that the input flood group showed reliable sensitivity to the mismatch at the post-test for ORC-I based on the

average residual RTs of the whole sentence, which indicated that the input flood group might have gained from the training. The inferential statistics also suggested that the input flood group read mismatched items significantly slower than the matched items relative to the test-only group at the third critical word, in the comparison between the pre- and the post-test. However, this statistically significant effect might due to the input flood read the matched items slower than the mismatched items at the pre-test, and the within-group effect size of the input flood group at the third critical word was negligible. Thus, the significant effect might be a chance finding.

4.2.3 To what extent, and at what point in the sentence, are effects observable in online comprehension as measured by eye movements?

The visual world eye-tracking tests were used to test the online comprehension. Each critical item in the test contained three critical words, a verb, an article 'the', and a noun (in a different order for SRCs and ORCs). The fixation proportions of looking at the targets from the onset of the first critical word to the offset of the third critical word were analysed.

Sections 4.2.3.1 to 4.2.3.4 present the results of SRC-A, SRC-I, ORC-A and ORC-I structures respectively. In each section, line plots of pre-, post- and delayed post-test including the fixation proportions of looking at the targets and distractors from 200ms before the onset of the first critical word and lasting for 3000ms is provided and analysed initially. In each plot, the offsets of the first, second and the third words are indicated with vertical lines.

Then, the mixed-effects growth curve analysis was used to conduct the inferential statistical analysis. For each target structure, the results for each critical word starting from the word onset and lasting until the onset of the next critical word are presented separately. The models with the baseline of the test-only group and the input flood group are reported separately. As illustrated in Section 3.4.1.3, in selecting the time vector, the primary model only included the first-order time vector, and the second-, third-time vectors were added step by step. However, for the critical word 'the', because the length of this word was less than 150 ms, the maximum of the second-time order vector could be added. The formulas for the models with different time vectors are shown below.

Linear time vector model: gca_model1 = lmer(e_log ~ (ot1)*phase*group + (1 | subject) + (1 | trial), control = lmerControl(optimizer="bobyqa"),data=eye_data)

Quadratic time vector model: gca_model2 = lmer(e_log ~ (ot1+ot2)*phase*group + (1 | subject) + (1 | trial), control = lmerControl(optimizer="bobyqa"),data=eye_data)

Cubic time vector model: gca_model3 = lmer(e_log ~ (ot1+ot2+ot3)*phase*group + (1 | subject) + (1 | trial), control = lmerControl(optimizer="bobyqa"),data=eye_data)

The AIC and LRT results for selecting time vectors are shown in Appendix 23. For the details for the model results see Appendix 29 (baseline of the test-only group) and Appendix 30 (baseline of the input flood group).

4.2.3.1 eye-tracking results for SRC-A structure

Analysis of fixation proportion

Figure 4.2.9 indicated that for SRC-A, the two training groups (i.e., the parsing and the input flood group) did not have substantial gains across the three test phases, in terms of the moment at which the proportion of looking at the targets diverged from looking at the distractors. For the parsing group, they fixated on the target picture before the end of the first critical word across the pre-, post- and delayed post-test, and the proportion of looking at the targets and distractors diverged at around 500 ms after the onset of the first critical word. For the input flood group, they even fixated on the target pictures later at the post- and delayed post-test compared to the pre-test. However, the test-only group indeed fixated on the targets earlier at the post-test compared to the pre-test. At the pre-test, they started to continuously looking at the targets at the third critical word (at around 700 ms after the onset of the first critical word). The divergent point moved forward to the first critical word (at around 400 ms after onset of the first critical word) at the post-test, but moved back to the third critical word at the delayed post-test.

Model analysis

For the first critical word, the AIC indicated that the model with the linear time vector fitted the data best, but the LRT suggested that the model with the combination of linear, quadratic and cubic time vector was the best fitting one (see Appendix 20). However, as the line charts shown, the fixation proportion of looking at the targets had more than one bend, so the model with the cubic time vector might fit the data better than the linear one. Thus, the cubic model was adopted (marginal R^2=.00 conditional

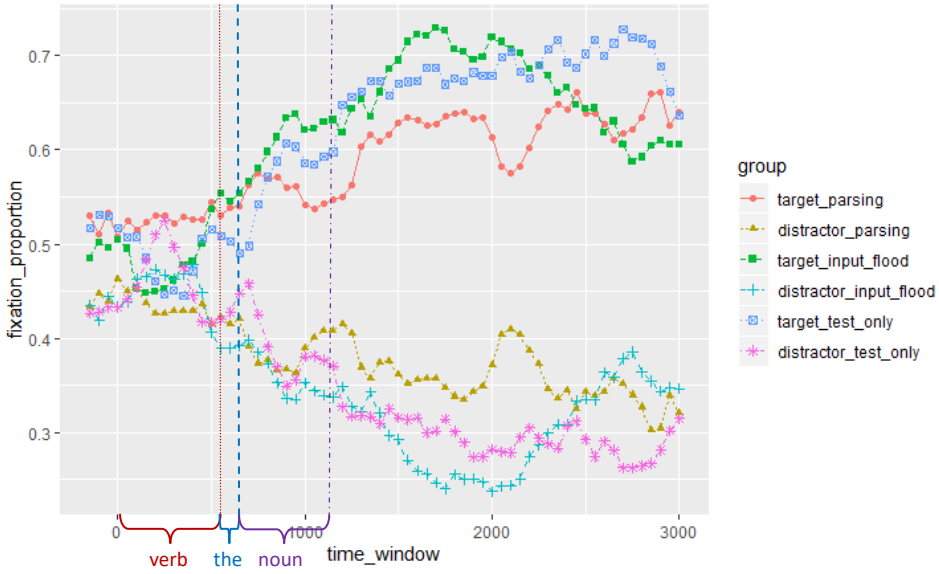
R^2=.01). Neither the model with the baseline of the test-only group nor with the baseline of the input flood group had statistically significant effects.

For the second critical word, the model with the linear time vector was the best fitting (marginal R^2 = .01 conditional R^2 = .01). A two-way interaction between test phase (pre- vs. delayed post-test) and group (test-only vs. input flood group) was found to be statistically significant (*b* = .35 [.13, .56], SE = .13, *t* = 2.66, *p* = .008). In addition in the model with the baseline of the input flood group, the interaction between test phase (pre- vs. delayed post-test) and group (input flood vs. parsing group) had reliable effects (*b* = -.23 [-.44, -.02], SE = .13, *t* = -1.83, *p* = .067), as the 95% CI around the estimate *b* did not pass through zero.

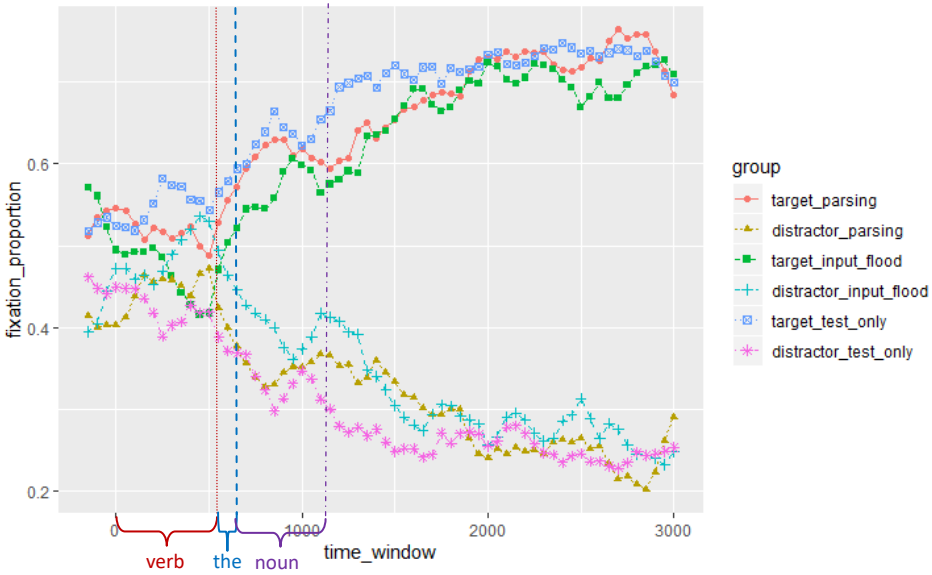
For the third critical word, the AIC and LRT results conflicted (see Appendix 20). The AIC indicated that the model with linear time vector fits the model best, while the LRT suggested that both the model with the quadratic and with the cubic fitted the data better than the one with linear. Based on the plots, the lines of the third critical words had more than one bend. Thus, the model with cubic time vector was adopted (marginal R^2 = .00 conditional R^2 = .01). No statistically significant effects could be observed in the model with test-only group baseline or with the input flood group baseline.

Figure 4. 2. 9 The fixation proportion of looking at the target and distractor for SRC-A

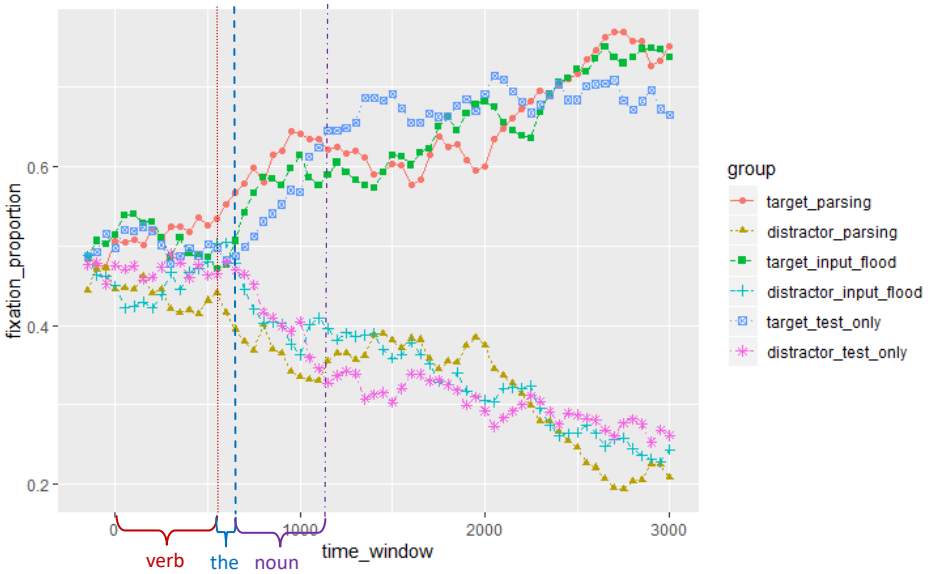
Example sentence (SRC-A):
The cat that chases (verb) the dog (noun) is big.



Pre-test



Post-test



Delayed post-test

Note: red vertical line: offset of the first critical word; blue vertical line: offset of the second critical word; purple vertical line: offset of the third critical word

4.2.3.2 eye-tracking results for SRC-I structure

Analysis of fixation proportion

Figure 4.2.10 presents the fixation proportions of looking at the targets and distractors for SRC-I structure. The line charts indicated that the three groups fixated on the target pictures earlier at the post- and/or at the delayed post-test compared to the pre-test. At the pre-test, all the three groups did not fixate on the target picture until the third critical word. At the post-test, it could be observed that the parsing group and the input flood group started to fixate on the target picture before the end of the first critical word. On the other hand, the test-only group could only start to continuously look at the targets during the third critical word. At the delayed post-test, the eye-movement pattern was similar for the three groups. They started to fixate on the target picture during the second critical word.

Model analysis

For the first critical word, the AIC suggested that the linear model fitted the data best, while the LRT indicated the model with cubic time vector was the best-fitting one. Considering the lines of the first critical word has more than one bend, the model with cubic time vector was selected (marginal R^2 = .00 conditional R^2 = .00). In the model of the test-only group as the baseline, the results indicated that as a whole, the input flood group had smaller proportion of looking at the targets compared to the test-only group (test only vs. input flood: *b* = -.10 [-.18, -.02], SE = .05, *t* = -2.00, *p* = .046). In addition, two two-way interactions between test phase and group (pre- vs. post-test: test-only vs. input flood group: *b* = .13 [.02, .24], SE = .07, *t* = 1.93, *p* = 0.053; pre- vs. delayed post-test: test-only vs. input flood group: *b* = .12 [.01, .23], SE = .07, *t* = 1.74, *p* = .084) and a three-way interaction between the first time vector, test phase and group (linear time vector: pre- vs. post-test: test-only vs. parsing group: *b* = -.43 [-.81, -.04], SE = .23, *t* = -1.84, *p* = .067) were found to be reliable. However, the three-way reliable effects found with the parsing and the test-only group might not have practical meaning. The lower fixation proportion of looking at the targets of the parsing group relative to the test-only group in the comparison between the pre- and the post-test

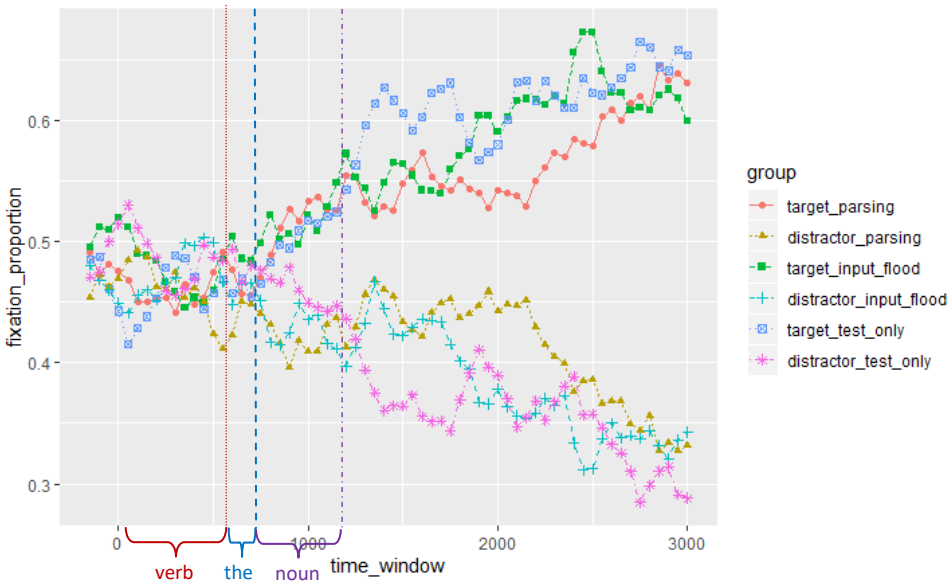
might reflect the proportions between 100 ms to 300 ms after the onset of the first critical word. During that time period, the two groups had the fixation proportions to the targets lower than .50 at both the pre- and the post-test phase. In the model with baseline of the input flood group, a comparison between the input flood and the parsing group (*b* = .09 [.01, .17], SE = .05, *t* = 1.90, *p* = .058) and a three-way interaction between the linear time vector, test phase and group (linear time vector: pre- vs. post-test: input flood group vs. parsing group: *b* = -.39 [-0.81, -.04], SE = .23, *t* = -1,70, *p* = .089) were found to have reliable effects.

For the second critical word, the linear time vector model was the best fitting model (marginal R^2 = .01 conditional R^2 = .01). The interaction between the first-order time vector and pre- to post-test (regardless of the group) was significant (*b* = -.37 [-.64, -.10], SE = .16, *t* = -2.24, *p* = .025). In the model with the input flood group as the baseline, no statistically significant effect could be found.

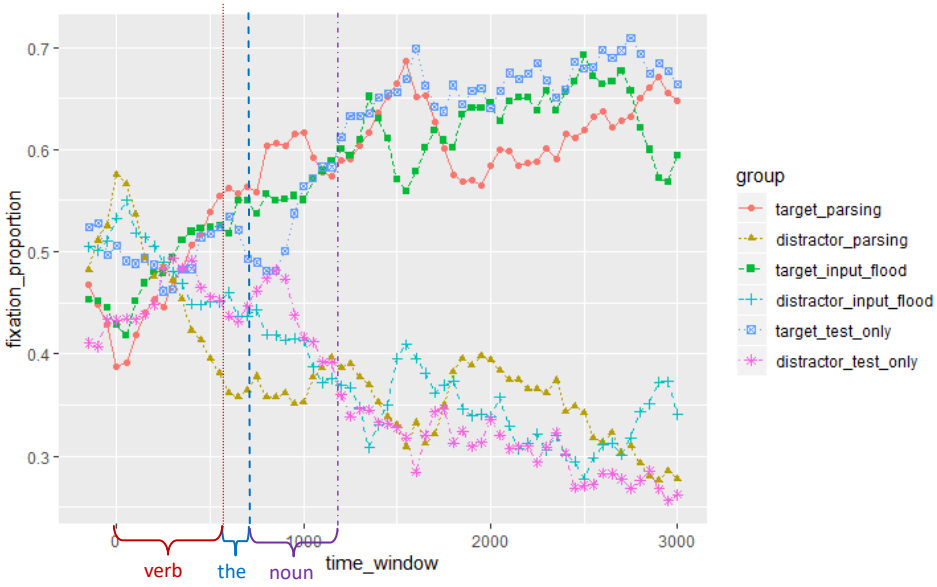
For the third critical word, the model with linear time vector was selected (marginal R^2 = .00 conditional R^2 = .00). No statistically significant effect was found in the model with the test-only group baseline or with the input flood group baseline.

Figure 4. 2. 10 The fixation proportion of looking at the target and distractor for SRC-I

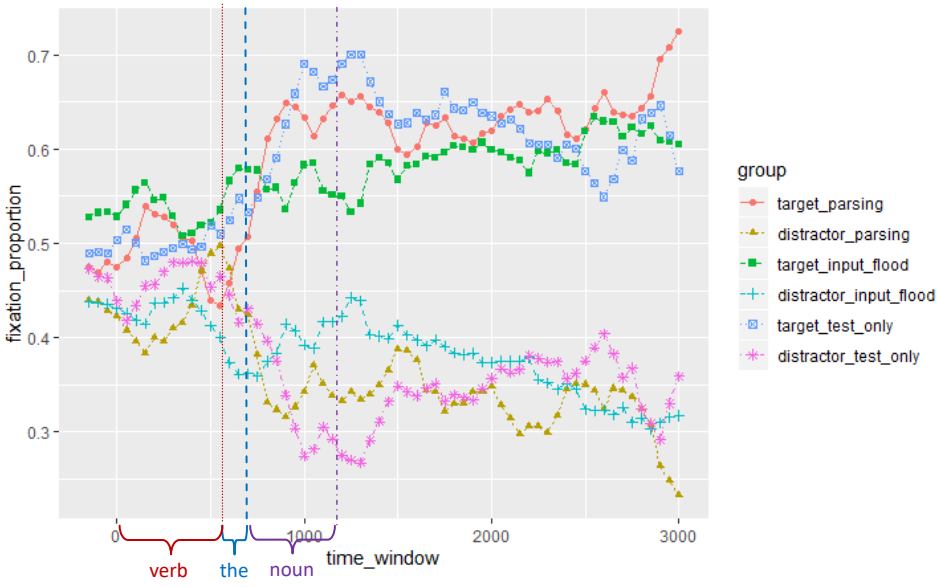
Example sentence (SRC-I):
The car that hits (verb) the bike (noun) is white.



Pre-test



Post-test



Delayed post-test

Note: red vertical line: offset of the first critical word; blue vertical line: offset of the second critical word; purple vertical line: offset of the third critical word

4.2.3.3 eye-tracking results for ORC-A structure

Analysis of fixation proportion

The fixation proportions of looking at the targets and distractors of ORC-A are shown in figure 4.2.11. In general, the three groups fixated on the targets slightly earlier at the post- and delayed post-test compared to the pre-test. At the pre-test, all the three groups started to fixate on the target pictures after hearing the third critical word, and the input flood group fixated on the targets earlier than the other two groups. At the post-test, the input flood group and the test-only group fixated on the targets at end of the second critical word. In addition, the divergent point of looking at the targets and distractors still occurred at the third critical word, but it showed around 300 ms earlier than at the pre-test. At the delayed post-test, all three groups started to fixate on the targets at the end of the second critical word.

Model analysis

For the first critical word, the AIC indicated that the model with the linear time vector fitted the model better, but the LRT suggested that the model with quadratic time vector was better than the linear one. As shown in the plots, during the first critical word, more than one bend could be observed in the lines. Thus, the model with the quadratic time vector was adopted (marginal R^2 = .01 conditional R^2 = .01). No statistically significant effects could be observed in both models (i.e., the model with the baseline of the test-only or the input flood group).

For the second critical word, the results of the AIC and LRT were conflicted. The AIC suggested that the model with linear time vector was the best-fitting one, while the model included cubic time vector was selected by the LRT. Looking at the plots, more than one bend could be observed in the lines, so the results of the cubic one would be reported (marginal R^2 = .00 conditional R^2 = .01). However, both the test-only group baseline model and the input flood group baseline model did not have statistically significant effects.

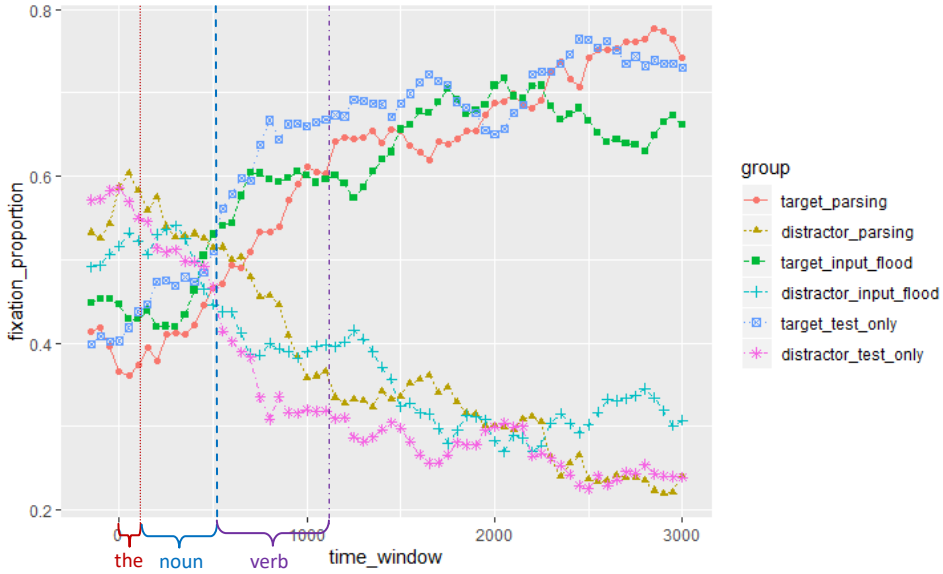
For the third critical word, still, the models suggested by the AIC (the linear time vector) and the LRT (the cubic time vector) were different. Considering the plots, the

model with the cubic time vector was selected (marginal R^2 = .00 conditional R^2 = .00). The negative significant effect was found in the interaction between the linear time vector and the group (Linear time vector: test-only vs. input flood group: *b* = -0.31 [-0.56, -0.07], SE = 0.15, *t* = -2.08, *p* = 0.038). The model with the input flood baseline did not have statistically significant effects.

Figure 4. 2. 11 The fixation proportion of looking at the target and distractor for ORC-A

Example sentence (ORC-A):
The cat that the dog (noun) chases (verb) is big.

Pre-test



Post-test

Delayed post-test

Note: red vertical line: offset of the first critical word; blue vertical line: offset of the second critical word; purple vertical line: offset of the third critical word

4.2.3.4 eye-tracking results for ORC-I structure

Analysis of fixation proportion

Figure 4.2.12 presents the fixation proportions of looking at target and distractor for the ORC-I structure. Compared to the pre-test, the three groups fixated on the targets earlier at the post- and the delayed post-test, and there was no salient difference between groups at each test phase. At the pre-test, the three groups started to fixate on the target picture at the third critical word. At the post- and the delayed post-test, they fixated on the target picture at around the end of the second critical word, which was about 200 ms earlier than at the pre-test.

Model analysis

For the first critical word, the linear time vector model was the best fitting one (marginal R^2 = .01 conditional R^2 = .01). A reliable effect was found with the interaction between the linear time vector and the group (test-only group vs. parsing group) (*b* = -.31 [-.59, -.03], SE = .17, *t* = -1.81, *p* = .070). In the model with the baseline of the input flood group, a two-way interaction (linear time vector: input flood vs. parsing group: *b* = -.57 [-.86, -.28], SE = .18, *t* = -3.19, *p*= .001) and two three-way interactions (linear time vector: pre- vs. post-test: input flood vs. parsing group: *b* = .54 [.12, .95], SE = .25, *t* = 2.17, *p* = .030; linear time vector: pre- vs. delayed post-test: input flood vs. parsing group: *b* = .50 [.08, .91], SE = .25, *t* = 1.97, *p* = .049) were found to be statistically significant. However, significant effects indicated the higher proportion of looking at the targets of the parsing group over the input flood group might not be meaningful. Looking at the line charts, it could be observed that across the three test-phases, the proportions of looking at the targets of the input flood were always under .50 at the first critical word. For the parsing group, at the post-test, although the proportion was higher than .50 during the first critical word, there was a sharp decline to slightly less than .50 at the second critical word. This indicated that the fixation on the targets at the first critical word might just have been by chance. At the delayed post-test, at the first critical word, the proportions of looking at the targets were below .50 for both the parsing and the input flood group.

For the second critical word, the linear time vector model was selected (marginal R^2 = .00 conditional R^2 = .01). No statistically significant effect could be found in the model with the test-only group baseline or with the input flood group baseline.

For the third critical word, the AIC suggested the linear model fitted the data best, while the LRT suggested the model with the quadratic time vector was the best-fitting one. Considering the bends of lines in the plots, the quadratic model was selected (marginal R^2 = .00 conditional R^2 = .01). A two-way interaction was found to be statistically significant (linear time vector: test-only vs. input flood group: *b* = -0.36 [-0.63, -0.09], SE = 0.16, *t* = -2.23, *p* = 0.026). In the model with the baseline of the input flood group, a two-way interaction (linear time vector: input flood vs. parsing group: *b* = .39 [-.85, -.10], SE = 0.16, *t* = 2.41, *p* = 0.016) and a three-way interaction (linear time vector: pre-test vs. post-test: input flood vs. parsing group: *b* = -.48 [.12, .66], SE = 0.23, *t* = -2.08, *p* = 0.037) had statistically significant effects.

Figure 4. 2. 12 The fixation proportion of looking at the target and distractor for ORC-I

Example sentence:
The car that the bike (noun) hits (verb) is black.

Pre-test

Post-test

Delayed post-test

Note: red vertical line: offset of the first critical word; blue vertical line: offset of the second critical word; purple vertical line: offset of the third critical word

4.2.3.5 Summary of the results in the eye-tracking tests

Descriptive analysis from visual inspection (proportion and timing of fixations on target)

In summary, the line charts of the four structures indicated that the changes of eye-movement pattern towards the targets across the time were similar for the three groups. Except for the SRC-A structure, where the participants already fixated on the targets at the first critical word at the pre-test, for the other three structures, all the three groups showed some improvement in the time of looking at the targets at the post- and delayed post-test. The effects of teaching parsing strategies or the input flood training were not substantial in the eye-tracking data.

Model analysis

In summary, for all the structures, although a few statistically significant effects were observed, most of them were not meaningful. The only seemed meaningful significant effect was the interaction between the time vector, the test phase, and the group for the ORC-I structure, which indicated that at the third critical word, the compared to the pre-test, the input flood group had higher proportion of fixation to the targets than the parsing group at the post-test. However, looking at the line charts, in fact, the input flood group did not fixate on the targets earlier than the parsing group at both pre- and the post-test. Thus, this significant effect might not be robust enough to demonstrate the input group had more gains than the parsing group.

4.2.4 To what extent are effects observable in the oral production?

The picture description tests were used to examine oral production, and the accuracy scores were calculated. In this test, the parsing group was expected to have higher accuracy scores than the input flood and test-only groups at post- and delayed post-test, and the input flood group was expected to outperform the test-only group.

4.2.4.1 Descriptive analysis

Overall, for all the structures, all the three groups showed some improvement from pre- to post-test, with the parsing group scoring much higher than the other two groups at the post- and the delayed post-test.

For SRCs, the parsing group showed substantial gains in the accuracy at the post-test and the delayed post-test, though a slight decline in accuracy could be observed at the delayed post-test compared to the post-test for the SRC-I structure. In addition, the input flood group and the test-only group had some gains across the time, but the gains were rather small.

For ORCs, the accuracy of the three groups steadily increased across the time. The post-test scores of the parsing group doubled compared to the pre-test, and at the delayed post-test, the scores showed the continuous improvement relative to the post-test. For the test-only group, the improvement from pre- to post- and delayed post-test also very salient, but the accuracy was still rather low at around .50.

Table 4. 2. 13 Mean (SDs) scores for oral sentence description tests

structure	test phase	parsing	input flood	test-only
SRC-A (k=4 or 6 [a])	pre-test	.54 (.50)	.48 (.50)	.52 (.50)
	post-test	.72 (.45)	.59 (.49)	.56 (.50)
	delayed post-test	.80 (.40)	.64 (.48)	.62 (.49)
SRC-I (k=4 or 6 [a])	pre-test	.55 (.50)	.53 (.50)	.50 (.50)
	post-test	.80 (.40)	.58 (.49)	.54 (.50)
	delayed post-test	.75 (.43)	.65 (.48)	.61 (.49)
ORC-A (k=5)	pre-test	.37 (.48)	.41 (.49)	.21 (.41)
	post-test	.76 (.43)	.52 (.50)	.41 (.49)
	delayed post-test	.81 (.39)	.71 (.46)	.48 (.50)
ORC-I (k=5)	pre-test	.36 (.48)	.43 (.50)	.20 (.40)
	post-test	.78 (.42)	.58 (.50)	.39 (.49)
	delayed post-test	.81 (.40)	.71 (.46)	.51 (.50)

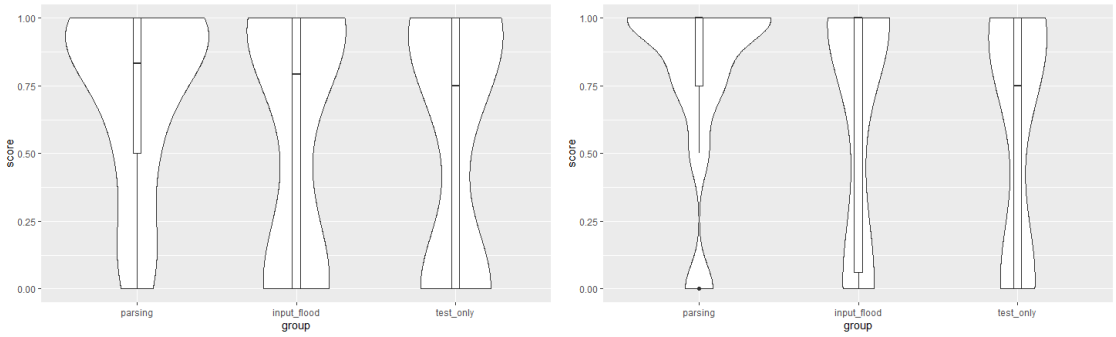
Note: [a] Numbers of items differed between version 1 and version 2, version 3 and version 4 of the tests, respectively; versions were counterbalanced across pre-, post-, and delayed post-test, within groups.

4.2.4.2 Plots

Figure 4.2.13 to 4.2.16 show the violin plots including boxplots about the mean scores of all the items of a target structure from each participant. For SRCs, the distribution of the mean scores at the post- and delayed post-test indicated that the parsing group had more participants scoring at ceiling than the other two groups, while the input flood and the test-only group a the similar distribution.

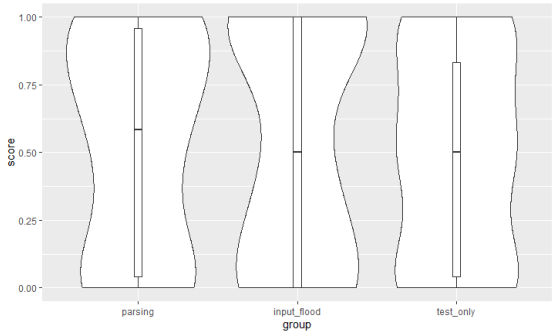
Figure 4. 2. 13 Comparison of accuracy scores of three learner groups in different test phase for SRC-A structure in oral sentence description test

Pre-test

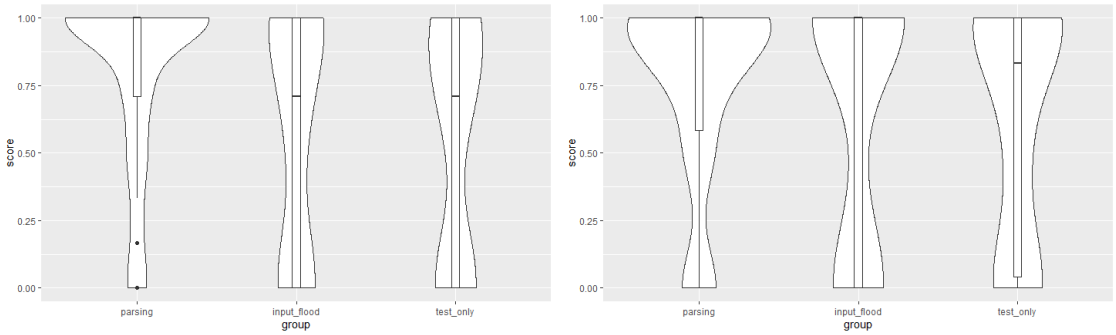


Post-test Delayed post-test

Figure 4. 2. 14 Comparison of accuracy scores of three learner groups in different test phase for SRC-I structure in oral sentence description test



Pre-test



Post-test Delayed post-test

For ORCs, the improvement in the accuracy of the parsing group across time was obvious. At the pre-test, the mean scores of the parsing group were almost equally distributed from 0 to 1 point, with slightly more individuals scored lower than .50. However, at the post- and delayed post-test, most participants in the parsing group scored over .75 or even higher. For the test-only group, the majority of the participants had very low scores at the pre-test, but at the post- and delayed post-tests, the distribution of scores in the test-only group was similar to that of the input flood group.

Figure 4. 2. 15 Comparison of accuracy scores of three learner groups in different test phase for ORC-A structure in oral sentence description test

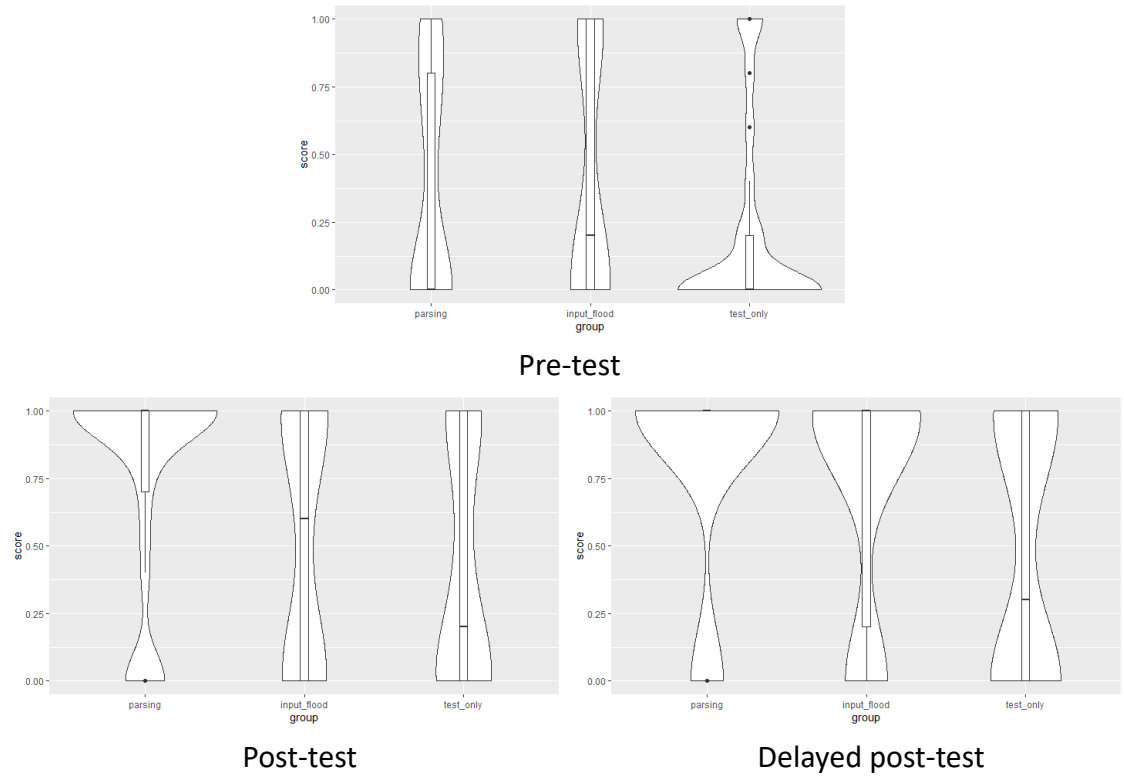
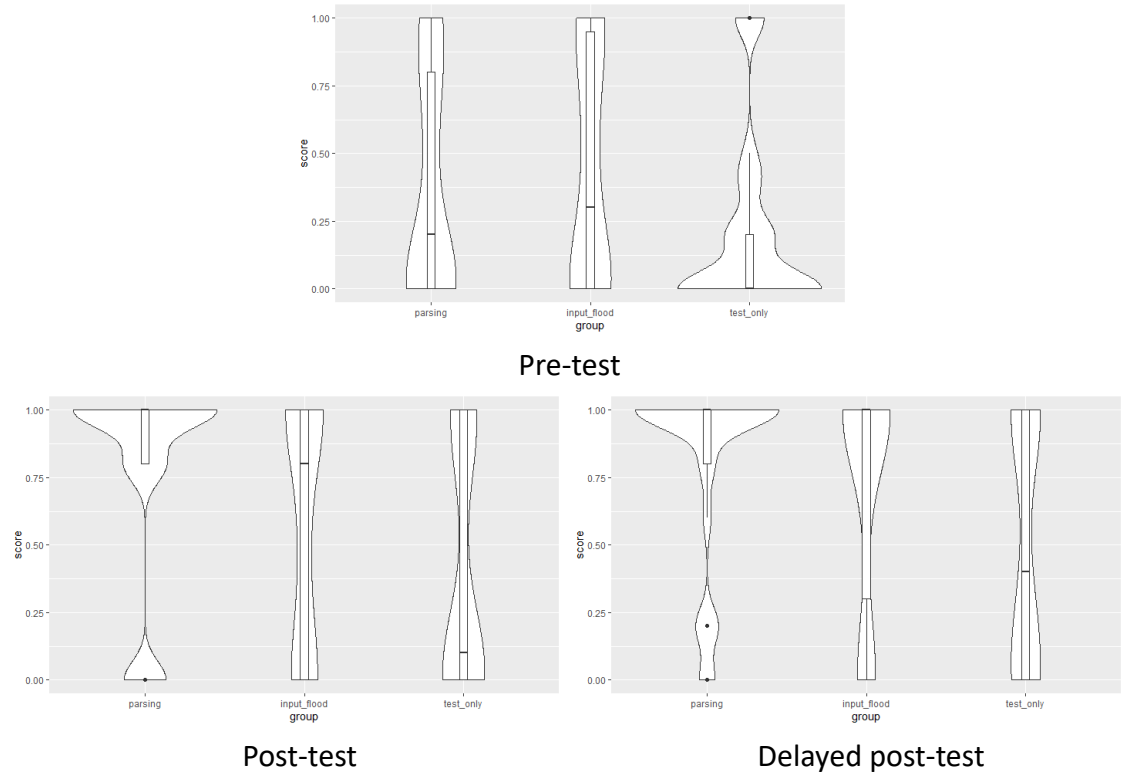


Figure 4. 2. 16 Comparison of accuracy scores of three learner groups in different test phase for ORC-A structure in oral sentence description test



4.2.4.3 Examination of effect size

For within-group contrasts, reflecting changes over time, (see Table 4.2.14), the effect sizes for SRCs were extremely small or negligible (all the reliable effect sizes are marked in bold typeface), and small effect sizes were found with ORC structures.

For the SRCs, the gains made by the parsing group were reliable, though extremely small (the CI did not pass through zero, but the upper limits of CI did not exceed the field-general benchmarks for a small effect). In addition, the input flood group showed extremely small but reliable improvements in the SRC-A at the post- and the delayed post-test and in the SRC-I at the delayed post-test. Moreover, the test-only group also had reliable gains in the SRC-A at the delayed post-test.

For the ORCs, most of the gains of the three groups were reliable. In the parsing group, small effects were observed in the pre- to the post- and the pre- to the delayed post-test comparisons. Moreover, the small effects were also found with the input flood group for the ORC-A structure and with the test-only groups in ORC-I structure for the comparison between the pre- and the delayed post-test.

Table 4. 2. 14 Within-group effect size (Cohen's *d*) [95% CI] for oral picture description test

structure	group	pre-post	pre-delayed	post-delayed
SRC-A (k=4 or 6)[a]	parsing	**.27 [.10, .45]**	**.38 [.20, .55]**	.15 [-.02, .32]
	input flood	**.21 [.03, .39]**	**.26 [.08, .43]**	.12 [-.06, .29]
	test-only	.10 [-.07, .27]	**.19 [.01, .36]**	.12 [-.05, .30]
SRC-I (k=4 or 6)[a]	parsing	**.41 [.23, .59]**	**.37 [.19, .54]**	-.09 [-.26, .08]
	input flood	.12 [-.05, .30]	**.18 [.01, .36]**	.10 [-.07, .28]
	test-only	.04 [-.13, .22]	.19 [-.02, .37]	.14 [-.03, .31]
ORC-A (k=5)	parsing	**.66 [.48, .85]**	**.80 [.60, 1.00]**	**.23 [.06, .41]**
	input flood	.27 [.10, .45]	**.63 [.44, .82]**	**.46 [.28, .65]**
	test-only	**.44 [.26, .62]**	**.56 [.38, .75]**	.17 [-.01, .34]
ORC-I (k=5)	parsing	**.80 [.60, .99]**	**.87 [.67, 1.07]**	.10 [-.07, .27]
	input flood	**.35 [.17, .53]**	**.58 [.40, .77]**	**.31 [.14, .49]**
	test-only	**.39 [.21, .57]**	**.61 [.42, .80]**	.26 [.09, .44]

Note: Bold typeface indicates that the CI did not pass through zero; [a] Numbers of items differed between version 1 and version 2, version 3 and version 4 of the tests, respectively; versions were counterbalanced across pre-, post-, and delayed post-test, within groups.

For between-group contrasts (see Table 4.2.15), for SRCs, small or very small effects were found at the post- and the delayed post-test for SRC-A in the comparison between the parsing and the input flood group, and the parsing and the test-only group. In addition, for SRC-I, the changes of Cohen's *d* from the pre- to the post-test showed small effects in the comparisons between the parsing and the input flood as well as between the parsing and the test only group.

For ORCs, small to medium effect sizes were observed in most contrasts. At the pre-test, the advantages for the parsing and the input flood groups over the test-only group were reliable, and the contrasts between input flood and test-only group the effects reached the small effect benchmark. At the post-test, small effects were observed in the contrasts between the parsing and the input flood group; in the parsing and test-only group contrasts, the effect sizes were medium. In terms of the effect size changes from the pre-test to the post-test, the small effect sizes could also be observed between the parsing and the other two groups. At the delayed post-test, both the parsing and the input flood group outperformed the test-only group, and had small to medium effects. The score differences between the parsing and the input flood group (for ORC-I) were negligible, as the CI pass through the zero. In addition, the scores changed from the pre-test to the delayed post-test did not reach the small effects in any comparison.

Table 4. 2. 15 Between-group effect size (Cohen's *d*) [95% CI] for oral picture description test

structure	test phase	parsing vs. input flood	parsing vs. test-only	input flood vs. test-only
SRC-A (k=4 or 6)[a]	pre-test	.13 [-.11, .37]	.04 [-.20, .28]	-.08 [-.33, .15]
	post-test	**.28 [.03, .52]**	**.32 [.08, .57]**	.05 [-.20, .29]
	delayed post-test	**.35 [.11, .59]**	**.40 [.15, .64]**	.04 [-.20, .29]
	pre-post *d* change	.15 [N/A]	.28 [N/A]	.13 [N/A]
	pre-delayed *d* change	.22 [N/A]	.36 [N/A]	.12 [N/A]
SRC-I (k=4 or 6)[a]	pre-test	.04 [-.20, .29]	.09 [-.15, .33]	.05 [-.20, .29]
	post-test	**.49 [.24, .73]**	**.58 [.34, .83]**	.09 [-.15, .33]
	delayed post-test	.23 [-.01, .47]	**.31 [.07, .55]**	.08 [-.16, .32]
	pre-post *d* change	**.45 [N/A]**	**.49 [N/A]**	.04 [N/A]
	pre-delayed *d* change	.19 [N/A]	.22 [N/A]	.03 [N/A]
ORC-A (k=5)	pre-test	-.08 [-.32, .16]	**.36 [.12, .61]**	**.45 [.20, .69]**
	post-test	**.53 [.29, .78]**	**.77 [.52, 1.02]**	.22 [-.03, .46]
	delayed post-test	**.25 [.01, .49]**	**.75 [.50, 1.00]**	**.48 [.23, .73]**
	pre-post *d* change	**.61 [N/A]**	**.41 [N/A]**	-.23 [N/A]
	pre-delayed *d* change	.33 [N/A]	.39 [N/A]	.03 [N/A]
ORC-I (k=5)	pre-test	-.14 [-.38, .10]	**.35 [.10, .59]**	**.49 [.24, .74]**
	post-test	**.43 [.19, .67]**	**.85 [.60, 1.10]**	**.38 [.14, .63]**
	delayed post-test	.22 [-.02, .47]	**.66 [.41, .91]**	**.42 [.18, .67]**
	pre-post *d* change	**.57 [N/A]**	**.50 [N/A]**	-.11 [N/A]
	pre-delayed *d* change	.36 [N/A]	.31 [N/A]	-.07 [N/A]

Note: Bold typeface indicates that the CI did not pass through zero; [a] Numbers of items differed between version 1 and version 2, version 3 and version 4 of the tests, respectively; versions were counterbalanced across pre-, post-, and delayed post-test, within groups.

4.2.4.4 Inferential statistical analysis

Mixed effects logistic regressions were carried out for SRC-A, SRC-I, ORC-A and ORC-I separately. Each model was run with the baseline of the test-only group and the input flood group separately. The results reported in this section are mainly based on the model with the baseline of the test-only group. For the model with the baseline of the input flood group, only the statistically significant or reliable effects related to the comparisons between the parsing and the input flood group are reported in this section (for the full results for models with input flood baseline see Appendix 31). AIC

and LRT results for model selection see Appendix 24.

SRC-A

The table 4.2.16 presents the model analysis of accuracy scores of SRC-A structure. The result showed no statistically significant effect in the model, and the only reliable interaction was between group and test phase as the 95% CI for the estimate *b* did not pass through zero. The parsing group was predicted to be 4.33 times more likely to correctly produce the target structure at the post-test than at the pre-tests relative to the test-only group. In the model with the baseline of the input flood group, no statistically significant effect could be found.

Table 4. 2. 16 The fixed effects of the model analysis of accuracy scores for SRC-A in oral sentence description test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR [CI]
Intercept	.14 [-.69, .97]	.50	.28	.779	1.15 [.50, 2.64]
test vs. parsing	.11 [-1.03, 1.25]	.69	.16	.872	1.12 [.36, 3.51]
test vs. input	-.39 [-1.56, .78]	.71	-.55	.582	.68 [.21, 2.18]
pre- vs. post-	.09 [-.91, 1.10]	.61	.15	.878	1.10 [.40, 2.99]
pre- vs. delayed-	.97 [-.19, 2.12]	.70	1.38	.169	2.63 [.83, 8.37]
test vs. parsing × pre- vs. post-	**1.47 [.08, 2.85]**	**.84**	**1.74**	**.083**	**4.33 [1.08, 17.34]**
test vs. input × pre- vs. post-	1.00 [-.43, 2.44]	.87	1.15	.250	2.73 [.65, 11.43]
test vs. parsing × pre- vs. delayed-	1.48 [-.12, 3.07]	.97	1.52	.128	4.38 [.89, 21.61]
test vs. input × pre- vs. delayed-	1.13 [-.54, 2.80]	1.01	1.11	.266	3.09 [.58, 16.41]

Note: Model formula: model4=glmer(score ~ group*stage + (1+stage|subject) + (1+stage|item), data=oral_production, family=binomial, control = glmerControl(optimizer = "bobyqa")); parsing = parsing group; input = input flood group; test = test-only group; Marginal R^2 = .06, conditional R^2 = .78; bold typeface indicates a reliable effect

SRC-I

Table 4.2.17 shows the model results for SRC-I structure. The statistically significant effects were found in a comparison between pre-test and delayed post-test (regardless of the group) and in an interaction between group (test-only group vs. parsing group) and test phase (pre-test vs. post-test). That is, the parsing group was predicted to be 16.64 times more likely to correctly produce an SRC-I sentence than the test-only group at the post-test relative to the pre-test. In addition, in the model with baseline of the input flood group, it was found that compared to the input flood group, the parsing

group was 12.52 more likely to produce SRC-I at the post-test relative to the pre-test (*b* [CI] = 2.53 [.59, 4.46], SE = 1.18, *z* = 2.15, *p* = .032, OR = 12.52 [1.81, 86.58]).

Table 4. 2. 17 The fixed effects of the model analysis of accuracy scores for SRC-I in oral sentence description test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR [CI]
Intercept	-.16 [-1.24, .92]	.66	-.24	.807	.85 [.29, 2.52]
test vs. parsing	.39 [-1.11, 1.90]	.92	.43	.667	1.48 [.33, 6.70]
test vs. input	.36 [-1.20, 1.92]	.95	.38	.706	1.43 [.30, 6.83]
pre- vs. post-	.54 [-.74, 1.83]	.78	.70	.486	1.72 [.48, 6.23]
pre- vs. delayed-	1.83 [.32, 3.33]	.91	2.00	.046*	6.20 [1.38, 27.91]
test vs. parsing × pre- vs. post-	2.81 [.84, 4.78]	1.20	2.35	.019*	16.64 [2.32, 119.57]
test vs. input × pre- vs. post-	.28 [-1.59, 2.15]	1.14	.25	.803	1.33 [.20, 8.61]
test vs. parsing × pre- vs. delayed-	1.04 [-1.01, 3.10]	1.25	.83	.404	2.84 [.36, 22.20]
test vs. input × pre- vs. delayed-	.23 [-1.94, 2.40]	1.32	.17	.863	1.25 [.14, 10.99]

Note: model5=glmer(score ~ group*stage + (1+stage|subject) + (1+group|item), data=oral_production, family=binomial, control = glmerControl(optimizer = "bobyqa")); Marginal R^2 = .08, conditional R^2 =.87; parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when α < .05

ORC-A

The model results of ORC-A structure were shown in the table 4.2.18. The statistically significant effects were found with the interactions between pre-test and post-test and between pre-test and delayed post-test, regardless of the group. In addition, the three two-way interactions between group and test phase were statistically significant. Compared to the pre-test, the parsing group was predicted to be 9.61 times and 27.16 times more likely to produce ORC-A correctly than the test-only group at the pre-test and the post-test respectively. The results also indicated that compared to the input flood group, the test-only group had more gains at the post-test compared to the pre-test. In the model with the baseline of the input flood group, the parsing group had more gains at the post- and the delayed post-test relative to the pre-test in the comparison between the input flood group (input flood vs. parsing: pre- vs. post-test: *b* [CI] = 3.97 [2.33, 5.62], SE = 1.00, *z* = 3.97, *p* < .001, OR [CI] = 53.21 [10.26, 275.91]; input flood vs. parsing: pre- vs. delayed post-test: *b* [CI] = 2.79 [.57, 5.01], SE = 1.35, *z* = 2.06, *p* = .039, OR [CI] = 16.28 [1.76, 150.38]).

Table 4. 2. 18 The fixed effects of the model analysis of accuracy scores for ORC-A in oral sentence description test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR [CI]
Intercept	-5.13 [-7.44, -2.83]	1.40	-3.66	<.001***	.01 [.00, .06]
test vs. parsing	2.83 [-.18, 5.85]	1.83	1.55	.122	17.02 [.84, 346.24]
test vs. input	**3.52 [.44, 6.60]**	**1.87**	**1.88**	**.060**	**33.81 [1.55, 736.06]**
pre- vs. post-	3.38 [2.24, 4.52]	.69	4.87	<.001***	29.43 [9.39, 92.24]
pre- vs. delayed-	4.37 [3.15, 5.58]	.74	5.91	<.001***	78.69 [23.36, 265.11]
test vs. parsing × pre- vs. post-	2.26 [.48, 4.04]	1.08	2.09	.037*	9.61 [1.62, 56.98]
test vs. input × pre- vs. post-	-1.71 [-3.15, -.28]	.87	-1.96	.049*	.18 [.04, .76]
test vs. parsing × pre- vs. delayed-	3.30 [1.08, 5.52]	1.35	2.44	.015*	27.16 [2.94, 250.75]
test vs. input × pre- vs. delayed-	.51 [-1.19, 2.22]	1.04	.49	.639	1.67 [.30, 9.19]

Note: model1=glmer(score ~ group*stage + (1|subject) + (1|item), data=oral_production, family=binomial, control = glmerControl(optimizer = "bobyqa")); Marginal R^2 = .22, Conditional R^2 = .93; parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect;* significantly differently from zero when $\alpha \leq .05$; *** significantly differently from zero when $\alpha < .001$

ORC-I

Table 4.2.19 presents the inferential statistical results of ORC-I structure. The results indicated that the participants as a whole significantly gained at the post- and delayed post-test (relative to the pre-test). In addition, two three-way interactions between the test phase (pre- vs. post-test and pre- vs. delayed post-test) and the group (test-only vs. parsing group) were found to be statistically significant. Compared to the pre-test, the parsing group was predicted to be 1067.04 times and 56.96 times more likely to produce correct ORC-I sentences than the test-only group at the post- and the delayed post-test respectively. In terms of the model with the input flood group as the baseline, it was found that the compared to the pre-test, the parsing group was 97.51 times more likely to correctly produce ORC-I sentences than the input flood group at the post-test (*b* [CI] = 4.58 [.80, 8.36], SE = 2.30, *z* = 1.99, *p* =.046, OR [CI] = 97.51 [2.23, 4269.33]).

Table 4. 2. 19 The fixed effects of the model analysis of accuracy scores for ORC-I in oral sentence description test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-3.93 [-5.90, -1.95]	1.20	-3.27	.001**	.02 [.00, .14]
test vs. parsing	1.98 [-.49, 4.46]	1.50	1.32	.188	7.26 [.61, 86.20]
test vs. input	**2.80 [.26, 5.34]**	**1.55**	**1.81**	**.070**	**16.44 [1.29, 209.10]**
pre- vs. post-	.91 [-1.85, 3.68]	1.68	.54	.587	2.49 [.16, 39.54]
pre- vs. delayed-	3.83 [1.73, 5.92]	1.27	3.00	.003**	45.84 [5.64, 372.70]
test vs. parsing × pre- vs. post-	6.97 [2.35, 11.60]	2.81	2.48	.013*	1067.04 [10.48, 108605.50]
test vs. input × pre- vs. post-	2.39 [-1.52, 6.31]	2.38	1.01	.314	10.94 [.22, 547.51]
test vs. parsing × pre- vs. delayed-	4.04 [.69, 7.39]	2.04	1.98	.047*	56.96 [2.00, 1624.67]
test vs. input × pre- vs. delayed-	1.26 [-1.75, 4.27]	1.83	.69	.490	3.54 [.17, 71.64]

Note: Model formula: model5=glmer(score ~ group*stage + (1+stage|subject) + (1+group|item), data=oral_production, family=binomial, control = glmerControl(optimizer = "bobyqa")); Marginal R^2 = .20, Conditional R^2 = .95; parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$

4.2.4.5 Summary of the results in the oral production test

Overall, all the participants gained at post- and delayed-post tests (relative to pre-test) in the accuracy of their oral production test. The gains of the parsing group were more than the other two groups in all the structures, and especially so for ORCs. In SRC-I and both types of ORCs, statistically significant effects between the parsing group and the other two groups could be found. The between-group effect sizes also indicated an advantage for the parsing group over the other two groups at the post- and the delayed post-test. However, the input flood training did not seem to facilitate accuracy of oral production, because there was no positive significant effect that could be found between the test-only and the input flood comparisons.

4.2.5 To what extent are effects observable in the metalinguistic knowledge test?

Metalinguistic knowledge was measured by sentence-picture matching tests. Each item of the test contained a picture and a sentence, and the sentence might match or mismatch the picture. The scores were collected from three tasks: 1) decide whether the sentence matched the picture, 2) correct the mismatched sentences to match the

picture, and 3) explain the reason for mismatch. The scores of each task were analysed separately. The parsing group was expected to outperform the input flood and the test only group at the post- and delayed post-test, on all three measures from this test.

4.2.5.1 Analysis of accuracy scores of deciding match or mismatch

a) Descriptive analysis

Table 4.2.20 presents the mean (*SDs*) of accuracy scores of deciding match or mismatch in metalinguistic knowledge test. For matched items, all the three groups scored at ceiling for all the structures from the pre- to the delayed post-test.

For mismatched items, the improvement of accuracy scores across the time could be observed in all the groups for all the structures. Compared to the input flood and the test-only group, the scores of the parsing group were the highest and were at ceiling at the post- and the delayed post-test. For ORC-I structure, the scores of the pre-test were lower than the other structures, yet at the post- and the delayed post-test, the scores did not have salient differences compared with the scores for other structures in each group.

Table 4. 2. 20 Mean (*SDs*) scores of deciding match or mismatch in metalinguistic knowledge test

type	test phase	parsing		input flood		test-only	
		match	mismatch	match	mismatch	match	mismatch
SRC-A (k=2)	pre-test	1.00 (.00)	.83 (.38)	.94 (.24)	.83 (.38)	.94 (.24)	.85 (.36)
	post-test	1.00 (.00)	.94 (.23)	.96 (.19)	.88 (.32)	1.00 (.00)	.87 (.34)
	delayed post-test	.98 (.14)	.96 (.19)	.92 (.27)	.88 (.32)	.98 (.14)	.87 (.34)
SRC-I (k=2)	pre-test	.96 (.19)	.81 (.39)	.96 (.19)	.79 (.41)	.98 (.14)	.62 (.49)
	post-test	.96 (.19)	.98 (.14)	.94 (.24)	.85 (.36)	.98 (.14)	.87 (.34)
	delayed post-test	1.00 (.00)	.91 (.29)	.92 (.27)	.85 (.36)	.96 (.19)	.81 (.40)
ORC-A (k=2)	pre-test	.93 (.26)	.85 (.36)	.94 (.24)	.75 (.44)	.98 (.14)	.81 (.40)
	post-test	.98 (.14)	.94 (.23)	.94 (.24)	.81 (.39)	1.00 (.00)	.90 (.30)
	delayed post-test	.94 (.23)	.96 (.19)	.98 (.14)	.83 (.38)	.96 (.19)	.94 (.24)

	pre-test	.98 (.14)	.63 (.49)	.94 (.24)	.71 (.46)	.94 (.21)	.48 (.50)
ORC-I	post-test	.94 (.23)	.83 (.38)	.96 (.14)	.73 (.45)	.98 (.19)	.76 (.43)
(k=2)	delayed post-test	.98 (.14)	.89 (.32)	.94 (.14)	.75 (.44)	.98 (.18)	.87 (.34)

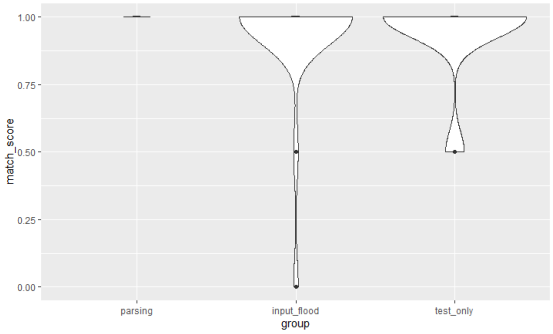
Note: The *type* refers to the structure of the sentence presented to the participants. For mismatched items, the SRCs were required to correct to the ORCs and the ORCs were required to correct to SRCs.

b) Plots

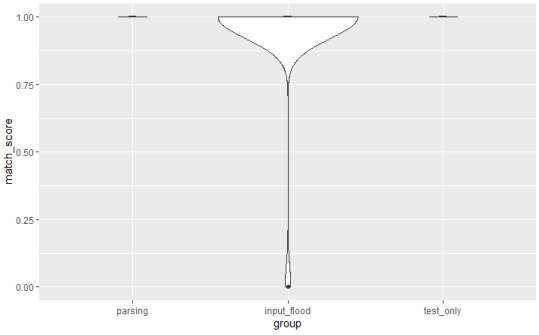
Matched items

Figures 4.2.17 to 4.2.20 present the accuracy scores of each target structure for the matched items (based on the mean scores of each participant in each type of relative clause) using the violin plots with Boxplots. For all the structures, the plots indicated that overall, the majority of the participants scored at ceiling across time, though with slight fluctuation. At each time phase, the group differences were very small.

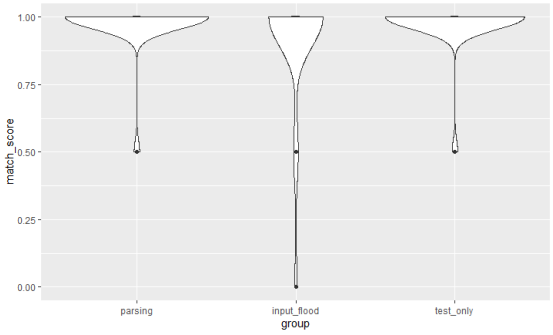
Figure 4. 2. 17 Comparison of accuracy scores (for matched items) of three learner groups in different test phase for SRC-A structure in the deciding match or mismatch task of metalinguistic knowledge test



Pre-test



Post-test



Delayed post-test

Figure 4. 2. 18 Comparison of accuracy scores (for matched items) of three learner groups in different test phase for SRC-I structure in the deciding match or mismatch task of metalinguistic knowledge test

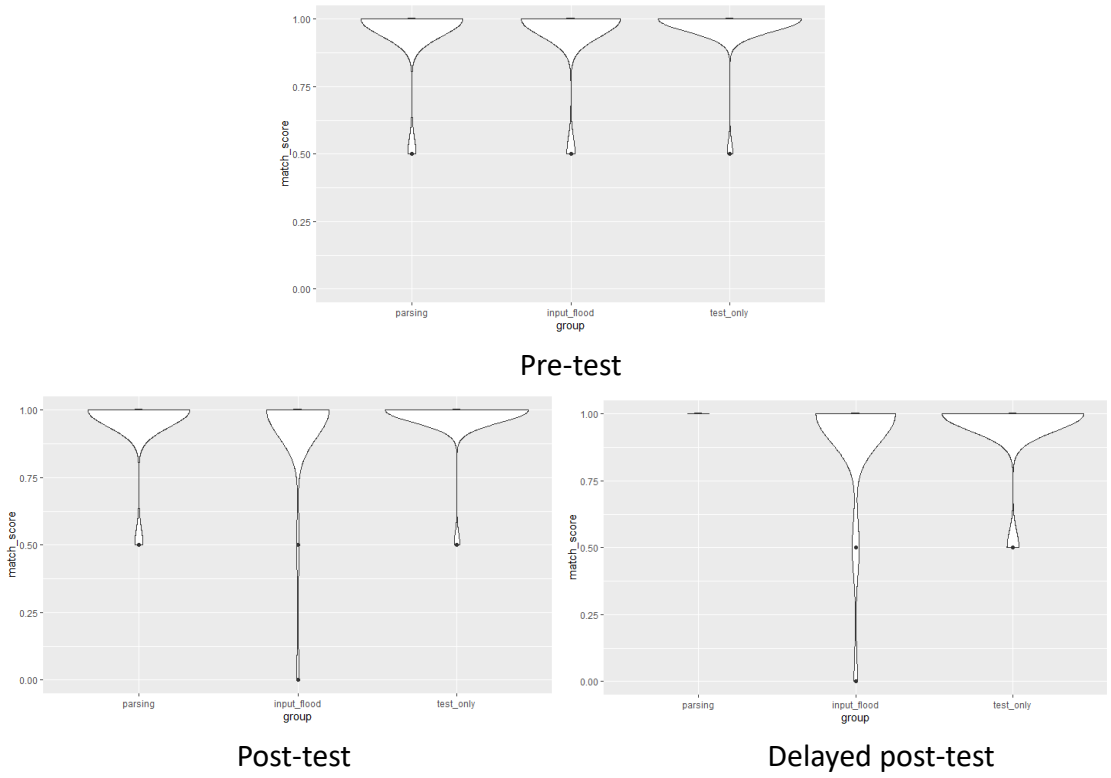


Figure 4. 2. 19 Comparison of accuracy scores (for matched items) of three learner groups in different test phase for ORC-A structure in the deciding match or mismatch task of metalinguistic knowledge test

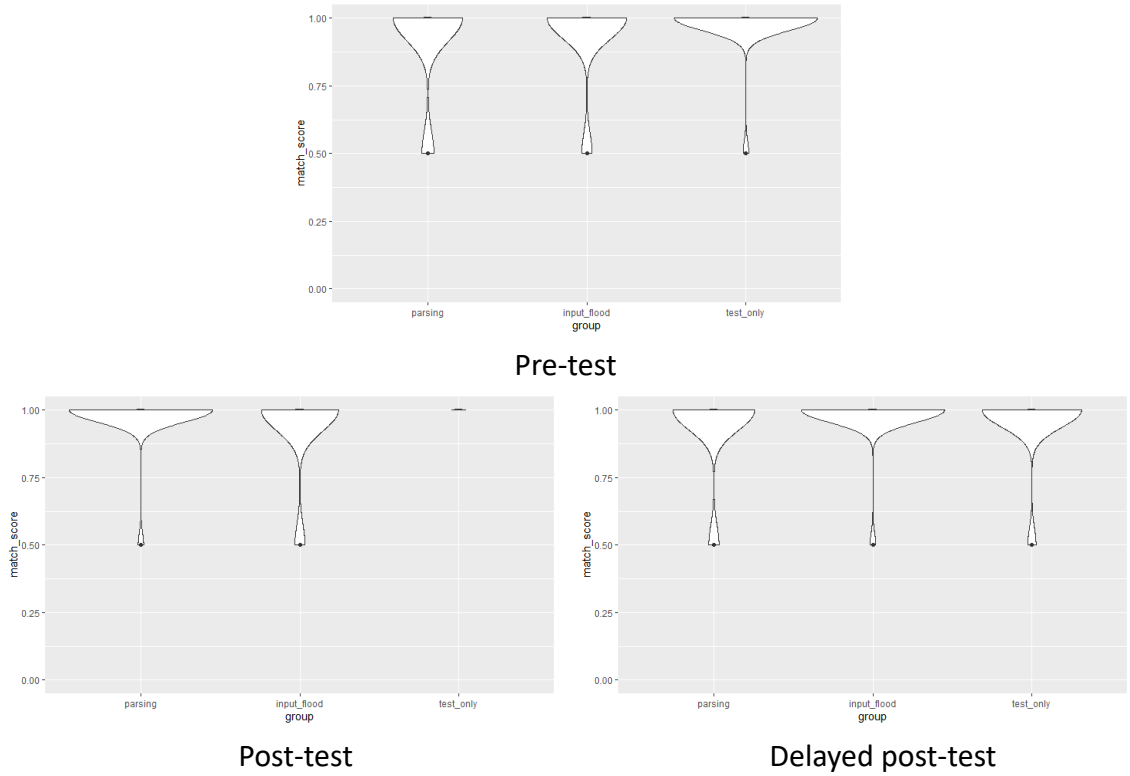
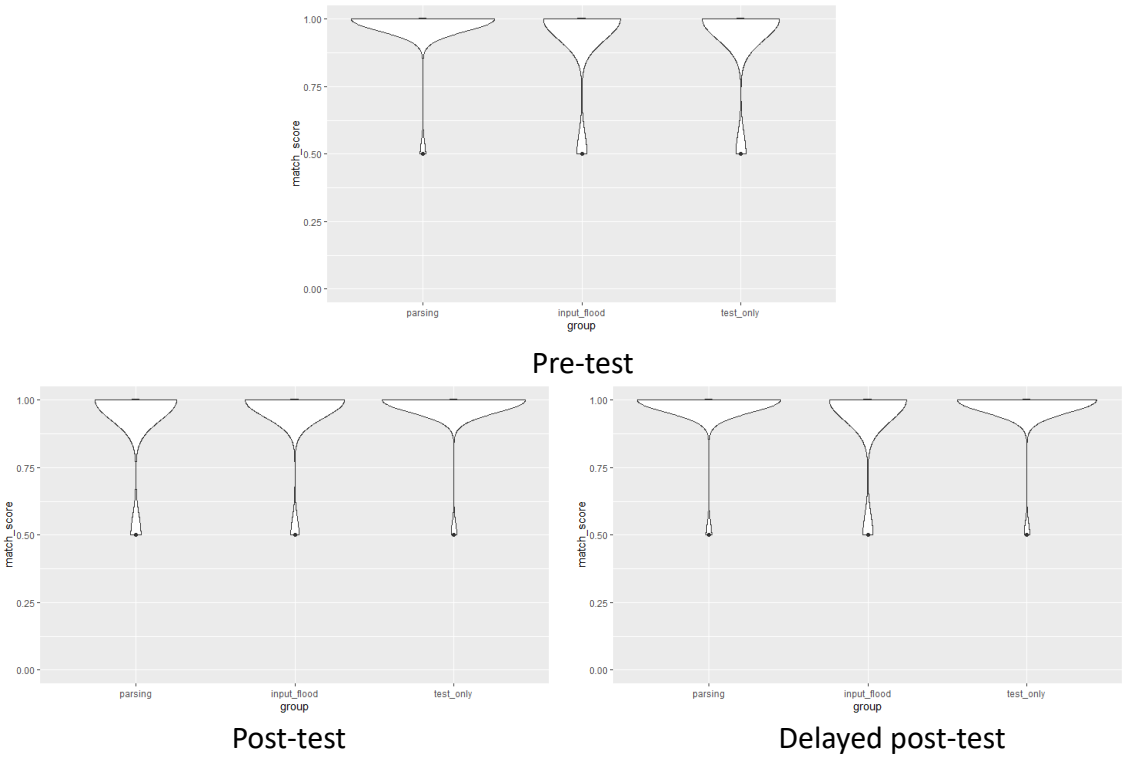


Figure 4. 2. 20 Comparison of accuracy scores (for matched items) of three learner groups in different test phase for ORC-I structure in the deciding match or mismatch task of metalinguistic knowledge test



Mismatched items

The accuracy scores of mismatched items in deciding match or mismatch task are shown in figures 4.2.21 to 4.2.24. For SRCs, advantages of the parsing group over the input flood and test-only groups were shown at post- and delayed post-test. In addition, for the input flood and the test-only group, the distributions of the scores were similar at the post- and the delayed post-test.

For ORC-A, the parsing group had more participants who scored at ceiling than the input flood and the test-only groups at the three test phases. For ORC-I, at the post-test and the delayed post-test, the three groups had the similar distributions of the scores, and the parsing group slightly outperformed the other two groups.

Figure 4. 2. 21 Comparison of accuracy scores (for mismatched items) of three learner groups in different test phase for SRC-A structure in the deciding match or mismatch task of metalinguistic knowledge test

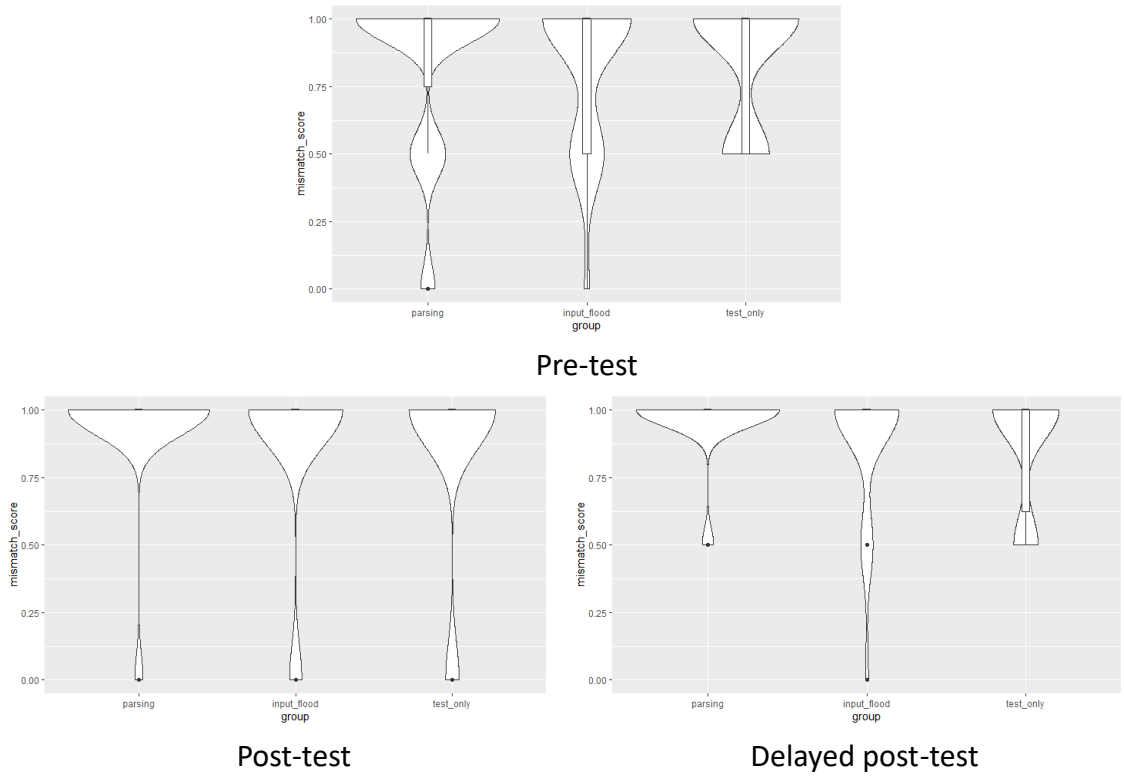


Figure 4. 2. 22 Comparison of accuracy scores (for mismatched items) of three learner groups in different test phase for SRC-I structure in the deciding match or mismatch task of metalinguistic knowledge test

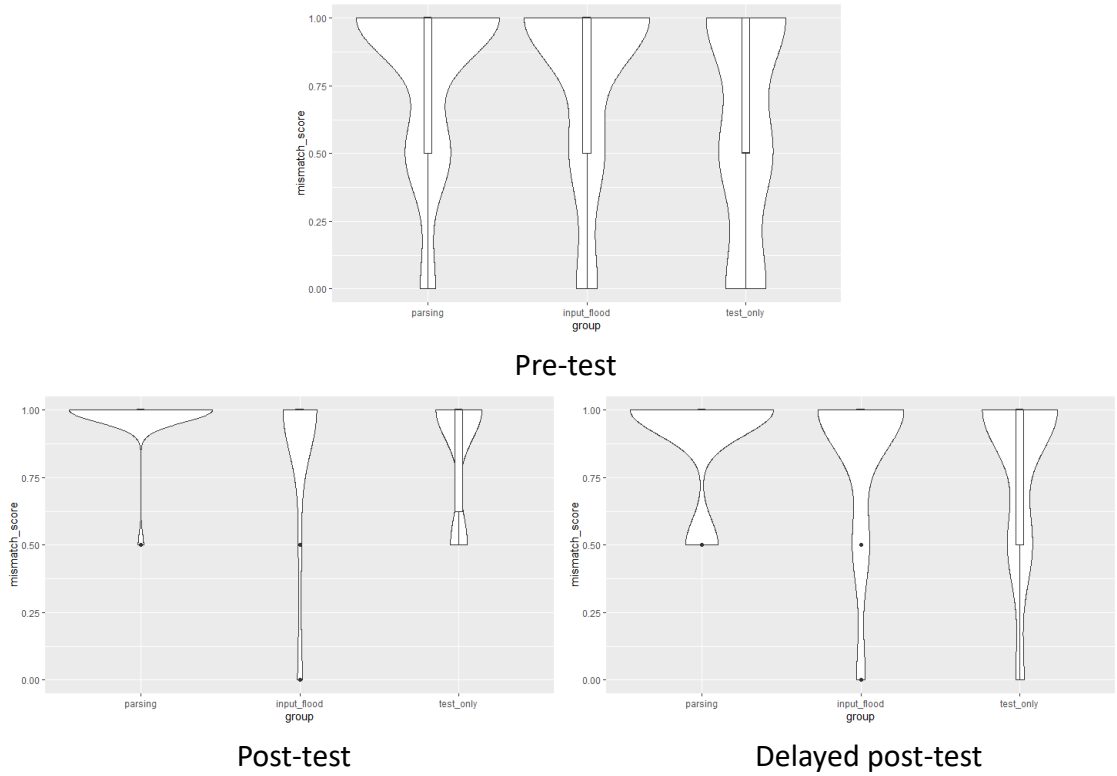


Figure 4. 2. 23 Comparison of accuracy scores (for mismatched items) of three learner groups in different test phase for ORC-A structure in the deciding match or mismatch task of metalinguistic knowledge test

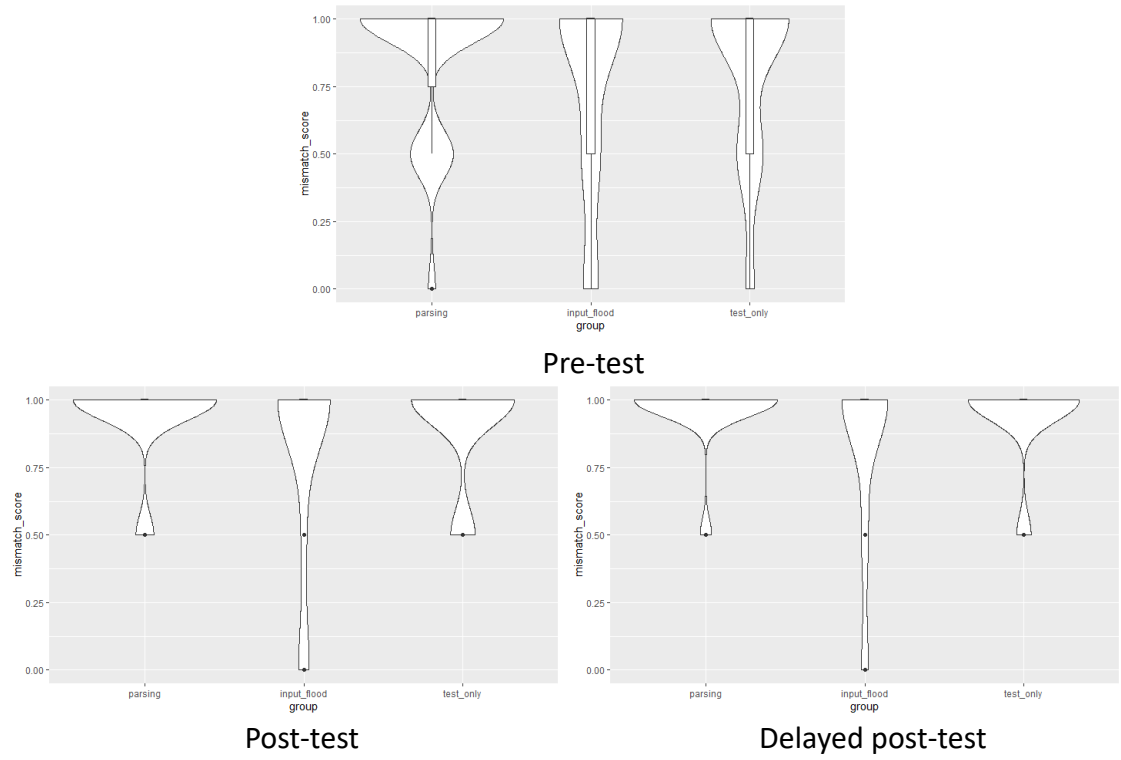
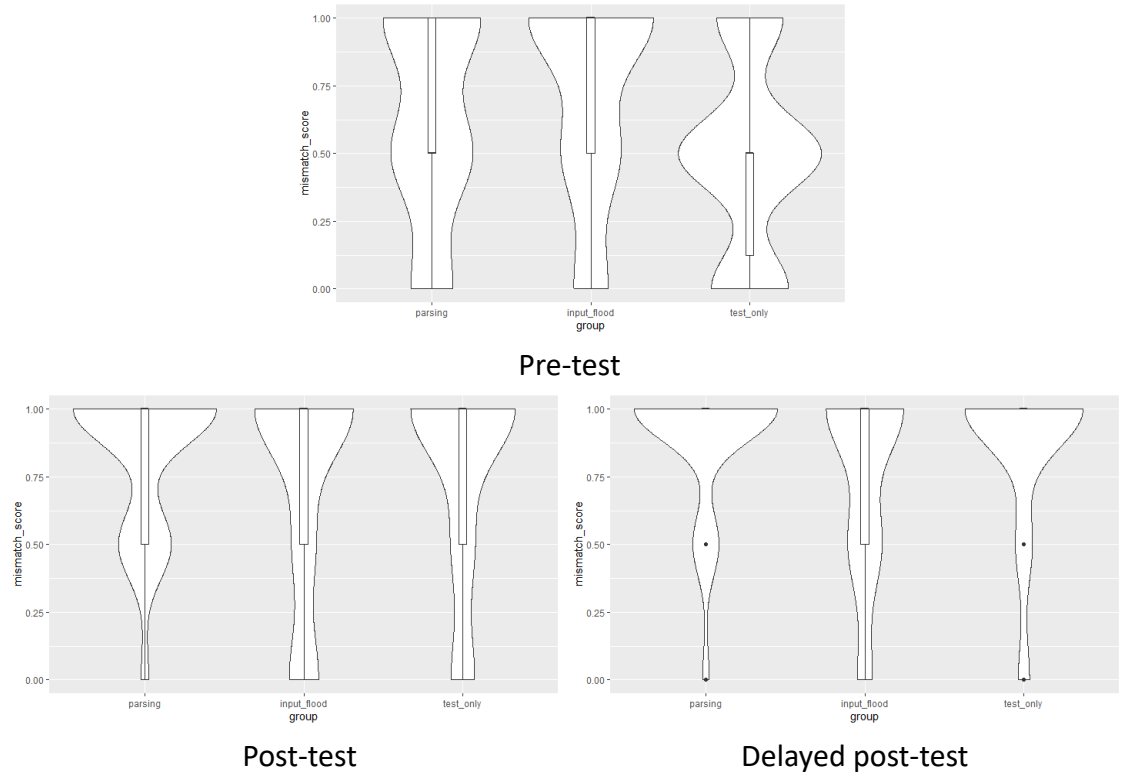


Figure 4. 2. 24 Comparison of accuracy scores (for mismatched items) of three learner groups in different test phase for ORC-I structure in the deciding match or mismatch task of metalinguistic knowledge test



c) Examination of effect sizes

Matched items

Table 4.2.21 showed the effect sizes of the within-group contracts, reflecting change over time. All the effect sizes were negligible, and all the 95% CI passed through 0.

Table 4. 2. 21 Within-group effect size (Cohen's *d*) [95% CI] for matched items in deciding match or mismatch task of metalinguistic knowledge test

structure	group	pre-post	pre-delayed	post-delayed
SRC-A (k=2)	parsing	N/A	-.14 [-.41, .13]	-.14 [-.41, .13]
	input flood	.14 [-.14, .42]	-.14 [-.42, .14]	-.20 [-.48, .08]
	test-only	.20 [-.08, .48]	.14 [-.14, .42]	-.14 [-.42, .14]
SRC-I (k=2)	parsing	.00 [-.27, .27]	.19 [-.08, .47]	.19 [-.08, .47]
	input flood	-.08 [-.35, .19]	-.14 [-.42, .14]	-.08 [-.35, .19]
	test-only	.00 [-.27, .27]	-.08 [-.35, .19]	-.08 [-.35, .19]
ORC-A (k=2)	parsing	.18 [-.09, .46]	.05 [-.22, .32]	-.14 [-.41, .13]
	input flood	.00 [-.27, .27]	.14 [-.14, .42]	.20 [-.08, .48]
	test-only	.14 [-.14, .42]	-.08 [-.35, .19]	-.14 [-.42, .14]
ORC-I (k=2)	parsing	-.14 [-.41, .13]	.00 [-.27, .27]	.14 [-.13, .41]
	input flood	.06 [-.21, .34]	.00 [-.27, .27]	-.06 [-.34, .21]
	test-only	.08 [-.20, .36]	.14 [-.14, .42]	.00 [-.28, .28]

Note: N/A refers to one or both of the *SDs* of the two test phases within a contrast was 0

For the between-group contrasts (see table 4.2.22), only the difference between the parsing and the input flood group for SRC-I structure at delayed post-test (and at the changes from the pre- to the delayed post-test) had small effects. The differences in other contrasts were negligible, because the 95% CI passed through 0.

Table 4. 2. 22 Between-group effect size (Cohen's *d*) [95% CI] for matched items in deciding match or mismatch task of metalinguistic knowledge test

structure	test phase	parsing vs. input flood	parsing vs. test-only	input flood vs. test-only
SRC-A (k=2)	pre-test	.35 [-.03, .73]	.35 [-.03, .73]	.00 [-.38, .38]
	post-test	.28 [-.10, .66]	N/A	-.28 [-.67, .11]
	delayed post-test	.28 [-.11, .66]	.01 [-.38, .39]	-.27 [-.66, .12]
	pre-post *d* change	-.07 [N/A]	N/A	-.28 [N/A]
	pre-delayed *d* change	-.07 [N/A]	-.34 [N/A]	-.27 [N/A]
SRC-I (k=2)	pre-test	.01 [-.37, .39]	-.11 [-.49, .27]	-.11 [-.50, .27]
	post-test	.10 [-.28, .48]	-.11 [-.49, .27]	-.20 [-.58, .19]
	delayed post-test	**.41 [.02, .79]**	.28 [-.10, .66]	-.16 [-.55, .22]
	pre-post *d* change	.09 [N/A]	.00 [N/A]	-.09 [N/A]
	pre-delayed *d* change	**.40 [N/A]**	.39 [N/A]	-.05 [N/A]
ORC-A (k=2)	pre-test	-.07 [-.45, .32]	-.26 [-.64, .12]	-.20 [-.58, .19]
	post-test	.20 [-.18, .59]	-.19 [-.57, .19]	-.34 [-.73, .05]
	delayed post-test	-.19 [-.57, .19]	-.08 [-.46, .30]	.11 [-.27, .50]
	pre-post *d* change	.27 [N/A]	.07 [N/A]	-.14 [N/A]
	pre-delayed *d* change	-.12 [N/A]	.18 [N/A]	.31 [N/A]
ORC-I (k=2)	pre-test	.20 [-.18, .59]	.20 [-.18, .59]	.00 [-.38, .38]
	post-test	-.08 [-.46, .30]	-.19 [-.57, .20]	-.11 [-.50, .28]
	delayed post-test	.20 [-.18, .59]	.01 [-.38, .39]	-.20 [-.58, .19]
	pre-post *d* change	-.28 [N/A]	-.39 [N/A]	-.11 [N/A]
	pre-delayed *d* change	.00 [N/A]	-.19 [N/A]	-.20 [N/A]

Note: N/A refers to one or both of the *SDs* of the two groups within a contrast was 0; Bold typeface refers to the CI did not pass through zero.

Mismatched items

The effect sizes of within-group contrasts (see table 4.2.23), reflecting change over time, indicated that almost all the changes were very small (95% CI did not pass through zero, and the upper limits reached or were near the benchmark of small effects) to negligible.

For the parsing and the test-only group, small or very small effects could be observed with all the structures in the comparison between the pre- and the post-test and in the comparison between the pre- and the delayed post-test. A small effect was found with the test-only group in the pre- and the post-test comparison (*d* [CI] = .73

[.42, 1.04]) for the ORC-I, but this might have been because the mean score of the pre-test was rather low (mean [*SD*] = .48 [.50]).

No reliable effect could be found with the input flood group, because all the CIs passed through zero.

Table 4. 2. 23 Within-group effect size (Cohen's *d*) [95% CI] for mismatched items in deciding match or mismatch task of metalinguistic knowledge test

structure	group	pre-post	pre-delayed	post-delayed
SRC-A (k=2)	parsing	.24 [-.03, .51]	**.33 [.06, .61]**	.06 [-.21, .33]
	input flood	.13 [-.15, .40]	.14 [-.14, .42]	.00 [-.27, .27]
	test-only	.04 [-.23, .32]	.05 [-.22, .33]	.00 [-.27, .27]
SRC-I (k=2)	parsing	**.39 [.12, .68]**	.19 [-.08, .46]	-.23 [-.50, .05]
	input flood	.14 [-.14, .42]	.13 [-.15, .40]	.00 [-.27, .27]
	test-only	**.40 [.12, .69]**	**.32 [.04, .61]**	-.14 [-.42, 14]
ORC-A (k=2)	parsing	.21 [-.06, .48]	**.30 [.02, .58]**	.06 [-.21, .33]
	input flood	.14 [-.14, .42]	.20 [-.08, .48]	.05 [-.23, .32]
	test-only	.19 [-.08, .47]	**.30 [.02, .59]**	.10 [-.18, .37]
ORC-I (k=2)	parsing	**.33 [.05, .60]**	**.50 [.21, .79]**	.11 [-.16, .38]
	input flood	.04 [-.24, .31]	.08 [-.19, .35]	.04 [-.23, .32]
	test-only	**.48 [.19, .78]**	**.73 [.42, 1.04]**	.20 [-.08, .48]

Note: Bold typeface indicates that the CI did not pass through zero.

Table 4.2.24 presents the effect sizes of the between-group contrasts, reflecting differences between groups. For SRC-A structure, all the differences between groups were negligible, because the CIs passed through zero. For SRC-I, at the pre-test, the test-only group had a lower score than the other groups, and the score differences compared with the parsing group had small effect. Small effects were also found at the post-test between the parsing and the input flood group, and in the comparison between the parsing and the test-only group. The changing of the scores from the pre-test to the post-test indicated that the gains made by the parsing group over the input flood group had small effect, but the parsing group did not make more gains than the test-only group. Moreover, the small effect of changing *d* from the pre- to the post-test was also found in with the input flood vs. test-only group comparison, which

means that the test-only group made more gains than the input flood group.

For ORC-A structure, the scores that the parsing group over the input flood group had small effects at the post- and delayed post-tests. For ORC-I, the small effects were found at the pre-test in the comparison between the input flood and the test-only group. In addition, the *d* changes indicated that the parsing and test-only groups made more gains than the input flood group at the post- and delayed post-test compared to the pre-test, and small to medium effects were observed (medium effect were observed in the comparison between the input flood and test-only group, changing from the pre-test to the delayed post-test).

Table 4. 2. 24 Between-group effect size (Cohen's *d*) [95% CI] for mismatched items in deciding match or mismatch task of metalinguistic knowledge test

structure	test phase	parsing vs. input flood	parsing vs. test-only	input flood vs. test-only
SRC-A (k=2)	pre-test	.02 [-.36, .40]	-.03 [-.42, .35]	-.05 [-.44, .33]
	post-test	.21 [-.17, .60]	.27 [-.11, .65]	.06 [-.33, .44]
	delayed post-test	.30 [-.09, .68]	.35 [-.03, .74]	.06 [-.33, .44]
	pre-post *d* change	.19 [N/A]	.30 [N/A]	.11 [N/A]
	pre-delayed *d* change	.28 [N/A]	.38 [N/A]	.11 [N/A]
SRC-I (k=2)	pre-test	.07 [-.32, .45]	**.45 [.06, .83]**	.38 [-.01, .77]
	post-test	**.50 [.11, .88]**	**.45 [.06, .83]**	-.05 [-.44, .33]
	delayed post-test	.19 [-.20, .57]	.29 [-.10, .67]	.10 [-.28, .49]
	pre-post *d* change	**.43 [N/A]**	.00 [N/A]	**-.43 [N/A]**
	pre-delayed *d* change	.12 [N/A]	-.16 [N/A]	-.28 [N/A]
ORC-A (k=2)	pre-test	.26 [-.13, .64]	.12 [-.26, .50]	-.14 [-.52, .25]
	post-test	**.42 [.04, .81]**	.15 [-.23, .53]	-.27 [-.66, .11]
	delayed post-test	**.45 [.07, .84]**	.10 [-.28, .48]	-.36 [-.75, .02]
	pre-post *d* change	.16 [N/A]	.03 [N/A]	-.13 [N/A]
	pre-delayed *d* change	.19 [N/A]	-.02 [N/A]	-.22 [N/A]
ORC-I (k=2)	pre-test	-.17 [-.55, .21]	.30 [-.08, .68]	**.48 [.09, .87]**
	post-test	.25 [-.13, .63]	.17 [-.21, .55]	-.08 [-.46, .31]
	delayed post-test	.36 [-.02, .75]	.07 [-.31, .45]	-.29 [-.68, .09]
	pre-post *d* change	**.42 [N/A]**	-.13 [N/A]	**-.56 [N/A]**
	pre-delayed *d* change	**.53 [N/A]**	-.23 [N/A]	**-.77 [N/A]**

Note: Bold typeface refers to the 95% CI did not pass through zero.

d) Inferential statistic analysis

The mixed effect logistic regression models were used to conduct the inferential statistical analysis. The models with the baseline of the test-only and the input flood group were run separately. For the model with the input flood group baseline, only the statistically significant effects related to the comparisons between the input flood and the parsing group were reported in the chapter (full results see Appendix 32). The AIC and LRT results for model selection see Appendix 25.

Matched items

The inferential statistical results for SRC-A and SRC-I structures are not provided because all the models did not converge. The means and the plots of the SRC-A and SRC-I showed that the score difference across the time and between groups was very small, and very unlikely to have any significant differences.

ORC-A

Table 4.2.25 presents the results of the mixed effects model analysis for ORC-A structure. No statistically significant effects were found in this analysis either with the baseline of the test-only group or the input flood group.

Table 4. 2. 25 The fixed effects of the model analysis of ORC-A structure (matched items) for deciding match or mismatch task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
intercept	4.03 [2.33, 5.72]	1.03	3.92	<.001***	56.02 [10.33, 303.90]
test vs. parsing	-1.42 [-3.28, 0.44]	1.13	-1.26	.209	.24 [.04, 1.55]
test vs. input	-1.15 [-3.07, 0.76]	1.16	-.99	.323	.32 [.05, 2.15]
pre- vs. post-	16.48 [-125.36, 158.33]	86.24	.19	.848	144E+7 [.00, 5.80E+68]
pre- vs. delayed-	-.70 [-2.73, 1.33]	1.23	-.57	.571	.50 [.07, 3.77]
test vs. parsing × pre- vs. post-	-15.02 [-156.88, 126.83]	86.24	-.17	.862	.00 [.00, 1.20E+55]
test vs. input × pre- vs. post-	-16.47 [-158.32, 125.38]	86.24	-.19	.849	.00 [.00, 2.83E+54]
test vs. parsing × pre- vs. delayed-	1.02 [-1.38, 3.43]	1.46	.70	.484	2.78 [.25, 30.83]
test vs. input × pre- vs. delayed-	1.86 [-.93, 4.65]	1.70	1.10	.272	6.44 [.40, 104.83]

Note: Model formula: model1=glmer(match_score ~ group*stage + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000)) (only this model was converged); Marginal R^2=0.51, conditional R^2=0.51; parsing = parsing group; input = input flood group; test = test-only group; *** significantly differently from zero when $\alpha < .001$

ORC-I

The model results for ORC-I structure are shown in table 4.2.26. Neither the model

with the test-only group baseline nor the model with the input flood baseline had statistically significant effects.

Table 4. 2. 26 The fixed effects of the model analysis of ORC-I structure (matched items) for deciding match or mismatch task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	3.55 [2.01, 5.09]	.94	3.78	<.001***	34.83 [7.44, 163.02]
test vs. parsing	1.30 [-.72, 3.31]	1.22	1.06	.289	3.66 [0.49, 27.40]
test vs. input	.04 [-1.45, 1.53]	.91	0.05	.963	1.04 [0.23, 4.63]
pre- vs. post-	1.38 [-.61, 3.37]	1.21	1.14	.254	3.98 [0.54, 29.22]
pre- vs. delayed-	1.21 [-.79, 3.20]	1.21	0.99	.321	3.34 [0.45, 24.62]
test vs. parsing × pre- vs. post-	-2.56 [-5.38, .25]	1.71	-1.50	.135	0.08 [0.00, 1.29]
test vs. input × pre- vs. post-	-.91 [-3.47, 1.64]	1.55	-0.59	.557	0.40 [0.03, 5.17]
test vs. parsing × pre- vs. delayed-	-1.26 [-4.38, 1.87]	1.90	-0.66	.508	0.28 [0.01, 6.47]
test vs. input × pre- vs. delayed-	-1.26 [-3.74, 1.22]	1.51	-0.84	.402	0.28 [0.02, 3.37]

Note: Model formula: model1=glmer(match_score ~ group*stage + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000)); Marginal R^2=0.01, conditional R^2=0.08; parsing = parsing group; input = input flood group; test = test-only group; *** significantly differently from zero when α < .001

Mismatched items

SRC-A

For SRC-A, the model results are shown in table 4.2.27. No statistically significant comparison or interaction could be found in the model with the baseline of the test-only group or with the baseline of the input flood group.

Table 4. 2. 27 The fixed effects of the model analysis of SRC-A structure (mismatched items) for deciding match or mismatch task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	2.47 [1.29, 3.66]	.72	3.43	.001**	11.87 [3.62, 38.90]
test vs. parsing	-.09 [-1.33, 1.14]	.75	-.12	.902	.91 [.27, 3.13]
test vs. input	-.09 [-1.34, 1.16]	.76	-.12	.907	.91 [.26, 3.20]
pre- vs. post-	.25 [-1.22, 1.72]	.89	.28	.777	1.29 [.30, 5.59]
pre- vs. delayed-	.22 [-1.20, 1.65]	.87	.26	.795	1.25 [.30, 5.20]
test vs. parsing × pre- vs. post-	1.29 [-.42, 3.00]	1.04	1.24	.216	3.62 [.66, 20.04]
test vs. input × pre- vs. post-	.36 [-1.23, 1.95]	.97	.37	.710	1.43 [.29, 7.03]
test vs. parsing × pre- vs. delayed-	1.73 [-.14, 3.59]	1.13	1.52	.128	5.61 [.87, 36.15]
test vs. input × pre- vs. delayed-	.49 [-1.13, 2.10]	.98	.50	.621	1.63 [.32, 8.20]

Note: Model formula: model6=glmer(mismatch_score ~ group*stage + (1|subject) +

(stage|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=0.07, conditional R^2=0.50; parsing = parsing group; input = input flood group; test = test-only group; ** significantly differently from zero when α < .01

SRC-I

Table 4.2.28 shows the referential statistical results for SRC-I structure. The statistical significant effects could be found with the contrasts between the test-only and the parsing group, between the test-only and the input flood group, between the pre-test and the post-test, and between pre-test and delayed post-test. This indicated that overall the parsing and the input flood group outperformed the test-only group, and the participants as a whole were more likely to correctly respond to SRC-I items at the post- and the delayed post-test compared to the pre-test. In addition, one reliable effect could be observed in the interaction between group (test-only vs. input flood) and test phase (pre-test vs. post-test), which suggested that the test-only group had more chance to correctly respond to the SRC-I items relative to the input flood group in the comparison between the pre- and the post-test. In addition, in the model with the baseline of the input flood group, the parsing group was predicted to be 10.90 times more likely to provide a correct response for SRC-I items than the input flood group at the post-test compared to the pre-test (*b* [CI] = 2.39[.08, 4.70], SE = 1.40, *z* = 1.70, *p* = .089, OR [CI] = 10.90 [1.08, 109.55]).

Table 4. 2. 28 The fixed effects of the model analysis of SRC-I structure (mismatched items) for deciding match or mismatch task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	.95 [-.40, 2.31]	.83	1.16	.248	2.60 [.67, 10.11]
test vs. parsing	2.04 [.58, 3.50]	.89	2.30	.022*	7.70 [1.78, 33.20]
test vs. input	1.69 [.28, 3.10]	.86	1.97	.049*	5.41 [1.32, 22.19]
pre- vs. post-	2.04 [.49, 3.60]	.94	2.16	.031*	7.71 [1.63, 36.43]
pre- vs. delayed-	2.32 [.49. 4.16]	1.12	2.08	.037*	10.21 [1.63, 63.96]
test vs. parsing × pre- vs. post-	.65 [-1.60, 2.90]	1.37	.48	.634	1.92 [.20, 18.22]
test vs. input × pre- vs. post-	**-1.74 [-3.31, -.16]**	**.96**	**-1.81**	**.070**	**.18 [.04, .85]**
test vs. parsing × pre- vs. delayed-	-.68 [-2.43, 1.08]	1.07	-.64	.525	.51 [.09, 2.93]
test vs. input × pre- vs. delayed-	-.85 [-2.58, .87]	1.05	-.81	.417	.43 [.08, 2.40]

Note: Model formula: model6=glmer(mismatch_score ~ group*stage + (1|subject) +

(stage|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=0.14, conditional R^2=0.72; parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates that the CI of estimate *b* did not pass through zero; * significantly differently from zero when α < .005

ORC-A

The model analysis results are shown in table 4.2.29. Statistically significant effects were found in the contrast between pre- and delayed post-test, which means that in general, the participants improved at the delayed post-test relative to the pre-test. However, the differences in the interactions were negligible. In the model with the input flood group baseline, no statistically significant effect could be found.

Table 4. 2. 29 The fixed effects of the model analysis of ORC-A structure (mismatched items) for deciding match or mismatch task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	2.06 [1.10, 3.02]	.58	3.53	<.001***	7.83 [3.01, 20.43]
test vs. parsing	.49 [-0.78, 1.77]	.77	.64	.524	1.64 [0.46, 5.85]
test vs. input	-.33 [-1.57, .91]	.75	-.44	.661	0.72 [0.21, 2.48]
pre- vs. post-	1.04 [-.06, 2.13]	.67	1.56	.118	2.83 [0.95, 8.45]
pre- vs. delayed-	1.71 [0.44, 2.98]	.77	2.21	.027*	5.52 [1.55, 19.67]
test vs. parsing × pre- vs. post-	.26 [-1.41, 1.94]	1.02	.26	.796	1.30 [0.24, 6.95]
test vs. input × pre- vs. post-	-.51 [-1.99, 0.97]	.90	-.57	.571	0.60 [0.14, 2.63]
test vs. parsing × pre- vs. delayed-	.08 [-1.84, 2.01]	1.17	.07	.944	1.09 [0.16, 7.46]
test vs. input × pre- vs. delayed-	-.95 [-2.57, 0.68]	.99	-.96	.339	0.39 [0.08, 1.98]

Note: Model formula: model1=glmer(mismatch_score ~ group*stage + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=0.11, conditional R^2=0.52; parsing = parsing group; input = input flood group; test = test-only group; *** significantly differently from zero when α < .001; * significantly differently from zero when α < .05

ORC-I

The results for ORC-I structure (see table 4.2.30) indicated that overall, the participants as a whole scored significantly higher at the post- and delayed post-test compared to the pre-test, and the input flood group significantly scored higher than the test-only group as a whole. However, two two-way interactions between group (test-only vs. input flood) and the test phase (pre-test vs. post-test; pre-test vs. delayed post-test)

were statistically significant. The results indicated that the input flood group was less likely to correctly respond to ORC-I items compared to the test-only group at the post- and the delayed post-test compared to the pre-test. Referring to the mean scores of the test-only and the input-flood group at the pre-, post-, and delayed post-test, the mean score of the test-only group at the pre-test was much lower that the input flood group (mean [SD]$_{\text{test-only}}$ = .48 [.50], mean [SD] $_{\text{input flood}}$ = .71[.46]), which might lead to the seemingly significantly more gains of the test-only group than the input flood group. In addition, in the model with the baseline of the input flood group, a two-way interaction was statistically significant. The parsing group was predicted to be 5.63 more likely to respond to the ORC-I items than the input flood at the delayed post-test relative to the pre-test (*b* [CI] = 1.73 [.35, 3.10], SE = 0.84, *z* = 2.07, *p* = .039, OR [CI] = 5.63 [1.42, 22.28])

Table 4. 2. 30 The fixed effects of the model analysis of ORC-I structure (mismatched items) for deciding match or mismatch task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-.05 [-.96, .86]	.55	-.09	.931	.95 [.38, 2.36]
test vs. parsing	.91 [-.21, 2.03]	.68	1.34	.180	2.49 [.81, 7.63]
test vs. input	1.46 [.29, 2.62]	.71	2.06	.039*	4.29 [1.34, 13.69]
pre- vs. post-	1.92 [.99, 2.85]	.57	3.39	<.001***	6.82 [2.69, 17.29]
pre- vs. delayed-	2.98 [1.87, 4.08]	.67	4.42	<.001***	19.62 [6.49, 59.32]
test vs. parsing × pre- vs. post-	-.38 [-1.67, .92]	.79	-.48	.634	.69 [.19, 2.51]
test vs. input × pre- vs. post-	-1.66 [-2.97, -.35]	.80	-2.09	.037*	.19 [.05, .70]
test vs. parsing × pre- vs. delayed-	-.91 [-2.38, .56]	.89	-1.02	.310	.40 [.09, 1.75]
test vs. input × pre- vs. delayed-	-2.64 [-4.07, -1.21]	.87	-3.03	.002**	.07 [.02, .30]

Note: Model formula: model1=glmer(mismatch_score ~ group*stage + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=0.11, conditional R^2=0.56; parsing = parsing group; input = input flood group; test = test-only group; *** significantly differently from zero when α < .001;** significantly differently from zero when α < .01; * significantly differently from zero when α < .05

e) Summary

In summary, in deciding match or mismatch task for the metalinguistic knowledge test, for the matched items, no statistically significant effects could be observed in all the

structures. The mean scores and the plots showed that all the three groups scored at ceiling from the pre- to the delayed post-test.

For mismatched items, all the three groups scored higher at the post- and the delayed post-test, relative to pre-test for all the structures. The between-group effect sizes and the inferential statistics indicated that for the SRC-I and ORC-I structures, the parsing and the test-only group had significantly more gains than the input flood group across time.

4.2.5.2 Analysis of accuracy scores of sentence correction task in the metalinguistic test

a) Descriptive analysis

The table 4.2.31 presents the means and the *SD*s of each group at each test phase. The gains at the post- and delayed post-tests could be observed with all the groups for all the structures, and the parsing group scored higher than both the input flood and the test-only group at the post- and delayed post-tests.

Table 4. 2. 31 Mean (*SDs*) scores of sentence correction task in metalinguistic knowledge test

Structure	test phase	parsing	input flood	test-only
SRC-A (k=2)	pre-test	.73 (.45)	.69 (.47)	.50 (.51)
	post-test	.94 (.23)	.75 (.44)	.73 (.45)
	delayed post-test	.94 (.24)	.80 (.40)	.79 (.41)
SRC-I (k=2)	pre-test	.67 (.48)	.67 (.47)	.43 (.50)
	post-test	.94 (.23)	.81 (.40)	.80 (.40)
	delayed post-test	.89 (.32)	.79 (.41)	.75 (.44)
ORC-A (k=2)	pre-test	.78 (.42)	.71 (.46)	.66 (.48)
	post-test	.94 (.23)	.79 (.41)	.88 (.33)
	delayed post-test	.96 (.19)	.73 (.45)	.92 (.27)
ORC-I (k=2)	pre-test	.58 (.50)	.65 (.48)	.39 (.49)
	post-test	.83 (.38)	.73 (.45)	.76 (.43)
	delayed post-test	.89 (.32)	.67 (.47)	.86 (.35)

Note: The *structure* refers to the structure of the sentence presented to the participants. For mismatched items, the SRCs were required to correct to the ORCs and the ORCs were required to correct to SRCs.

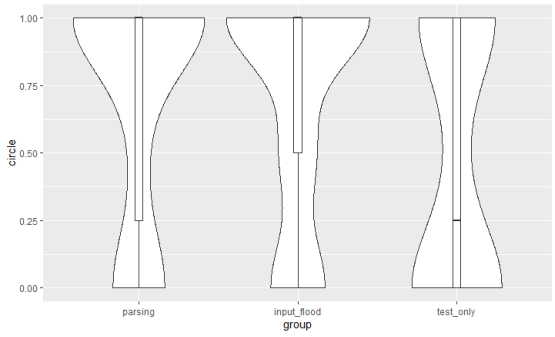
b) Plots

Figures 4.2.25 to 4.2.28 present the violin plots with boxplots of mean scores for each participant in sentence correction task of the metalinguistic knowledge test. For the SRCs and ORC-A, the advantages of the parsing group over the input flood and the test-only group could be observed. At the post- and delayed post-test, the parsing group had more participants scoring at ceiling than the other two groups. In addition, in the two types of ORCs, the test-only group showed improvements, as more people scored at ceiling at the post- and the delayed post-test compared to the pre-test.

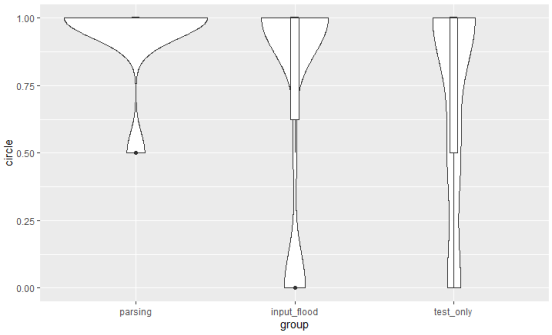
In sum, the plots indicated that the parsing strategies training facilitated the performance in the sentence correction task of metalinguistic knowledge for all the structures, and the test-only group also had improvements in ORCs.

SRC-A

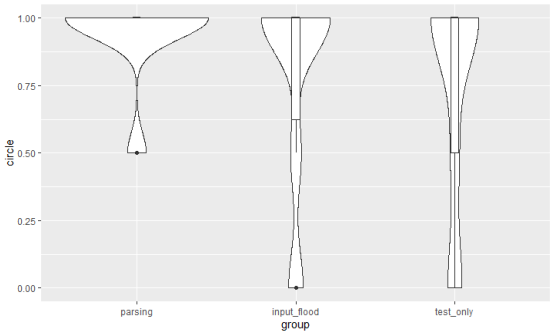
Figure 4. 2. 25 Comparison of accuracy scores of three learner groups in different test phase for SRC-A structure in the sentence correction task of metalinguistic knowledge test



Pre-test



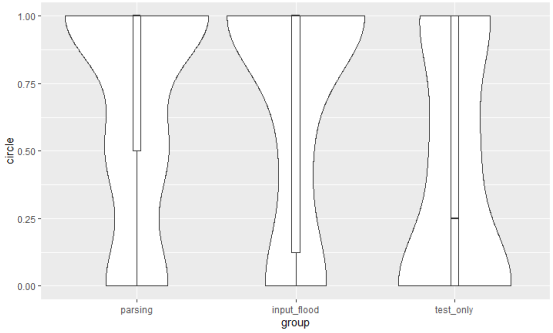
Post-test



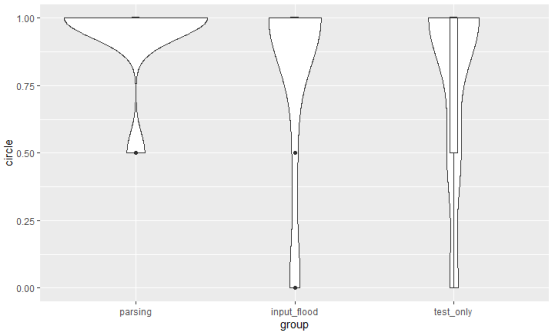
Delayed post-test

SRC-I

Figure 4. 2. 26 Comparison of accuracy scores of three learner groups in different test phase for SRC-I structure in the sentence correction task of metalinguistic knowledge test



Pre-test

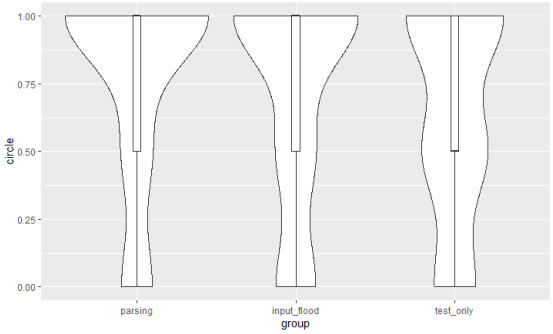


Post-test

Delayed post-test

ORC-A

Figure 4. 2. 27 Comparison of accuracy scores of three learner groups in different test phase for ORC-A structure in the sentence correction task of metalinguistic knowledge test



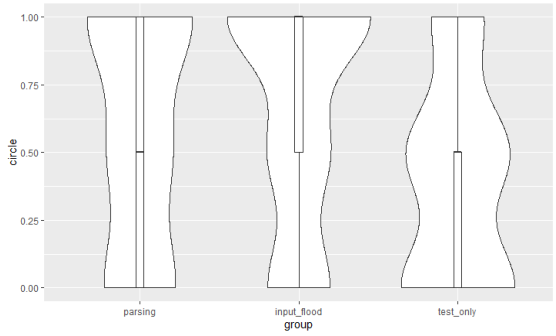
Pre-test

Post-test

Delayed post-test

ORC-I

Figure 4. 2. 28 Comparison of accuracy scores of three learner groups in different test phase for ORC-I structure in the sentence correction task of metalinguistic knowledge test



Pre-test

Post-test

Delayed post-test

c) Examination of effect sizes

Table 4.2.32 presents the effect sizes of within-group contrasts, reflecting changes over time. The results showed that for all the structures, the parsing and the test-only group had reliable gains (the 95% CI did not pass through zero) over time, and the changes of scores for the input flood group were negligible. For the parsing group, the small effects were observed in the SRC-I structure, for the pre- to post-test comparison; for the test-only group, the improvement between pre- and post-, delayed post-test had small effects.

Table 4. 2. 32 Within-group effect size (Cohen's *d*) [95% CI] for sentence correction task of metalinguistic knowledge test

structure	group	pre-post	pre-delayed	post-delayed
SRC-A (k=2)	parsing	**.38 [.09, .67]**	**.45 [.15, .75]**	.00 [-.27, .27]
	input flood	.11 [-.17, .38]	.17 [-.11, .45]	.09 [-.19, .36]
	test-only	**.37 [.08, .67]**	**.54 [.24, .86]**	.09 [-.20, .38]
SRC-I (k=2)	parsing	**.75 [.35, 1.14]**	**.40 [.11, .69]**	-.15 [-.43, .12]
	input flood	.28 [.00, .56]	.21 [-.07, .49]	-.05 [-.33, .22]
	test-only	**.58 [.28, .89]**	**.45 [.15, .75]**	-.14 [-.42, .14]
ORC-A (k=2)	parsing	**.48 [.09, .87]**	**.37 [.08, .66]**	.06 [-.21, .34]
	input flood	.18 [-.10, .45]	.04 [-.23, .32]	-.12 [-.39, .16]
	test-only	**.34 [.05, .64]**	**.42 [.12, .72]**	.09 [-.19, .37]
ORC-I (k=2)	parsing	**.58 [.19, .97]**	**.59 [.30, .90]**	.11 [-.16, .38]
	input flood	.15 [-.13, .43]	.03 [-.24, .31]	-.13 [-.40, .15]
	test-only	**.60 [.30, .91]**	**.88 [.56, 1.22]**	.20 [-.08, .48]

Note: Bold typeface indicates that the CI did not pass through zero.

The effect sizes of between-group contrasts are presented in table 4.2.33. The results indicated that the test-only group performed worse than the other two groups at the pre-test, and the differences had small effects for the SRCs structures (compared to the parsing and the input flood groups) and the ORC-I structure (compared to the input flood group). At the post-test, the advantages of the parsing group over the input flood and the test-only group had small effects in SRCs structures; and in the ORC-A, the small effects also could be observed in the comparison between the parsing and the input flood group. At the delayed post-test, except SRC-I, the gains for the parsing group over the input flood group had small effects, and the score differences between the parsing and the test-only group had small effects for SRC-A structure.

However, taking account of the influence of the pre-test (see the *d* changes from the pre-test to the post- and delayed post-test), at the post-test, the gains of the paring group over the input flood group had small effects for SRC-A, SRC-I and ORC-I structure. At the delayed post-test, the small effect was found with ORC-A structure, and the medium effect could be observed in ORC-I structure. Moreover, the test-only group

made more gains at the post- and delayed post-test compared with the input flood group, and some of the gains had small to medium effects.

Table 4. 2. 33 Between-group effect size (Cohen's *d*) [95% CI] for sentence correction task of metalinguistic knowledge test

structure	test phase	parsing vs. input flood	parsing vs. test-only	input flood vs. test-only
SRC-A (k=2)	pre-test	.09 [-.30, .48]	**.49 [.09, .90]**	**.40 [.00, .79]**
	post-test	**.56 [.17, .95]**	**.62 [.22, 1.01]**	.06 [-.33, .44]
	delayed post-test	**.46 [.07, .85]**	**.47 [.07, .87]**	.00 [-.39, .40]
	pre-post *d* change	**.47 [N/A]**	.13 [N/A]	-.34 [N/A]
	pre-delayed *d* change	.37 [N/A]	-.02 [N/A]	**-.40 [N/A]**
SRC-I (k=2)	pre-test	-.01 [-.40, .37]	**.49 [.09, .88]**	**.50 [.10, .90]**
	post-test	**.42 [.04, .81]**	**.43 [.04, .82]**	.01 [-.38, .40]
	delayed post-test	.27 [-.11, .66]	.36 [-.02, .75]	.09 [-.29, .47]
	pre-post *d* change	**.43 [N/A]**	-.06 [N/A]	**-.49 [N/A]**
	pre-delayed *d* change	.28 [N/A]	-.13 [N/A]	**-.41 [N/A]**
ORC-A (k=2)	pre-test	.17 [-.22, .55]	.28 [-.12, .68]	.11 [-.28, .51]
	post-test	**.47 [.08, .85]**	.24 [-.15, .62]	-.24 [-.63, .15]
	delayed post-test	**.67 [.27, 1.06]**	.17 [-.22, .56]	-.51 [-.91, -.21]
	pre-post *d* change	.30 [N/A]	-.04 [N/A]	-.35 [N/A]
	pre-delayed *d* change	**.50 [N/A]**	-.11 [N/A]	**-.62 [N/A]**
ORC-I (k=2)	pre-test	-.16 [.54, .23]	.37 [-.02, .76]	**.54 [.14, .93]**
	post-test	.25 [-.13, .63]	.18 [-.20, .57]	-.07 [-.45, .32]
	delayed post-test	**.54 [.15, .92]**	.09 [-.30, .47]	-.45 [-.84, -.05]
	pre-post *d* change	**.41 [N/A]**	-.19 [N/A]	**-.61 [N/A]**
	pre-delayed *d* change	**.70 [N/A]**	-.28 [N/A]	**-.99 [N/A]**

Note: Bold typeface indicates that the CI did not pass through zero.

d) Inferential statistical analysis

The inferential statistical analyses were conducted using mixed logistic regression models, run separately for SRC-A, SRC-I, ORC-A and ORC-I structures. The results reported below were based on model with the baseline of the test-only group. The statistically significant effects, related to the comparisons between the input flood and the parsing group, in the model with the baseline of the input flood group would be reported in this section (for full results see Appendix 32). The AIC and LRT results for

model selections see Appendix25.

SRC-A

Table 4.2.34 presents the inferential statistical results of SRC-A structure. The statistically significant effects were found in four comparisons: the test-only and the parsing group comparison, the pre-test and the post-test comparison, and the pre-test and the delayed post-test comparison. The results indicated that the parsing group statistically scored higher than the test-only group in general, and the participants as a whole statistically scored higher at the post- and delayed post-test than the pre-test. In addition, a comparison between the test-only and the input flood group, and a two-way interaction between group (test-only vs. input flood) and test phase (pre-test vs. post-test) were found to be reliable (CI of estimate *b* did not pass through zero). This indicated as a whole, the input flood outperformed the test-only group, but the test-only group showed more gains than the input flood group at the post-test compared to the pre-test. In the model with the baseline of the input flood group, the interaction between group (input flood vs. parsing group) and test phase (pre-test vs. post-test) was found to be statistically significant ((*b* [CI] = 2.23 [.54, 3.91], SE = 1.02, *z* = 2.18, p = .030, OR [CI] = 9.29 [1.72, 50.08]). This suggested that compared to the pre-test, the parsing group was predicted to be 9.29 times more likely to correctly respond to the question than the input flood group at the post-test.

Table 4. 2. 34 The fixed effects of the model analysis of SRC-A structure for sentence correction task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-0.19 [-1.28, 0.91]	.66	-.28	.779	.83 [0.28, 2.48]
test vs. parsing	2.06 [0.52, 3.60]	.94	2.20	.028*	7.85 [1.68, 36.60]
test vs. input	1.68 [0.19, 3.17	.91	1.85	.064	5.35 [1.21, 23.71]
pre- vs. post-	2.05 [0.98, 3.12]	.65	3.15	.002**	7.75 [2.66, 22.63]
pre- vs. delayed-	2.35 [1.16, 3.55]	.72	3.25	.001**	10.52 [3.19, 34.65]
test vs. parsing × pre- vs. post-	.73 [-0.97, 2.42]	1.03	0.71	.479	2.07 [0.38, 11.27]
test vs. input × pre- vs. post-	**-1.50 [-2.93, -0.07]**	**.87**	**-1.73**	**.084**	**.22 [0.05, 0.93]**
test vs. parsing × pre- vs. delayed-	.27 [-1.50, 2.05]	1.08	.25	.800	1.31 [0.22, 7.78]
test vs. input × pre- vs. delayed-	-1.32 [-2.88, 0.23]	.94	-1.40	.161	.27 [0.06, 1.26]

Note: Model formula: model1=glmer(circle ~ group*stage + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer =

"bobyqa",optCtrl=list(maxfun=100000))); marginal R^2=0.18, conditional R^2=0.71; parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when α < .05; ** significantly differently from zero when α < .01;

SRC-I

The statistical results are presented in table 4.2.35. The results indicated that overall, the parsing group and the input flood group scored significantly higher than the test-only group, and the three groups as a whole had significant higher score at the post- and delayed post-test compared to the pre-test. In addition, the interaction between the group (test-only vs. input flood group) and the test phase (pre-test vs. post-test) was statistically significant, which indicated that the test-only group was more likely to correctly respond to the SRC-I items than the input flood group at the post-test relative to the pre-test. In the model with the baseline of the input flood group, a two-way interaction was found to be statistically significant (input flood group vs. parsing group: pre-test vs. post-test: *b* [CI] = 2.13 [.35, 3.91], SE = 1.08, *z* = 1.97, *p* = .049, OR [CI] = 8.42 [1.42, 49.87]). The parsing group was predicted to be 8.42 times more likely to provide correct answers than the input flood group at the post-test compared to the pre-test.

Table 4. 2. 35 The fixed effects of the model analysis of SRC-I structure for sentence correction task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-.72 [-1.92, 0.47]	.73	-1.00	.320	.48 [0.15, 1.61]
test vs. parsing	2.09 [.59, 3.59]	.91	2.29	.022*	8.10 [1.81, 36.32]
test vs. input	2.15 [.65, 3.65]	.91	2.36	.018**	8.60 [1.92, 38.49]
pre- vs. post-	3.17 [1.72, 4.63]	.89	3.58	<.001***	23.87 [5.56, 102.49]
pre- vs. delayed-	3.23 [1.52, 4.94]	1.04	3.11	.002**	25.28 [4.57, 139.82]
test vs. parsing × pre- vs. post-	.09 [-1.70, 1.89]	1.09	.09	.931	1.10 [0.18, 6.60]
test vs. input × pre- vs. post-	-2.04 [-3.61, -.46]	.96	-2.13	.033*	.13 [0.03, 0.63]
test vs. parsing × pre- vs. delayed-	-.41 [-2.11, 1.28]	1.03	-.40	.688	.66 [0.12, 3.61]
test vs. input × pre- vs. delayed-	-1.54 [-3.21, .14]	1.02	-1.51	.132	.22 [0.04, 1.15]

Note: Model formula: model6=glmer(circle ~ group*stage + (1|subject) + (stage|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); marginal R^2=0.18, conditional R^2=0.74; parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from

zero when α < .05; ** significantly differently from zero when α < .01; *** significantly differently from zero when α < .001

ORC-A

Table 4.2.36 shows the model results of ORC-A structure. The statistically significant effects were found with the pre-test vs. post-test comparison and the pre-test vs. delayed post-test comparison, which means that the participants as a whole scored significantly higher than at the post- and delayed post-test than at the pre-test. In addition, the test-only group had significant more gains than the input flood group at the delayed post-test compared to the pre-test. In the model with the input flood group baseline, a two-way interaction between the group (input flood vs. parsing group) and the test phase (pre-test vs. delayed post-test) was found statistically significant (*b* [CI] = 2.22 [.52, 3.92], SE = 1.03, *z* = 2.15, *p* = .032, OR [CI] = 9.20 [1.68, 50.38]). The odds ratio suggested that the parsing group was 9.20 times more likely to provide correct answers than the input flood group at the delayed post-test compared to the pre-test.

Table 4. 2. 36 The fixed effects of the model analysis of ORC-A structure for sentence correction task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	**.95 [.09, 1.82]**	**.52**	**1.83**	**.068**	**2.60 [1.10, 6.14]**
test vs. parsing	.97 [-.26, 2.19]	.74	1.30	.194	2.63 [0.77, 8.95]
test vs. input	.45 [-.75, 1.66]	.73	.62	.536	1.57 [0.47, 5.26]
pre- vs. post-	1.78 [.73, 2.83]	.64	2.80	.005**	5.94 [2.08, 16.95]
pre- vs. delayed-	2.46 [1.27, 3.64]	.72	3.42	.001**	11.65 [3.57, 38.00]
test vs. parsing × pre- vs. post-	.15 [-1.47, 1.77]	.99	.16	.876	1.17 [0.23, 5.90]
test vs. input × pre- vs. post-	-1.13 [-2.54, .28]	.86	-1.32	.187	0.32 [0.08, 1.32]
test vs. parsing × pre- vs. delayed-	-.08 [-1.91, 1.76]	1.12	-.07	.944	0.92 [0.15, 5.80]
test vs. input × pre- vs. delayed-	-2.30 [-3.79, -.81]	.91	-2.54	.011*	0.10 [0.02, 0.45]

Note: Model formula: model1=glmer(circle ~ group*stage + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=0.16, conditional R^2=0.56; parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when α < .05; ** significantly differently from zero when α < .01;

ORC-I

The model results for ORC-I (see table 4.2.37) showed the statistically significant effects

in comparisons between the test-only group and the input flood group, between the pre-test and the post-test, between the pre-test and the delayed post-test, and in an interaction between group (test-only vs. input flood) and test phase (pre-test vs. post- & delayed post-test). In addition, the comparison between the test-only group and the parsing group was also reliable. The results indicated that the participants as a whole scored significantly higher at the post- and the delayed post-test compared to the pre-test. The input flood group had significantly higher score than the test-only group; however, the input flood group had smaller gains than the test-only group at the post- and delayed post-test relative to the test-only group. In the model with the baseline of the input flood group, the interaction between the group (input flood vs. parsing group) and the test phase (pre-test vs. delayed post-test) was statistically significant. The odds ratio predicted that the the parsing group was 10.54 times more likely to correctly respond to the item than the input flood group at the delayed post-test compared to the pre-test (*b* [CI] = 2.22 [.52, 3.92], SE = 1.03, *z* = 2.15, *p* = .032, OR [CI] = 9.20 [1.68, 50.38]).

Table 4. 2. 37 The fixed effects of the model analysis of ORC-I structure for sentence correction task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-.65 [-1.63, 0.32]	.59	-1.11	.269	.52 [0.20, 1.38]
test vs. parsing	**1.26 [0.05, 2.48]**	**.74**	**1.71**	**.088**	**3.54 [1.05, 11.98]**
test vs. input	1.71 [0.46, 2.96]	.76	2.26	.024*	5.53 [1.59, 19.24]
pre- vs. post-	2.63 [1.61, 3.65]	.62	4.24	<.001***	13.84 [4.99, 38.40]
pre- vs. delayed-	3.65 [2.43, 4.86]	.74	4.95	<.001***	38.32 [11.41, 128.72]
test vs. parsing × pre- vs. post-	-.65 [-2.03, 0.72]	.83	-.78	.433	.52 [0.13, 2.05]
test vs. input × pre- vs. post-	-1.90 [-3.27, -0.52]	.84	-2.27	.023*	.15 [0.04, 0.59]
test vs. parsing × pre- vs. delayed-	-1.17 [-2.73, 0.38]	.95	-1.24	.215	.31 [0.07, 1.47]
test vs. input × pre- vs. delayed-	-3.53 [-5.04, -2.01]	.92	-3.83	<.001***	.03 [0.01, 0.13]

Note: Model formula: model1=glmer(circle ~ group*stage + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=0.15 conditional R^2=0.62; parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when α < .05; *** significantly differently from zero when α < .001;

e) Summary

The statistical results indicated that the participants as a whole gained at the post- and delayed post-test. Referring to the descriptive results and the plots, this improvement was mainly attributed to the gains of the parsing and the test-only group. The gains of the input flood group across the time were very small according to the means of the group at each test phase, and there was no reliable effect that could be found with the input flood group with the changes over time.

The model results indicated that in SRC-I, ORC-A and ORC-I structures, the input flood group had smaller gains than the other two groups. In ORCs, the advantages of the test-only group over the input flood group were statistically significant. However, the greater gains of the test-only group might because the test-only group had much lower pre-test scores than the other groups. Yet, at the post- and the delayed post-test, the test-only group did not have significant higher score than the other groups. Moreover, the parsing group scored at ceiling at the post- and delayed post-test, though the range of improvement might not be statistically different than the test-only group (as noted, this might be because the test-only group had lower pre-test scores than the other groups).

In summary, teaching parsing strategies seemed to facilitate sentence correction in the metalinguistic test, but there was no evidence of any such advantages of input flood training. In addition, it seemed that the test-only group gained from doing the tests, especially for the ORC-A and ORC-I structures.

4.2.5.3 Analysis of accuracy scores of reason explanation of the metalinguistic test

a) Descriptive analysis

Table 4.2.38 presents the mean (*SD*s) scores of each group at each test phase. The results suggested that for all the structures, the three groups scored extremely low at the pre-test, especially the parsing and the test-only group. The parsing group showed the significant improvement at the post-test for all the structures, though they had some decrease at the delayed post-test (though these scores were still much higher than the pre-test scores). The range of improvement was larger for the SRCs than the

ORCs. The input flood group almost had same score across the time. The test-only group showed small gains at the delayed post-test for SRC-A and ORC-I, but not for the SRC-I and ORC-A structures.

Table 4. 2. 38 Mean (*SDs*) scores of reason explanation task in metalinguistic knowledge test

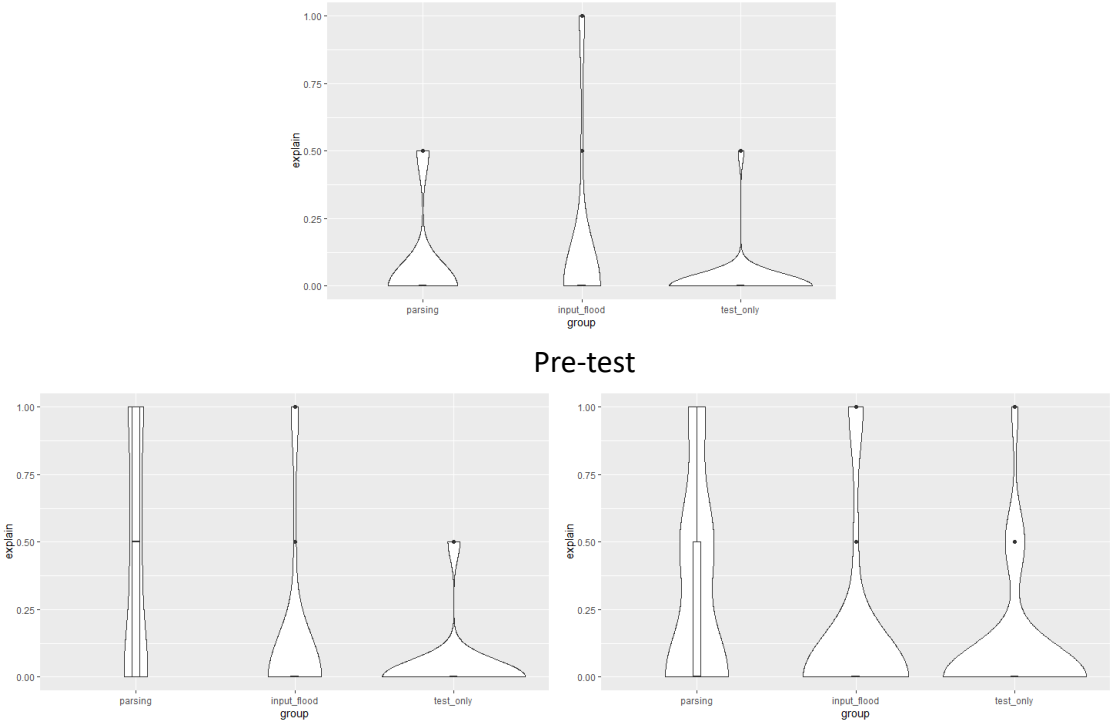
Structure	test phase	parsing	input flood	test-only
SRC-A (k=2)	pre-test	.07 (.26)	.13 (.34)	.02 (.14)
	post-test	.41 (.50)	.13 (.34)	.04 (.19)
	delayed post-test	.30 (.46)	.13 (.34)	.10 (.30)
SRC-I (k=2)	pre-test	.04 (.19)	.12 (.32)	.04 (.19)
	post-test	.41 (.50)	.12 (.32)	.02 (.14)
	delayed post-test	.33 (.48)	.15 (.36)	.04 (.19)
ORC-A (k=2)	pre-test	.09 (.29)	.10 (.30)	.04 (.19)
	post-test	.33 (.48)	.11 (.32)	.04 (.19)
	delayed post-test	.31 (.47)	.12 (.32)	.08 (.27)
ORC-I (k=2)	pre-test	.04 (.19)	.12 (.32)	.04 (.19)
	post-test	.35 (.48)	.08 (.27)	.02 (.14)
	delayed post-test	.24 (.43)	.10 (.30)	.12 (.32)

b) Plots

Figures 4.2.29 to 4.2.32 presents the violin plots with the boxplots for SRC-A, SRC-I, ORC-A and ORC-I separately. For all the structures, in general, all the participants had extremely low score at the pre-test. For the parsing group, the proportions of low scores decreased and the proportion of the high scores increased at the post- and delayed post-test. The distribution of scores for the input flood group did not change across time. The test-only group showed some improvement at delayed post-test for SRC-A and ORC-I structures, though the increase of the proportion of the participants who had high score was rather small.

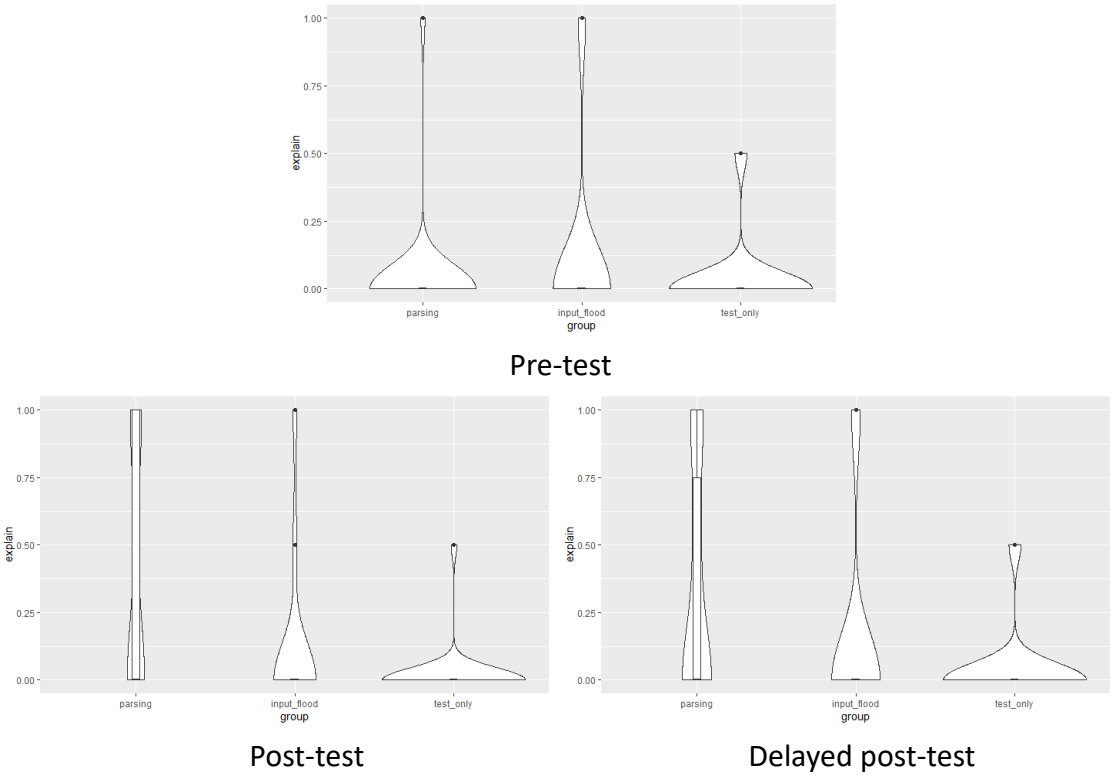
SRC-A

Figure 4. 2. 29 Comparison of accuracy scores of three learner groups in different test phase for SRC-A structure in the reason explanation task of metalinguistic knowledge test



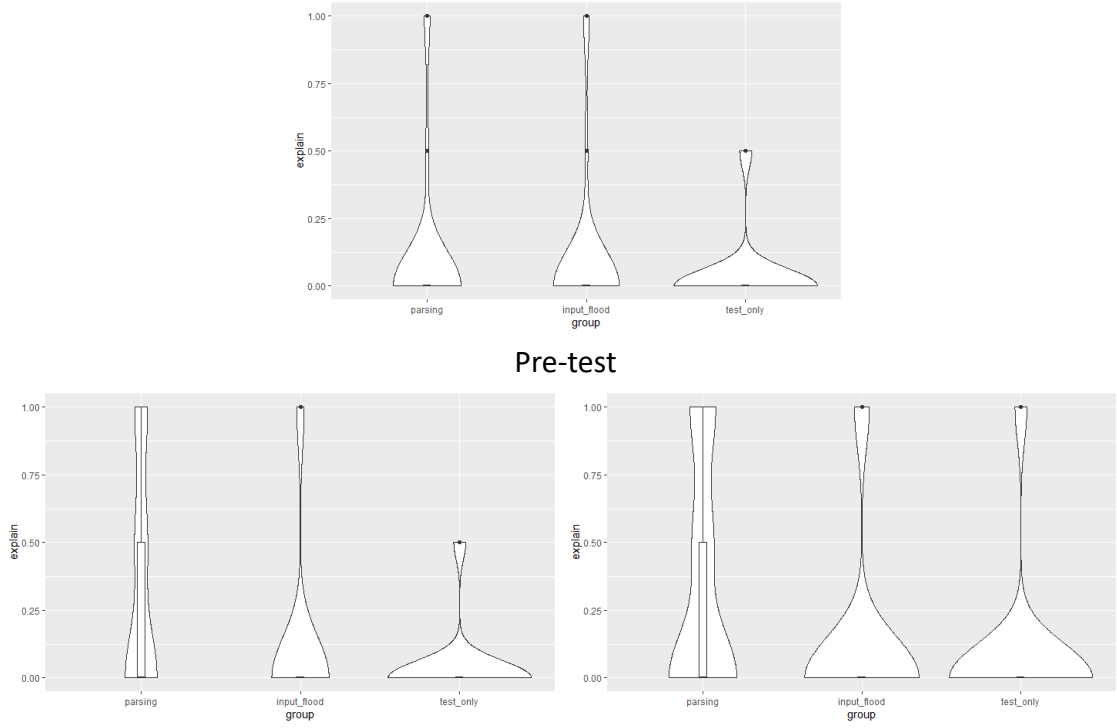
SRC-I

Figure 4. 2. 30 Comparison of accuracy scores of three learner groups in different test phase for SRC-I structure in the reason explanation task of metalinguistic knowledge test



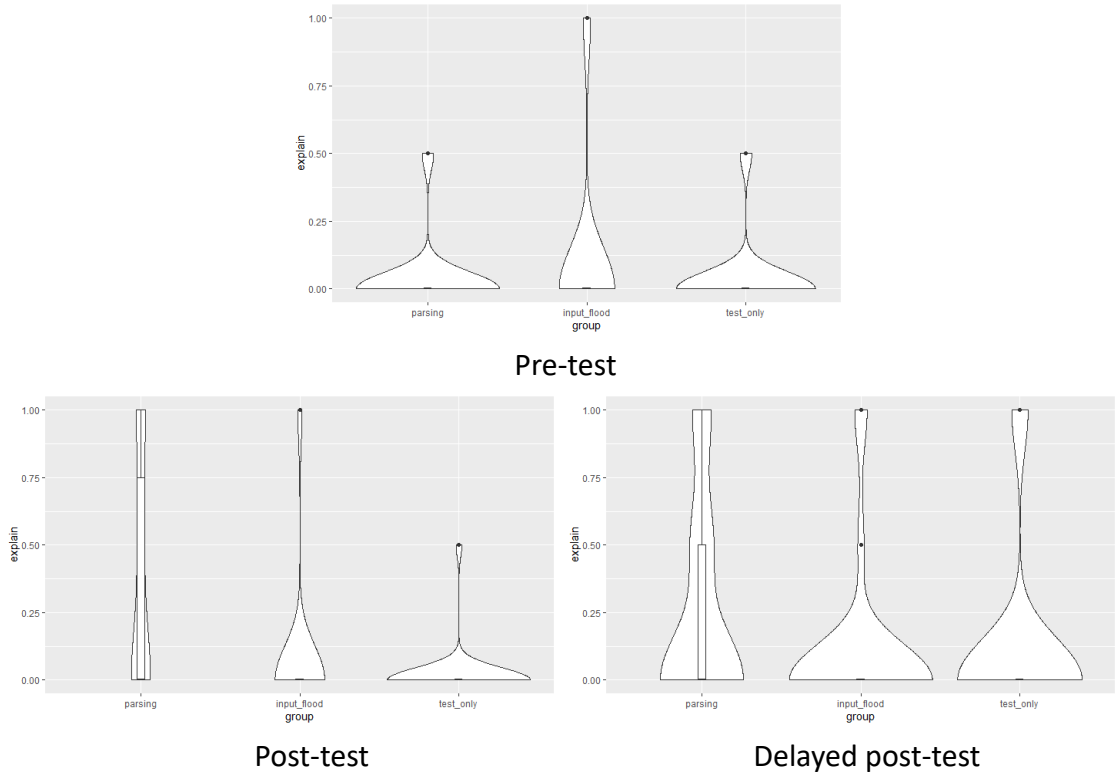
ORC-A

Figure 4. 2. 31 Comparison of accuracy scores of three learner groups in different test phase for ORC-A structure in the reason explanation task of metalinguistic knowledge test



ORC-I

Figure 4. 2. 32 Comparison of accuracy scores of three learner groups in different test phase for ORC-I structure in the reason explanation task of metalinguistic knowledge test



c) Examination of effect sizes

The within-group effect sizes, reflecting change over time, are presented in table 4.2.39. Reliable gains were found in the parsing group at post- and delayed post-test for all the structures. All the pre- to post-test comparisons of the parsing group had small effects, and the comparison between the pre-test and delayed post-test for the SRC-I structure also had small effects. The score changes of the input flood and the test-only group were negligible, because the 95% CIs passed through zero.

Table 4. 2. 39 Within-group effect size (Cohen's *d*) [95% CI] for reason explanation task of metalinguistic knowledge test

structure	group	pre-post	pre-delayed	post-delayed
SRC-A (k=2)	parsing	**.84 [.44, 1.23]**	**.48 [.20, .77]**	-.22 [-.50, .05]
	input flood	.00 [-.27, .27]	.00 [-.27, .27]	.00 [-.27, .27]
	test-only	.08 [-.19, .35]	.23 [-.05, .51]	.25 [-.03, .52]
SRC-I (k=2)	parsing	**.99 [.58, 1.38]**	**.64 [.35, .94]**	-.17 [-.45, .10]
	input flood	.00 [-.27, .27]	.11 [-.16, .39]	.14 [-.14, .42]
	test-only	-.08 [-.35, .19]	.00 [-.27, .27]	.14 [-.14, .42]
ORC-A (k=2)	parsing	**.61 [.22, .99]**	**.48 [.20, .77]**	-.04 [-.31, .23]
	input flood	.05 [-.22, .33]	.06 [-.21, .34]	.00 [-.27, .27]
	test-only	.00 [-.27, .27]	.11 [-.16, .39]	.14 [-.14, .42]
ORC-I (k=2)	parsing	**.86 [.46, 1.25]**	**.45 [.17, .74]**	-.22 [-.50, .05]
	input flood	-.11 [-.39, .16]	-.08 [-.35, .19]	.06 [-.21, .34]
	test-only	-.08 [-.36, .20]	.20 [-.08, .48]	.33 [.04, .61]

Note: Bold typeface indicates that the CI did not pass through zero.

The effect sizes of between group contrasts, reflecting the differences among groups, were shown in table 4.2.40. At the pre-test, the higher scores of the input flood group over the test-only group had the small effects for SRC-A structure.

At the post-test, the reliable effect sizes were found with the parsing versus input flood and test-only groups for all the structures, and most of them had small to medium effects. In addition, for SRC-I, the scores comparison between the parsing and the test-only group had the large effect.

At the delayed post-test, the advantages of the parsing group over the input flood

and the test-only group were reliable in SRC-A, SRC-I and ORC-A structure, and had small to medium effects. Moreover, another small effect was found with the input flood and the test-only group comparison for the SRC-I structure. The other differences between groups were negligible because the 95% CI passed through zero.

The changes of *d* from the pre-test to the post- and delayed post-test indicated that the parsing group made more gains than the input flood and the test-only groups across all the structures at the post- and/or the delayed post-test with small to large effects. In addition, the all of the changes of *d* did not reach the small benchmark with the comparison between the input flood and the test-only group.

Table 4. 2. 40 Within-group effect size (Cohen's *d*) [95% CI] for reason explanation task of metalinguistic knowledge test

structure	test phase	parsing vs. input flood	parsing vs. test-only	input flood vs. test-only
SRC-A (k=2)	pre-test	-.20 [-.58, .18]	.26 [-.12, .64]	**.44 [.05, .83]**
	post-test	**.64 [.24, 1.03]**	**.97 [.57, 1.37]**	.34 [-.04, .73]
	delayed post-test	**.40 [.01, .78]**	**.51 [.13, .09]**	.12 [-.27, .50]
	pre-post *d* change	**.84 [N/A]**	**.71 [N/A]**	-.10 [N/A]
	pre-delayed *d* change	**.60 [N/A]**	.25 [N/A]	-.32 [N/A]
SRC-I (k=2)	pre-test	-.30 [-.68, .09]	-.01 [-.39, .37]	.29 [-.10, .67]
	post-test	**.70 [.30, 1.09]**	**1.06 [.65, 1.46]**	.39 [.00, .77]
	delayed post-test	**.42 [.04, .81]**	**.81 [.41, 1.20]**	**.40 [.01, .78]**
	pre-post *d* change	**1.00 [N/A]**	**1.07 [N/A]**	.10 [N/A]
	pre-delayed *d* change	**.72 [N/A]**	**.82 [N/A]**	.11 [N/A]
ORC-A (k=2)	pre-test	-.01 [-.39, .37]	.22 [-.17, .60]	.23 [-.16, .61]
	post-test	**.53 [.15, .92]**	**.81 [.41, 1.20]**	.29 [-.10, .67]
	delayed post-test	**.49 [.11, .88]**	**.62 [.23, 1.01]**	.13 [-.26, .51]
	pre-post *d* change	**.54 [N/A]**	**.59 [N/A]**	.06 [N/A]
	pre-delayed *d* change	**.50 [N/A]**	**.40 [N/A]**	-.10 [N/A]
ORC-I (k=2)	pre-test	-.30 [.-68, .09]	-.01 [-.39, .37]	.29 [-.10, .67]
	post-test	**.70 [.30, 1.09]**	**.92 [.52, 1.33]**	.27 [-.12, .65]
	delayed post-test	.39 [.00, .77]	.33 [.-06, .71]	-.06 [-.45, .32]
	pre-post *d* change	**1.00 [N/A]**	**.93 [N/A]**	-.02 [N/A]
	pre-delayed *d* change	**.69 [N/A]**	.34 [N/A]	-.35 [N/A]

Note: Bold typeface indicates that the CI did not pass through zero.

d) Inferential statistical analysis

The inferential statistical analyses of reason explanation task of metalinguistic task were conducted using mixed effect logistic regression models, and models for each structure were run separately. The models with the baseline of the test-only group and the baseline of the input flood group were calculated separately. For the input flood baseline model, only the significant effects related to the comparison between the input flood and the parsing group would be reported in this section (for full results see appendix 32). The AIC and LRT results for model selections see Appendix25.

SRC-A

The model results for SRC-A are shown in table 4.2.41. The statistically significant effect was found with the comparison between pre- and delayed post-test, which indicated that the participants as a whole performed better at the delayed post-test than at the pre-test. In addition, the interaction between the group (test-only vs. parsing group) and test phase (pre-test vs. post-test) was found to be reliable, as the CI around the estimate *b* did not pass through zero. The parsing group was predicted to be 2.94 times more likely to provide metalinguistic knowledge than the test-only group at the post-test compared to the pre-test. In the model with the input flood baseline, the statistically significant effects could be found in the two interactions between group and test phase. It was predicted that the parsing group was 51.51 and 17.41 times more likely to provide metalinguistic knowledge than the input flood group at the post- and the delayed post-test compared to the pre-test respectively (input flood vs. parsing: pre-test vs. post-test: *b* [CI] = 3.94 [1.78, 6.10], SE = 1.31, *z* = 3.00, *p* = .003, OR [CI] = 51.51 [5.93, 447.80]; input flood vs. parsing: pre-test vs. delayed post-test: *b* [CI] = 2.86 [.73, 4.98], SE = 1.29, *z* = 2.21, *p* = .027, OR [CI] = 17.41 [2.08, 145.84]).

Table 4. 2. 41 The fixed effects of the model analysis of SRC-A structure for reason explanation task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-8.49 [-11.79, -5.19]	2.01	-4.23	<.001***	.00 [.00, 0.01]
test vs. parsing	2.69 [-0.59, 5.96]	1.99	1.35	.178	14.66 [14.66, 389.21]
test vs. input	2.65 [-0.45, 5.74]	1.88	1.41	.159	14.09 [0.64, 310.27]
pre- vs. post-	1.21 [-1.27, 3.69]	1.51	.81	.421	3.37 [0.28, 40.16]
pre- vs. delayed-	**2.82 [0.43, 5.22]**	**1.46**	**1.94**	**.053**	**16.80 [1.53, 184.45]**
test vs. parsing × pre- vs. post-	**2.94 [0.06, 5.83]**	**1.76**	**1.68**	**.094**	**18.99 [1.06, 340.68]**
test vs. input × pre- vs. post-	-1.00 [-3.87, 1.88]	1.75	-.57	.568	.37 [0.02, 6.53]
test vs. parsing × pre- vs. delayed-	.18 [-2.58, 2.95]	1.68	.11	.913	1.20 [0.08, 19.02]
test vs. input × pre- vs. delayed-	-2.67 [-5.48, 0.14]	1.71	-1.57	.117	.07 [0.00, 1.14]

Note: Model formula: model1=glmer(explain ~ group*stage + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=0.16, conditional R^2=0.87; parisng = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; *** significantly differently from zero when α < .001

SRC-I

Table 4.2.42 presents the model results of SRC-I structure. The statistical significant effects were found with one comparison and two interactions. The results indicated that compared to the pre-test, the parsing group was more likely to provide metalinguistic knowledge than the test-only group at the post- and delayed post-test with extremely high odds ratio. However, comparison between the test-only and the parsing group suggested that the test-only group significantly outperformed the parsing group as a whole. This effect was conflicted with the descriptive results and the effect size. It is hard to explain why this effect occur, it might because the sample size for each items was rather small (k = 2). In addition, in the model with the input flood baseline, the input flood was predicted to be more likely to provide metalinguistic knowledge than the parsing group as a whole (input flood vs. parsing: *b* [CI] = -7.57 [-13.40, -1.74], SE = 3.54, *z* = -2.14, *p* = .033, OR [CI] = .00 [.00, .18]). Compared to the pre-test, the parsing group was predicted to be more likely to provide metalinguistic knowledge than the input flood group at the post- (input flood vs. parsing: pre-test vs. post-test: *b* [CI] = 10.80 [5.60, 16.01], SE = 3.16, *z* = 3.42, *p* = .001, OR [CI] = 4.91E+04

[270.16, 8.93E+06]) and the delayed post-test (input flood vs. parsing: pre-test vs. delayed post-test: *b* [CI] = 8.20 [3.31, 13.10], SE = 2.97, *z* = 2.76, *p* = .006, OR [CI] = 3685.86 [27.39, 4.89E+05]).

Table 4. 2. 42 The fixed effects of the model analysis of SRC-I structure for reason explanation task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-10.21 [-13.96, -6.46]	2.28	-4.48	<.001***	.00 [0.00, 0.00]
test vs. parsing	-7.50 [-13.68, -1.32]	3.76	-2.00	.046*	.00 [0.00, 0.27]
test vs. input	0.07 [-4.17, 4.31]	2.58	.03	.979	1.07 [0.02, 74.11]
pre- vs. post-	-.90 [-3.47, 1.67]	1.56	-.58	.564	0.41 [0.03, 5.30]
pre- vs. delayed-	0.11 [-2.12, 2.34]	1.36	.08	.936	1.12 [0.12, 10.42]
test vs. parsing × pre- vs. post-	11.79 [6.34, 17.25]	3.32	3.56	<.001***	1.32E+05 [564.74, 3.10E+07]
test vs. input × pre- vs. post-	.99 [-2.23, 4.21]	1.96	.51	.612	2.70 [0.11, 67.37]
test vs. parsing × pre- vs. delayed-	9.44 [4.34, 14.54]	3.10	3.05	.002**	1.26E+04 [76.90, 2.06E+06]
test vs. input × pre- vs. delayed-	1.23 [-1.68, 4.15]	1.77	.70	.486	3.44 [.19, 63.32]

Note: Model formula: model1=glmer(explain ~ group*stage + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.06, conditional R^2=.98; parisng = parsing group; input = input flood group; test = test-only group; *** significantly differently from zero when α < .001; ** significantly differently from zero when α < .01; * significantly differently from zero when α≤.05

ORC-A

The model results of ORC-A structure (shown in table 4.2.43) had a statistically significant effect in the interaction between the group and the test phase. The results suggested that the parsing group was more likely to provide metalinguistic knowledge than the test-only group at the post-test compared to pre-test with very high odds ratio 53.23. Moreover, in the model with the baseline of the input flood group, the parsing group was predicted to be 28.06 times and 17.72 times more likely to provide metalinguistic knowledge than the input flood group at the post- and the delayed post-test respectively compared to the pre-test (input flood vs. parsing: pre-test vs. post-test: *b* [CI] = 3.33 [.85, 5.82], SE = 1.51, *z* = 2.21, *p* = .027, OR [CI] = 28.06 [2.34,

335.7]; input flood vs. parsing: pre-test vs. delayed post-test: *b* [CI] = 2.87 [.49, 5.26], SE = 1.45, *z* = 1.99, *p* = .047, OR [CI] = 17.72 [1.64, 191.88]).

Table 4. 2. 43 The fixed effects of the model analysis of ORC-A structure for reason explanation task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-8.23 [-11.69, -4.77]	2.10	-3.91	<.001***	.00 [0.00, 0.01]
test vs. parsing	.47 [-3.48, 4.41]	2.40	.20	.846	1.59 [0.03, 82.40]
test vs. input	.45 [-2.55, 3.45]	1.82	.25	.804	1.57 [0.08, 31.57]
pre- vs. post-	-.24 [-2.33, 1.84]	1.27	-.19	.847	0.78 [0.10, 6.32]
pre- vs. delayed-	1.04 [-0.85, 2.92]	1.15	0.91	.365	2.82 [0.43, 18.58]
test vs. parsing × pre- vs. post-	3.97 [1.21, 6.74]	1.68	2.36	.018*	53.23 [3.35, 846.09]
test vs. input × pre- vs. post-	.64 [-2.07, 3.35]	1.65	.39	.698	1.90 [0.13, 28.60]
test vs. parsing × pre- vs. delayed-	2.38 [-0.12, 4.88]	1.52	1.57	.117	10.83 [0.89, 131.88]
test vs. input × pre- vs. delayed-	-.49 [-3.04, 2.06]	1.55	-.32	.751	.61 [0.05, 7.82]

Note: Model formula: model1=glmer(explain ~ group*stage + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.06, conditional R^2=.92; parisng = parsing group; input = input flood group; test = test-only group; *** significantly differently from zero when α < .001; * significantly differently from zero when α≤.05

ORC-I

The statistical results for ORC-I structure are shown in table 4.2.44. An interaction between the group and test phase was found to be statistically significant. It suggested that the parsing group had significantly higher probability of providing metalinguistic knowledge than the test-only group at the post-test relative to the pre-test. In addition, a comparison between the pre-test and the delayed post-test was found to be reliable, which indicated that the participants as whole were more likely to provide metalinguistic knowledge at the delayed post-test compared to the pre-test. In the model with the baseline of the input flood group, two two-way interactions were found to be statistically significant (input flood vs. parsing: pre-test vs. post-test: *b* [CI] = 5.80 [3.14, 8.45], SE = 1.61, *z* = 3.59, *p* < .001, OR [CI] = 329.37 [23.17, 4682.74]; input flood vs. parsing: pre-test vs. delayed post-test: *b* [CI] = 4.08 [1.61, 6.55], SE = 1.50, *z* = 2.72, *p* = .007, OR [CI] = 59.00 [4.99, 697.11]).

Table 4. 2. 44 The fixed effects of the model analysis of ORC-I structure for reason explanation task in metalinguistic knowledge test

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-7.55 [-10.48, -4.62]	1.78	-4.23	<.001***	.00 [.00, .01]
test vs. parsing	-.02 [-3.24, 3.20]	1.96	-.01	.992	.98 [.04, 24.50]
test vs. input	1.71 [-1.03, 4.44]	1.66	1.03	.304	5.52 [0.36, 84.98]
pre- vs. post-	-.93 [-3.24, 1.39]	1.41	-.66	.511	.40 [0.04, 4.01]
pre- vs. delayed-	**2.00 [0.15, 3.85]**	**1.13**	**1.78**	**.076**	**7.39 [1.16, 47.00]**
test vs. parsing × pre- vs. post-	5.85 [2.77, 8.93]	1.87	3.12	.002**	347.52 [15.95, 7571.61]
test vs. input × pre- vs. post-	-.14 [-3.04, 2.77]	1.77	-.08	.937	.87 [0.05, 15.89]
test vs. parsing × pre- vs. delayed-	1.72 [-.90, 4.34]	1.59	1.08	.281	5.57 [.41, 76.50]
test vs. input × pre- vs. delayed-	-2.49 [-4.97, .00]	1.51	-1.65	.099	.08 [.01, 1.00]

Note: Model formula: model1=glmer(explain ~ group*stage + (1|subject) + (1|item), data=meta_score, family=binomial, control = glmerControl(optimizer = "bobyqa",optCtrl=list(maxfun=100000))); Marginal R^2=.12, conditional R^2=.87; parisng = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; *** significantly differently from zero when α < .001; ** significantly differently from zero when α < .01

e) Summary

In summary, for the reason explanation task, the benefits of teaching parsing strategies, revealed by both descriptive and statistical results, were very substantial. The participants of the parsing group showed significant improvement at the post- and delayed post-test for every structure. In addition, the input flood group almost kept the same scores across the time, which means that the input flood training did not facilitate the learning of metalinguistic knowledge. The mean scores of the test-only group showed some gains at the delayed post-test for the SRC-A and ORC-I structures, but the effect sizes results indicated that the gains were negligible.

4.2.5.4 Summary of the three tasks of the metalinguistic test

In the first task, deciding whether the sentence matched the picture or not, for the matched items, all the three groups performed at ceiling for all the structures across the time. For the mismatched items, the parsing group and the test-only group showed significantly more gains than the input flood group for SRC-I and ORC-I at the post- and

at the delayed post-test compared to the pre-test.

In the second task, correct the mismatched sentence to make it match the picture, the parsing and the test-only group had more gains than the input flood group at the post- and/or the delayed post-test compared to the pre-test.

In the third task, explain the reason for why the sentence mismatched the picture, only the parsing group showed statistically significant improvements at the post- and the delayed post-test. The other two groups could not provide metalinguistic knowledge across the time.

Chapter 5 Discussion

In this chapter, the findings of the current study will be discussed in the light of previous studies. The discussion will be structured around the results of the two research questions.

Section 5.1 will discuss RQ 1: Which type of relative clause (SRC vs. ORC) is more difficult in online and offline comprehension, oral production and metalinguistic knowledge tests?

Section 5.2 will discuss RQ 2: To what extent can teaching parsing strategies (with explicit information and practice), exposure alone, or no exposure (tests alone), develop the learning of relative clauses?

5.1 Which type of relative clause (SRC vs. ORC) is more difficult in online, offline comprehension, production, metalinguistic knowledge, and the role of animacy of main clause noun?

5.1.1 Summary of findings

5.1.1.1 SRCs vs. ORCs

To answer this question, we examined the data from the native speakers and from the L2 learners' pre-test data only (i.e., before they had undertaken the training). The expectation that the ORCs would be more difficult compared to SRCs for both native speakers and L2 learners was partially supported. The native speakers consistently performed better in SRCs compared to ORCs across all the online and offline measures except in the metalinguistic knowledge test. It was found that the native speakers could not provide metalinguistic knowledge for either type of relative clause (i.e., in the task where they had to provide a reason for their answer). However, in the other two tasks of the metalinguistic knowledge test (i.e., decide whether the sentence matches the picture or not; correct the mismatched sentences to match the picture), they were able to respond correctly, and, moreover, they had higher accuracy scores in SRCs than ORCs.

The L2 learners generally scored higher in SRC items than ORC items in the offline

comprehension test, the SPR test, and the oral production test. In the eye-tracking test, only the SRCs with animate heads were easier than the ORCs. That is, the L2 learners fixated on the target pictures earlier in SRC-A compared to SRC-I, ORC-A and ORC-I items. In the metalinguistic knowledge test, the L2 learners were not able to provide metalinguistic knowledge for either structure, similar to the native speakers. For the other two tasks of the metalinguistic knowledge test (decide about a match or mismatch; correct the sentence to match the picture), the L2 learners scored higher in both types of relative clauses with animate heads. In these two tasks, the accuracy scores of SRCs were not significantly higher than those of ORCs.

5.1.1.2 Animacy in relative clause processing and production

In terms of the influence of animacy in processing and producing relative clauses, ORCs with inanimate heads were predicted to be easier than the ORCs with animate heads for both native speakers and L2 learners. It was indeed found that the native speakers had higher accuracy scores in ORC-I items compared to ORC-A items in some of the tests, including the oral production and the metalinguistic knowledge tests (the deciding 'match or mismatch' task). However, it was only the difference in scores between ORC-I and ORC-A in the oral production test that was actually statistically significant. In fact, in the offline and the online *comprehension* measures (i.e., SPR, eye-tracking), there seemed to be a numerical trend towards the opposite finding (i.e. higher scores, more sensitive to the mismatch and earlier fixations for ORC-A relative to ORC-I), though neither trend was statistically significant.

For the L2 learners, contrary to the prediction, they did not perform better in ORC-I relative to ORC-A. Instead, they scored higher in ORC-A relative to ORC-I across all the measures except for the eye-tracking and the oral production tests, where the performance for ORC-A and ORC-I was roughly equal.

These findings will now be discussed in the light of previous studies.

5.1.2 Discussion of asymmetry between SRC and ORC in L1 and L2 processing and production

It was observed that in general, SRCs were easier to comprehend and produce than

ORCs for both NSs and L2 learners. This finding was in line with previous studies (e.g. Diessel & Tomasello, 2005; Keenan & Comrie, 1977; Kim & O'Grady, 2016; Traxler, Morris & Seely, 2002).

5.1.2.1 Offline comprehension

There were two tests that were used to measure offline comprehension; the aural sentence-picture matching and the first task (the task of deciding whether the sentence matched the picture) of the metalinguistic knowledge test. For both offline comprehension measures, the NSs scored at ceiling for SRCs (in aural sentence-picture matching test: mean [SD] $_{SRC\text{-}A}$ = .94 [.25], mean [SD] $_{SRC\text{-}I}$ = .93 [.26]; in the first task of metalinguistic knowledge test (mismatched items): mean [SD] $_{SRC\text{-}A}$ = .93 [.26], mean [SD] $_{SRC\text{-}I}$ = .98 [.15]). Compared to SRCs, in both measures they scored lower in **one type** of ORCs (in the aural sentence-picture matching test: mean [SD] $_{ORC\text{-}A}$ = .91 [.29], **mean [SD] $_{ORC\text{-}I}$ = .84 [.37]**; in the first task of the metalinguistic knowledge test (mismatched items): **mean [SD] $_{ORC\text{-}A}$ = .81 [.40]**, mean [SD] $_{ORC\text{-}I}$ = .93 [.26]). This was statistically significant (i.e., for ORC-A in aural sentence-picture matching test and ORC-I in the first task of the metalinguistic knowledge test).

Similar findings were observed with L2 learners. In the aural sentence-picture matching test, the L2 learners also performed better in SRCs (mean [SD] $_{SRC\text{-}A}$ = .86 [.35], mean [SD] $_{SRC\text{-}I}$ = .86 [.34]) than **one type** of ORCs (mean [SD] $_{ORC\text{-}A}$ = .84 [.37], **mean [SD] $_{ORC\text{-}I}$ = .77 [.42]**). The advantage for SRCs was statistically significant. In addition, the same tendency also could be found in the first task of the metalinguistic knowledge test (for mismatched items: mean [SD] $_{SRC\text{-}A}$ = .84 [.37], mean [SD] $_{SRC\text{-}I}$ = .74 [.44], mean [SD] $_{ORC\text{-}A}$ = .80 [.40], **mean [SD] $_{ORC\text{-}I}$ = .61 [.49]**).

In sum, the current study demonstrated an asymmetry between SRC and ORC in offline comprehension. Although the NSs were expected to score at ceiling for all types of relative clauses, they still showed some deficits in ORCs.

5.1.2.2 Oral production

In the oral production test, the native speakers scored at ceiling for SRCs (mean [SD] $_{SRC\text{-}A}$ = .94 [.23], mean [SD] $_{SRC\text{-}I}$ = .93 [.31]), but had rather low scores in ORCs (mean [SD]

$_{ORC\text{-}A}$ = .44 [.50], mean [SD] $_{ORC\text{-}I}$ = .55 [.50]). The L2 learners had lower scores when producing relative clauses compared to comprehending them, regardless of the type of relative clauses. They could only correctly produce around half of the SRCs, and the rate of accuracy was lower for ORCs (for both types of ORCs: mean [SD] = .33 [.47]).

These findings from oral production suggest a striking advantage of SRCs over ORCs for both native speakers and L2 learners, which is in line with previous studies about relative clauses production (e.g., Diessel & Tomasello, 2005; Kim & O'Grady, 2016). Both native speakers and L2 learners tended to produce passive SRCs when they were expected to produce ORCs. In the designing of the test, in order to elicit ORCs rather than passive SRCs, the salient cues (i.e., description of each picture in simple sentences with active voice) were provided. However, the participants still made the endeavour to produce passive SRCs, and this accounted for 77% and 88% of the 'errors' for native speakers and L2 learners respectively in producing ORCs. To some extent, this preference could also demonstrate that SRCs are easier than ORCs, and the participants are more likely to insert a subject gap than an object gap when they produce relative clauses.

5.1.2.3 Metalinguistic knowledge

In the metalinguistic knowledge test, the second and third tasks asked the participants to correct the mismatched sentences to match the picture and explain why there was a mismatch. It was found that the asymmetry between SRCs and ORCs was not statistically significant, though numerically the native speakers could correct more SRC sentences (mean [SD] $_{SRC\text{-}A}$ = .75 [.44], mean [SD] $_{SRC\text{-}I}$ = .78 [.42]) than ORC sentences (mean [SD] $_{ORC\text{-}A}$ = .64 [.49], mean [SD] $_{ORC\text{-}I}$ = .68 [.47]). Contrary to expectations, the L2 learners, did not, numerically, show superior competence in correcting SRCs (mean [SD] $_{SRC\text{-}A}$ = .64 [.48], mean [SD] $_{SRC\text{-}I}$ = .59 [.49]) to ORCs (mean [SD] $_{ORC\text{-}A}$ = .72 [.45], mean [SD] $_{ORC\text{-}I}$ = .54 [.50]). In fact, they performed better in **one type** of ORC. These numerical findings for L2 learners were inconsistent with the idea that ORCs are more difficult than SRCs (as found by, e.g. Keenan & Comrie, 1977). However, our finding might just be a chance finding because the higher accuracy score of ORC-A compared to SRCs was

not in fact statistically significant. Thus, it was found that there was not a reliably, substantial difference between ORCs and SRCs in the metalinguistic knowledge test, which can neither confirm nor disconfirm previous research.

The findings of native speakers and L2 learners suggested that in the sentence correction task, ORC is no more difficult than SRC, and at least the advantage of SRC is not as salient as that in offline comprehension and production. This might be due to the nature of the task itself. In this task, the participants were required to move only one word to make the sentence match the picture. It is possible that the participants noticed that for every mismatched sentence (whether ORC or SRC), moving the verb of the clause forward or backward would achieve the match of the sentence and the picture. When they were aware of this pattern, they are potentially able to apply it to every mismatched item, which makes ORC seems to have a similar difficulty level with SRCs.

However, in the task of having to give a reason to explain their correction, neither the native speakers nor the L2 learners demonstrated that they had metalinguistic knowledge about either type of relative clause. The findings of NSs are in line with Green and Hecht (1992) that argued that NSs were could correct sentences without knowing the explicit rules. In addition, the findings of L2 learners are consistent with those of Kasprowicz and Marsden (2018) (though with a very different type of participant). They suggested that the learners did not have metalinguistic knowledge about the target structures before receiving the intervention.

5.1.2.4 Online comprehension

Online comprehension was measured by SPR and visual-world eye-tracking tests. The native speakers and L2 learners were more sensitive to the syntactic cues during reading and listening to SRCs compared to ORCs, though this pattern was not statistically significant.

Native speakers

In the SPR test, the native speakers did not show sensitivity to the syntactic cue as soon as the first critical word came. They started to become sensitive to the

sentence-picture anomaly at the third critical word (the third word after the disambiguating cue) for SRCs and **one type** of ORCs (e.g., SRC: *The cat that chases the **dog** is big*; ORC: *The cat that the dog **chases** is big*), while for **the other type** of ORCs, sensitivity was observed at the third critical word **and** the next word following the critical region (e.g., ORC: *The car that the bike hits **is** red*). In addition, at the third critical word, the RT differences between the mismatched items and the matched items was bigger in SRCs compared to ORCs. Notably, in SPR tests, a spill-over effect can happen. That is, the effects of sensitivity could be observed at the word *following* the word where sensitivity might be expected to appear (Keating & Jegerski, 2015). Hence, in the current study, sensitivity could occur as early as the second critical word (i.e., in SRCs '*the*'; in ORC *noun*). This finding is in line with those of Roberts and Liszka (2013) about detecting sensitivity to anomalies.

In the eye-tracking test, the native speakers demonstrated the ability of using syntactic cue in interpreting the meaning of sentences, which is line with previous studies (e.g., Altmann & Kamide, 1999; Chen, 2010; Hopp, 2015; Hopp & Lemmerth, 2018; Lew-Williams & Fernald, 2010). Relating to the processing difficulty of SRCs and ORCs, the native speakers started to fixate on the target pictures earlier in SRCs compared to ORCs. For SRCs, they could use the language cue to interpret the meaning of the sentence during the first critical words (i.e., the verb), and this time point was around 400 ms after the onset of the first critical word. On the other hand, for ORCs, the native speakers started to look at the targets between the end of the second critical word (i.e., the noun) and during the third critical word (i.e., the verb), between 600 ms and 800 ms after the onset of the first critical word. Thus, the native speakers started to use the language cue to process the meaning of SRCs 200 ms to 400 ms earlier than that of ORCs. However, it is usually assumed that launching an eye-movement to the intended region would take around 200 ms (Barr, 2008), and this time period could be as short as 100 ms (Altmann, 2011). Hence, the earliest time when the native speakers started to comprehend the meaning of the ORCs might be during the second critical word (i.e., the noun).

L2 Learners

In the SPR, the L2 learners started to become sensitive to the sentence-picture mismatch at the third critical word for SRCs and one type of ORCs, and the RT differences between mismatched items and matched items were bigger in SRCs compared to the ORC. Considering the possibility of the spill-over effect, sensitivity of the mismatch could take place one word before being observed. Thus, the L2 learners might start to be sensitive to the mismatch at the second critical word. In addition, notably, for one type of SRCs and one type of ORCs, the L2 learners were still sensitive to the mismatch after the critical words, as the RT differences increased at the next word following the critical region. For another type of ORCs, the L2 learners did not show sensitivity to the mismatch across the sentence. Therefore, in general, the L2 learners were more sensitive to mismatch in SRCs relative to ORCs.

For the eye-tracking test, the L2 learners could fixate on the target picture before the end of the first critical word for **one type** of SRCs, but for **the other type** of SRCs and ORCs, the L2 learners could only start to look at the target picture during the third critical word. To some extent, these results were consistent with the idea that SRCs were easier than ORCs in online comprehension.

The findings from the results of the L2 learners might indicate the modified Shallow Structure Hypothesis (SSH) (Clahsen & Felser, 2018). However, as the current study did not aim to direct compare the native speakers with L2 processing, it was just a tentative suggestion. The SSH states that compare to native speakers, L2 learners tend to rely more on the non-grammatical information (relative to grammar information) in sentence processing. Thus, L2 learners might be less sensitive to the syntactic cues and might have reduced ability in using the cues compare to native speakers. The current study found that L2 learners could use the syntactic cues (at least for some structures) to process relative clauses to some extent. The results of the SPR and the eye-tracking suggested that the L2 learners had a similar pattern of processing with the native speakers; the L2 learners were indeed *less sensitive* to the syntactic cues, but *not insensitive*. In the SPR test, both the native speakers and the L2 learners

started to be sensitive to the anomaly at the third critical word. Nevertheless, the native speakers recovered from the slower RTs in reading mismatched items relative to the matched items at the next word straight after the critical region (except **for one type** of ORCs), while the L2 learners continued reading the mismatched items slower than matched items at the next word after the critical region. For **one type** of SRCs and **one type of** ORCs, the L2 learners showed the biggest RT differences between the mismatched and matched items at the word *after* the critical region (after the third critical word), which indicated that the L2 learners might be more sensitive to the anomaly after the critical regions. This finding is in line with that of Fujita and Cunnings (2021) and Williams et al. (2001), which found that L2 learners could process sentences in the similar way with native speakers but L2 learners were perhaps slower that native speakers.

In the eye-tracking test, the L2 learners showed a similar pattern of eye-movements with the native speakers. The L2 learners could fixate on the target picture before the end of the critical regions for all the structures, though, for **one type** of SRCs and ORCs, the L2 learners started to correctly process sentences later than the native speakers. The findings demonstrated that the L2 learners showed they were able to use syntactic cues to interpret the meaning of the upcoming information, and the ability of using the cues was influenced by the structures (i.e., they showed the anticipatory behaviour in **one type** of SRCs but not in the other structures). This is broadly consistent with the findings of Hopp and Lemmerth (2016), Foucart and Frenck-Mestre (2011) and Weber and Paris (2004).

5.1.2.4 Summary

In general, the findings demonstrated that the SRCs are easier than ORCs in offline, online comprehension and oral production for both native speakers and L2 learners. For online comprehension, the L2 learners showed a similar processing pattern as that of the native speakers but were perhaps less efficient in using the syntactic cues compared to the native speakers in processing the sentences. However, neither of the groups had metalinguistic knowledge about any type of relative clauses.

5.1.3 Discussion of the influence of head animacy

The findings of the native speakers are partially in line with Macdonald et al. (2020) which suggested that ORCs with inanimate heads were likely to be easier than ORCs with animate heads, while SRCs were easier than ORCs regardless of the animacy of the head nouns. In the current study, as discussed in Section 5.1.2, the native speakers constantly performed better in SRCs relative to ORCs. However, the influence of animacy of head noun on the difficulty of ORCs varied across the tests. Only in the oral production test, the native speakers produced statistically significantly more ORC-I than ORC-A, and in other tests the difficulty of ORC-A and ORC-I was not statistically different. In fact, in the online measures and the offline comprehension test, the native speakers tended to have *better* performance in ORC-A compared to ORC-I. Four possible reasons might be used to explain the findings. First, the current study did not investigate the influence of the animacy of the *second* noun to the relative clauses processing (although the number of items that had animate or inanimate second noun was balanced). Previous studies found that ORCs with an inanimate first (head) noun and animate *second* noun are used more frequently than the ORCs with two animate heads (Fox & Thompson, 1990; Kidd et al., 2007; Reali & Christiansen, 2007); and Macdonald et al. (2020) also suggested that when the second noun was inanimate, the ORCs with inanimate heads were not easier than those with animate heads. Thus, in the current study, the influence of the animacy of the head noun might be small, because animacy of the second noun was mixed (the numbers of animate and inanimate second nouns were roughly equal). Second, for offline comprehension, because the native speakers scored at or near ceiling for all the structures, the lower accuracy scores in ORC-I relative to ORC-A might have happened by chance. Third, as suggested by Macdonald et al. (2020), the influence of head-animacy is bigger in offline comprehension than online comprehension. Thus, in the online measures in the current study, the earlier fixation and greater sensitivity to ORC-A compared to ORC-I might be a chance finding. Fourth, in the online tests, the participants were allowed to observe the picture or pictures and the verb used in the sentence before they heard (or

read) the sentence stimuli. This might decrease the influence of the animacy of the head noun, because the participants already knew the agent and patient before processing sentences.

For the L2 learners, these results suggested that the L2 learners in this study contradicted the tendency observed in previous studies. The L2 learners showed statistically significantly (or at least had reliable effect sizes) higher accuracy scores in offline comprehension and two tasks of metalinguistic knowledge (i.e., deciding whether the sentence match the picture, correct the sentence to match the picture). It is hard to explain why the L2 learners were likely to have better performance in ORC-A rather than ORC-I. One possible reason might be that the L2 learners did not commonly use ORCs in English, thus they were unfamiliar with ORCs as a whole, and they might just prefer the relative clauses with animate heads.

In sum, the findings suggest that in general, ORC-I is easier than ORC-A for native speakers especially in oral production, but not in online measures. On the other hand, it is found that the L2 learners had better performance in ORC-A rather than ORC-I in all measures.

5.2 To what extent can teaching parsing strategies (with explicit information and practice), exposure alone, or no exposure (tests alone), develop the learning of relative clauses?

5.2.1 Summary of findings

The expectation that teaching parsing strategies with practice would facilitate the learning of relative clauses was partially supported, and the effects could mainly be found in offline measures. The parsing group showed significant improvements after training in the offline comprehension, oral production and metalinguistic knowledge test and the effects were durable. However, in the online measures, the parsing group only had limited gains in the SPR test for SRC structures, and the gains could only be observed at the immediate post-test. In the eye-tracking test, the parsing group did not make a substantial improvement in the time spent on fixating on the target structures

after training.

In addition, the input flood group also made some improvements in the offline tests, though most of them were not statistically significant and their magnitude was less than that of the parsing group.

For the test-only group, very limited improvement could be found, except in the oral production test for ORC-A structure and in the offline comprehension test for ORC structures and one task in the metalinguistic knowledge test (i.e., correct the mismatched sentence to match the picture) where significant improvements were indeed observed.

These findings will be discussed in light of the previous studies in the following sections.

5.2.2 Effects of training on offline tests

The findings revealed that teaching parsing strategies with practice facilitated offline comprehension, oral production, and metalinguistic knowledge of relative clauses, and in most cases, the gains of the parsing group were significantly more than those of the two other groups. These findings were in line with the previous studies which had demonstrated the effects of explicit training, like processing instruction (PI), on offline L2 learning outcomes (e.g., Andringa & Curcic, 2015; Benati, 2005; Kasprowicz & Marsden, 2018; Marsden, 2006; Marsden & Chen, 2011; VanPatten & Wong, 2004).

In addition, the input flood group and test-only group also showed some improvements over time. However, in most of cases, the gains of the input flood group were not significantly more than those of the test-only group. Because the current study had a large test battery, the participants had the opportunity to learn the target structures from simply taking part in the tests. Test effects could indeed be observed in the current study, and for the groups that did make gains, at least *some* test effect could be the cause – that is, taking the test could interact with the intervention. The results indicated that the input flood training was not as effective as the training of parsing strategies, and the improvements of the input flood group might be due to the test effects. This finding was in line with Marsden and Chen (2011). In their study, the

group which received affective activities (similar to input flood training) did not gain in offline comprehension. However, the finding was inconsistent with that of Kasprowicz and Marsden (2018) who found that the input enhancement training had equivalent effects with PI in learning German case-marking. The different findings might be due to Kasprowicz and Marsden (2018) provided EI prior to the input enhancement training and asked the learners to notice (spot) the form each time (by clicking on the article on the screen), but the current study did not. The difference between the current finding and that of Kasprowicz and Marsden finding could provide evidence supporting the role of EI and explicit noticing practice during input flood training.

In the following sections, the findings of each offline test will be discussed in detail.

5.2.2.1 Effects of training on offline comprehension

In the offline comprehension test, all the three groups showed some improvements across time. For SRCs, the gains in accuracy scores of the parsing group were statistically more than those of the input flood (at the post-test for SRC-A) and of the test-only group (at the delayed post-test for SRC-A and SRC-I). For ORCs, the improvements of the parsing group at the post- and the delayed post-tests were extremely small, but reliable (shown by within-group effect sizes). However, the gains of the three groups in ORCs were not statistically different between each pair of groups. Thus, the advantage of teaching parsing strategies with practice over the input flood training and no training was limited and was shown only in SRCs.

In addition, the results also revealed that the input flood group did not significantly outperform the test-only group at the post- and delayed post-test compared to the pre-test for all the structures, which indicated the input flood training might not contribute to the development of offline comprehension.

Two possible reasons might be able to explain the limited effects of teaching parsing strategies. First, the three groups already scored near ceiling at the pre-test, so the room for improvement was limited. For ORC-A structure, although the parsing group did not significantly outperform the other two groups, they already scored at

ceiling at the post- (mean [SD] = .93 [.25]) and then also at the delayed post-test (mean [SD] = .96 [.19]). Second, for SRCs, although the inferential statistics showed the parsing group had more gains than the other two groups at post- or delayed post-test compared to the pre-test, the numerical score differences between groups were relatively small at every test phase (i.e., mean difference no more than .06). Hence, the statistically significant effects might not have too much practical meaning. The three groups were likely to have similar amount of gains over time.

Given that the first task of the metalinguistic knowledge test (decide whether the sentence matched the picture) also measured offline comprehension, the results of this task are also discussed here. For the matched items, all the groups scored at ceiling at the pre-test, so no substantial change of accuracy scores could be observed across the time. For the mismatched items, all the groups showed some improvements at the post- and the delayed post-test compared to the pre-test. The statistical results indicated that the parsing and the test-only group significantly had more improvements in SRC-I and ORC-I compared to the input flood group. In SRC-A and ORC-A, the improvements of the three groups did not have statistically significant difference. The findings suggest that the participants were able to gain in accuracy through taking part in the tests. In addition, the greater gains of the test-only group relative to the input flood group in SRC-I and ORC-I might be due to the lower accuracy scores of the test-only group at the pre-test and the fact that, the two groups then scored similarly at the post- and the delayed post-test. The low score of the test-only group at pre-test might a chance finding, because in most cases – across the other tests in the battery – the three groups performed similarly at the pre-test.

In sum, teaching parsing strategies with practice facilitated offline comprehension, but the effects were limited. It seems that the test itself also developed offline comprehension. This may be because the participants involved in the current study had upper-intermediate language proficiency, and they scored relatively high at the pre-test. Thus, being more familiar to the targets through attending tests might be sufficient for the learners to achieve at ceiling at post- and the delayed post-test.

5.2.2.2 Effects of training on oral production

In oral production test, all the three groups showed some improvement in the accuracy scores at the post- and the delayed post-test for all the structures. The gains of the parsing group were statistically more than those of the other two groups, especially for ORCs, and the gains of the other two groups were not statistically significant. The findings demonstrated striking effects of teaching parsing strategies on the oral production of relative clauses.

In terms of the greater gains in ORCs than in SRCs, two possible reasons might be used to explain this asymmetry. First, as illustrated in Section 5.1.2.2, the reason that accounted for 88% of the 'mistakes' in ORCs was that the participants tended to produce a passive SRC when they were expected to generate an ORC sentence. Note that during training, the parsing group was provided with EI that introduced ORCs at the beginning of their practice sessions, and they were forced to comprehend the meaning of the ORC sentences in the practice activities. Coumel et al. (2020) mentioned that people tended to use "a previously-dispreferred syntactic structure" in production after being exposed to a large amount of this structure, and this is known as syntactic priming (p. 3). In the current study, such priming effects could perhaps explain the observations from the higher scores among the parsing group after their training. They started to abandon the passive SRCs and started to produce ORCs at the post- and the delayed post-test. Thus, striking improvements in ORCs were observed. Second, at the pre-test, the scores of the SRCs were much higher than those of ORCs. It means that the room for improvement was bigger in ORCs relative to SRCs.

In addition, the participants who received input flood training did not have significantly more gains than the participants who only undertook the tests. This finding is in line with Marsden and Chen (2011) and is partially in line with Marsden (2006). Marsden and Chen (2011) found that exposure to the targets alone did not develop either comprehension or production of the target form. Marsden (2006) found that learners that had rather low proficiency could not benefit from enriched input. However, for the learners with a higher initial proficiency, Marsden found that the

enriched input group *could* make similar gains in oral and written production relative to the PI (that is, the + EI + input practice) group. The difference of the current study and Marsden (2006) was that Marsden (2006) provided EI prior to the input activities while the current study did not. This could suggest that L2 learners might not be able to develop production of target structures through mere exposure to the targets.

Moreover, it should be mentioned that for ORC-A, the test-only group even statisitically gained more than the input flood group at the post-test relative to the pre-test. Since the gains of the test-only group were indeed observed, at least some test effects might have happened. It might suggest that taking part into the tests could develop production of relative clauses. However, the more gains of the test-only group than the input flood group in ORC-A might just due to the test-only group had lower pre-test scores (mean [SD] = .21 [.41]) than the input flood group (mean [SD] = .41 [.49]). It is unknown whether the low pre-test scores of the test-only group were attributed to the lack of ability in producing ORC-A, or just happened by chance. It is possible that the more improvement of the test-only group just because the test-only group accidentally scored very low at the pre-test, and scored normally at the post-test. Thus, it is difficult to know 'how much' test effects occurred.

5.2.2.3 Effects of training on metalinguistic knowledge

In the metalinguistic knowledge test, the parsing and the test-only group showed almost identical improvements at the post- and the delayed post-test compared to the pre-test in the task of sentence correction. The input flood group did not show significant gains across the time in this task.

However, in the task of reason explanation, the three groups all scored at floor at the pre-test. Only the parsing group showed improvements in providing metalinguistic knowledge at the post- and the delayed post-test, though slight decline could be observed at the delayed post-test compared to the post-test.

The findings suggested that teaching parsing strategies could facilitate the competence in sentence correction, and the learners were also able to gain this competence through the test itself. However, the finding that the test-only group

statistically had greater gains than the input flood group was contrary to expectations. This finding could be attributed to two possible reasons. First, for SRCs, the input flood group had similar accuracy scores as the test only group at the post- and the delayed post-test (see table 4.2.31). However, the input flood group scored much higher than the test-only group at the pre-test (the difference between the two groups even had a small but meaning and reliable effect size SRC-A: *d* = .40 [.00, .79], SRC-I: *d* = .50 [.10, .90]). This probably led to the significantly better gains from the test-only group compared to the input flood group for the SRCs. Second, the characteristics of the task might account for the finding. The metalinguistic knowledge test only had two items for each type of relative clause, and the way of correcting the sentence was to move the verb before or after the noun. The significant gains of the test-only group might be due to individual differences, which could not be measured in the current study. Some participants of the test-only group may have just known or worked out (though better language analytic ability, see Kasprowicz, Marsden & Sephton, 2019) the rule of sentence correction, so the improvements could be observed.

The finding that only the parsing group made improvements in providing metalinguistic knowledge at the post- and delayed post-test was in line with Kasprowicz and Marsden (2018) who found that the explicit instruction benefited metalinguistic awareness. In the current study, only the parsing group received explicit information about the target structures, and the results suggested that some of them indeed learned the knowledge from the training. However, although the parsing group had significant gains at the post- and the delayed post-test, the mean scores of this task were still rather low for all the structures (lower than .50, see table 4.2.38). The results support the finding of Alderson et al. (1997), who found that the relation between the language proficiency and metalinguistic knowledge can be weak. In the current study, the parsing group scored at or near ceiling at the post- and the delayed post-test in offline comprehension, production, and sentence correction, but only a few of participants demonstrated they had metalinguistic knowledge to the target forms. In addition, it was found that metalinguistic knowledge decayed, as

improvements showed at the post-test decreased at the delayed post-test. This is because remembering explicit rules can place heavy demands on working memory (R. Ellis, 2009), and not all learners are equipped to understand, retain and use them (at least not without a lot of practice) (see e.g., DeKeyser, 2015, 2017). As in the current study, the parsing strategies training only involved brief introduction to the EI and only have about one-hour practice in total, the learners are likely to forget the rules as time goes by.

5.2.3 Effects of training on online tests

In the current study, SPR and visual world eye-tracking tests were used to measure online comprehension. The focuses of the two online tests were different. The SPR tests investigated whether the participants were sensitive to the language cues, while the eye-tracking tests explored the extent to which (and when) participants could use the cue to interpret the meaning of the sentence. Using the cue in the eye-tracking test could either be done by retrospectively assigning part of speech to the words, already heard, and/or by using the cue to *predict* upcoming language – both work in order to assign sentential meaning to select the appropriate picture.

In the SPR tests, the results indicated that the parsing group had some limited improvements in SRCs (reflected by the effect sizes), but the improvements were not durable. It seemed that the parsing group was not sensitive to the mismatch between the sentence and picture during the critical regions at the pre-test, but they showed reliable sensitivity to the mismatch at the third critical word at the post-test. However, sensitivity was not observed at the delayed post-test. The other two groups did not show significant improvements across the time. This finding is inconsistent with those of both Dracos and Henry (2021) and McManus and Marsden (2017). Dracos and Henry (2021) did not find evidence supporting the idea that task-essential training could facilitate sensitivity to violations in Spanish verbal inflections. In McManus and Marsden (2017), the group that received L2 + L1 EI (with task-essential practice) increased their sensitivity to the violation of French imperfect tense inflections at the post- and the delayed post-test, but the group that *only* received L2 explicit

information (with task-essential practice) – which is more like the training in the *current study* – did not show improvement in the online test (i.e., SPR test). This is in contrast to the current study, in which gains in sensitivity of the sentence-picture mismatch in SRCs were indeed observed, though the improvements were rather small.

In addition, the finding that the parsing training improved sensitivity to the mismatch at the post-test for SRCs was partially in line with the VanPatten and Smith (2019). VanPatten and Smith (2019) found that native English speakers could be trained to become sensitive to case-marking cues in Latin SOV sentences but not in SVO sentences. Thus, they claimed that whether sensitivity to the language cues could be trained was influenced by the type of the specific structure. The current study broadly aligns with this general point, as the training effects were observed in, specifically, SRCs and not ORCs. However, VanPatten and Smith (2019) found that the *non-canonical* word order sentences were more likely to be sensitive to training than the canonical ones, which contrasts with the current study's finding of the opposite tendency (the canonical structure, SRC, showed sensitivity, whereas the non-canonical ORC did not). One possible reason might be that the L2 proficiency of the participants was different between the two studies. The participants in VanPatten and Smith (2019) had no existing knowledge to the target language. Although their participants showed sensitivity to the language cue, they could not use the cue in comprehension. On the other hand, the participants in the current study were regarded to have upper-intermediate L2 proficiency, and the offline tests demonstrated they could score at or near ceiling in offline comprehension and in production of the target structures. Thus, the improvements might be more likely to be observed in the easier structure (i.e., SRC) compared to the more difficult structure (i.e., ORC).

Moreover, one unexpected phenomenon was found with the test-only group. Reflected by the effect sizes, the test-only group showed reliable sensitivity to the mismatch at the second and/or third critical word at the pre-test. However, they lost sensitivity at the post- and the delayed post-test. Two possible reasons might account for this phenomenon. First, sensitivity presenting at the pre-test might just have been

by chance. Second, the SPR tests did not instruct the participants to decide whether the sentence matched the picture or not, and the comprehension questions were based on the meaning of the sentences. Some participants might assume what they needed to do in the test was correctly respond to the comprehension questions. In order to reduce the influence of the mismatched pictures (so they could focus attention on the comprehension questions), they may have tried to *not* look at the pictures and orient their attentions on the sentences. However, it is not easy to explain why this only happened to the test-only group. It might be because this group did not attend to any training between tests, so they were more likely to guess the purpose of the study.

In the eye-tracking tests, the plots indicated all three groups fixated on the targets earlier at the post- and the delayed post-test compared to at the pre-test for all the structures except SRC-A, though the improvements were not statistically significant. For the SRC-A, at the pre-test, the participants looked at the target pictures during the first critical word, so no improvement could be observed. The improvements in the point at which the targets were fixated might be due to the participants being more familiar with the test items, and so test effects were generated. Neither the parsing strategies (EI with practice) nor the input flood training seemed to benefit the participants in using the syntactic cue to interpret the meaning of the sentence.

This finding was in line with Andriga and Curcic (2015), who found that teaching explicit knowledge could not facilitate online processing (measured by visual world eye-tracking test) of direct object assignment in an artificial language. However, some studies found evidence supporting the effects of explicit training on online comprehension through eye-tracking test. Hopp (2016) suggested that intermediate L2 German learners whose first language was English showed anticipatory behaviour in determining grammatical gender of nouns after receiving training activities on the specific determiner – noun sequences. In addition, in Wong and Ito (2018), the intermediate L2 learners of French who received PI changed eye-movement pattern in processing French causative structures, though the participants still could not process

the sentences in the target way. The main reason that the finding of the current study was inconsistent with Hopp (2016) and Wong and Ito (2018) might be because the target structure was different. The structures (i.e., German grammatical gender, French causative structure) used in the previous studies were based on morphosyntax, but the current study focused on the effects of explicit training on syntax (i.e., word order for English relative clauses). The mechanisms involved in or the burden of processing syntax and morphology might be different, and/or training effects may be more difficult to observe in online measures (i.e., eye-tracking test) when processing syntax relative to morphology.

In sum, the findings suggest that teaching parsing strategies with practice had very little effect on online processing of relative clauses; the limited effects observed could only be reliably recorded in SPR tests for SRCs. In the eye-tracking test, the parsing group did not have significantly more gains than the other two groups in using the cue to predict or retrospectively assign the meaning of two nouns in the sentence, even though they were forced to use the cue whilst comprehending (i.e., *during* a sentential parse) the meaning of relative clauses during the training. Comparing the findings of the two online measures, it was perhaps that training effects were slightly easier to be observed in sensitivity to the syntactic cues relative to using the cue to predictively or retrospectively interpret the meaning of the sentence (eye-tracking). It might be because the parsing group received EI about the target structures, so they started to show sensitivity to the cue. However, more training (more than two 30-minute sessions) might be needed to develop the ability to use the cue in the online processing (and yet, the high scores on the offline, production, and metalinguistic knowledge tests suggest that the learners did have some reliable representations or knowledge about the target structure). In terms of the input flood group, no training effects could be found in either test. This is partially in line with Issa and Morgan-Short (2019). They found that input enhancement training did not contribute to the *learning* of Spanish direct-object pronouns, though it helped the learners to allocate attention to the target forms.

5.2.4 Summary of effects of training

The current study demonstrated the effects of teaching parsing strategies on offline comprehension, production and metalinguistic knowledge of English relative clauses. However, it barely influenced the online processing of the target structures. The test-only group also showed significant gains in the sentence correction task of the metalinguistic knowledge test. It is not completely clear why this would be particularly the case for the test-only group gained more than the input flood group, which may be due to the individual differences. The input flood training did not seem to benefit the online processing, and offline comprehension, production, or metalinguistic knowledge of either type of relative clause, which might be due to the fact that no EI was provided before the training practice.

5.3 General discussion

The implication of the findings for syntax learning, explicit and implicit processing, and L1 influence on L2 learning will be discussed respectively in the following sections.

5.3.1 Implication for syntax learning

The current study investigated the effectiveness of a type of innovative instruction, teaching parsing strategies, on learning English syntax (i.e., English relative clauses). This type of instruction originated from the referential activity of PI, but was different from it. PI aims to push learners away from incorrect processing strategies by training them to connect forms to the meanings in real world (VanPatten 2005, 2015; VanPatten & Cadierno, 1993a, 1993b). Thus, the target structures used in PI studies were morphosyntax which had referential meanings (see for a small collection: Kasprowicz & Marsden, 2018; Marsden, 2006; Marsden & Chen, 2011; VanPatten & Cadierno, 1993a; Wong & Ito, 2018). The parsing strategies instruction used in the current study attempted to train learners to use part of speech (i.e., word order in relative clauses) to process abstract syntax. During the training, the stimuli provided to the learners were stopped after the syntactic cue, which forced the learners to use the cue to interpret the meaning. To the best of our knowledge, only the current study and an unpublished PhD thesis (Thompson-Lee, 2021) investigated this type of instruction, but Thompson-Lee (2021) still

focused on its effects on learning morphosyntax (i.e., English passive voice).

Previous studies have demonstrated that PI facilitates offline comprehension and production (e.g., Kasprowicz & Marsden, 2018; Marsden, 2006; Marsden & Chen, 2011; VanPatten & Cadierno, 1993a; Wong & Ito, 2018); yet, the effects of PI and other explicit training on online processing of morphosyntax were controversial. Some studies found the evidence supporting explicit training develops online processing to some extent (Hopp, 2016; McManus & Marsden, 2017; VanPatten & Smith, 2019; Wong & Ito, 2018), while a few studies indicated that online processing might not be altered via training (Andringa & Curcic, 2015; Dracos & Henry, 2021). The occurrence of the inconsistent results might be due to the difference of linguistic structures, language backgrounds, L2 proficiency, and so on. However, to the best of our knowledge, the literature related to the effects of explicit training on syntax learning is blank. The current study attempted to fill this knowledge gap. It was found that teaching parsing strategies promoted offline comprehension and metalinguistic knowledge, and promoted oral production to a large extent; nevertheless, the effects on online processing were limited. The learners became more sensitive to the syntactic cue (measured by SPR test) after receiving the training session. However, the gains could only be observed at the post-test for SRCs, and the gains were lost at the delayed post-test. It seems that small amount of parsing strategies training might be able to trigger the sensitivity to the syntactic cue for easier structures (i.e., SRCs), but the learners still could not use it in real-time sentence processing (measured by the eye-tracking test). Admittedly, the training sessions in the current study were rather short (lasting for around one hour in total). If more training items were provided, the more robust online effects might be observed.

The current study made the first attempt to explore the effects of teaching parsing strategies (using the word order to assign the roles of two nouns in relative clauses) on syntax learning. Further studies in this agenda are worth to be conducted. This type of explicit training might be used to study other syntactic phenomena (e.g., English *wh*-questions). For example, in subject extraction (e.g., *Who did Ann say likes her friend*?, citied from Juffs & Harrington, 1995), after the verb (*say*), there is a verb + noun phrase

construction; in object extraction (e.g., *Which man did Jane say her friend likes*?, citied from Juffs & Harrington, 1995), noun phrase + verb construction follows the verb (*say*). In other words, the word order after the verb could assign the roles of the two nouns in a *wh*-question. In addition, in embedded *wh*-question (e.g., *What do you think the man is eating?*), the word order of the question is noun + verb, while in non-embedded question (e.g., *What is the man eating*?), the question has a verb + noun word order. Therefore, the effects of teaching parsing strategies on syntax learning could be further studied by using those structures as the target forms.

5.3.2 Implication for explicit and implicit processing in the L2

The current study confirmed that small amount (one hour in total) of parsing strategies training could develop explicit comprehension (represented by the results of offline comprehension and the deciding match or mismatch test of the metalinguistic knowledge test), which was in line with previous studies (see a small collection, Andringa & Curcic, 2015; Marsden, 2006; McManus & Marsden, 2017; VanPatten & Cardiano, 1993a).

In addition, the results suggested that the parsing group had some limited gains in online processing (measured by the SPR tests) for the SRCs. It is generally agreed that online techniques (e.g., eye-tracking and SPR) tap into implicit knowledge, as they measure the real-time comprehension and allow little or no time for using explicit knowledge (Keating & Jegerski, 2015). Thus, the finding might indicate that teaching parsing strategies was potentially able to develop implicit processing of the L2 syntax, and the effects were more likely to be observed with easier structures (i.e., SRCs) relative to the more difficult ones (i.e., ORCs). However, it is should be reiterated that the training sessions of the current study were rather brief, and the learners might not establish automatization of the knowledge of the more difficult structure (i.e., ORCs). McManus and Marsden (2017, 2018, 2019) provided L2 learners 3.5 hours explicit training across four weeks, and observed the substantial gains in online comprehension (measured by SPR). MacManus and Marsden (2019) found that the learners showed automatization of the knowledge after 1.5 hours training. It is possible that increasing

the amount of training, the improvements in implicit processing might be shown in ORCs as well as SRCs, and might be in the eye-tracking tests. However, how much training could general more robust effects on implicit processing still need further study.

The results of the input flood group indicated that exposure only to the target structure (incidental learning) benefit neither explicit nor implicit processing. Although the input flood group showed some numerical gains in offline comprehension, they were not statistically significant. In addition, the gains of the input flood group were not significantly different from those of the test-only group. Thus, the improvement of the two groups might just due to the test effects. In the current study, the aim of including input flood training was to examine whether the learners could pick up the language and gain from exposure only. Thus, the learners were not required to understand the meaning of the sentence. It seems that the less effectiveness of the input flood training was because the training material did not trigger the learner's noticing to the target structures. It might be because the structures were not salient enough to be noticed (Hernández, 2018), the amount of the training stimuli were not sufficient (Uchihara et al., 2019), or the learners' L2 proficiency was not high enough to pick up the language (Hernández, 2008, 2011; Marsden, 2006). To make the input flood training more effective in L2 learning, besides increasing the training items, EI of the target structures could be provided prior to the training activities to help the learners to induce noticing. In addition, the learners could be required to spot the target structure to raise their awareness. Kasprowicz and Marsden (2018) has combined these two methods (i.e., pre-practice EI and spotting the forms) with the input flood training, and found that the 'improved input-flood' training was equally effective as the 'form-meaning connection' training (EI + referential activities of PI). Furthermore, in the future study, it might be worth investigating that whether requiring learners to understand the meaning of the sentence (by providing them complete sentences used in the parsing strategies group) could have identical effects to training them to use syntactic cues to process. This type of training is similar to the referential activities of PI, and the learners might be able to notice the

structures from the activities and pick up the knowledge inductively.

5.3.3 The influence of learners' L1

The current study was conducted with the Chinese-speaking L2 learners of English and native English speakers. The structure of Chinese relative clauses is different from English, since English relative clauses are head-initial while those of Chinese are head-final. However, the findings suggested that the L2 learners processed English relative clauses in the similar way to that of the native English speakers, though they were less sensitive to the syntactic cues and had reduced ability in using them relative to the native speakers. Thus, it seems that the L1 transfer did not have substantial influence on L2 processing.

However, as the current study did not test the cross-linguistic effects on L2 learning, no participant from L1 background other than Chinese was involved. Nevertheless, it is possible that the learners whose L1 shares the grammar features with L2 might have more gains than the learners whose L1 differ from the L2. This prediction is made based on Tolentino and Tokowicz (2014) which investigated the influence of language similarity on L2 instruction. They found that when the grammar features of the two languages were similar, the learners improved more than in the condition where the grammar features were dissimilar. The gains of the similar features were less likely to be influenced by the instruction type relative to the dissimilar features. In addition, their study also found that compared to the dissimilar features, the similar features were more likely to be gained from simply exposure to the targets or exposure of targets with highlighting differences. They explained it might be because for similar features, learners L1 could be positively transferred to their L2 (Tolentino & Tokowicz, 2014). In light of Tolentino and Tokowicz (2014), it could be hypothesized that in the current study, if the learners' L1 had similar relative clause structures to English (e.g., French relative clause is fundamentally identical to English relative clause, Labelle, 1990), they might generally have more gains than Chinese-speaking L2 English learners, especially in the input flood training condition.

Chapter 6 Conclusion

6.1 Summary of the study

This thesis has presented the findings of a study that investigated the effects of teaching parsing strategies (with EI and practice) on learning English relative clauses. 79 upper-intermediate Chinese-speaking L2 learners of English took part in the study. The training effects were examined through a pre-, one-week post-, and three-week delayed post-test design. In addition, 21 native English speakers were also involved to test how native speakers comprehend and produce relative clauses. The native speakers performed in a way that was generally expected, which suggest that the test measures used in the study worked.

The L2 learners were randomly assigned into the parsing group, the input flood group and the test-only group. Between the pre- and the post-test, the parsing group (EI with practice about parsing SRCs and ORCs whilst they were reading or hearing them) and the input flood group (exposure to the targets) received two training sessions (around 30 minutes per session). The activities in the parsing strategies training forced the learners to use the syntactic cue, the word order after the relative pronoun, to comprehend relative clauses. In SRCs (e.g., *The cat that **chases** the dog is big*), a verb is straight after the relative pronoun, while in ORCs (e.g., *The cat that **the dog** chases is big*), a noun phrase occurs after the relative pronoun. In the activities, the words after the verb (in SRC) or the noun phrase (in ORC) was not provided. Thus, the participants needed to anticipate the rest of the sentence based on the syntactic cue. The input flood group received the exact same number of training items as the parsing group, but no EI was provided. In the input flood training activities, the complete sentences were presented, and the participants were not required to comprehend the meaning of the full sentences, but had to make a decision about the meaning of the nouns in the sentences.

Five measures were included in the test battery. The aural sentence-picture matching test was used to test the offline comprehension; the SPR and visual world

eye-tracking test were utilised to measure online comprehension; oral production was examined through a picture description task; and metalinguistic knowledge was also tested in a type of picture-based grammaticality judgement and anomaly correction test.

6.2 Summary of the findings

With regard to the first research question, which type of relative clause is more difficult, the results generally confirmed the asymmetry between SRC and ORC in L1 and L2 online and offline comprehension and production. This is in line with the hierarchy put forward by Keenan and Comrie (1977) that the SRC is easier than the ORC. The finding is also in line the previous studies (e.g. Diessel & Tomasello, 2005; Kim & O'Grady, 2016; Traxler, Morris & Seely, 2002). In addition, in the findings of the native speakers, there are three noteworthy features. First, the native speakers were potentially sensitive to the syntactic cue in SPR test, and sensitivity was observed only at the third critical word for SRCs and for **one type** of ORCs (for another type of ORC, sensitivity was observed at the word after the critical regions). Notably, because a spill-over effect could take place, the native speaker might already be sensitive to the mismatch at the second critical word. Second, in the eye-tracking test, native speakers demonstrated that they could use the syntactic cue to interpret the meaning (predict and/or retrospectively assign part of speech) of the sentence during the first critical word for SRCs, while for ORCs, the most reliable eye movements took place later compared to the SRCs (i.e., at the second or the third critical word). Third, the native speakers did not have metalinguistic knowledge about either type of relative clause, which is consistent with Green and Hecht (1992).

In terms of the influence of head animacy on the difficulty of relative clause processing, previous studies had suggested that the ORCs with inanimate heads tended to be easier than those with animate heads (Kidd et al., 2007; Macdonald et al., 2020; Traxler et al., 2005), but the animacy of the head noun could not alter the asymmetry between SRC and ORC (Macdonald et al., 2020). The finding of the current study demonstrated that ORCs are always more difficult for both native speakers and L2

learners in all the measures. In addition, the findings with the native speakers partially support the pattern that ORC-I is easier than ORC-A. It was found that the native speakers significantly *produced* more ORC-I than ORC-A in the oral production test. In other tests, however, the difference between ORC-A and ORC-I was not significantly different. However, the L2 learners did not show any reliable preference to the ORCs with inanimate heads; instead, in fact, they had better performance in the ORC-A items compared to the ORC-I in all the measures.

With regard to the second research question, the extent to which teaching parsing strategies with explicit information and practice, exposure alone, and test-only develops the learning of relative clauses, the findings have demonstrated the effects of teaching parsing strategies on offline comprehension, production and metalinguistic knowledge. This is in line with the previous studies that are related to explicit training, like PI (e.g., Andringa & Curcic, 2015; Benati, 2005; Kasprowicz & Marsden, 2018; Marsden, 2006; Marsden & Chen, 2011; VanPatten & Wong, 2004). However, the current findings suggested that although teaching parsing strategies could facilitate offline comprehension, the participants also had the opportunity to gain from the test alone.

The improvements of the parsing group in oral production were striking, and the gains in this measure in ORCs were bigger than those of SRCs. At pre-test, the participants preferred to produce passive SRCs when they were expected to produce ORCs. After the training, the parsing group started to produce the target ORCs at the post- and the delayed post-test. This may have demonstrated the occurrence of priming effects (Coumel et al., 2020). In addition, the parsing group gained metalinguistic knowledge about the target structure at the post- and the delayed post-test, likely due to the provision of EI during training. This finding is in line with Kasprowicz and Marsden (2018).

However, the findings indicate that the effects of teaching parsing strategies on online processing were limited, which is partially consistent with Andringa and Curcic (2015) as well as VanPatten and Smith (2019). The parsing group showed some gains in

sensitivity to the syntactic cue (measured by SPR) at the post-test, but the gains only could be observed for SRCs. Also, the gains were not durable, as they were not maintained at the delayed post-test. Moreover, in the eye-tracking test, the parsing group did not have significant more improvements in the point at which they started to fixate on the targets compared to the input flood and the test-only group. The findings may tentatively suggest that improvements in sensitivity to the syntactic cue might be easier to be observed (in SPRs, as found by McManus & Marsden, 2017, who also observed effects of training) relative to using the cue to interpret the meaning of the sentence in online processing (as used by Andringa & Curcic, 2015, who did not find reliable effects of training).

The results of the input flood group suggested that exposure to the target alone could not contribute to the learning of relative clauses. The input flood group did not show significant improvements across time in online or offline comprehension, oral production or metalinguistic knowledge. This is in line with Marsden and Chen (2011) and partially consistent with Issa and Morgan-Short (2019).

In addition, some test effects were found in the current study (i.e., in the sentence correction task of the metalinguistic knowledge test), which might be because a large test battery was involved. Thus, the participants might have gained from taking part in the tests alone, which may have been due to them processing some prior knowledge and/or ability to use the target feature).

6.3 Limitations and future research

There are several possible limitations of the current study that should be identified. First, the duration of the training sessions was relatively short. For both the parsing and the input flood group, only two 30-minute (approximately) training sessions were provided, which was shorter than some studies. For example, in McManus and Marsden (2017), four 45-minute training sessions were involved. In the current study, the training effects on sensitivity to the syntactic cue were observed for SRCs in SPR tests. It is possible that if longer training sessions were provided, more online effects might be able to be found in the SPR and the eye-tracking tests.

A second limitation relates to the control of animacy of two nouns in the relative clauses. In the current study, only the animacy of the head noun was controlled (experimentally manipulated). For the second noun in the relative clause, the animacy of nouns was balanced but was not experimentally manipulated. However, Macdonald et al. (2020) suggested that only the ORCs with inanimate head nouns *and* animate second nouns tended to be easier than the ORCs with animate heads. In the current study, the L2 learners were found to have better performance in ORC-A relative to ORC-I, which was different from previous findings, but if the animacy of the second noun had been manipulated, a different asymmetry might have been found.

A third limitation is about balancing the items of each target structure in two of the tests. In the offline comprehension and the oral production tests, the numbers of each structure within a version (all four versions were affected) of a test were different. Although the versions were counterbalanced across pre-, post-, and delayed post-test, within groups, the results would be more accurate if the items within a version of a test were equal.

A fourth limitation relates to the data cleaning of the eye-tracking test. Only very few visual world eye-tracking studies have reported how they cleaned the data before analysis. For instance, Altmann and Kamide (2007) removed the fixation durations that were below 100ms. In order to avoid the influence of overly long fixations, the current study removed the fixation durations that were below and over the boundary of mean ± 2.5 SD, and also adopted the lower boundary of 50ms. This is in line with the general suggestions for *reading-based* eye-tracking (Godfroid, 2019). However, whether the way of cleaning visual world eye-tracking data is the same as that of reading-based data is worthy of study. The eye-tracking results might be different if different data cleaning methods had been adopted.

A fifth limitation related to the instrument reliability of offline comprehension and metalinguistic knowledge test. Plonsky and Derrick (2016) conducted a meta-analysis to investigate the reported reliability in L2 research. They suggested that regardless of language proficiency, the median instrument reliability of L2 learners was .81, and that

of native speakers was .87. In the current study, for native speakers, the instrument reliability of offline comprehension and metalinguistic knowledge tests (all three tasks) were less than .87. For the L2 learners, the instrument reliability of the offline comprehension test and *two tasks* of the metalinguistic knowledge test (i.e., deciding match or mismatch; sentence correction) did not reach .81 across the three test phases. It should be acknowledged that these two measures were not highly reliable.

Finally, further research could be co**n**ducted to analyse whether automatization took place during the training. McManus and Marsden (2019) found that their participants gradually became automatized, by analysing the performance *during* training. In the current study, limited online effects might be because the automatization did not happen during training. Coefficient of variation analyses could be conducted with the training data to detect whether the automatization had occurred or not.

6.4 Contributions of the study

The current study has made several contributions to the agenda of research into explicit training and L2 learning of English relative clauses as well as the understanding of L1 comprehension and use of relative clauses.

First, the current study looked into the online processing of relative clauses in a new perspective. The study investigated whether native English speakers and L2 learners were sensitive to the syntactic cue that disambiguates subject from object relative clauses (i.e., word order after the relative pronoun), and whether they could use the cue to interpret the meaning of sentence in real time. The findings of the *SPR test* suggested that the native speakers were potentially sensitive to the cue, but sensitivity was only observed at the third critical word for SRCs and **one type** of ORCs (because a spill-over effect could happen, the native speakers might start to be sensitive to the cue at the *second* critical word). However, the L2 learners did not show sensitivity to the syntactic cue during the three critical words at the pre-test. The findings of the *eye-tracking test* suggested that the native speakers used the syntactic cue to interpret the meaning of sentences for SRCs (started to fixate on the targets

before the end of the first critical word) and **one type** of ORCs (started to fixate on the targets during the second critical word). The L2 learners only could use the cue to interpret **one type of** SRCs where the fixations to the targets presented at the first critical word.

Second, the current study contributed to research that evaluates the online effects of explicit training. Only few published studies have investigated this agenda, and the results have been inconsistent. Some studies found that explicit training could facilitate online comprehension (Hopp, 2016; McManus & Marsden, 2017; VanPatten & Smith, 2019; Wong & Ito, 2018) while the others not (Andringa & Curcic, 2015; Dracos & Henry, 2021). The current study utilised two online measures, SPR and visual world eye-tracking. The limited online effects of the explicit training were found in the SPR tests, but were not observed in the eye-tracking test.

Third, the design of the explicit training used in the current study was informed from PI, but had differences compared to PI. PI is normally used to teach structures that have direct referential meaning, such as animacy, tense (e.g., VanPatten & Cadierno, 1993a, 1993b), but the structure used in the current study, relative clause, is syntax that does not, arguable, have referential meaning, as the training focused on word order and part of speech. The current study attempted to teach L2 learners to use this syntactic cue to predictively and retrospectively interpret the meaning of the sentence. The sentences in the training items were stopped after the cue, so the learners were forced to comprehend the sentences based on the cue. The findings suggest that the training effects were mainly observed in offline comprehension, production and metalinguistic knowledge. Limited online effects were found with SRCs.

Finally, the current study contributes to methodological issues in examining online effects of explicit training. Some previous studies measured online effects using trials to criterion (e.g. Fernández, 2008; Henry, Culman & VanPatten, 2009; VanPatten & Borst, 2012), but this method lacked external validation and the criterion to indicate 'correct processing' was arbitrary (Fernández, 2008). Followed Keating and Jegerski (2015) SPR and the visual world eye-tracking tests were involved in the current study.

The SPR measured sensitivity to the syntactic cue by introducing an anomaly between the picture and the sentence being read. The eye-tracking test allowed the participants to comprehend sentences in a natural way, and it evaluated whether the learners could use the cue – or the word after it – to interpret the sentence meaning. The results of the two tests were different. After receiving explicit training, the participants demonstrated some improvements in sensitivity to the anomaly at the post-test, but did not demonstrate relevant changes in their eye-movements.

Appendices

Appendix 1: In formation page and consent form

Information Page
To what extent can teaching parsing strategies help second language syntax learning?

Dear Students,

My name is Niu Xiaoran. I am a PhD student in the Centre for Language Learning and Use in the Department of Education. I used to be an English teacher. I would like to invite you to take part in my research project. You will receive some money on completion of the activities, and we hope that you will have fun and learn English!

Before agreeing to take part, please read this information sheet carefully and let us know if anything is unclear or you would like further information. Please also read the information about General Data Protection Regulation (GDPR) that is provided on a separate sheet.
(Link: https://www.york.ac.uk/education/research/gdpr_information/)

Purpose of the study
The study is investigating the extent to which a new type of instruction, based on listening and reading through computer-based activities, can help second language learners learning English grammar.

Who can take part?
We would like to recruit participants whose native language is Chinese and the IELTS score is between 6.0 and 7.0 to take part in the study.
We would also like to recruit participants who are native speakers of English.

Your performance in the research project will not be reported to your course tutors – the research project is entirely separate to your courses at the University.

What would this mean for you?
For the participants whose native language is Chinese:
If you take part in the study, you will be randomly divided into two groups, a “training group” and a “non-active comparison group” who will only do some of the activities (the tests) but will receive the training materials after the final set of test activities.

For the training group, you will take activities (tests, for research purposes) which include an eye-tracking activity and a series of other language tasks lasting in total of

approximately 1.5 hours. Then, in the following week, you will receive two training sessions about a complex grammar feature. Each training session will last about 30 minutes. On the three day after the second training session, you will take similar tests as before (post-tests), and three weeks later, you will take a final set (delayed post-tests). After the delayed post-tests, you will be awarded £20 if you have completed all the training and tests in the time scale as arranged with the researcher.

For test only group, you will only take the test activities (including eye tracking and the other tests), just like the participants in the other group. After the first set of activities, you with then do another set one week later, and the final set will take place in three weeks. After that, you will receive the training materials and will be awarded £15 if you have completed all the activities in the time scale as arranged with the researcher.

For the participants whose native language is English:
You will take the test activities including eye-tracking and a series of other language tasks lasting in total of approximately 1.5 hours. You only need to come for one time. You will be awarded £6 if you have completed all the activities.

Participation is voluntary
Participation is optional. If you do decide to take part, you will be given a copy of this information sheet for your records and will be asked to complete a consent form. If you change your mind at any point during the study, you will be able to withdraw your participation during the study without having to provide a reason. You will be able to withdraw your data until six weeks after the end of your final set of activities. After that time, it won't be possible to withdraw your data as it will have been included in the analysis, anonymised and prepared to be made openly available (see below).

Anonymity and confidentiality
The data that you provide (e.g. test responses, eye-movements during the tests) will be stored by code number. Any information that identifies you will be stored separately from the data. You are free to withdraw from the study at any time during data collection and up to six weeks after the data is collected.

Storing and using your data
Data will be stored securely on a password protected computer.

The file linking your name with the numerical identifier will be kept encrypted and on a password protected computer. This file will be destroyed once the thesis has been submitted and papers have been published (up to about five years after the end of the study)

Once the datasets have been completely anonymised and prepared, they could be made openly available for others to use and stored indefinitely on an 'open repository',

such as IRIS (www.iris-database.org) and/or the Open Science Framework (osf.io). The data that I collect (test responses) may be used in *anonymous* format in different ways. Please indicate on the consent form attached with a ☑ if you are happy for this anonymised data to be used in the ways listed.

Questions or concerns

If you have any questions about this participant information sheet or concerns about how your data is being processed, please feel free to contact Niu Xiaoran by email (xn548@york.ac.uk), or the Chair of Ethics Committee via email education-research-administrator@york.ac.uk. If you are dissatisfied with the responses you receive, please contact the University's Data Protection Officer at dataprotection@york.ac.uk

I hope that you will agree to take part. If you are happy to participate, please complete the form attached and return it to the researcher.

Please keep this information sheet for your own records.

Thank you for taking the time to read this information.

Yours sincerely

Niu Xiaoran & Professor Emma Marsden (my PhD Supervisor)

To what extent can teaching parsing strategies help second language syntax learning?
Consent Form

Please tick each box if you are happy to take part in this research.

I confirm that I have read and understood the information above and I understand that this will involve me taking part as described above.	
I understand that participation in this study is voluntary.	
I understand that my data will not be identifiable and the anonymous data may be used in publications, presentations and made freely available online.	
I confirm that I have read the information about GDPR	

NAME_____

SIGNATURE_____

DATE_____

Appendix 2: Questionnaire of background information

Q1: Your name: _____

Q2: Your age: _____

Q3: Your gender:

Male

Female

Prefer not to say

Q4: Study Level:

Undergraduate – year 1

Undergraduate – year 2

Undergraduate – year 3

Undergraduate – year 4

Master

PhD – year 1

PhD – year 2

PhD – year 3

PhD – year 4

PhD – year 5

PhD – year 6

Q5: The language that is used in the home when you were a child:

Chinese

English

Others: _____

Q6: How long have you been in the UK? Please give your answer to the nearest month (e.g., 3 months, 9 months, 1 year and 2 months, etc.)

_____

Q7: Have you been to an English-speaking country for one month or more before your current stay in the UK? If yes, please insert how long did you stay in that country (please give your answer to the nearest month).

Yes: _____

No

Q8: Please give your highest IELTS score:

	1	2	3	4	5	6	7	8	9
Overall									
Listening									
Reading									
Speaking									
Writing									

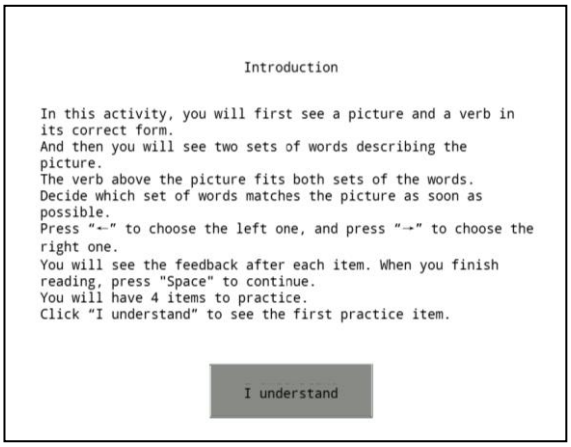
Q9: Please rate your own proficiency with English grammar. "1" indicates "not at all proficient" while "5" indicates "very proficient":

Not at all proficient				Very proficient
1	2	3	4	5

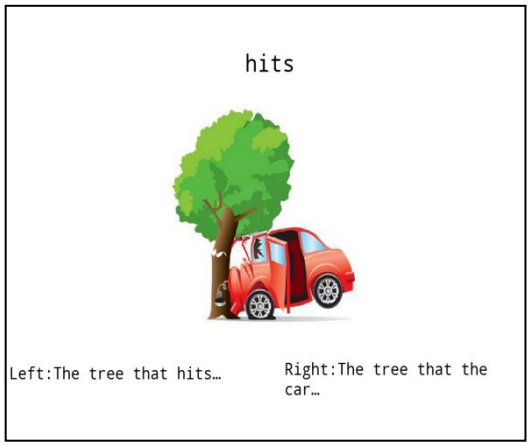
Note: The native English speakers only answered questions 1 to 5.

Appendix 3: Example of parsing strategies training activities 1 & 2

Activity 1 (reading based)

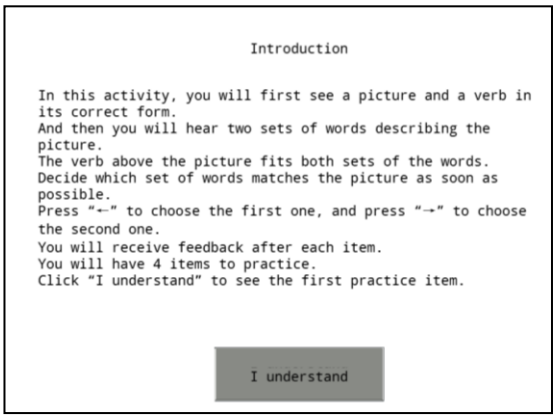


Instruction page

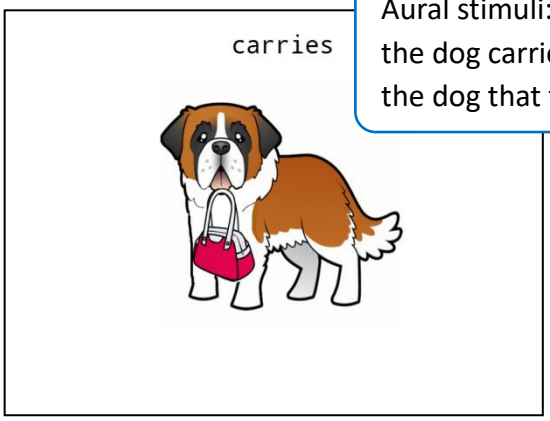


Example item

Activity 2 (listening based)



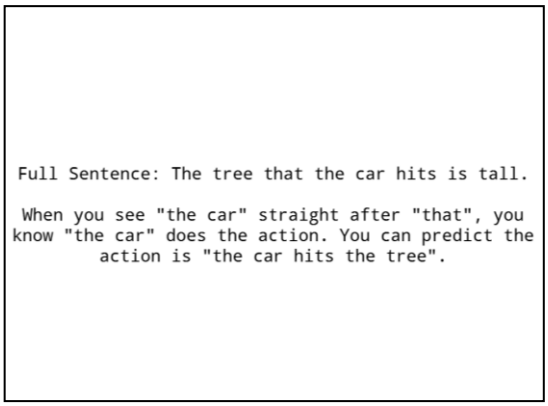
Instruction page



Example item

Feedback

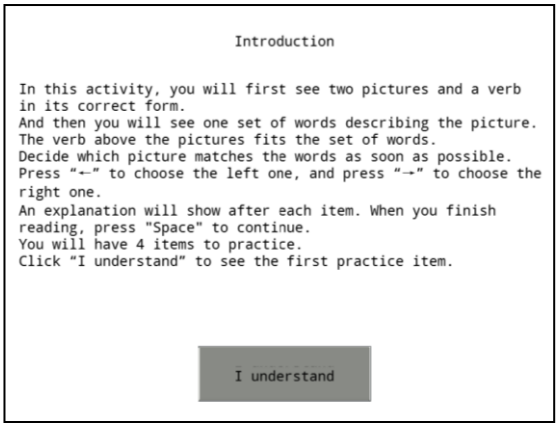
Feedback for correct response



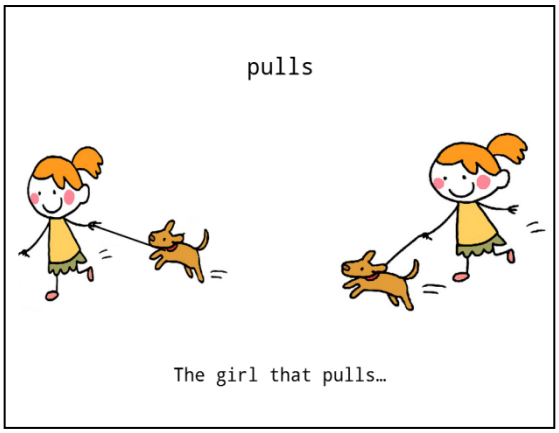
Feedback for incorrect feedback
(Played aurally in the activity 2)

Appendix 4: Example of parsing strategies training activities 3 & 4

Activity 3 (reading based)

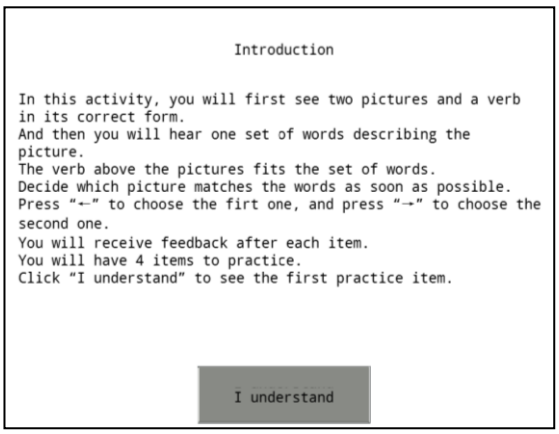


Instruction page



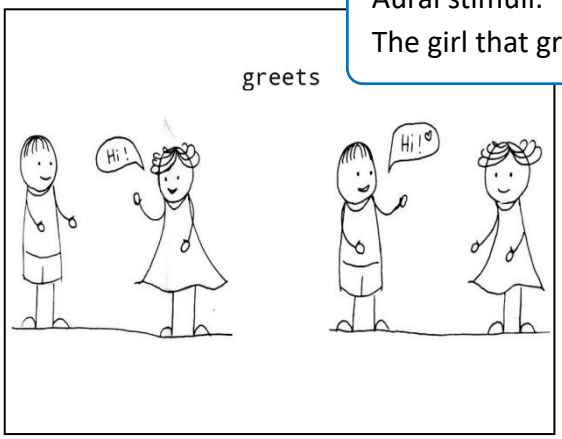
Example item

Activity 4 (listening based)



Instruction page

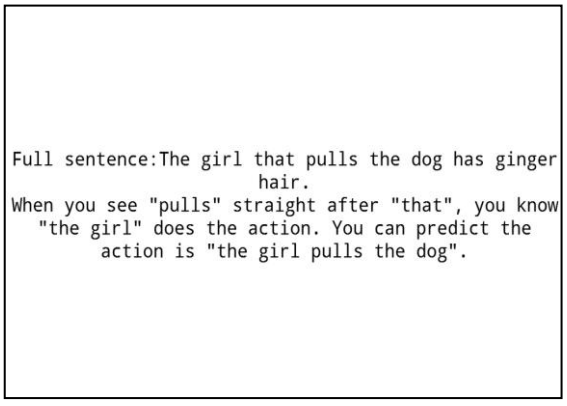
Aural stimuli:
The girl that greets…



Example item

Feedback

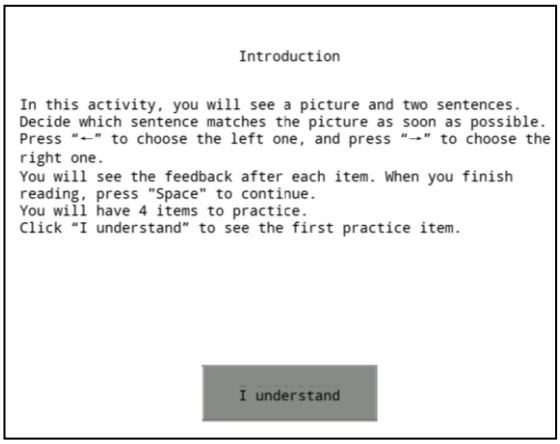
Feedback for correct response



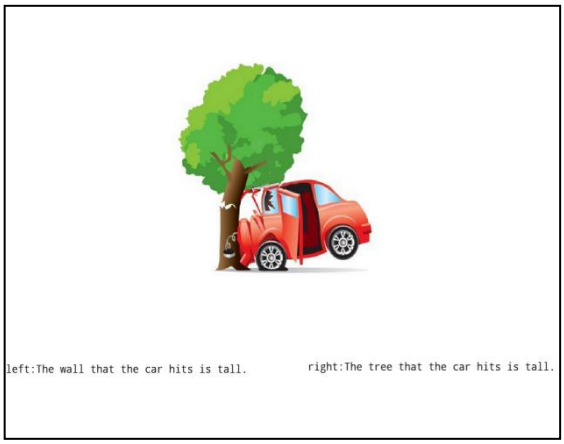
Feedback for incorrect feedback
(Played aurally in the activity 4)

Appendix 5: Example of input flood training activities 1 & 2

Activity 1 (reading based)

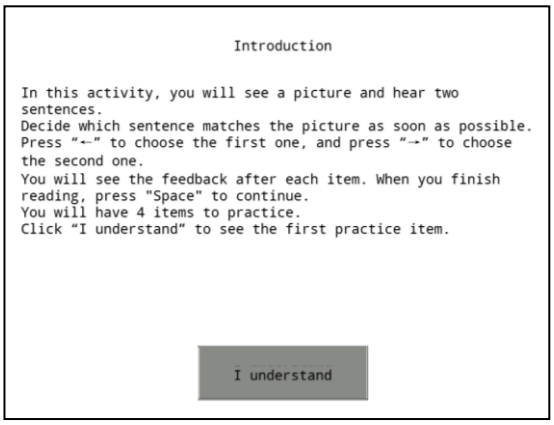


Instruction page

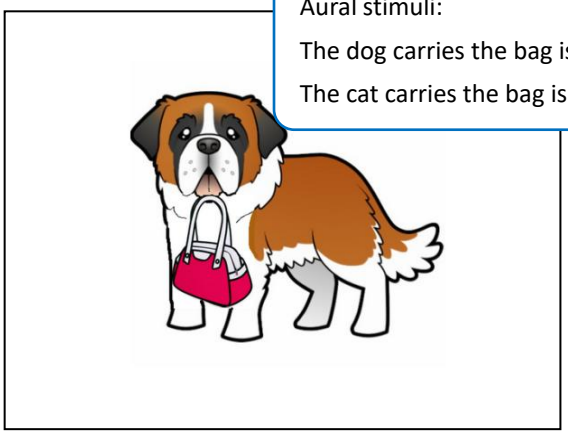


Example item

Activity 2 (listening based)



Instruction page



Aural stimuli:

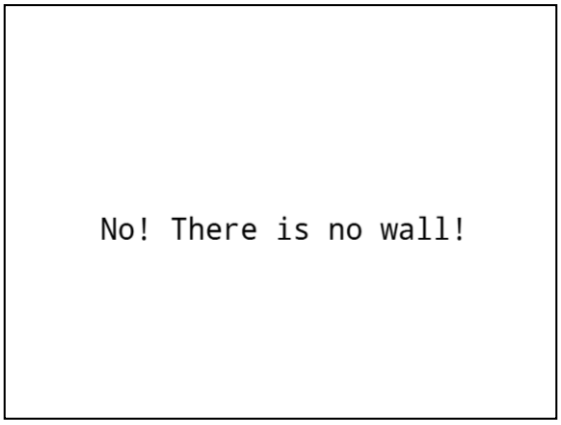
The dog carries the bag is brown.

The cat carries the bag is brown.

Example item

Feedback

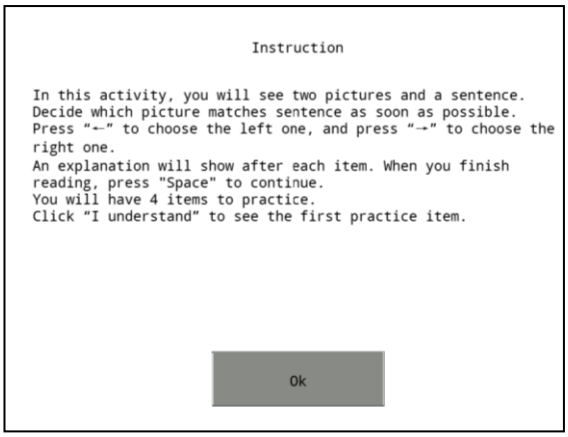
Feedback for correct response



Feedback for incorrect response

Appendix 6: Example of input flood training activities 3 & 4

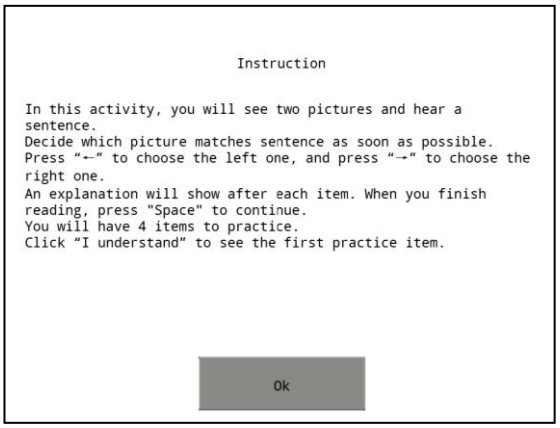
Activity 3 (reading based)



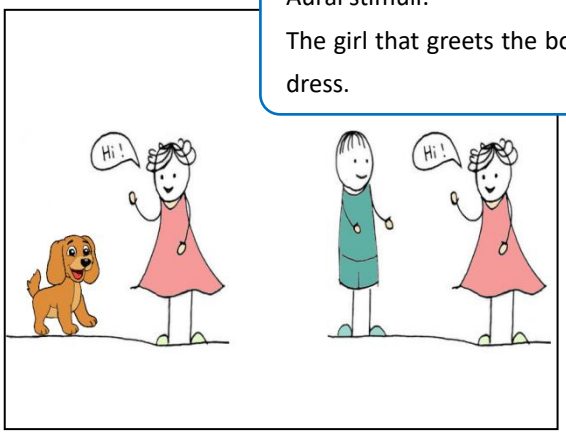
Instruction page

Example item

Activity 4 (listening based)



Instruction page

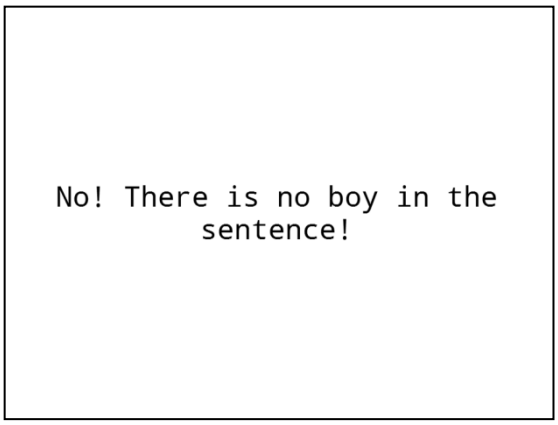


Example item

Aural stimuli:
The girl that greets the boy wears a dress.

Feedback

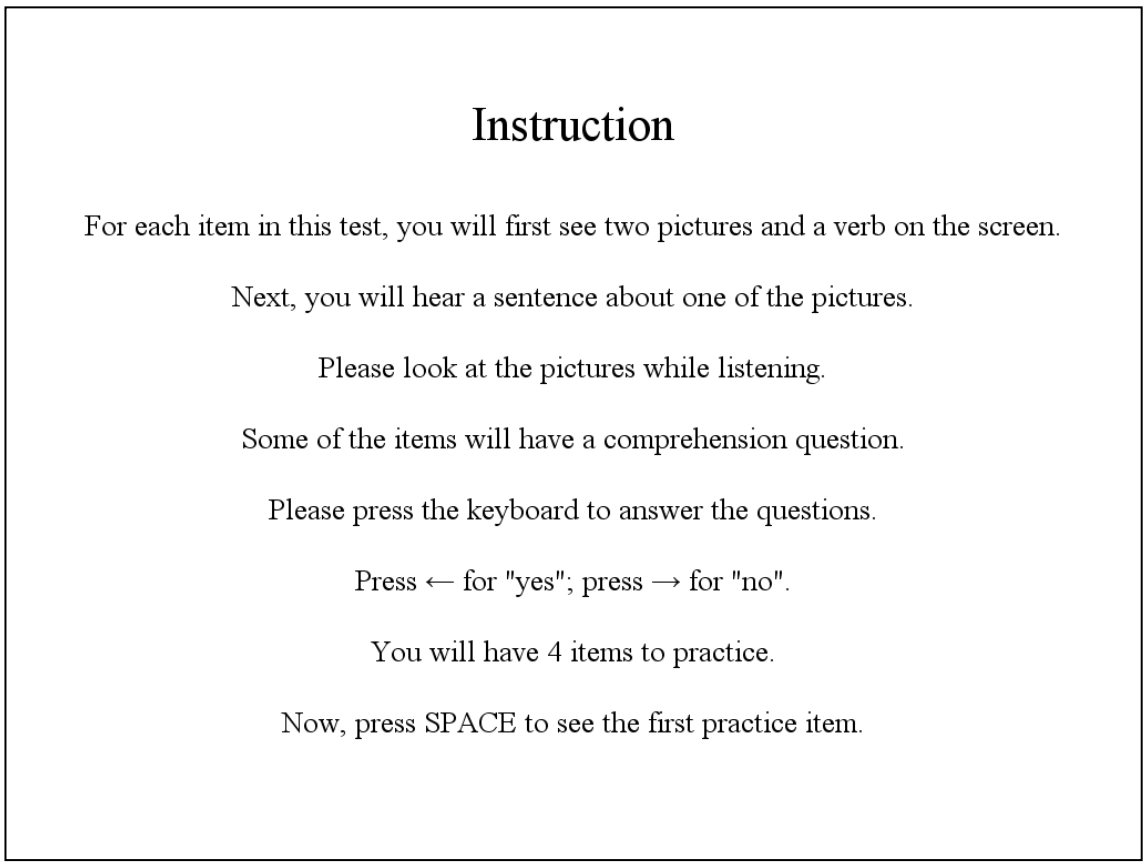
Feedback for correct response



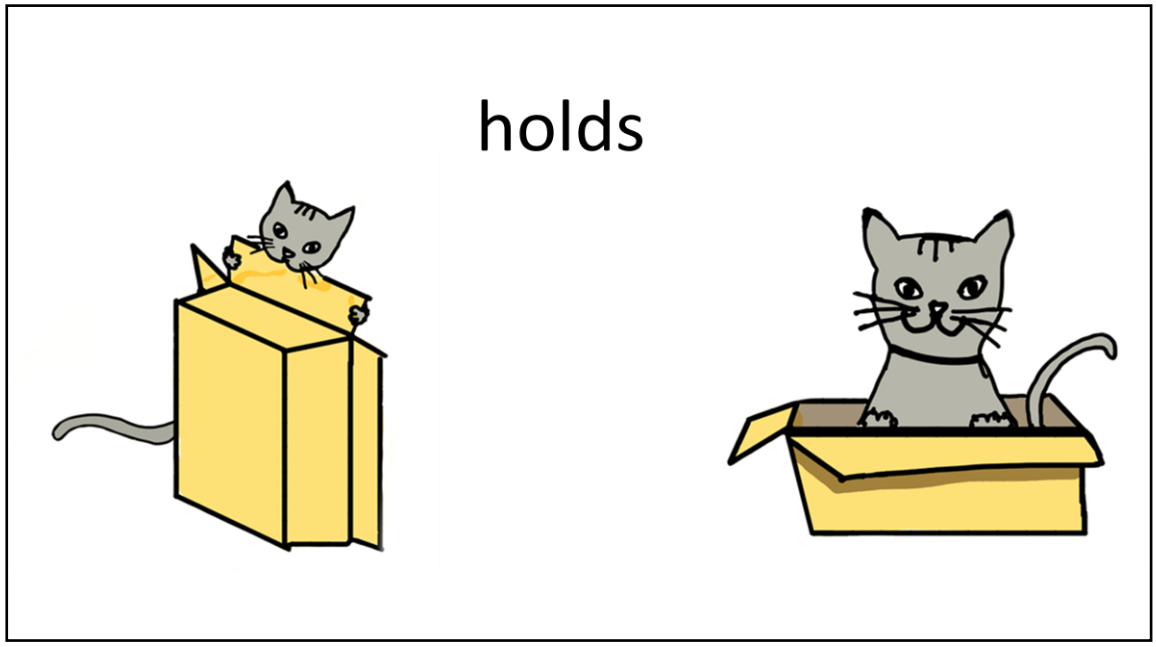
Feedback for incorrect response

Appendix 7: Example of visual word eye-tracking test

Instruction page



Example item



Aural stimuli: The cat that the box holds is grey.

Appendix 8: Example of self-paced reading test

Instruction page

Instruction

In each item, you will first see a picture and a verb in its correct form. Then, you need to press SPACE BAR to read a sentence word by word. Please keep clicking until a dot appears. Please make sure you try to understand each sentence. After some items, you need to answer a question about the meaning of the SENTENCE.
You will have four items to practice.
Click "I understand" to see the first practice item.

I understand

Example item

Stimuli: The boy that the camera films is happy.

films

films

The

films

boy

films

that

films

the

films

camera

films

films

films

is

films

happy

The sentence shows word by word by each key pressing.

Appendix 9: Example of offline comprehension test (aural sentence-picture matching)

Instruction page

Introduction

For each item, you will first see two pictures and then hear a sentence.
You need to decide which picture matches the sentence.
Press ← to choose the left picture, and press → to choose the right one.
Please choose the mathced pictures as fast as you can.
Now, click "I understand" to see the first practice item.

I understand

Example item

Aural stimuli: The water that the man freezes is cold.

Appendix 10: Example of oral production test (picture description)

Instruction page

Introduction

For each item, you will first see two pictures. Then, you will hear and read a description of each of the two pictures. And then you will hear and read a question about one of the people. You must answer the question out loud.
The beginning of the answer is given for you.
Please do NOT use words such as “and”, “but”, “so”, “because”, “when” and so on.
Please do NOT describe the positions of the things: “on the left”, “on the right”, “the first one”, and “the second one”.
When you have finished, press any key to move on.
Now, please click "I understand" to see the first item.

I understand

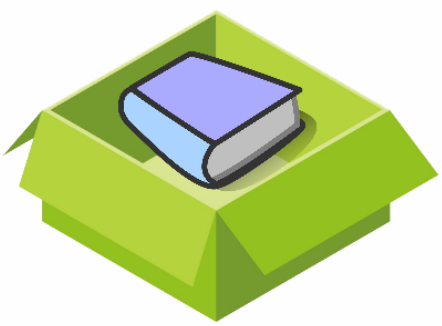
Example item

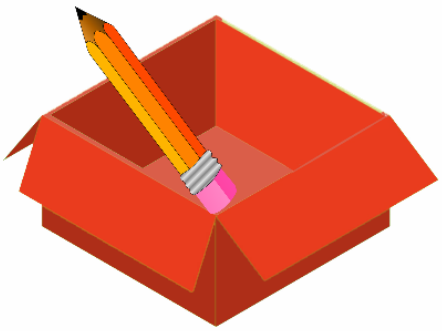
In the first picture, the box is green. The box holds the book.

In the second picture, the box is red. The box holds the pencil.

Which box is green?

The box…





Appendix 11: Example of metalinguistic knowledge test

Instruction page

Does the sentence match the picture?

Instruction: For each item, you will see one picture and one sentence. Decide whether the sentence matches or mismatches the picture by ticking the box ☑. If you think there is a **mismatch**,

1) **Circle** the word or words in the sentence that do not match the picture.

2) **Explain the reason**, **as fully as possible**, why the sentence does not match the picture.

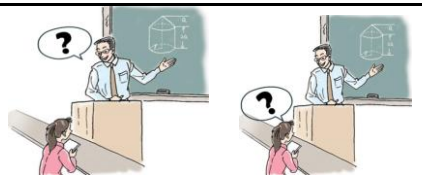
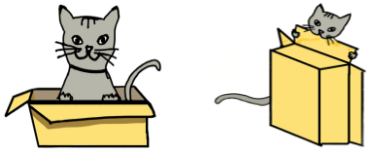
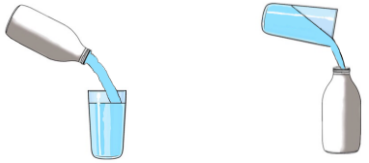
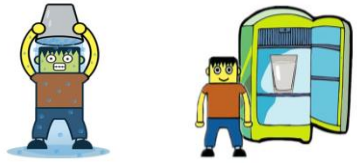
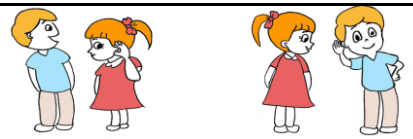
3) Can you **move one word** to make it match? You **cannot exchange** (swap round) two or more words! Show this move with an arrow on the sentence.

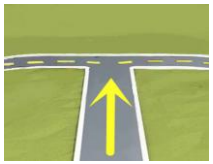
Example item

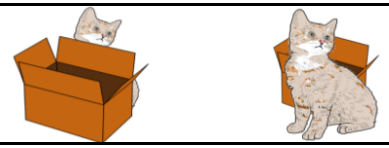
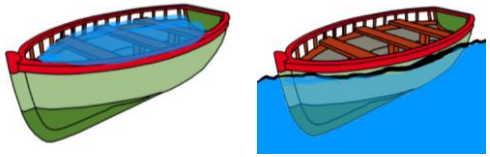
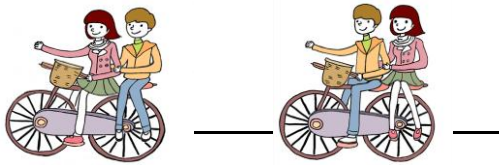
The boy that amuses the girl is cute.

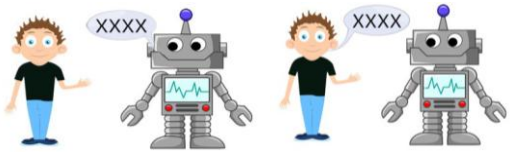
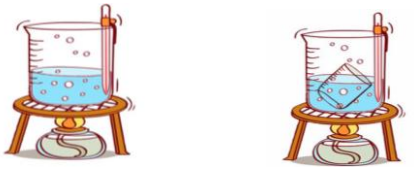
Match ☐ Mismatch☐

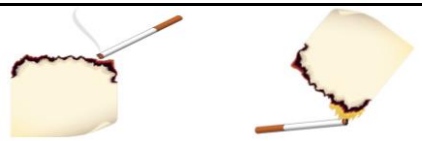
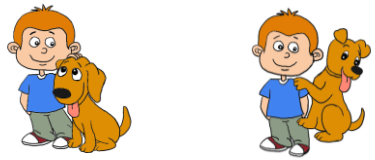
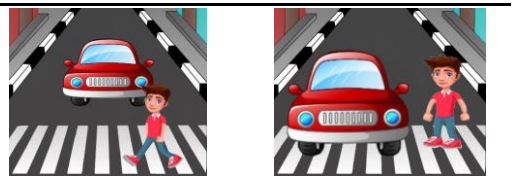
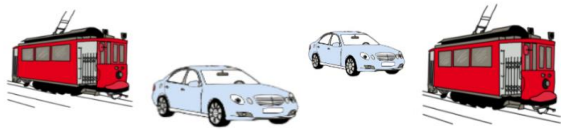
Appendix 12: Critical items of the visual world eye-tracking test

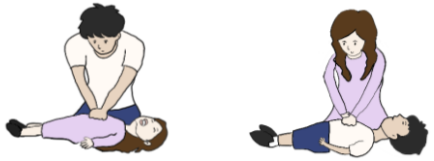
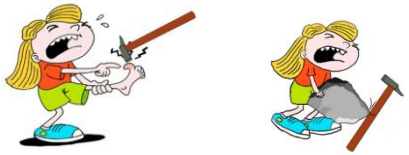
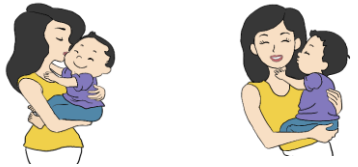
NO.	Ver-sion	Verb	Type	Animacy	Stimuli	pictures
1	1	asks	SRC	animate	The man that asks the woman has short hair.	
2	2	asks	ORC	animate	The man that the woman asks has short hair.	
3	3	asks	SRC	animate	The woman that asks the man has long hair.	
4	4	asks	ORC	animate	The woman that the man asks has long hair.	
5	1	holds	SRC	animate	The cat that holds the box is grey.	
6	2	holds	ORC	animate	The cat that the box holds is grey.	
7	3	holds	SRC	inanimate	The box that holds the cat is big.	
8	4	holds	ORC	inanimate	The box that the cat holds is big.	
9	4	cleans	SRC	animate	The man that cleans the shower is tall.	
10	3	cleans	ORC	animate	The man that the shower cleans is tall.	
11	2	cleans	SRC	inanimate	The shower that cleans the man is new.	
12	1	cleans	ORC	inanimate	The shower that the man cleans is new.	
13	1	fills	SRC	inanimate	The bottle that fills the glass is new.	
14	2	fills	ORC	inanimate	The bottle that the glass fills is new.	
15	3	fills	SRC	inanimate	The glass that fills the bottle is new.	
16	4	fills	ORC	inanimate	The glass that the bottle fills is new.	
17	2	freezes	SRC	animate	The man that freezes the water is cold.	
18	4	freezes	ORC	animate	The man that the water freezes is cold.	
19	1	freezes	SRC	inanimate	The water that freezes the man is cold.	
20	3	freezes	ORC	inanimate	The water that the man freezes is cold.	
21	2	hears	SRC	animate	The boy that hears the girl has blond hair.	
22	3	hears	ORC	animate	The boy that the girl hears has blond hair.	
23	4	hears	SRC	animate	The girl that hears the boy has blond hair.	

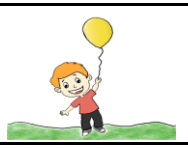
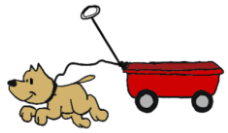
24	1	hears	ORC	animate	The girl that the boy hears has blond hair.	
25	1	lifts	SRC	animate	The man that lifts the woman has blond hair.	 
26	2	lifts	ORC	animate	The man that the woman lifts has blond hair.	
27	3	lifts	SRC	animate	The woman that lifts the man has blond hair.	
28	4	lifts	ORC	animate	The woman that the man lifts has blond hair.	
29	3	presents	SRC	animate	The girl that presents the picture is cute.	 
30	1	presents	ORC	animate	The girl that the picture presents is cute.	
31	4	presents	SRC	inanimate	The picture that presents the girl is big.	
32	2	presents	ORC	inanimate	The picture that the girl presents is big.	
33	2	presses	SRC	inanimate	The book that presses the letter is big.	 
34	3	presses	ORC	inanimate	The book that the letter presses is big.	
35	4	presses	SRC	inanimate	The letter that presses the book is pink.	
36	1	presses	ORC	inanimate	The letter that the book presses is pink.	
37	1	shows	SRC	inanimate	The road that shows the arrow is straight.	 
38	2	shows	ORC	inanimate	The road that the arrow shows is straight.	
39	3	shows	SRC	inanimate	The arrow that shows the road is straight.	
40	4	shows	ORC	inanimate	The arrow that the road shows is straight.	
41	2	calls	SRC	animate	The man that calls the girl is happy.	 
42	3	calls	ORC	animate	The man that the girl calls is happy.	
43	4	calls	SRC	animate	The girl that calls the man is happy.	
44	1	calls	ORC	animate	The girl that the man calls is happy.	
45	2	covers	SRC	inanimate	The fruit that covers the cream is sweet.	 
46	3	covers	ORC	inanimate	The fruit that the cream covers is sweet.	
47	4	covers	SRC	inanimate	The cream that covers the fruit is sweet.	
48	1	covers	ORC	inanimate	The cream that the fruit covers is sweet.	
49	1	hides	SRC	animate	The cat that hides the box is cute.	

50	2	hides	ORC	animate	The cat that the box hides is cute.	
51	3	hides	SRC	inanimate	The box that hides the cat is big.	
52	4	hides	ORC	inanimate	The box that the cat hides is big.	
53	1	contains	SRC	inanimate	The water that contains the boat is deep.	
54	2	contains	ORC	inanimate	The water that the boat contains is deep.	
55	3	contains	SRC	inanimate	The boat that holds the water is small.	
56	4	contains	ORC	inanimate	The boat that the water contains is small.	
57	2	messages	SRC	inanimate	The phone that messages the computer is new.	
58	3	messages	ORC	inanimate	The phone that the computer messages is new.	
59	4	messages	SRC	inanimate	The computer that messages the phone is new.	
60	1	messages	ORC	inanimate	The computer that the phone messages is new.	
61	4	moves	SRC	animate	The man that moves the truck is short.	
62	3	moves	ORC	animate	The man that the truck moves is short.	
63	2	moves	SRC	inanimate	The truck that moves the girl is yellow.	
64	1	moves	ORC	inanimate	The truck that the man moves is yellow.	
65	1	serves	SRC	animate	The woman that serves the man has big eyes.	
66	2	serves	ORC	animate	The woman that the man serves has big eyes.	
67	3	serves	SRC	animate	The man that serves the woman has big eyes	
68	4	serves	ORC	animate	The man that the woman serves has big eyes.	
69	2	supports	SRC	animate	The woman that supports the ball is strong.	
70	4	supports	ORC	animate	The woman that the ball supports is strong.	
71	1	supports	SRC	inanimate	The ball that supports the woman is yellow.	
72	3	supports	ORC	inanimate	The ball that the woman supports is yellow.	
73	2	takes	SRC	animate	The girl that takes the boy is slim.	
74	3	takes	ORC	animate	The girl that the boy takes is slim.	
75	4	takes	SRC	animate	The boy that takes the girl is slim.	

76	1	takes	ORC	animate	The boy that the girl takes is slim.	
77	3	tells	SRC	animate	The boy that tells the robot is clever.	
78	1	tells	ORC	animate	The boy that the robot tells is clever.	
79	4	tells	SRC	inanimate	The robot that tells the boy is clever.	
80	2	tells	ORC	inanimate	The robot that the boy tells is clever.	
81	1	boils	SRC	inanimate	The cup that boils the water is old.	
82	2	boils	ORC	inanimate	The cup that the water boils is old.	
83	3	boils	SRC	inanimate	The water that boils the cup is hot.	
84	4	boils	ORC	inanimate	The water that the cup boils is hot.	
85	1	follows	SRC	animate	The boy that follows the car runs fast.	
86	2	follows	ORC	animate	The boy that the car follows runs fast.	
87	3	follows	SRC	inanimate	The car that follows the boy is yellow.	
88	4	follows	ORC	inanimate	The car that the boy follows is yellow.	
89	1	pushes	SRC	animate	The man that pushes the woman is short.	
90	2	pushes	ORC	animate	The man that the woman pushes is short.	
91	3	pushes	SRC	animate	The woman that pushes the man is short.	
92	4	pushes	ORC	animate	The woman that the man pushes is short.	
93	4	hits	SRC	animate	The girl that hits the basketball is tall.	
94	3	hits	ORC	animate	The girl that the basketball hits is tall.	
95	2	hits	SRC	inanimate	The basketball that hits the girl is new.	
96	1	hits	ORC	inanimate	The basketball that the girl hits is new.	
97	2	holds	SRC	animate	The boy that holds the skateboard is cute.	
98	4	holds	ORC	animate	The boy that the skateboard holds is cute.	
99	1	holds	SRC	inanimate	The skateboard that holds the boy is red.	
100	3	holds	ORC	inanimate	The skateboard that the boy holds is red.	
101	2	lights	SRC	inanimate	The cigarette that lights the paper is cheap.	

102	3	lights	ORC	inanimate	The cigarette that the paper lights is cheap.	
103	4	lights	SRC	inanimate	The paper that lights the cigarette is thin.	
104	1	lights	ORC	inanimate	The paper that the cigarette lights is thin.	
105	2	touches	SRC	animate	The dog that touches the man is brown.	
106	3	touches	ORC	animate	The dog that the man touches is brown.	
107	4	touches	SRC	animate	The man that touches the dog is short.	
108	1	touches	ORC	animate	The man that the dog touches is short.	
109	1	photographs	SRC	animate	The girl that photographs the boy has brown hair.	
110	2	photographs	ORC	animate	The girl that the boy photographs has brown hair.	
111	3	photographs	SRC	animate	The boy that photographs the girl has blond hair.	
112	4	photographs	ORC	animate	The boy that the girl photographs has blond hair.	
113	1	shoots	SRC	inanimate	The tank that shoots the helicopter is green.	
114	2	shoots	ORC	inanimate	The tank that the helicopter shoots is green.	
115	3	shoots	SRC	inanimate	The helicopter that shoots the tank is green.	
116	4	shoots	ORC	inanimate	The helicopter that the tank shoots is green.	
117	3	stops	SRC	animate	The man that stops the car is thin.	
118	1	stops	ORC	animate	The man that the car stops is thin.	
119	4	stops	SRC	inanimate	The car that stops the man is red.	
120	2	stops	ORC	inanimate	The car that the man stops is red.	
121	2	answers	SRC	animate	The woman that answers the man has long hair.	
122	3	answers	ORC	animate	The woman that the man answers has long hair.	
123	4	answers	SRC	animate	The man that answers the woman has short hair.	
124	1	answers	ORC	animate	The man that the woman answers has short hair.	
125	2	chases	SRC	inanimate	The tram that chases the car is fast.	
126	3	chases	ORC	inanimate	The tram that the car chases is fast.	
127	4	chases	SRC	inanimate	The car that chases the tram is fast.	

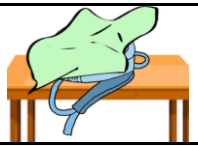
128	1	chases	ORC	inanimate	The car that the tram chases is fast.	
129	1	destroys	SRC	inanimate	The helicopter that destroys the cannon is big.	
130	2	destroys	ORC	inanimate	The helicopter that the cannon destroys is big.	
131	3	destroys	SRC	inanimate	The cannon that destroys the helicopter is big.	
132	4	destroys	ORC	inanimate	The cannon that the helicopter destroys is big.	
133	2	hangs	SRC	inanimate	The rope that hangs the hook is long.	
134	3	hangs	ORC	inanimate	The rope that the hook hangs is long.	
135	4	hangs	SRC	inanimate	The hook that hangs the rope is sharp.	
136	1	hangs	ORC	inanimate	The hook that the rope hangs is sharp.	
137	1	helps	SRC	animate	The man that helps the woman is kind.	
138	2	helps	ORC	animate	The man that the woman helps is kind.	
139	3	helps	SRC	animate	The woman that helps the man is kind.	
140	4	helps	ORC	animate	The woman that the man helps is kind.	
141	1	hits	SRC	animate	The woman that hits the hammer is tall.	
142	2	hits	ORC	animate	The woman that the hammer hits is tall.	
143	3	hits	SRC	inanimate	The hammer that hits the woman is new.	
144	4	hits	ORC	inanimate	The hammer that the woman hits is new.	
145	2	kisses	SRC	animate	The woman that kisses the baby is nice.	
146	3	kisses	ORC	animate	The woman that the baby kisses is nice.	
147	4	kisses	SRC	animate	The baby that kisses the woman is cute.	
148	1	kisses	ORC	animate	The baby that the woman kisses is cute.	
149	4	leaves	SRC	animate	The man that leaves the car is busy.	
150	3	leaves	ORC	animate	The man that the car leaves is busy.	
151	2	leaves	SRC	inanimate	The car that leaves the man is small.	
152	1	leaves	ORC	inanimate	The car that the man leaves is small.	
153	2	pulls	SRC	animate	The boy that pulls the balloon is happy.	

154	4	pulls	ORC	animate	The boy that the balloon pulls is happy.	 
155	1	pulls	SRC	inanimate	The balloon that pulls the boy is yellow.	
156	3	pulls	ORC	inanimate	The balloon that the boy pulls is yellow.	
157	3	carries	SRC	animate	The dog that carries the cart is brown.	 
158	1	carries	ORC	animate	The dog that the cart carries is brown.	
159	4	carries	SRC	inanimate	The cart that carries the dog is red.	
160	2	carries	ORC	inanimate	The cart that the dog carries is red.	

Appendix 13: Critical items of the self-paced reading test

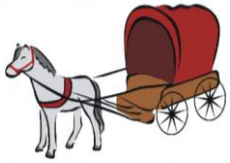
NO.	Ver-sion	Verb	Type	Animacy	Stimuli	Match /Mismatch	pictures
1	1	follows	SRC	animate	The tiger that follows the lion is happy.	Match	
2	2	follows	ORC	animate	The tiger that the lion follows is happy.	Mismatch	
3	3	follows	SRC	animate	The lion that follows the tiger is yellow.	Mismatch	
4	4	follows	ORC	animate	The lion that the tiger follows is yellow.	Match	
5	2	finds	SRC	animate	The boy that finds the girl is thin.	Match	
6	1	finds	ORC	animate	The boy that the girl finds is thin.	Mismatch	
7	4	finds	SRC	animate	The girl that finds the boy is thin.	Mismatch	
8	3	finds	ORC	animate	The girl that the boy finds is thin.	Match	
9	3	amuses	SRC	animate	The girl that amuses the boy is cute.	Match	
10	4	amuses	ORC	animate	The girl that the boy amuses is cute.	Mismatch	
11	1	amuses	SRC	animate	The boy that amuses the girl is cute.	Mismatch	
12	2	amuses	ORC	animate	The boy that the girl amuses is cute.	Match	
13	4	disturbs	SRC	animate	The man that disturbs the woman is relaxed.	Match	
14	3	disturbs	ORC	animate	The man that the woman disturbs is annoyed.	Mismatch	
15	2	disturbs	SRC	animate	The woman that disturbs the man is busy.	Mismatch	
16	1	disturbs	ORC	animate	The woman that the man disturbs is busy.	Match	
17	1	examines	SRC	animate	The man that examines the woman is worried.	Match	
18	2	examines	ORC	animate	The man that the woman examines is worried.	Mismatch	
19	3	examines	SRC	animate	The woman that examines the man is worried.	Mismatch	
20	4	examines	ORC	animate	The woman that the man examines is worried.	Match	
21	2	greets	SRC	animate	The woman that greets the girl is happy.	Match	
22	1	greets	ORC	animate	The woman that the girl greets is tall.	Mismatch	
23	4	greets	SRC	animate	The girl that greets the woman is short.	Mismatch	

24	3	greets	ORC	animate	The girl that the woman greets is happy.	Match	
25	3	hears	SRC	animate	The man that hears the woman is old.	Match	
26	4	hears	ORC	animate	The man that the woman hears is old.	Mismatch	
27	1	hears	SRC	animate	The woman that hears the man is old.	Mismatch	
28	2	hears	ORC	animate	The woman that the man hears is old.	Match	
29	4	helps	SRC	animate	The dog that helps the man is yellow.	Match	
30	3	helps	ORC	animate	The dog that the man helps is yellow.	Mismatch	
31	2	helps	SRC	animate	The man that helps the dog is kind.	Mismatch	
32	1	helps	ORC	animate	The man that the dog helps is tall.	Match	
33	1	kills	SRC	animate	The lion that kills the man is hungry.	Match	
34	2	kills	ORC	animate	The lion that the man kills is poor.	Mismatch	
35	3	kills	SRC	animate	The man that kills the lion is strong.	Mismatch	
36	4	kills	ORC	animate	The man that the lion kills is poor.	Match	
37	2	leaves	SRC	animate	The man that leaves the girl is kind.	Match	
38	1	leaves	ORC	animate	The man that the girl leaves is kind.	Mismatch	
39	4	leaves	SRC	animate	The girl that leaves the man is cute.	Mismatch	
40	3	leaves	ORC	animate	The girl that the man leaves is cute.	Match	
41	3	scares	SRC	animate	The dog that scares the boy is fierce.	Match	
42	4	scares	ORC	animate	The dog that the boy scares is poor.	Mismatch	
43	1	scares	SRC	animate	The boy that scares the dog is fierce.	Mismatch	
44	2	scares	ORC	animate	The boy that the dog scares is poor.	Match	
45	4	touches	SRC	animate	The woman that touches the girl is kind.	Match	
46	3	touches	ORC	animate	The woman that the girl touches is kind.	Mismatch	
47	2	touches	SRC	animate	The girl that touches the woman is tired.	Mismatch	
48	1	touches	ORC	animate	The girl that the woman touches is tired.	Match	
49	1	fills	SRC	inanimate	The cup that fills the bowl is green.	Match	

50	2	fills	ORC	inanimate	The cup that the bowl fills is green.	Mismatch	
51	3	fills	SRC	inanimate	The bowl that fills the cup is blue.	Mismatch	
52	4	fills	ORC	inanimate	The bowl that the cup fills is blue.	Match	
53	2	carries	SRC	inanimate	The chair that carries the box is wooden.	Match	
54	1	carries	ORC	inanimate	The chair that the box carries is wooden.	Mismatch	
55	4	carries	SRC	inanimate	The box that carries the chair is brown.	Mismatch	
56	3	carries	ORC	inanimate	The box that the chair carries is brown.	Match	
57	3	charges	SRC	inanimate	The battery that charges the phone is green.	Match	
58	4	charges	ORC	inanimate	The battery that the phone charges is green.	Mismatch	
59	1	charges	SRC	inanimate	The phone that charges the battery is new.	Mismatch	
60	2	charges	ORC	inanimate	The phone that the battery charges is new.	Match	
61	4	chases	SRC	inanimate	The helicopter that chases the plane is grey.	Match	
62	3	chases	ORC	inanimate	The helicopter that the plane chases is grey.	Mismatch	
63	2	chases	SRC	inanimate	The plane that chases the helicopter is white.	Mismatch	
64	1	chases	ORC	inanimate	The plane that the helicopter chases is white.	Match	
65	1	stops	SRC	inanimate	The train that stops the car is old.	Match	
66	2	stops	ORC	inanimate	The train that the car stops is old.	Mismatch	
67	3	stops	SRC	inanimate	The car that stops the train is yellow.	Mismatch	
68	4	stops	ORC	inanimate	The car that the train stops is yellow.	Match	
69	2	contains	SRC	inanimate	The bottle that contains the paper is clear.	Match	
70	1	contains	ORC	inanimate	The bottle that the paper contains is clear.	Mismatch	
71	4	contains	SRC	inanimate	The paper that contains the bottle is yellow.	Mismatch	
72	3	contains	ORC	inanimate	The paper that the bottle contains is yellow.	Match	
73	3	covers	SRC	inanimate	The cloth that covers the bag is green.	Match	
74	4	covers	ORC	inanimate	The cloth that the bag covers is green.	Mismatch	
75	1	covers	SRC	inanimate	The bag that covers the cloth is blue.	Mismatch	

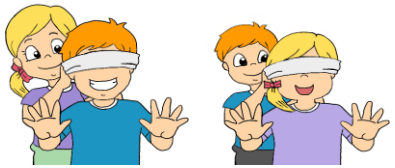
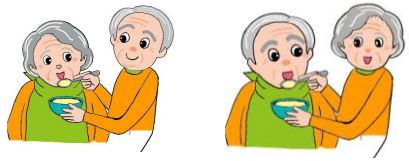
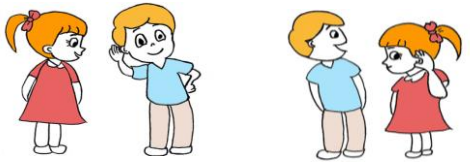
76	2	covers	ORC	inanimate	The bag that the cloth covers is blue.	Match	
77	4	holds	SRC	inanimate	The bathtub that holds the bucket is grey.	Match	
78	3	holds	ORC	inanimate	The bathtub that the bucket holds is grey.	Mismatch	
79	2	holds	SRC	inanimate	The bucket that holds the bathtub is pink.	Mismatch	
80	1	holds	ORC	inanimate	The bucket that the bathtub holds is pink.	Match	
81	1	messes	SRC	inanimate	The ice-cream that messes the soup is sweet.	Match	
82	2	messes	ORC	inanimate	The ice-cream that the soup messes is sweet.	Mismatch	
83	3	messes	SRC	inanimate	The soup that messes the ice-cream is tasty.	Mismatch	
84	4	messes	ORC	inanimate	The soup that the ice-cream messes is tasty.	Match	
85	2	presents	SRC	inanimate	The computer that presents the phone is new.	Match	
86	1	presents	ORC	inanimate	The computer that the phone presents is new.	Mismatch	
87	4	presents	SRC	inanimate	The phone that presents the computer is new.	Mismatch	
88	3	presents	ORC	inanimate	The phone that the computer presents is new.	Match	
89	3	ruins	SRC	inanimate	The cola that ruins the phone is tasty.	Match	
90	4	ruins	ORC	inanimate	The cola that the phone ruins is tasty.	Mismatch	
91	1	ruins	SRC	inanimate	The phone that ruins the cola is new.	Mismatch	
92	2	ruins	ORC	inanimate	The phone that the cola ruins is new.	Match	
93	4	surrounds	SRC	inanimate	The river that surrounds the castle is clear.	Match	
94	3	surrounds	ORC	inanimate	The river that the castle surrounds is clear.	Mismatch	
95	2	surrounds	SRC	inanimate	The castle surrounds the river is ancient.	Mismatch	
96	1	surrounds	ORC	inanimate	The castle that the river surrounds is ancient.	Match	
97	1	holds	SRC	inanimate	The table that holds the boy is brown.	Match	
98	2	holds	ORC	inanimate	The table that the boy holds is brown.	Mismatch	
99	3	holds	SRC	animate	The boy that holds the table is cute.	Mismatch	
100	4	holds	ORC	animate	The boy that the table holds is cute.	Match	
101	2	hides	SRC	inanimate	The box that hides the boy is grey.	Match	

102	1	hides	ORC	inanimate	The box that the boy hides is grey.	Mismatch	
103	4	hides	SRC	animate	The man that hides the box is scared.	Mismatch	
104	3	hides	ORC	animate	The man that the box hides is scared.	Match	
105	3	dries	SRC	inanimate	The towel that dries the man is green.	Match	
106	4	dries	ORC	inanimate	The towel that the man dries is green.	Mismatch	
107	1	dries	SRC	animate	The man that dries the towel is relaxed.	Mismatch	
108	2	dries	ORC	animate	The man that the towel dries is relaxed.	Match	
109	4	films	SRC	inanimate	The video that films the boy is new.	Match	
110	3	films	ORC	inanimate	The video that the boy films is new.	Mismatch	
111	2	films	SRC	animate	The boy that films the video is happy.	Mismatch	
112	1	films	ORC	animate	The boy that the video films is happy.	Match	
113	1	hits	SRC	inanimate	The car that hits the boy is green.	Match	
114	2	hits	ORC	inanimate	The car that the boy hits is green.	Mismatch	
115	3	hits	SRC	animate	The boy that hits the car is naughty.	Mismatch	
116	4	hits	ORC	animate	The boy that the car hits is naughty.	Match	
117	2	moves	SRC	inanimate	The truck that moves the boy is yellow.	Match	
118	1	moves	ORC	inanimate	The truck that the boy moves is yellow.	Mismatch	
119	4	moves	SRC	animate	The boy that moves the truck is cute.	Mismatch	
120	3	moves	ORC	animate	The boy that the truck moves is cute.	Match	
121	3	protects	SRC	inanimate	The umbrella that protects the girl is yellow.	Match	
122	4	protects	ORC	inanimate	The umbrella that the girl protects is yellow.	Mismatch	
123	1	protects	SRC	animate	The girl that protects the umbrella is thin.	Mismatch	
124	2	protects	ORC	animate	The girl that the umbrella protects is thin.	Match	
125	4	wraps	SRC	inanimate	The quilt that wraps the man is blue.	Match	
126	3	wraps	ORC	inanimate	The quilt that the man wraps is blue.	Mismatch	
127	2	wraps	SRC	animate	The man that wraps the quilt is cold.	Mismatch	

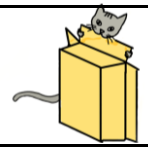
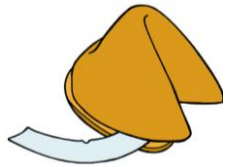
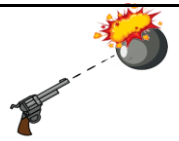
128	1	wraps	ORC	animate	The man that the quilt wraps is cold.	Match	
129	1	cuts	SRC	animate	The woman that cuts the paper is thin.	Match	
130	2	cuts	ORC	animate	The woman that the paper cuts is thin.	Mismatch	
131	3	cuts	SRC	inanimate	The paper that cuts the woman is blue.	Mismatch	
132	4	cuts	ORC	inanimate	The paper that the woman cuts is blue.	Match	
133	2	photographs	SRC	animate	The woman that photographs the camera is tall.	Match	
134	1	photographs	ORC	animate	The woman that the camera photographs is thin.	Mismatch	
135	4	photographs	SRC	inanimate	The camera that photographs the woman is new.	Mismatch	
136	3	photographs	ORC	inanimate	The camera that the woman photographs is new.	Match	
137	3	introduces	SRC	animate	The woman that introduces the news is happy.	Match	
138	4	introduces	ORC	animate	The woman that the news introduces is happy.	Mismatch	
139	1	introduces	SRC	inanimate	The news that introduces the woman is objective.	Mismatch	
140	2	introduces	ORC	inanimate	The news that the woman introduces is objective.	Match	
141	4	pulls	SRC	animate	The horse that pulls the cart is grey.	Match	
142	3	pulls	ORC	animate	The horse that the cart pulls is grey.	Mismatch	
143	2	pulls	SRC	inanimate	The cart that pulls the horse is red.	Mismatch	
144	1	pulls	ORC	inanimate	The cart that the horse pulls is red.	Match	
145	1	pushes	SRC	animate	The man that pushes the door is happy.	Match	
146	2	pushes	ORC	animate	The man that the door pushes is happy.	Mismatch	
147	3	pushes	SRC	inanimate	The door that pushes the man is new.	Mismatch	
148	4	pushes	ORC	inanimate	The door that the man pushes is new.	Match	
149	2	shows	SRC	animate	The man that shows the picture is wise.	Match	
150	1	shows	ORC	animate	The man that the picture shows is wise.	Mismatch	
151	4	shows	SRC	inanimate	The picture that shows the man is colourful.	Mismatch	
152	3	shows	ORC	inanimate	The picture that the man shows is colourful.	Match	
153	3	tells	SRC	animate	The woman that tells the phone is smart	Match	

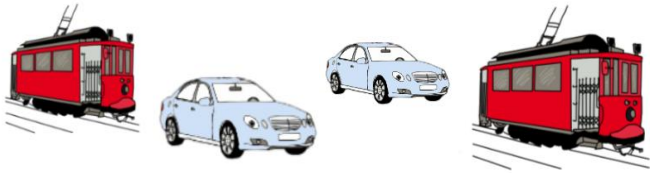
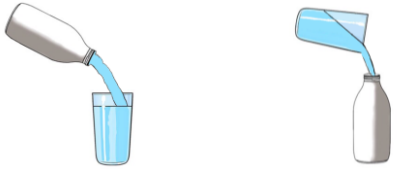
154	4	tells	ORC	animate	The woman that the phone tells is smart.	Mismatch	
155	1	tells	SRC	inanimate	The phone that tells the woman is old.	Mismatch	
156	2	tells	ORC	inanimate	The phone that the woman tells is old.	Match	
157	4	wets	SRC	animate	The woman that wets the t-shirt is happy.	Match	
158	3	wets	ORC	animate	The woman that the t-shirt wets is happy.	Mismatch	
159	2	wets	SRC	inanimate	The t-shirt that wets the woman is blue.	Mismatch	
160	1	wets	ORC	inanimate	The t-shirt that the woman wets is blue.	Match	

Appendix 14: Test items of the offline comprehension test

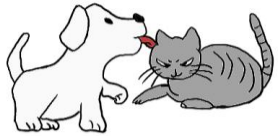
NO.	Version	Verb	Type	Animacy	Stimuli	Pictures
1	1	Blindfolds	SRC	animate	The boy that blindfolds the girl is cute.	
2	2	Blindfolds	ORC	animate	The boy that the girl blindfolds is cute.	
3	3	Blindfolds	SRC	animate	The girl that blindfolds the boy is cute.	
4	4	Blindfolds	ORC	animate	The girl that the boy blindfolds is cute.	
5	2	Feeds	SRC	animate	The man that feeds the woman has grey hair.	
6	1	Feeds	ORC	animate	The man that the woman feeds has grey hair.	
7	4	Feeds	SRC	animate	The woman that feeds the man has grey hair.	
8	3	Feeds	ORC	animate	The woman that the man feeds has grey hair.	
9	1	Greets	SRC	animate	The man that greets the boy is tall.	
10	2	Greets	ORC	animate	The man that the boy greets is tall.	
11	3	Greets	SRC	animate	The boy that greets the man is short.	
12	4	Greets	ORC	animate	The boy that the man greets is short.	
13	2	Helps	SRC	animate	The girl that helps the boy has brown hair.	
14	1	Helps	ORC	animate	The girl that the boy helps has brown hair.	
15	4	Helps	SRC	animate	The boy that helps the girl has brown hair.	
16	3	Helps	ORC	animate	The boy that the girl helps has brown hair.	
17	1	Calls	SRC	animate	The man that calls the girl is happy.	
18	2	Calls	ORC	animate	The man that the girl calls is happy.	
19	3	Calls	SRC	animate	The girl that calls the man is happy.	
20	4	Calls	ORC	animate	The girl that the man calls is happy.	
21	2	Hears	SRC	animate	The boy that hears the girl has blond hair.	
22	1	Hears	ORC	animate	The boy that the girl hears has blond hair.	
23	4	Hears	SRC	animate	The girl that hears the boy has blond hair.	

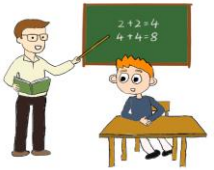
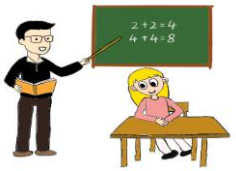
24	3	Hears	ORC	animate	The girl that the boy hears has blond hair.		
25	1	Kisses	SRC	animate	The woman that kisses the boy has black hair.		
26	2	Kisses	ORC	animate	The woman that the boy kisses has black hair.		
27	3	Kisses	SRC	animate	The boy that kisses the woman has black hair.		
28	4	Kisses	ORC	animate	The boy that the woman kisses has black hair.		
29	2	Pushes	SRC	animate	The man that pushes the woman is short.		
30	1	Pushes	ORC	animate	The man that the woman pushes is short.		
31	4	Pushes	SRC	animate	The woman that pushes the man is short.		
32	3	Pushes	ORC	animate	The woman that the man pushes is short.		
33	1	Burns	SRC	animate	The man that burns the paper is scared.		
34	2	Burns	ORC	animate	The man that the paper burns is scared.		
35	3	Burns	SRC	inanimate	The paper that burns the man is hot.		
36	4	Burns	ORC	inanimate	The paper that the man burns is hot.		
37	4	Chases	SRC	animate	The dog that chases the wheel is cute.		
38	3	Chases	ORC	animate	The dog that the wheel chases is cute.		
39	2	Chases	SRC	inanimate	The wheel that chases the dog is big.		
40	1	Chases	ORC	inanimate	The wheel that the dog chases is big.		
41	1	Cleans	SRC	animate	The girl that cleans the towel is in pink.		
42	2	Cleans	ORC	animate	The girl that the towel cleans is in pink.		
43	3	Cleans	SRC	inanimate	The towel that cleans the girl is purple.		
44	4	Cleans	ORC	inanimate	The towel that the girl cleans is purple.		
45	4	Destroys	SRC	animate	The man that destroys the helicopter is brave.		
46	3	Destroys	ORC	animate	The man that the helicopter destroys is brave.		
47	2	Destroys	SRC	inanimate	The helicopter that destroys the man is new.		
48	1	Destroys	ORC	inanimate	The helicopter that the man destroys is new.		
49	1	Carries	SRC	animate	The cat that carries the box is grey.		

50	2	carries	ORC	animate	The cat that the box carries is grey.	 
51	3	Carries	SRC	inanimate	The box that carries the cat is big.	
52	4	Carries	ORC	inanimate	The box that the cat carries is big.	
53	4	Freezes	SRC	animate	The man that freezes the water is cold.	 
54	3	Freezes	ORC	animate	The man that the water freezes is cold.	
55	2	Freezes	SRC	inanimate	The water that freezes the man is cold.	
56	1	Freezes	ORC	inanimate	The water that the man freezes is cold.	
57	1	Hides	SRC	animate	The cat that hides the box is cute.	 
58	2	Hides	ORC	animate	The cat that the box hides is cute.	
59	3	Hides	SRC	inanimate	The box that hides the cat is big.	
60	4	Hides	ORC	inanimate	The box that the cat hides is big.	
61	4	Pulls	SRC	animate	The boy that pulls the balloon is happy.	 
62	3	Pulls	ORC	animate	The boy that the balloon pulls is happy.	
63	2	Pulls	SRC	inanimate	The balloon that pulls the boy is yellow.	
64	1	Pulls	ORC	inanimate	The balloon that the boy pulls is yellow.	
65	1	Holds	SRC	inanimate	The bag that contains the suitcase is green.	 
66	2	Holds	ORC	inanimate	The bag that the suitcase contains is green.	
67	3	Holds	SRC	inanimate	The suitcase that contains the bag is brown.	
68	4	Holds	ORC	inanimate	The suitcase that the bag contains is brown.	
69	2	Covers	SRC	inanimate	The paper that covers the cookie is clean.	 
70	1	Covers	ORC	inanimate	The paper that the cookie covers is clean.	
71	4	Covers	SRC	inanimate	The cookie that covers the paper is tasty.	
72	3	Covers	ORC	inanimate	The cookie that the paper covers is tasty.	
73	1	Explodes	SRC	inanimate	The gun that explodes the bomb is small.	 
74	2	Explodes	ORC	inanimate	The gun that the bomb explodes is small.	
75	3	Explodes	SRC	inanimate	The bomb that explodes the gun is big.	

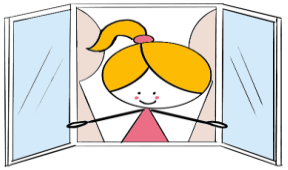
76	4	Explodes	ORC	inanimate	The bomb that the gun explodes is big.	
77	2	Hits	SRC	inanimate	The hammer that hits the stone is big.	
78	1	Hits	ORC	inanimate	The hammer that the stone hits is big.	
79	4	Hits	SRC	inanimate	The stone that hits the hammer is hard.	
80	3	Hits	ORC	inanimate	The stone that the hammer hits is hard.	
81	1	Chases	SRC	inanimate	The tram that chases the car is fast.	
82	2	Chases	ORC	inanimate	The tram that the car chases is fast.	
83	3	Chases	SRC	inanimate	The car that chases the tram is fast.	
84	4	Chases	ORC	inanimate	The car that the tram chases is fast.	
85	2	Fills	SRC	inanimate	The bottle that fills the glass is new.	
86	1	Fills	ORC	inanimate	The bottle that the glass fills is new.	
87	4	Fills	SRC	inanimate	The glass that fills the bottle is new.	
88	3	Fills	ORC	inanimate	The glass that the bottle fills is new.	
89	1	Hangs	SRC	inanimate	The rope that hangs the hook is long.	
90	2	Hangs	ORC	inanimate	The rope that the hook hangs is long.	
91	3	Hangs	SRC	inanimate	The hook that hangs the rope is sharp.	
92	4	Hangs	ORC	inanimate	The hook that the rope hangs is sharp.	
93	2	Contains	SRC	inanimate	The water that contains the boat is deep.	
94	1	Contains	ORC	inanimate	The water that the boat contains is deep.	
95	4	Contains	SRC	inanimate	The boat that holds the water is small.	
96	3	Contains	ORC	inanimate	The boat that the water contains is small.	

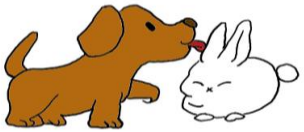
Appendix 15: Test items of the oral production test

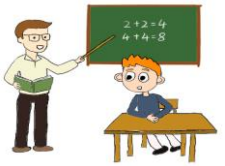
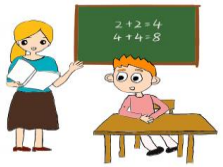
NO.	Version	Verb	Type	Animacy	Stimuli	Answer	Pictures
1	1	helps	SRC	animate	In the first picture, the girl has blond hair. The girl helps the woman. In the second picture, the girl has black hair. The girl helps the man. Which girl has blond hair?	The girl that helps the woman has blond hair.	
2	2	helps	SRC	animate	In the first picture, the girl has blond hair. The girl helps the woman. In the second picture, the girl has black hair. The girl helps the man. Which girl has black hair?	The girl that helps the man has black hair.	
3	1	licks	SRC	animate	In the first picture, the dog is brown. The dog licks the rabbit. In the second picture, the dog is white. The dog licks the cat. Which dog is brown?	The dog that licks the rabbit is brown.	
4	2	licks	SRC	animate	In the first picture, the dog is brown. The dog licks the rabbit. In the second picture, the dog is white. The dog licks the cat. Which dog is white?	The dog that licks the cat is white.	
5	1	lifts	SRC	animate	In the first picture, the man wears glasses. The man lifts the boy. In the second picture, the man wears a jacket. The man lifts the girl. Which man wears glasses?	The man that lifts the boy wears glasses?	

6	2	lifts	SRC	animate	In the first picture, the man wears glasses. The man lifts the boy. In the second picture, the man wears a jacket. The man lifts the girl. Which man wears a jacket?	The man that lifts the girl wears a jacket?	
7	3	teaches	SRC	animate	In the first picture, the man has brown hair. The man teaches the boy. In the second picture, the man has black hair. The man teaches the girl. Which man has brown hair?	The man that teaches the boy has brown hair.	
8	4	teaches	SRC	animate	In the first picture, the man has brown hair. The man teaches the boy. In the second picture, the man has black hair. The man teaches the girl. Which man has black hair?	The man that teaches the girl has black hair.	
9	3	washes	SRC	animate	In the first picture, the girl has long hair. The girl washes the dog. In the second picture, the girl has short hair. The girl washes the cat. Which girl has long hair?	The girl that washes the dog has long hair.	
10	4	washes	SRC	animate	In the first picture, the girl has long hair. The girl washes the dog. In the second picture, the girl has short hair. The girl washes the cat. Which girl has short hair?	The girl that washes the cat has short hair.	

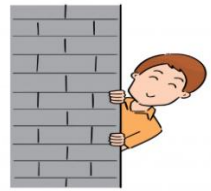
11	1	bites	SRC	animate	In the first picture, the girl has blond hair. The girl bites the apple. In the second picture, the girl has black hair. The girl bites the candy. Which girl has blond hair?	The girl that bites the apple has blond hair.	
12	2	bites	SRC	animate	In the first picture, the girl has blond hair. The girl bites the apple. In the second picture, the girl has black hair. The girl bites the candy. Which girl has black hair?	The girl that bites the candy has black hair.	
13	1	catches	SRC	animate	In the first picture, the boy wears a cap. The boy catches the baseball. In the second picture, the boy has blond hair. The boy catches the football. Which boy wears a cap?	The boy that catches the baseball wears a cap.	
14	2	catches	SRC	animate	In the first picture, the boy wears a cap. The boy catches the baseball. In the second picture, the boy has blond hair. The boy catches the football. Which boy has blond hair?	The boy that catches the football has blond hair.	
15	1	drives	SRC	animate	In the first picture, the man wears a cap. The man drives a truck. In the second picture, the man has brown hair. The man drives a car. Which man wears a cap?	The man that drives the truck wears a cap.	

16	2	drives	SRC	animate	In the first picture, the man wears a cap. The man drives a truck. In the second picture, the man has brown hair. The man drives a car. Which man has brown hair?	The man that drives the car has brown hair.	
17	3	eats	SRC	animate	In the first picture, the boy has brown hair. The boy eats the ice-cream. In the second picture, the boy has blond hair. The boy eats the corns. Which boy has brown hair?	The boy that eats the ice-cream has brown hair.	
18	4	eats	SRC	animate	In the first picture, the boy has brown hair. The boy eats the ice-cream. In the second picture, the boy has blond hair. The boy eats the corns. Which boy has blond hair?	The boy that eats the corns has blond hair.	
19	3	opens	SRC	animate	In the first picture, the girl has short hair. The girl opens the door. In the second picture, the girl has long hair. The girl opens the window. Which girl has short hair?	The girl that opens the door has short hair.	
20	4	opens	SRC	animate	In the first picture, the girl has short hair. The girl opens the door. In the second picture, the girl has long hair. The girl opens the window. Which girl has long hair?	The girl that opens the window has long hair.	

21	3	helps	ORC	animate	In the first picture, the woman has long hair. The boy helps the woman. In the second picture, the woman has short hair. The girl helps the woman. Which woman has long hair?	The woman that the boy helps has long hair.	
22	4	helps	ORC	animate	In the first picture, the woman has long hair. The boy helps the woman. In the second picture, the woman has short hair. The girl helps the woman. Which woman has short hair?	The woman that the girl helps has short hair.	
23	3	licks	ORC	animate	In the first picture, the rabbit is white. The dog licks the rabbit. In the second picture, the rabbit is grey. The snake licks the rabbit. Which rabbit is white?	The rabbit that the dog licks is white.	
24	4	licks	ORC	animate	In the first picture, the rabbit is white. The dog licks the rabbit. In the second picture, the rabbit is grey. The snake licks the rabbit. Which rabbit is grey?	The rabbit that the snake licks is grey.	
25	3	lifts	ORC	animate	In the first picture, the boy has blond hair. The woman lifts the boy. In the second picture, the boy has black hair. The man lifts the boy. Which boy has blond hair?	The boy that the woman lifts has blond hair.	

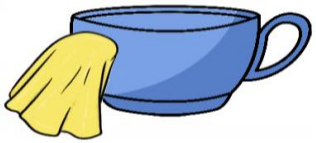
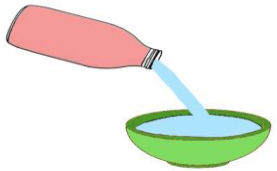
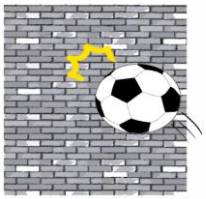
26	4	lifts	ORC	animate	In the first picture, the boy has blond hair. The woman lifts the boy. In the second picture, the boy has black hair. The man lifts the boy. Which boy has black hair?	The boy that the man lifts has black hair.	
27	1	teaches	ORC	animate	In the first picture, the boy wears a blue t-shirt. The man teaches the boy. In the second picture, the boy wears a pink t-shirt. The woman teaches the boy. Which boy wears a blue t-shirt?	The boy that the man teaches wears a blue t-shirt.	
28	2	teaches	ORC	animate	In the first picture, the boy wears a blue t-shirt. The man teaches the boy. In the second picture, the boy wears a pink t-shirt. The woman teaches the boy. Which boy wears a pink t-shirt?	The boy that the woman teaches wears a pink t-shirt.	
29	1	washes	ORC	animate	In the first picture, the dog is grey. The boy washes the dog. In the second picture, the dog is brown. The girl washes the dog. Which dog is grey?	The dog that the boy washes is grey.	
30	2	washes	ORC	animate	In the first picture, the dog is grey. The boy washes the dog. In the second picture, the dog is brown. The girl washes the dog. Which dog is brown?	The dog that the girl washes is brown.	

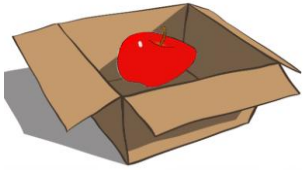
31	1	burns	ORC	animate	In the first picture, the girl has blond hair. The pot burns the girl. In the second picture, the girl has brown hair. The candle burns the girl. Which girl has blond hair?	The girl that the pot burns has blond hair.	
32	2	burns	ORC	animate	In the first picture, the girl has blond hair. The pot burns the girl. In the second picture, the girl has brown hair. The candle burns the girl. Which girl has brown hair?	The girl that the candle burns has brown hair.	
33	1	carries	ORC	animate	In the first picture, the man wears a green t-shirt. The balloon carries the man. In the second picture, the man wears a hat. The truck carries the man. Which man wears a green t-shirt?	The man that the balloon carries wears a green t-shirt.	
34	2	carries	ORC	animate	In the first picture, the man wears a green t-shirt. The balloon carries the man. In the second picture, the man wears a hat. The truck carries the man. Which man wears a hat?	The man that the truck carries wears a hat.	
35	1	cools	ORC	animate	In the first picture, the man wears blue shorts. The fan cools the man. In the second picture, the man wears white shorts. The water cools the man. Which man wears blue shorts?	The man that the fan cools wears blue shorts.	

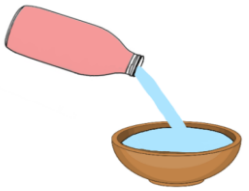
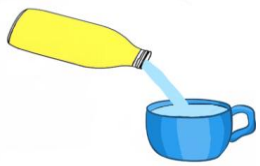
36	2	cools	ORC	animate	In the first picture, the man wears blue shorts. The fan cools the man. In the second picture, the man wears white shorts. The water cools the man. Which man wears white shorts?	The man that the water cools wears white shorts.	
37	3	hides	ORC	animate	In the first picture, the man is scared. The table hides the man. In the second picture, the man is happy. The wall hides the man. Which the man is scared?	The man that the table hides is scared.	
38	4	hides	ORC	animate	In the first picture, the man is scared. The table hides the man. In the second picture, the man is happy. The wall hides the man. Which the man is happy?	The man that the wall hides is happy.	
39	3	presents	ORC	animate	In the first picture, the woman has short hair. The computer presents the woman. In the second picture, the woman has long hair. The television presents the woman. Which woman has short hair?	The woman that the computer presents has short hair.	
40	4	presents	ORC	animate	In the first picture, the woman has short hair. The computer presents the woman. In the second picture, the woman has long hair. The television presents the woman. Which woman has long hair?	The woman that the television presents has long hair.	

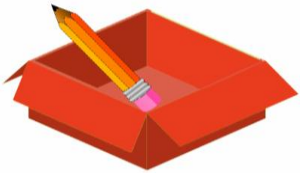
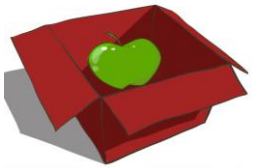
41	3	bites	ORC	inanimate	In the first picture, the apple is green. The girl bites the apple. In the second picture, the apple is red. The worm bites the apple. Which apple is green?	The apple that the girl bites is green.	
42	4	bites	ORC	inanimate	In the first picture, the apple is green. The girl bites the apple. In the second picture, the apple is red. The worm bites the apple. Which apple is red?	The apple that the worm bites is red.	
43	3	catches	ORC	inanimate	In the first picture, the ball is a baseball. The woman catches the ball. In the second picture, the ball is a football. The man catches the ball. Which ball is a baseball?	The ball that the woman catches is a baseball.	
44	4	catches	ORC	inanimate	In the first picture, the ball is a baseball. The woman catches the ball. In the second picture, the ball is a football. The man catches the ball. Which ball is a football?	The ball that the man catches is a football.	
45	3	drives	ORC	inanimate	In the first picture, the car is blue. The man drives the car. In the second picture, the car is red. The woman drives the car. Which car is blue?	The car that the man drives is blue.	

46	4	drives	ORC	inanimate	In the first picture, the car is blue. The man drives the car. In the second picture, the car is red. The woman drives the car. Which car is red?	The car that the woman drives is red.	
47	1	eats	ORC	inanimate	In the first picture, the ice-cream is pink. The girl eats the ice-cream. In the second picture, the ice-cream is yellow. The boy eats the ice-cream. Which ice-cream is pink?	The ice-cream that the girl eats is pink.	
48	2	eats	ORC	inanimate	In the first picture, the ice-cream is pink. The girl eats the ice-cream. In the second picture, the ice-cream is yellow. The boy eats the ice-cream. Which ice-cream is yellow?	The ice-cream that the boy eats is yellow.	
49	1	opens	ORC	inanimate	In the first picture, the door is brown. The boy opens the door. In the second picture, the door is blue. The girl opens the door. Which door is brown?	The door that the boy opens is brown.	
50	2	opens	ORC	inanimate	In the first picture, the door is brown. The boy opens the door. In the second picture, the door is blue. The girl opens the door. Which door is blue?	The door that the girl opens is blue.	

51	1	cleans	ORC	inanimate	In the first picture, the cup is white. The brush cleans the cup. In the second picture, the cup is blue. The cloth cleans the cup. Which cup is white?	The cup that the brush cleans is white.	
52	2	cleans	ORC	inanimate	In the first picture, the cup is white. The brush cleans the cup. In the second picture, the cup is blue. The cloth cleans the cup. Which cup is blue?	The cup that the cloth cleans is blue.	
53	1	fills	ORC	inanimate	In the first picture, the bowl is brown. The kettle fills the bowl. In the second picture, the bowl is green. The bottle fills the bowl. Which bowl is brown?	The bowl that the kettle fills is brown.	
54	2	fills	ORC	inanimate	In the first picture, the bowl is brown. The kettle fills the bowl. In the second picture, the bowl is green. The bottle fills the bowl. Which bowl is green?	The bowl that the bottle fills is green.	
55	1	hits	ORC	inanimate	In the first picture, the wall is grey. The ball hits the wall. In the second picture, the wall is red. The car hits the wall. Which wall is grey?	The wall that the ball hits is grey.	

56	2	hits	ORC	inanimate	In the first picture, the wall is grey. The ball hits the wall. In the second picture, the wall is red. The car hits the wall. Which wall is red?	The wall that the car hits is red.	
57	3	holds	ORC	inanimate	In the first picture, the book is blue. The box holds the book. In the second picture, the book is yellow. The bag holds the book. Which book is blue?	The book that the box holds is blue.	
58	4	holds	ORC	inanimate	In the first picture, the book is blue. The box holds the book. In the second picture, the book is yellow. The bag holds the book. Which book is yellow?	The book that the bag holds is yellow.	
59	3	keeps	ORC	inanimate	In the first picture, the apple is green. The bag keeps the apple. In the second picture, the apple is red. The box keeps the apple. Which apple is green?	The apple that the bag keeps is green.	
60	4	keeps	ORC	inanimate	In the first picture, the apple is green. The bag keeps the apple. In the second picture, the apple is red. The box keeps the apple. Which apple is red?	The apple that the box keeps is red.	

61	3	cleans	SRC	inanimate	In the first picture, the brush is green. The brush cleans the cup. In the second picture, the brush is blue. The brush cleans the bowl. Which brush is green?	The brush that cleans the cup is green.	
62	4	cleans	SRC	inanimate	In the first picture, the brush is green. The brush cleans the cup. In the second picture, the brush is blue. The brush cleans the bowl. Which brush is blue?	The brush that cleans the bowl is blue.	
63	3	fills	SRC	inanimate	In the first picture, the bottle is pink. The bottle fills the bowl. In the second picture, the bottle is yellow. The bottle fills the cup. Which bottle is pink?	The bottle that fills the bowl is pink.	
64	4	fills	SRC	inanimate	In the first picture, the bottle is pink. The bottle fills the bowl. In the second picture, the bottle is yellow. The bottle fills the cup. Which bottle is yellow?	The bottle that fills the cup is yellow.	
65	3	hits	SRC	inanimate	In the first picture, the car is orange. The car hits the tree. In the second picture, the car is green. The car hits the wall. Which car is orange?	The car that hits the tree is orange.	

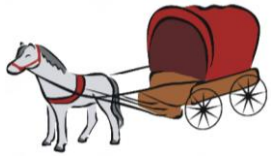
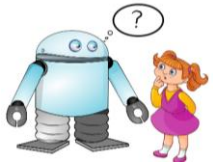
66	4	hits	SRC	inanimate	In the first picture, the car is orange. The car hits the tree. In the second picture, the car is green. The car hits the wall. Which car is green?	The car that hits the wall is green.	
67	1	holds	SRC	inanimate	In the first picture, the box is green. The box holds the book. In the second picture, the box is red. The box holds the pencil. Which box is green?	The box that holds the book is green.	
8	2	holds	SRC	inanimate	In the first picture, the box is green. The box holds the book. In the second picture, the box is red. The box holds the pencil. Which box is red?	The box that holds the pencil is red.	
69	1	keeps	SRC	inanimate	In the first picture, the box is brown. The box keeps the orange. In the second picture, the box is red. The box keeps the apple. Which box is brown?	The box that keeps the orange is brown.	
70	2	keeps	SRC	inanimate	In the first picture, the box is brown. The box keeps the orange. In the second picture, the box is red. The box keeps the apple. Which box is red?	The box that keeps the apple is red.	

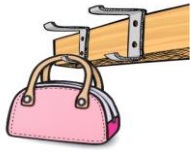
71	3	burns	SRC	inanimate	In the first picture, the pot is grey. The pot burns the girl. In the second picture, the pot is red. The pit burns the boy. Which pot is grey?	The pot that burns the girl is grey.	
72	4	burns	SRC	inanimate	In the first picture, the pot is grey. The pot burns the girl. In the second picture, the pot is red. The pit burns the boy. Which pot is red?	The pot that burns the boy is red.	
73	3	carries	SRC	inanimate	In the first picture, the balloon is red. The balloon carries the man. In the second picture, the balloon is yellow. The balloon carries the girl. Which balloon is red?	The balloon that carries the man is red.	
74	4	carries	SRC	inanimate	In the first picture, the balloon is red. The balloon carries the man. In the second picture, the balloon is yellow. The balloon carries the girl. Which balloon is yellow?	The balloon that carries the woman is yellow.	
75	3	cools	SRC	inanimate	In the first picture, the fan is pink. The fan cools the woman. In the second picture, the fan is orange. The fan cools the man. Which fan is pink?	The fan that cools the woman is pink.	

76	4	cools	SRC	inanimate	In the first picture, the fan is pink. The fan cools the woman. In the second picture, the fan is orange. The fan cools the man. Which fan is orange?	The fan that cools the man is orange.	
77	1	hides	SRC	inanimate	In the first picture, the table is brown. The table hides the girl. In the second picture, the table is grey. The table hides the boy. Which table is brown?	The table that hides the girl is brown.	
78	2	hides	SRC	inanimate	In the first picture, the table is brown. The table hides the girl. In the second picture, the table is grey. The table hides the boy. Which table is grey?	The table that hides the boy is grey.	
79	1	presents	SRC	inanimate	In the first picture, the television is yellow. The television presents the man. In the second picture, the television is grey. The television presents the woman. Which television is yellow?	The television that presents the man is yellow.	
80	2	presents	SRC	inanimate	In the first picture, the television is yellow. The television presents the man. In the second picture, the television is grey. The television presents the woman. Which television is grey?	The television that presents the woman is grey.	

Appendix 16: Test items of the metalinguistic knowledge test

NO.	Ver-sion	Verb	Type	Animacy	Stimuli	Match/ Mismatch	Pictures
1	3	amuses	SRC	animate	The girl that amuses the boy is cute.	Mismatch	
2	4	amuses	ORC	animate	The girl that the boy amuses is cute.	Match	
3	1	amuses	SRC	animate	The boy that amuses the girl is cute.	Match	
4	2	amuses	ORC	animate	The boy that the girl amuses is cute.	Mismatch	
5	3	calls	SRC	animate	The man that calls the woman has short hair.	Match	
6	4	calls	ORC	animate	The man that the woman calls has short hair.	Mismatch	
7	1	calls	SRC	animate	The woman that calls the man has long hair.	Mismatch	
8	2	calls	ORC	animate	The woman that the man calls has long hair.	Match	
9	2	finds	SRC	animate	The boy that finds the girl is thin.	Match	
10	3	finds	ORC	animate	The boy that the girl finds is thin.	Mismatch	
11	4	finds	SRC	animate	The girl that finds the boy is thin.	Mismatch	
12	1	finds	ORC	animate	The girl that the boy finds is thin.	Match	
13	4	visits	SRC	animate	The woman that visits the girl wears a dress.	Match	
14	1	visits	ORC	animate	The woman that the girl visits wears a dress.	Mismatch	
15	2	visits	SRC	animate	The girl that visits the woman wears a dress.	Mismatch	
16	3	visits	ORC	animate	The girl that the woman visits wears a dress.	Match	
17	1	photographs	SRC	animate	The woman that photographs the camera is thin.	Match	
18	2	photographs	ORC	animate	The woman that the camera photographs is thin.	Mismatch	
19	3	photographs	SRC	inanimate	The camera that photographs the woman is new.	Mismatch	
20	4	photographs	ORC	inanimate	The camera that the woman photographs is new.	Match	

21	3	carries	SRC	animate	The dog that carries the basket is brown.	Match	
22	4	carries	ORC	animate	The dog that the basket carries is brown.	Mismatch	
23	1	carries	SRC	inanimate	The basket that carries the dog is big.	Mismatch	
24	2	carries	ORC	inanimate	The basket that the dog carries is big.	Match	
25	2	pulls	SRC	animate	The horse that pulls the cart is tall.	Match	
26	3	pulls	ORC	animate	The horse that the cart pulls is tall.	Mismatch	
27	4	pulls	SRC	inanimate	The cart that pulls the horse is red.	Mismatch	
28	1	pulls	ORC	inanimate	The cart that the horse pulls is red.	Match	
29	4	wets	SRC	animate	The woman that wets the t-shirt is tall.	Match	
30	1	wets	ORC	animate	The woman that the t-shirt wets is tall.	Mismatch	
31	2	wets	SRC	inanimate	The t-shirt that wets the woman is blue.	Mismatch	
32	3	wets	ORC	inanimate	The t-shirt that the woman wets is blue.	Match	
33	1	asks	SRC	inanimate	The robot that asks the girl is big.	Match	
34	2	asks	ORC	inanimate	The robot that the girl asks is big.	Mismatch	
35	3	asks	SRC	animate	The girl that asks the robot is tall.	Mismatch	
36	4	asks	ORC	animate	The girl that the robot asks is tall.	Match	
37	3	follows	SRC	inanimate	The ship that follows the whale is red.	Match	
38	4	follows	ORC	inanimate	The ship that the whale follows is red.	Mismatch	
39	1	follows	SRC	animate	The whale that follows the ship is blue.	Mismatch	
40	2	follows	ORC	animate	The whale that the ship follows is blue.	Match	
41	2	dries	SRC	inanimate	The towel that dries the man is green.	Match	
42	3	dries	ORC	inanimate	The towel that the man dries is green.	Mismatch	
43	4	dries	SRC	animate	The man that dries the towel is tall.	Mismatch	
44	1	dries	ORC	animate	The man that the towel dries is tall.	Match	

45	4	hides	SRC	inanimate	The box that hides the boy is brown.	Match	
46	1	hides	ORC	inanimate	The box that the boy hides is brown.	Mismatch	
47	2	hides	SRC	animate	The man that hides the box is scared.	Mismatch	
48	3	hides	ORC	animate	The man that the box hides is scared.	Match	
49	1	chases	SRC	inanimate	The boat that chases the ship is small.	Match	
50	2	chases	ORC	inanimate	The boat that the ship chases is small.	Mismatch	
51	3	chases	SRC	inanimate	The ship that chases the boat is big.	Mismatch	
52	4	chases	ORC	inanimate	The ship that the boat chases is big.	Match	
53	3	cleans	SRC	inanimate	The sponge cleans the water is yellow.	Match	
54	4	cleans	ORC	inanimate	The sponge that the water cleans is yellow.	Mismatch	
55	1	cleans	SRC	inanimate	The water that cleans the sponge is cold.	Mismatch	
56	2	cleans	ORC	inanimate	The water that the sponge cleans is cold.	Match	
57	2	hangs	SRC	inanimate	The hook that hangs the bag is new.	Match	
58	3	hangs	ORC	inanimate	The hook that the bag hangs is new.	Mismatch	
59	4	hangs	SRC	inanimate	The bag that hangs the hook is pink.	Mismatch	
60	1	hangs	ORC	inanimate	The bag that the hook hangs is pink.	Match	
61	4	messes	SRC	inanimate	The ice-cream that messes the soup is sweet.	Match	
62	1	messes	ORC	inanimate	The ice-cream that the soup messes is sweet.	Mismatch	
63	2	messes	SRC	inanimate	The soup that messes the ice-cream is sweet.	Mismatch	
64	3	messes	ORC	inanimate	The soup that the ice-cream messes is sweet.	Match	

Appendix 17 RQ1: AIC and LRT results for the offline comprehension test

Native speaker

Table Appx. 1 AIC results (converged models only) for the offline comprehension test for native speakers

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type; by-item random slope of relative clause type	324.16
Model2	relative clause type	by-subject and by-item random intercepts	by-item random slope of relative clause type	335.89
Model3	relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type;	323.85
Model4	relative clause type	by-subject and by-item random intercepts	N/A	321.37

Table Appx. 2 LRT results (converged models only) for the offline comprehension test for native speakers

Model comparisons	LRT
Model1:Model2	$X^2(-9)=29.74$, $p<.001$
Model1:Model3	$X^2(-9)=17.69$, $p=.039$
Model1:Model4	$X^2(-18)=33.21$, $p=.016$
Model2:Model4	$X^2(-9)=3.48$, $p=.942$
Model3:Model4	$X^2(-9)=15.52$, $p=.078$

L2 learners

Table Appx. 3 AIC results (converged models only) for the offline comprehension test for L2 learners

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type; by-item random slope of relative clause type	1589.63
Model2	relative clause type	by-subject and by-item random intercepts	by-item random slope of relative clause type	1575.77
Model3	relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type	1573.81
Model4	relative clause type	by-subject and by-item random intercepts	N/A	failed to converge

Table Appx. 4 LRT results (converged models only) for the offline comprehension test for L2 learners

Model comparisons	LRT
Model1:Model2	X^2 (-9)=4.14, *p*=.902
Model1:Model3	X^2 (-9)=2.17, *p*=.988

Appendix 18 RQ1: AIC and LRT results for the self-paced reading test

Native speakers

First critical word

Table Appx. 5 AIC results (converged models only) for the SPR test at the first critical word for native speakers

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	N/A	7138.31
Model4_1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type; by-item random slope of match or mismatch	7145.34
Model5_2	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of match or mismatch; by-item random slope of match or mismatch	7143.82

Table Appx. 6 LRT results (converged models only) for the SPR test at the first critical word for native speakers

Model comparisons	LRT
model1:model4_1	X^2 (11)=14.97, *p*=.184
model1:model5_2	X^2 (4)=2.49, *p*=.648

Second critical word

Table Appx. 7 AIC results (converged models only) for the SPR test at the second critical word for native speakers

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	N/A	7106.69
Model3_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch	7126.92
Model4_1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type; by-item random slope of match or mismatch	7126.47
Model4_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type	7122.96
Model5_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of match or mismatch	7109.92
Model6	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-item random slope of the relative clause type and match or mismatch	7110.18

Table Appx. 8 LRT results (converged models only) for the SPR test at the second critical word for native speakers

Model comparisons	LRT
model1:model3_3	X^2 (14)=7.77, *p*=.901
model1:model4_1	X^2 (11)=2.22, *p*=.998
model1:model4_3	X^2 (9)=1.73, *p*=.995
model4_1:model4_3	X^2 (-2)=.49, *p*=.783
model1:model5_3	X^2 (2)=.76, *p*=.683
model1:model6_1	X^2 (2)=.51, *p*=.776

Third critical word

Table Appx. 9 AIC results (converged models only) for the SPR test at the third critical word for native speakers

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	N/A	7472.09
Model5_2	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of match or mismatch; by-item random slope of match or mismatch	7466.53
Model5_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of match or mismatch	7462.77
Model6_1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-item random slope of match or mismatch	7475.83

Table Appx. 10 LRT results (converged models only) for the SPR test at the third critical word for native speakers

Model comparisons	LRT
model1:model5_2	X^2 (4)=13.56, *p*=.009
model1:model5_3	X^2 (2)=13.32, *p*=.001
model5_2:model5_3	X^2 (-2)=.24, *p*=.887
model1:model6_1	X^2 (4)=.27, *p*=.874
model5_2:model6_1	X^2 (-2)=13.29, *p*=.001

Whole sentence

Table Appx. 11 AIC results (converged models only) for the SPR test for whole sentence for native speakers

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	N/A	7209.60
Model3_2	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch; by-item random slope of match or mismatch	7234.17

Table Appx. 12 LRT results (converged models only) for the SPR test for whole sentence for native speakers

Model comparisons	LRT
model1:model3_2	X^2 (16)=7.43, *p*=.964
model1:model5_3	X^2 (2)=2.03, *p*=.362

L2 learners

First critical word

Table Appx. 13 AIC results (converged models only) for the SPR test at the first critical word for L2 learners

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	N/A	27877.16
Model3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch; by-item random slope of the relative clause type and match or mismatch	27919.56
Model3_1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch; by-item random slope of the relative clause type	27909.94
Model3_2	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch; by-item random slope of match or mismatch	27905.69
Model3_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch	27904.11

Model4_2	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type; by-item random slope of the relative clause type	27900.62
Model4_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type	27894.99
Model5	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of match or mismatch; by-item random slope of the relative clause type and match or mismatch	27896.00
Model5_1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of match or mismatch; by-item random slope of the relative clause type	27886.39
Model5_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of match or mismatch	27880.6
Model6	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-item random slope of the relative clause type and match or mismatch	27892.41

Table Appx. 14 LRT results (converged models only) for the SPR test at the first critical word for L2 learners

Model comparisons	LRT
model1:model3	X^2 (28)=13.60, *p*=.990
model1:model3_1	X^2 (23)=13.22, *p*=.947
model3:model3_1	X^2 (-5)=.38, *p*=.996
model1:model3_2	X^2 (16)=3.46, *p*=.999
model3:model3_2	X^2 (-12)=10.14, *p*=.604

model3_1:model3_2	X^2 (-7)=9.75, *p*=.203
model1:model3_3	X^2 (14)=1.05, *p*=1.000
model3:model3_3	X^2 (-14)=12.56, *p*=.562
model3_1:model3_3	X^2 (-9)=12.17, *p*=.204
model3_2:model3_3	X^2 (-2)=2.42, *p*=.298
model1:model4_2	X^2 (18)=12.54, *p*=.818
model3:model4_2	X^2 (-10)=1.07, *p*=.999
model3_1:model4_2	X^2 (-5)=.68, *p*=.984
model1:model4_3	X^2 (9)=.17, *p*=1.000
model3:model4_3	X^2 (-19)=13.43, *p*=.816
model4_2:model4_3	X^2 (-9)=12.37, *p*=.193
model1:model5	X^2 (16)=13.16, *p*=.661
model3:model5	X^2 (-12)=.44, *p*=1.000
model1:model5_1	X^2 (11)=12.77, *p*=.309
model3:model5_1	X^2 (-17)=.83, *p*=1.000
model5:model5_1	X^2 (-5)=.39, *p*=.996
model1:model5_3	X^2 (2)=.56, *p*=.756
model3:model5_3	X^2 (-26)=13.04, *p*=.984
model5:model5_3	X^2 (-14)=12.60, *p*=.558
model5_1:model5_3	X^2 (-9)=12.21, *p*=.202
model1:model6	X^2 (14)=12.75, *p*=.546
model3:model6	X^2 (-14)=.95, *p*=1.000
model5:model6	X^2 (-2)=.41, *p*=.815

Second critical word

Table Appx. 15 AIC results (converged models only) for the SPR test at the second critical word for L2 learners

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	N/A	27838.95
Model3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch; by-item random slope of	27863.73

			the relative clause type and match or mismatch	
Model3_2	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch; by-item random slope of match or mismatch	27843.02
Model3_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch	27839.17
Model4	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type; by-item random slope of the relative clause type and match or mismatch	27853.9
Model4_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type	27829.31
Model5	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of match or mismatch; by-item random slope of the relative clause type and match or mismatch	27867.73
Model5_2	interaction between relative clause type and match or	by-subject and by-item random intercepts	by-subject random slope of match or mismatch; by-item	27846.78

	mismatch		random slope of match or mismatch	
Model5_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of match or mismatch	27842.95
Model6	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-item random slope of the relative clause type and match or mismatch	27863.73
Model6_1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-item random slope of match or mismatch	27842.78

Table Appx. 16 LRT results (converged models only) for the SPR test at the second critical word for L2 learners

Model comparisons	LRT
model1:model3	X^2 (28)=31.22, *p*=.308
model1:model3_2	X^2 (16)=27.93, *p*=.032
model3:model3_2	X^2 (-12)=3.29, *p*=.993
model1:model3_3	X^2 (14)=27.79, *p*=.015
model3:model3_3	X^2 (-14)=3.43, *p*=.998
model3_1:model3_3	X^2 (-9)=12.17, *p*=.204
model3_2:model3_3	X^2 (-2)=.15, *p*=.930
model1:model4	X^2 (23)=31.05, *p*=.122
model3:model4	X^2 (-5)=.16, *p*=.999
model3_2:model4	X^2 (7)=3.12, *p*=.873
model3_3:model4	X^2 (9)=3.27, *p*=.952
model1:model4_3	X^2 (9)=27.64, *p*=.001
model3:model4_3	X^2 (-19)=3.58, *p*=1.000
model3_2:model4_3	X^2 (-7)=.29, *p*=.999
model3_3:model4_3	X^2 (-5)=.14, *p*=.999
model4:model4_3	X^2 (-14)=3.41, *p*=.998
model1:modell5	X^2 (16)=3.22, *p*=.999
model3:modell5	X^2 (-12)=27.99, *p*=.005
model3_2:model5	X^2 (0)=24.71, *p*<.001

model3_3:model5	X^2 (2)=24.56, *p*<.001
model4:model5	X^2 (-7)=27.84, *p*<.001
model4_3:model5	X^2 (7)=24.42, *p*<.001
model1:model5_2	X^2 (4)=.17, *p*=.997
model3:model5_2	X^2 (-24)=31.05, *p*=.153
model3_2:model5_2	X^2 (-12)=27.76, *p*=.006
model3_3:model5_2	X^2 (-10)=27.61, *p*=.002
model4:model5_2	X^2 (-19)=30.88, *p*=.042
model4_3:model5_2	X^2 (-5)=27.47, *p*<.001
model5:model5_2	X^2 (-12)=3.05, *p*=.995
model1:model5_3	X^2 (2)=.00, *p*=1.000
model3:model5_3	X^2 (-26)=31.22, *p*=.220
model3_2:model5_3	X^2 (-14)=27.93, *p*=.015
model3_3:model5_3	X^2 (-12)=27.78, *p*=.006
model4:model5_3	X^2 (-21)=31.05, *p*=.073
model4_3:model5_3	X^2 (-7)=27.64, *p*<.001
model5:model5_3	X^2 (-14)=3.22, *p*=.999
model5_2:model5_3	X^2 (-2)=.18, *p*=.918
model1:model6	X^2 (14)=3.22, *p*=.999
model3:model6	X^2 (-14)=27.99, *p*=.014
model3_2:model6	X^2 (-2)=24.71, *p*<.001
model3_3:model6	X^2 (0)=24.56, *p*<.001
model4:model6	X^2 (-9)=27.83, *p*=.001
model4_3:model6	X^2 (5)=24.42, *p*<.001
model5:model6	X^2 (-2)=.00, *p*=1
model5_2:model6	X^2 (10)=3.05, *p*=.980
model5_3:model6	X^2 (12)=3.22, *p*=.994
model1:model6_1	X^2 (2)=.17, *p*=.918
model3:model6_1	X^2 (-26)=31.05, *p*=.227
model3_2:model6_1	X^2 (-14)=27.76, *p*=.015
model3_3:model6_1	X^2 (-12)=27.61, *p*=.006
model4:model6_1	X^2 (-21)=30.88, *p*=.076
model4_3:model6_1	X^2 (7)=27.47, *p*<.001
model5:model6_1	X^2 (-14)=3.05, *p*=.999
model5_2:model6_1	X^2 (-2)=.00, *p*=1.000
model5_3:model6_1	X^2 (0)=.17, *p*<.001
model6:model6_1	X^2 (-12)=3.05, *p*=.995

Third critical word

Table Appx. 17 AIC results (converged models only) for the SPR test at the third critical word for L2 learners

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	N/A	28232.54
Model3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch; by-item random slope of the relative clause type and match or mismatch	28259.95
Model4_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type	28246.02
Model5_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of match or mismatch	28223.77

Table Appx. 18 LRT results (converged models only) for the SPR test at the third critical word for L2 learners

Model comparisons	LRT
model1:model3	X^2 (28)=28.59, p=.434
model1:model4_3	X^2 (9)=4.52, p=.874
model3:model4_3	X^2 (-19)=24.07, p=.194
model1:model5_3	X^2 (2)=12.76, p=.002
model3:model5_3	X^2 (-26)=15.82, p=.940

Whole sentence

Table Appx. 19 AIC results (converged models only) for the SPR test for the whole sentence for L2 learners

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	N/A	27389.41
Model3_2	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch; by-item random slope of match or mismatch	27402.44
Model3_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type and match or mismatch	27403.48
Model4_3	interaction between relative clause type and match or mismatch	by-subject and by-item random intercepts	by-subject random slope of the relative clause type	27402.00

Table Appx. 20 LRT results (converged models only) for the SPR test for the whole sentence for L2 learners

Model comparisons	LRT
model1:model3_2	X^2 (16)=18.97, *p*=.270
model1:model3_3	X^2 (14)=13.93, *p*=.455
model3_2:model3_3	X^2 (-2)=5.04, *p*=.080
model1:model4_3	X^2 (9)=5.41, *p*=.797
model3_2:model4_3	X^2 (-7)=13.57, *p*=.059
model3_3:model4_3	X^2 (-5)=8.52, *p*=.130

Appendix 19 RQ1: AIC and LRT results for the eye-tracking test

Native speakers

Select time order vector

Table Appx. 21 AIC results (converged models only) for the eye-tracking test for the native speakers in selecting time order vector

Models	Fixed effects	Random effects	AIC
model1	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	28247.59
model2	interaction between liner + quadratic time vector and relative clause type	by-subject and by-item random intercepts	28261.30
model3	interaction between liner + quadratic + cubic time vector and relative clause type	by-subject and by-item random intercepts	28274.93

Table Appx. 22 LRT results (converged models only) for the eye-tracking test for the native speakers in selecting time order vector

Model comparisons	LRT
Model1:Model2	X^2 (4)=5.72, p=.221
Model1:Model3	X^2 (8)=11.34, p=.183
Model2:Model3	X^2 (4)=5.62, p=.229

Select random effects

Table Appx. 23 AIC results (converged models only) for the eye-tracking test for the native speakers in selecting random effects

Models	fixed effects	random intercept	random slope	AIC
Model1	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	N/A	28247.59
Model4	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of linear time vector and relative clause	28277.69

			type; by-item random slope of linear time vector	
Model5	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of linear time vector and relative clause type; by-item random slope of relative clause type	28291.69
Model6	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of linear time vector and relative clause type	28273.69
Model7	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of linear time vector; by-item random slope of linear time vector and relative clause type	28278.87
Model8	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of linear time vector; by-item random slope of linear time vector	28254.87
Model9	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of linear time	28268.87

			vector; by-item random slope of relative clause type	
Model10	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of linear time vector	28250.87
Model12	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type; by-item random slope of linear time vector	28271.74
Model13	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type; by-item random slope of relative clause type	28285.74
Model15	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-item random slope of linear time vector and relative clause type	28275.59
Model16	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-item random slope of linear time vector	28251.59
Model17	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	by-item random slope of relative clause type	28265.59

Table Appx. 24 LRT results (converged models only) for the eye-tracking test for the native speakers in selecting random effects

Model comparisons	LRT
Model1:Model4	X^2 (16)=1.89, *p*=1.000
Model1:Model5	X^2 (23)=1.89, *p*=1.000
Model1:Model6	X^2 (14)=1.89, *p*=.999
model4:Model6	X^2 (-2)=.00, *p*=1.000
model5:Model6	X^2 (-9)=.00, *p*=1.000
Model1:Model7	X^2 (16)=.71, *p*=1.000
Model1:Model8	X^2 (4)=.71, *p*=.950
Model7:Model8	X^2 (-12)=.00, *p*=1.000
Model1:Model9	X^2 (11)=.71, *p*=1.000
Model7:Model9	X^2 (-5)=.00, *p*=1.000
Model1:Model10	X^2 (2)=.71, *p*=.699
Model7:Model10	X^2 (-14)=.00, *p*=1.00
Model8:Model10	X^2 (-2)=.00, *p*=1.000
Model9:Model10	X^2 (-9)=.00, *p*=1.000
Model1:Model12	X^2 (11)=2.15, *p*=.998
Model4:Model12	X^2 (5)=4.04, *p*=.543
Model1:Model13	X^2 (18)=2.15, *p*=1.000
Model5:Model13	X^2 (-5)=4.04, *p*=.543
Model1:Model15	X^2 (14)=.00, *p*=1.000
Model7:Model15	X^2 (-2)=.71, *p*=.699
Model1:Model16	X^2 (2)=.00, *p*=1.000
Model15:Model16	X^2 (-12)=.00, *p*=1.000
Model1:Model17	X^2 (9)=.00, *p*=1.000
Model15:Model17	X^2 (-5)=.00, *p*=1.000

L2 learners

Select time order vector

Table Appx. 25 AIC results (converged models only) for the eye-tracking test for the L2 learners in selecting time order vector

Models	fixed effects	random effects	AIC
model1	interaction between liner time vector and relative clause type	by-subject and by-item random intercepts	97456.13
model2	interaction between liner + quadratic time vector and relative clause type	by-subject and by-item random intercepts	97475.21

model3	interaction between liner + quadratic + cubic time vector and relative clause type	by-subject and by-item random intercepts	97495.67

Table Appx. 26 LRT results (converged models only) for the eye-tracking test for the L2 learners in selecting time order vector

Model comparisons	LRT
Model1:Model2	X^2 (4)=11.08, p=.026
Model1:Model3	X^2 (8)=23.54, p=.003
Model2:Model3	X^2 (4)=12.46, p=.014

Select random effects

Table Appx. 27 AIC results (converged models only) for the eye-tracking test for the L2 learners in selecting random efffects

Models	Fixed effects	Random intercept	Random slope	AIC
Model1	interaction between linear + quadratic + cubic time vector and relative clause type	by-subject and by-item random intercepts	N/A	97495.67
Model4	interaction between linear + quadratic + cubic time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of linear time vector and relative clause type; by-item random slope of linear time vector and relative clause type	97550.82
Model5	interaction between linear + quadratic + cubic time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type; by-item random slope of relative clause type	97530.96

Model6	interaction between linear + quadratic + cubic time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of linear time vector; by-item random slope of linear time vector	97503.53
Model7	interaction between linear + quadratic + cubic time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of quadratic time vector and relative clause type; by-item random slope of quadratic time vector and relative clause type	97550.31
Model8	interaction between linear + quadratic + cubic time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of cubic time vector and relative clause type; by-item random slope of cubic time vector and relative clause type	97550.95
Model9	interaction between linear + quadratic + cubic time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of quadratic time vector; by-item random slope of quadratic time vector	97502.99
Model10	interaction between linear + quadratic + cubic time vector and relative clause type	by-subject and by-item random intercepts	by-subject random slope of cubic time vector; by-item random slope of cubic time vector	97503.66

Table Appx. 28 LRT results (converged models only) for the eye-tracking test for the L2 learners in selecting random efffects

Model comparisons	LRT
Model1:Model4	X^2 (28)=.84, *p*=1.000
Model1:Model5	X^2 (18)=.71, *p*=1.000
Model4:Model5	X^2 (-10)=.14, *p*=1.000
Model1:Model5	X^2 (4)=.14, *p*=.998
Model4:Model6	X^2 (-24)=.70, *p*=1.000
Model1:Model7	X^2 (28)=1.36, *p*=1.000
Model1:Model8	X^2 (28)=.71, *p*=1.000
Model1:Model9	X^2 (4)=.67, *p*=.954
Model7:Model9	X^2 (-24)=.68, *p*=1.000
Model1:Model10	X^2 (4)=.01, *p*=1.000
Model8:Model10	X^2 (-24)=.71, *p*=1.000

Appendix 20 RQ1: AIC and LRT results for the oral production test

Native speakers

Table Appx. 29 AIC results (converged models only) for the oral production test for native speakers

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type; by-item random slope of relative clause type	314.95
Model2	relative clause type	by-subject and by-item random intercepts	by-item random slope of relative clause type	308.77
Model3	relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type	306.1
Model4	relative clause type	by-subject and by-item random intercepts	N/A	296.22

Table Appx. 30 LRT results (converged models only) for the oral production test for native speakers

Model comparisons	LRT
Model1:Model2	X^2 (-9)=11.81, *p*=.224
Model1:Model3	X^2 (-9)=9.15, *p*=.424
Model4:Model1	X^2 (18)=17.27, *p*=.505
Model2:Model4	X^2 (-9)=5.46, *p*=.793
Model3:Model4	X^2 (-9)=8.12, *p*=.522

L2 learners

Table Appx. 31 AIC results (converged models only) for the oral production test for L2 learners

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	relative clause type	by-subject and by-item random	by-subject random slope of relative clause type;	1347.56

		intercepts	by-item random slope of relative clause type	
Model2	relative clause type	by-subject and by-item random intercepts	by-item random slope of relative clause type	1674.79
Model3	relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type	1330.57
Model4	relative clause type	by-subject and by-item random intercepts	N/A	1658.4

Table Appx. 32 LRT results (converged models only) for the oral production test for L2 learners

Model comparisons	LRT
Model1:Model2	X^2 (-9)=345.23, *p*<.001
Model1:Model3	X^2 (-9)=1.01, *p*=.999
Model4:Model1	X^2 (18)=346.84, *p*<.001
Model2:Model4	X^2 (-9)=1.61, *p*=.996
Model3:Model4	X^2 (-9)=345.83, *p*<.001

Appendix 21 RQ1: AIC and LRT results for the metalinguistic knowledge test

Task of deciding match or mismatch

Native speakers

Mismatched items

Table Appx. 33 AIC results (converged models only) for task of deciding match or mismatch (mismatched items) in metalinguistic knowledge test for native speakers

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model2	relative clause type	by-subject and by-item random intercepts	by-item random slope of relative clause type	118.86
Model4	relative clause type	by-subject and by-item random intercepts	N/A	100.87

Table Appx. 34 LRT results (converged models only) for task of deciding match or mismatch (mismatched items) in metalinguistic knowledge test for native speakers

Model comparisons	LRT
Model2:Model4	X^2 (-9)=.01, *p*=1.000

L2 learners

Matched items

Table Appx. 35 AIC results (converged models only) for task of deciding match or mismatch (matched items) in metalinguistic knowledge test for L2 learners

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model4	relative clause type	by-subject and by-item random intercepts	N/A	214.47

Mismatched items

Table Appx. 36 AIC results (converged models only) for task of deciding match or mismatch (mismatched items) in metalinguistic knowledge test for L2 learners

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type; by-item random slope of relative clause type	637.47
Model3	relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type	621.96
Model4	relative clause type	by-subject and by-item random intercepts	N/A	605.43

Table Appx. 37 LRT results (converged models only) for task of deciding match or mismatch (mismatched items) in metalinguistic knowledge test for L2 learners

Model comparisons	LRT
Model1:Model3	X^2 (-9)=2.48, *p*=.981
Model4:Model1	X^2 (18)=3.96, *p*=.999
Model3:Model4	X^2 (-9)=1.47, *p*=.997

Task of sentence correction

Native speakers

Table Appx. 38 AIC results (converged models only) for task of sentence correction in metalinguistic knowledge test for native speakers

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model2	relative clause type	by-subject and by-item random intercepts	by-item random slope of relative clause type	175.52
Model4	relative clause type	by-subject and by-item random intercepts	N/A	163.46

Table Appx. 39 LRT results (converged models only) for task of sentence correction in metalinguistic knowledge test for native speakers

Model comparison	LRT
Model2:Model4	X^2 (-9)=5.94, *p*=.746

L2 learners

Table Appx. 40 AIC results (converged models only) for task of sentence correction in metalinguistic knowledge test for L2 learners

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model3	relative clause type	by-subject and by-item random intercepts	by-subject random slope of relative clause type	606.77
Model4	relative clause type	by-subject and by-item random intercepts	N/A	613.35

Table Appx. 41 LRT results (converged models only) for task of sentence correction in metalinguistic knowledge test for L2 learners

Model comparison	LRT
Model3:Model4	X^2 (-9)=24.58, *p*=.003

Task of reason explanation

Native speakers

Table Appx. 42 AIC results (converged models only) for task of reason explanation in metalinguistic knowledge test for native speakers

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model2	relative clause type	by-subject and by-item random intercepts	by-item random slope of relative clause type	50.85
Model4	relative clause type	by-subject and by-item random intercepts	N/A	42.78

Table Appx. 43 LRT results (converged models only) for task of reason explanation in metalinguistic knowledge test for native speakers

Model comparisons	LRT
Model2:Model4	X^2 (-9)=9.93, *p*=.356
Model3:Model4	X^2 (-9)=3.60, *p*=.936

L2 learners

Table Appx. 44 AIC results (converged models only) for task of reason explanation in metalinguistic knowledge test for L2 learners

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model2	relative clause type	by-subject and by-item random intercepts	by-item random slope of relative clause type	197.01
Model4	relative clause type	by-subject and by-item random intercepts	N/A	180.09

Table Appx. 45 AIC results (converged models only) for task of reason explanation in metalinguistic knowledge test for L2 learners

Model comparison	LRT
Model2:Model4	X^2 (-9)=1.08, *p*=.999

Appendix 22 RQ2: AIC and LRT results for the offline comprehension test

SRC-A

Table Appx. 46 AIC results for SRC-A structure for the offline comprehension test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	855.5
Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	880.17
Model4	interaction between group and test phase	by-subject and by-item random intercepts	by-subject and by-item random slopes of the test phase	867.94
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	864.58
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	863.69
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	860.76
Model8	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase	860.54

Table Appx. 47 LRT results for SRC-A structure for the offline comprehension test

Model comparisons	LRT
Model1:Model3	X^2 (19)=13.36, *p*=.820
Model1:Model4	X^2 (10)=7.59, *p*=.669
Model3:Model4	X^2 (-9)=5.77, *p*=.763
Model1:Model5	X^2 (10)=10.95, *p*=.362
Model3:Model5	X^2 (-9)=2.41, *p*=.983
Mode1:Model6	X^2 (5)=1.84, *p*=.871
Model3:Model6	X^2 (-14)=11.52, *p*=.645
Model4:Model6	X^2 (-5)=5.75, *p*=.331
Model1:Model7	X^2 (5)=4.77, *p*=.445
Model3:Model7	X^2 (-14)=8.59, *p*=.857
Model5:Model7	X^2 (-5)=6.18, *p*=.290
Model1:Model8	X^2 (5)=4.99, *p*=.417
Model3:Model8	X^2 (-14)=8.37, *p*=.869
Model4:Model8	X^2 (-5)=2.60, *p*=.762
Model5:Model8	X^2 (-5)=5.96, *p*=.310

SRC-I

Table Appx. 48 AIC results for SRC-I structure for the offline comprehension test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	887.35
Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	911.09
Model4	interaction between group and test phase	by-subject and by-item random intercepts	by-subject and by-item random slopes of the test phase	897.14
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	899.35

Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	890.87
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	893.6
Model8	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase	892.97

Table Appx. 49 LRT results for SRC-I structure for the offline comprehension test

Model comparisons	LRT
Model1:Model3	X^2 (19)=14.26, *p*=.769
Model1:Model4	X^2 (10)=10.21, *p*=.422
Model3:Model4	X^2 (-9)=4.04, *p*=.909
Mode1:Model5	X^2 (10)=7.10, *p*=.629
Model3:Model5	X^2 (-9)=6.26, *p*=.714
Model1:Model6	X^2 (5)=6.47, *p*=.263
Model3:Model6	X^2 (-14)=7.78, *p*=.900
Model4:Model6	X^2 (-5)=3.74, *p*=.587
Model1:Model7	X^2 (5)=3.75, *p*=.445
Model3:Model7	X^2 (-14)=10.50, *p*=.724
Model5:Model7	X^2 (-5)=4.24, *p*=.515
Model1:Model8	X^2 (5)=4.38, *p*=.496
Model3:Model8	X^2 (-14)=9.88, *p*=.771
Model4:Model8	X^2 (-5)=5.83, *p*=.323
Model5:Model8	X^2 (-5)=3.62, *p*=.606

ORC-A

Table Appx. 50 AIC results for ORC-A structure for the offline comprehension test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	881.67

Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	896.64
Model4	interaction between group and test phase	by-subject and by-item random intercepts	by-subject and by-item random slopes of the test phase	891.3
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	893.58
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	884.53
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	887.23
Model8	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase	888.15

Table Appx. 51 LRT results for ORC-A structure for the offline comprehension test

Model comparisons	LRT
Model1:Model3	X^2 (19)=23.03, *p*=.236
Model1:Model4	X^2 (10)=10.36, *p*=.409
Model3:Model4	X^2 (-9)=12.66, *p*=.179
Model1:Model5	X^2 (10)=8.09, *p*=.620
Model3:Model5	X^2 (-9)=14.94, *p*=.093
Model1:Model6	X^2 (5)=7.14, *p*=.211
Model3:Model6	X^2 (-14)=15.89, *p*=.320
Model4:Model6	X^2 (-5)=3.23, *p*=.665
Model1:Model7	X^2 (5)=4.44, *p*=.488
Model3:Model7	X^2 (-14)=18.59, *p*=.181
Model5:Model7	X^2 (-5)=3.65, *p*=.600
Model1:Model8	X^2 (5)=3.51, *p*=.621
Model3:Model8	X^2 (-14)=19.51, *p*=.146
Model4:Model8	X^2 (-5)=6.85, *p*=.232
Model5:Model8	X^2 (-5)=4.57, *p*=.470

ORC-I

Table Appx. 52 AIC results for ORC-I structure for the offline comprehension test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	1160.97
Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	1186.87
Model4	interaction between group and test phase	by-subject and by-item random intercepts	by-subject and by-item random slopes of the test phase	1172.74
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	1176.95
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	1165.59
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	1168.61
Model8	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase	1169.12

Table Appx. 53 LRT results for ORC-I structure for the offline comprehension test

Model comparisons	LRT
Model1:Model3	X^2 (19)=12.09, *p*=.882
Model1:Model4	X^2 (10)=8.23, *p*=.606

Model3:Model4	X^2 (-9)=3.86, *p*=.920
Mode1:Model5	X^2 (10)=4.02, *p*=.947
Model3:Model5	X^2 (-9)=8.07, *p*=.527
Model1:Model6	X^2 (5)=5.37, *p*=.372
Model3:Model6	X^2 (-14)=6.72, *p*=.945
Model4:Model6	X^2 (-5)=2.86, *p*=.722
Mode1:Model7	X^2 (5)=2.36, *p*=.798
Model3:Model7	X^2 (-14)=9.73, *p*=.781
Model5:Model7	X^2 (-5)=1.66, *p*=.894
Model1:Model8	X^2 (5)=1.85, *p*=.870
Model3:Model8	X^2 (-14)=10.24, *p*=.744
Model4:Model8	X^2 (-5)=6.38, *p*=.271
Model5:Model8	X^2 (-5)=2.17, *p*=.825

Appendix 23 RQ2: AIC and LRT results for the eye-tracking test

SRC-A

First critical word

Table Appx. 54 AIC results for SRC-A structure at the first critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	32687.06
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	32721.35
model3	interaction between linear + quadratic + cubic time vector, group and test phase	by-subject and by-item random intercepts	32753.66

Table Appx. 55 LRT results for SRC-A structure at the first critical word in the eye-tracking test

Model comparisons	LRT
Model1:Model2	X^2 (9)=16.29, *p*=.061
Model1:Model3	X^2 (18)=30.60, *p*=.032
Model2:Model3	X^2 (9)=14.31, *p*=.112

Second critical word

Table Appx. 56 AIC results for SRC-A structure at the second critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	9448.93
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	9477.69

Table Appx. 57 LRT results for SRC-A structure at the second critical word in the eye-tracking test

Model comparison	LRT
Model1:Model2	X^2 (9)=10.76, *p*=.292

Third critical word

Table Appx. 58 AIC results for SRC-A structure at the third critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	33420.78
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	33456.94
model3	interaction between linear + quadratic + cubic time vector, group and test phase	by-subject and by-item random intercepts	33485.60

Table Appx. 59 LRT results for SRC-A structure at the third critical word in the eye-tracking test

Model comparisons	LRT
Mode1:Model2	X^2 (9)=18.16, *p*=.003
Mode1:Model3	X^2 (18)=28.82, *p*=.051
Mode2:Model3	X^2 (9)=10.67, *p*=.299

SRC-I

First critical word

Table Appx. 60 AIC results for SRC-I structure at the first critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	31082.95
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	31116.17
model3	interaction between linear + quadratic + cubic time vector, group and test phase	by-subject and by-item random intercepts	31151.08

Table Appx. 61 LRT results for SRC-I structure at the first critical word in the eye-tracking test

Model comparisons	LRT
Model1:Model2	X2 (9)=15.22, *p*=.085
Model1:Model3	X2 (18)=32.13, *p*=.021
Model2:Model3	X2 (9)=16.91, *p*=.050

Second critical word

Table Appx. 62 AIC results for SRC-I structure at the second critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	8850.09
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	8880.04

Table Appx. 63 LRT results for SRC-I structure at the second critical word in the eye-tracking test

Model comparison	LRT
Model1:Model2	X^2 (9)=11.05, *p*=.216

Third critical word

Table Appx. 64 AIC results for SRC-I structure at the third critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	30810.00
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	30837.72
model3	interaction between linear + quadratic + cubic time vector, group and test phase	by-subject and by-item random intercepts	30853.20

Table Appx. 65 LRT results for SRC-I structure at the third critical word in the eye-tracking test

Model comparisons	LRT
Model1:Model2	X^2 (9)=9.72, *p*=.374
Model1:Model3	X^2 (18)=7.20, *p*=.988
Model2:Model3	X^2 (9)=2.52, *p*=.980

ORC-A

First critical word

Table Appx. 66 AIC results for ORC-A structure at the first critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	7215.12
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	7249.52

Table Appx. 67 LRT results for ORC-A structure at the first critical word in the eye-tracking test

Model comparison	LRT
Model1:Model2	X^2 (9)=16.40, *p*=.059

Second critical word

Table Appx. 68 AIC results for ORC-A structure at the second critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	23945.02
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	23976.26
model3	interaction between linear + quadratic + cubic time vector, group and test phase	by-subject and by-item random intercepts	24011.74

Table Appx. 69 LRT results for ORC-A structure at the second critical word in the eye-tracking test

Model comparisons	LRT
Model1:Model2	X^2 (9)=13.24, *p*=.152
Model1:Model3	X^2 (18)=30.71, *p*=.031
Model2:Model3	X^2 (9)=17.48, *p*=.042

Third critical word

Table Appx. 70 AIC results for ORC-A structure at the third critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	43738.78
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	43777.73
model3	interaction between linear + quadratic + cubic time vector, group and test phase	by-subject and by-item random intercepts	43809.00

Table Appx. 71 LRT results for ORC-A structure at the third critical word in the eye-tracking test

Model comparisons	LRT
Model1:Model2	X^2 (9) =20.95, *p*=.013
Model1:Model3	X^2 (18) =34.22, *p*=.012
Model2:Model3	X^2 (9) =13.27, *p*=.151

ORC-I

First critical word

Table Appx. 72 AIC results for ORC-I structure at the first critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	6937.63
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	6965.05

Table Appx. 73 LRT results for ORC-I structure at the first critical word in the eye-tracking test

Model comparison	LRT
Model1:Model2	X^2 (9)=9.42, *p*=.399

Second critical word

Table Appx. 74 AIC results for ORC-I structure at the second critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	22189.07
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	22216.38
model3	interaction between linear + quadratic + cubic time vector, group and test phase	by-subject and by-item random intercepts	22244.75

Table Appx. 75 LRT results for ORC-I structure at the second critical word in the eye-tracking test

Model comparisons	LRT
Model1:Model2	X^2 (9)=9.30, *p*=.410
Model1:Model3	X^2 (18)=19.68, *p*=.351
Model2:Model3	X^2 (9)=10.38, *p*=.321

Third critical word

Table Appx. 76 AIC results for ORC-I structure at the third critical word in the eye-tracking test

Models	Fixed effects	Random effects	AIC
model1	interaction between linear time vector, group and test phase	by-subject and by-item random intercepts	37905.49
model2	interaction between linear + quadratic time vector, group and test phase	by-subject and by-item random intercepts	37936.63
model3	interaction between linear + quadratic + cubic time vector, group and test phase	by-subject and by-item random intercepts	37959.32

Table Appx. 77 LRT results for ORC-I structure at the third critical word in the eye-tracking test

Model comparisons	LRT
Model1:Model2	X^2 (9) =13.14, *p*=.156
Model1:Model3	X^2 (18) =17.83, *p*=.467
Model2:Model3	X^2 (9) =4.68, *p*=.861

Appendix 24 RQ2: AIC and LRT results for the oral production test

SRC-A

Table Appx. 78 AIC results for SRC-A structure in the oral production test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	1014.23
Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	1021.55
Model4	interaction between group and test phase	by-subject and by-item random intercepts	by-subject and by-item random slopes of the test phase	1003.65
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	1003.65
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	1024.18
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	1024.23

Table Appx. 79 LRT results for SRC-A structure in the oral production test

Model comparisons	LRT
Model1:Model3	X^2 (19)=30.68, p=.044
Model1:Model4	X^2 (10)=30.58, p<.001
Model3:Model4	X^2 (-9)=.09, p=1.00
Model1:Model5	X^2 (10)=30.58, p<.001
Model3:Model5	X^2 (-9)=.09, p=1.00

Mode1:Model6	X^2 (5)=.05, *p*=1.00
Model3:Model6	X^2 (-14)=30.63, *p*=.006
Model4:Model6	X^2 (-5)=30.54, *p*<.001
Model1:Model7	X^2 (5)=.00, *p*=1.00
Model3:Model7	X^2 (-14)=30.68, *p*=.006
Model5:Model7	X^2 (-5)=30.58, *p*<.001

SRC-I

Table Appx. 80 AIC results for SRC-I structure in the oral production test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	964.63
Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	955.64
Model4	interaction between group and test phase	by-subject and by-item random intercepts	by-subject and by-item random slopes of the test phase	940.21
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	939.24
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	974.63
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	974.43

Table Appx. 81 LRT results for SRC-I structure in the oral production test

Model comparisons	LRT
Model1:Model3	X^2 (19)=46.98, *p*<.001
Model1:Model4	X^2 (10)=44.41, *p*<.001
Model3:Model4	X^2 (-9)=2.57, *p*=.979
Model1:Model5	X^2 (10)=45.38, *p*<.001
Model3:Model5	X^2 (-9)=1.60, *p*=.996
Model1:Model6	X^2 (5)=.00, *p*=1.00
Model3:Model6	X^2 (-14)=46.98, *p*<.001
Model4:Model6	X^2 (-5)=44.41, *p*<.001
Model1:Model7	X^2 (5)=.20, *p*=.999
Model3:Model7	X^2 (-14)=46.79, *p*<.001
Model5:Model7	X^2 (-5)=45.19, *p*<.001

ORC-A

Table Appx. 82 AIC results for ORC-A structure in the oral production test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	670.02
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	623.26
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	679.77
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	680.02

Table Appx. 83 LRT results for ORC-A structure in the oral production test

Model comparisons	LRT
Model1:Model5	X^2 (10)=66.77, *p*<.001
Model1:Model6	X^2 (5)=.25, *p*=.998

Model1:Model7	X^2 (5)=.00, *p*=1.00

Note: In model 5, the model for calculating the marginal and conditional R^2 did not converge. Thus, the primary model was adopted.

ORC-I

Table Appx. 84 AIC results for ORC-I structure in the oral production test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	776.92
Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	750.66
Model4	interaction between group and test phase	by-subject and by-item random intercepts	by-subject and by-item random slopes of the test phase	741.12
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	738.42
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	871.63
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	776.99

Table Appx. 85 LRT results for ORC-I structure in the oral production test

Model comparisons	LRT
Model1:Model3	X^2 (19)=64.26, *p*<.001
Model1:Model4	X^2 (10)=55.80, *p*<.001
Model3:Model4	X^2 (-9)=8.46, *p*=.489

Model1:Model5	X^2 (10)=58.50, *p*<.001
Model3:Model5	X^2 (-9)=5.75, *p*=.764
Model1:Model6	X^2 (5)=5.29, *p*=.382
Model3:Model6	X^2 (-14)=58.97, *p*<.001
Model4:Model6	X^2 (-5)=50.51, *p*<.001
Model1:Model7	X^2 (5)=9.93, *p*=.077
Model3:Model7	X^2 (-14)=54.33, *p*<.001
Model5:Model7	X^2 (-5)=48.58, *p*<.001

Appendix 25 RQ2: AIC and LRT results for the metalinguistic knowledge test

Task of deciding match or mismatch

Matched items

ORC-A

Table Appx. 86 AIC results for ORC-A structure in the task of deciding match or mismatch (matched items) of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	165.70

ORC-I

Table Appx. 87 AIC results for ORC-I structure in the task of deciding match or mismatch (matched items) of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	161.00

Mismatched items

SRC-A

Table Appx. 88 AIC results for SRC-A structure in the task of deciding match or mismatch (mismatched items) of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	328.12

Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	333.2
Model4	interaction between group and test phase	by-subject and by-item random intercepts	by-subject and by-item random slopes of the test phase	319.56
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	327.85
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	332.88
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	337.19
Model8	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase	351.54

Table Appx. 89 LRT results for SRC-A structure in the task of deciding match or mismatch (mismatched items) of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model3	X^2 (19)=32.92, *p*=.025
Model1:Model4	X^2 (10)=28.56, *p*=.001
Model3:Model4	X^2 (-9)=4.36, *p*=.886
Model1:Model5	X^2 (10)=20.27, *p*=.027
Model3:Model5	X^2 (-9)=12.66, *p*=.178
Model1:Model6	X^2 (5)=5.23, *p*=.388
Model3:Model6	X^2 (-14)=27.69, *p*=.016
Model4:Model6	$X^2$5)=23.33, *p*<.001
Model1:Model7	X^2 (5)=.93, *p*=.968
Model3:Model7	X^2 (-14)=32.00, *p*=.004
Model5:Model7	X^2 (-5)=19.34, *p*=.002
Model1:Model8	X^2 (5)=17.23, *p*=.004
Model3:Model8	X^2 (-14)=15.69, *p*=.333
Model4:Model8	X^2 (-5)=11.33, *p*=.045
Model5:Model8	X^2 (-5)=3.03, *p*=.695

SRC-I

Table Appx. 90 AIC results for SRC-I structure in the task of deciding match or mismatch (mismatched items) of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	350.97
Model2	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the interaction between group and test phase	409.22
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	360.62
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	349.28
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	359.29
Model8	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase	351.54

Table Appx. 91 LRT results for SRC-I structure in the task of deciding match or mismatch (mismatched items) of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model2	X^2 (49)=39.76, *p*=.824
Model1:Model5	X^2 (10)=10.36, *p*=.410
Model2:Model5	X^2 (-39)=29.40, *p*=.868
Model1:Model6	X^2 (5)=11.70, *p*=.039

Model2:Model6	X^2 (-44)=28.06, *p*=.971
Model1:Model7	X^2 (5)=1.68, *p*=.891
Model2:Model7	X^2 (-44)=38.08, *p*=.723
Model5:Model7	X^2 (-5)=8.68, *p*=.123
Model1:Model8	X^2 (5)=9.44, *p*=.093
Model2:Model8	X^2 (-44)=30.32, *p*=.942
Model5:Model8	X^2 (-5)=.924, *p*=.968

ORC-A

Table Appx. 92 AIC results for ORC-A structure in the task of deciding match or mismatch (mismatched items) of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	334.72
Model2	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the interaction between group and test phase	341.95
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	338.08

Table Appx. 93 LRT results for ORC-A structure in the task of deciding match or mismatch (mismatched items) of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model2	X^2 (49)=90.76, *p*<.001
Model1:Model6	X^2 (5)=6.64, *p*=.249
Model2:Model6	X^2 (-44)=84.13, *p*<.001
Model1:Model7	X^2 (5)=.76, *p*=.979
Model2:Model7	X^2 (-44)=90.00, *p*<.001

ORC-I

Table Appx. 94 AIC results for ORC-I structure in the task of deciding match or mismatch (mismatched items) of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	467.15
Model2	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the interaction between group and test phase	550.00
Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	498.20
Model4	interaction between group and test phase	by-subject and by-item random intercepts	by-subject and by-item random slopes of the test phase	484.62
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	482.15
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	476.71
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	474.10
Model8	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase	475.38

Table Appx. 95 LRT results for ORC-I structure in the task of deciding match or mismatch (mismatched items) of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model2	X^2 (49)=15.15, *p*=1.000
Model1:Model3	X^2 (19)=6.95, *p*=.995
Model2:Model3	X^2 (-30)=8.20, *p*=1.000
Model1:Model4	X^2 (10)=2.93, *p*=.990
Model2:Model4	X^2 (-39)=12.62, *p*=1
Model3:Model4	X^2 (-9)=4.42, *p*=.882
Model1:Model5	X^2 (10)=5.00, *p*=.891
Model2:Model5	X^2 (-39)=10.14, *p*=1.000
Model3:Model5	X^2 (-9)=1.95, *p*=.992
Model1:Model6	X^2 (5)=.44, *p*=.994
Model2:Model6	X^2 (-44)=14.71, *p*=1.000
Model3:Model6	X^2 (-14)=6.51, *p*=.952
Model4:Model6	X^2 (-5)=2.09, *p*=.837
Model1:Model7	X^2 (5)=3.05, *p*=.692
Model2:Model7	X^2 (-44)=12.10, *p*=1.000
Model3:Model7	X^2 (-14)=3.90, *p*=.996
Model5:Model7	X^2 (-5)=1.95, *p*=.855
Model1:Model8	X^2 (5)=1.77, *p*=.880
Model2:Model8	X^2 (-44)=13.38, *p*=1.000
Model3:Model8	X^2 (-14)=5.18, *p*=.983
Model5:Model8	X^2 (-5)=3.24, *p*=.664

Task of sentence correction

SRC-A

Table Appx. 96 AIC results for SRC-A structure in the task of sentence correction of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	401.78
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	378.78

Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	409.82
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	406.87
Model8	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase	376.90

Table Appx. 97 LRT results for SRC-A structure in the task of sentence correction of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model5	X^2 (10)=43.00, *p* <.001
Model1:Model6	X^2 (5)=1.97, *p* =.854
Model1:Model7	X^2 (5)=4.92, *p* =.426
Model5:Model7	X^2 (-5)=38.08, *p* <.001
Model1:Model8	X^2 (5)=34.88, *p* <.001
Model5:Model8	X^2 (-5)=8.12, *p*=.150

SRC-I

Table Appx. 98 AIC results for SRC-I structure in the task of sentence correction of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	404.41
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	410.21
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	403.55

Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	412.05
Model8	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase	402.71

Table Appx. 99 LRT results for SRC-I structure in the task of sentence correction of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model5	X^2 (10)=14.20, *p*=.164
Model1:Model6	X^2 (5)=10.86, *p*=.054
Model1:Model7	X^2 (5)=2.36, *p*=.798
Model5:Model7	X^2 (-5)=11.85, *p*=.037
Model1:Model8	X^2 (5)=11.70, *p*=.039
Model5:Model8	X^2 (-5)=2.50, *p*=.776

ORC-A

Table Appx. 100 AIC results for ORC-A structure in the task of sentence correction of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	375.30
Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	393.67
Model4	interaction between group and test phase	by-subject and by-item random intercepts	by-subject and by-item random slopes of the test phase	382.17
Model5	interaction between group	by-subject and by-item random	by-subject random slope of the test	378.16

	and test phase	intercepts	phase; by-item random slope of the group	
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	383.48
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	379.83
Model8	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase	370.95

Table Appx. 101 LRT results for ORC-A structure in the task of sentence correction of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model3	X^2 (19)=19.63, *p*=.417
Model1:Model4	X^2 (10)=13.13, *p*=.217
Model3:Model4	X^2 (-9)=6.50, *p*=.689
Model1:Model5	X^2 (10)=17.14, *p*=.071
Model3:Model5	X^2 (-9)=2.49, *p*=.981
Model1:Model6	X^2 (5)=1.82, *p*=.874
Model3:Model6	X^2 (-14)=17.81, *p*=.215
Model4:Model6	X^2 (-5)=11.31, *p*=.046
Model1:Model7	X^2 (5)=5.46, *p*=.3.62
Model3:Model7	X^2 (-14)=14.17, *p*=.437
Model5:Model7	X^2 (-5)=11.68, *p*=.039
Model1:Model8	X^2 (5)=14.35, *p*=.014
Model3:Model8	X^2 (-14)=5.28, *p*=.981
Model5:Model8	X^2 (-5)=2.80, *p*=.731

ORC-I

Table Appx. 102 AIC results for ORC-I structure in the task of sentence correction of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	462.72
Model2	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the interaction between group and test phase	547.44
Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	494.07
Model4	interaction between group and test phase	by-subject and by-item random intercepts	by-subject and by-item random slopes of the test phase	478.96
Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	479.73
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	471.58
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	470.69
Model8	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase	471.76

Table Appx. 103 LRT results for ORC-I structure in the task of sentence correction of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model2	X^2 (49)=13.28, *p*=1.000
Model1:Model3	X^2 (19)=6.65, *p*=.996
Model2:Model3	X^2 (-30)=6.63, *p*=1.000
Model1:Model4	X^2 (10)=3.76, *p*=.958
Model2:Model4	X^2 (-39)=9.52, *p*=1.000
Model3:Model4	X^2 (-9)=2.89, *p*=.968
Model1:Model5	X^2 (10)=2.99, *p*=.982
Model2:Model5	X^2 (-39)=10.92, *p*=1.000
Model3:Model5	X^2 (-9)=3.66, *p*=.932
Model1:Model6	X^2 (5)=1.15, *p*=.950
Model2:Model6	X^2 (-44)=12.14, *p*=1.000
Model3:Model6	X^2 (-14)=5.51, *p*=.977
Model4:Model6	X^2 (-5)=2.62, *p*=.759
Model1:Model7	X^2 (5)=2.03, *p*=.844
Model2:Model7	X^2 (-44)=11.25, *p*=1.000.
Model3:Model7	X^2 (-14)=4.62, *p*=.991
Model5:Model7	X^2 (-5)=.96, *p*=.966
Model1:Model8	X^2 (5)=.96, *p*=.966
Model2:Model8	X^2 (-44)=12.32, *p*=1.000
Model3:Model8	X^2 (-14)=5.69, *p*=.974
Model5:Model8	X^2 (-5)=2.03, *p*=.845

Task of reason explanation

SRC-A

Table Appx. 104 AIC results for SRC-A structure in the task of reason explanation of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	283.95
Model3	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of test phase; by-item random slope of the group and test phase	278.53

Model5	interaction between group and test phase	by-subject and by-item random intercepts	by-subject random slope of the test phase; by-item random slope of the group	271.24
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	293.09
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	290.11

Table Appx. 105 LRT results for SRC-A structure in the task of reason explanation of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model3	X^2 (19)=43.42, *p*=.001
Model1:Model5	X^2 (10)=32.71, *p*<.001
Model3:Model5	X^2 (-9)=10.71, *p*=.296
Model1:Model6	X^2 (5)=.86, *p*=.973
Model3:Model6	X^2 (-14)=42.56, *p*<.001
Model4:Model6	X^2 (-5)=17.11, *p*=.004
Model1:Model7	X^2 (5)=3.84, *p*=.573
Model3:Model7	X^2 (-14)=39.58, *p*<.001
Model5:Model7	X^2 (-5)=28.87, *p*<.001

SRC-I

Table Appx. 106 AIC results for SRC-I structure in the task of reason explanation of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	222.0
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	229.2

Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	231.5

Table Appx. 107 LRT results for SRC-I structure in the task of reason explanation of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model6	X^2 (5)=2.75, *p*=.738
Model1:Model7	X^2 (5)=.47, *p*=.993

ORC-A

Table Appx. 108 AIC results for ORC-A structure in the task of reason explanation of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	266.5
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	273.5
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	274.7

Table Appx. 109 LRT results for ORC-A structure in the task of reason explanation of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model6	X^2 (5)=2.98, *p*=.704
Model1:Model7	X^2 (5)=1.83, *p*=.872

ORC-I

Table Appx. 110 AIC results for ORC-I structure in the task of reason explanation of the metalinguistic knowledge test

Models	Fixed effects	Random effects		AIC
		random intercept	random slope	
Model1	interaction between group and test phase	by-subject and by-item random intercepts	N/A	254.50
Model6	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the test phase	260.20
Model7	interaction between group and test phase	by-subject and by-item random intercepts	by-item random slope of the group	264.50

Table Appx. 111 LRT results for ORC-I structure in the task of reason explanation of the metalinguistic knowledge test

Model comparisons	LRT
Model1:Model6	X^2 (5)=4.33, *p*=.503
Model1:Model7	X^2 (5)=.00, *p*=1.000

Appendix 26 RQ2: Fixed effects of model analysis in the offline comprehension test (baseline of the input flood group)

SRC-A

Table Appx. 112 The fixed effects of modela analysis in the offline comprehension test for the SRC-A structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	2.47 [1.86, 3.08]	.37	6.68	<.001***	11.86 [6.45, 21.80]
input vs. parsing	-.32 [-1.00, .36]	.41	-.78	.437	.73 [.37, 1.43]
input vs. test	-.29 [-.97, .39]	.42	-.70	.487	.75 [.38, 1.48]
pre- vs. post-	-.25 [-.85, .35]	.36	-.68	.496	.78 [.43, 1.42]
pre- vs. delayed-	1.20 [.42, 1.98]	.47	2.53	.011*	3.31 [1.52, 7.22]
input vs. parsing × pre- vs. post-	1.28 [.38, 2.17]	.54	2.35	.019*	3.59 [1.47, 8.78]
input vs. test × pre- vs. post-	.73 [-.12, 1.58]	.52	1.42	.157	2.08 [.89, 4.87]
input vs. parsing × pre- vs. delayed-	.46 [-.64, 1.57]	.67	.69	.491	1.59 [.53, 4.80]
input vs. test × pre- vs. delayed-	-.69 [-1.68, .29]	.60	-1.15	.249	.50 [.19, 1.34]

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when $\alpha \leq .05$; *** significantly different from zero when $\alpha<.001$

SRC-I

Table Appx. 113 The fixed effects of modela analysis in the offline comprehension test for the SRC-I structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	2.30 [1.68, 2.93]	.38	6.08	<.001***	1.02 [5.37, 18.70]
input vs. parsing	.04 [-.62, .69]	.40	.09	.929	1.04 [.54, 1.98]
input vs. test	.03 [-.63, .69]	.40	.07	.948	1.03 [.53, 1.98]
pre- vs. post-	.17 [-.43, .78]	.37	.47	.637	1.19 [.65, 2.17]
pre- vs. delayed-	.42 [-.21, 1.05]	.39	1.09	.275	1.52 [.81, 2.87]
input vs. parsing × pre- vs. post-	.35 [-.52, 1.22]	.53	.66	.509	1.42 [.59, 3.39]
input vs. test × pre- vs. post-	-.49 [-1.33, .34]	.51	-.97	.332	.61[.27, 1.41]
input vs. parsing × pre- vs. delayed-	.94 [-.05, 1.93]	.60	1.57	.117	2.57 [.96, 6.91]
input vs. test × pre- vs. delayed-	-.21 [-1.09, .68]	.54	-.39	.700	.81 [.34, 1.97]

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when $\alpha<.001$

ORC-A

Table Appx. 114 The fixed effects of modela analysis in the offline comprehension test for the ORC-A structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	2.13 [1.60, 2.65]	.32	6.60	<.001***	8.37 [4.93, 14.22]
input vs. parsing	-.18 [-.81, .44]	.38	-.49	.625	.83 [.45, 1.55]
input vs. test	-.18 [-.80, .45]	.38	-.46	.645	.84 [.45, 1.57]
pre- vs. post-	.95 [.29, 1.61]	.40	2.36	.018*	2.59 [1.34, 5.02]
pre- vs. delayed-	.91 [.26, 1.57]	.40	2.29	.022*	2.49 [1.29, 4.79]
input vs. parsing × pre- vs. post-	.15 [-.76, 1.06]	.55	.27	.786	1.16 [.47, 2.90]
input vs. test × pre- vs. post-	-.34 [-1.22, .53]	.53	-.65	.518	.71 [.29, 1.70]
input vs. parsing × pre- vs. delayed-	.81 [-.20, 1.81]	.61	1.32	.187	2.24 [.82, 6.11]
input vs. test × pre- vs. delayed-	-.17 [-1.06, .71]	.54	-.32	.746	.84 [.35, 2.04]

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when $\alpha \leq .05$; *** significantly different from zero when $\alpha<.001$

ORC-I

Table Appx. 115 The fixed effects of modela analysis in the offline comprehension test for the ORC-I structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	1.49 [.94, 2.03]	.33	4.47	<.001***	4.43 [2.56, 7.65]
input vs. parsing	.09 [-.55, .73]	.39	.22	.826	1.09 [.57, 2.06]
input vs. test	.21 [-.43, .86]	.39	.54	.586	1.24 [.65, 2.37]
pre- vs. post-	.64 [.12, 1.16]	.32	2.02	.044	1.89 [1.12, 3.18]
pre- vs. delayed-	**.59 [.06, 1.13]**	**.32**	**1.83**	**.067**	**1.81 [1.06, 3.09]**
input vs. parsing × pre- vs. post-	.21 [-.54, .97]	.46	.46	.643	1.24 [.58, 2.63]
input vs. test × pre- vs. post-	-.06 [-.81, .69]	.45	-.14	.890	.94 [.44, 1.98]
input vs. parsing × pre- vs. delayed-	.21 [-.55, .97]	.46	.45	.653	1.23 [.58, 2.63]
input vs. test × pre- vs. delayed-	.35 [-.44, 1.14]	.48	.74	.462	1.42 [.65, 3.13]

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect;*** significantly different from zero when $\alpha<.001$

Appendix 27 RQ2: Fixed effects of model analysis in the SPR test (baseline of the test-only group)

SRC-A

First critical word

Table Appx. 116 Fixed effects of model analysis in the SPR test for the SRC-A structure at the first critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	-3.96 [-8.82, 0.9]	2.90	-1.34	.181
test vs. parsing	3.72 [-3.02, 10.45]	4.09	.91	.364
test vs. input	-4.78 [-11.7, 2.14]	4.21	-1.14	.256
pre- vs. post-	3.15 [-3.74, 10.03]	4.19	.75	.452
pre- vs. delayed post-	.69 [-6.2, 7.57]	4.19	.16	.870
matched vs. mismatched	**9.02 [2.07, 15.97]**	**4.23**	**2.14**	**.033***
test vs. parsing × pre- vs. post-	-4.87 [-14.42, 4.68]	5.81	-.84	.402
test vs. input × pre- vs. post-	-.05 [-9.94, 9.84]	6.01	-.01	.994
test vs. parsing × pre- vs. delayed-	-.18 [-9.85, 9.50]	5.88	-.03	.976
test vs. input × pre- vs. delayed-	3.78 [-6.1, 13.65]	6.00	.63	.529
test vs. parsing × matched vs. mismatched	**-11.19 [-20.79, -1.60]**	**5.83**	**-1.92**	**.055**
test vs. input × matched vs. mismatched	.73 [-9.06, 10.53]	5.95	.12	.902
pre- vs. post- × matched vs. mismatched	-9.12 [-18.91, 0.67]	5.95	-1.53	.126
pre- vs. delayed- × matched vs. mismatched	-8.48 [-18.31, 1.34]	5.97	-1.42	.156
test vs. parsing × pre- vs. post- × matched vs. mismatched	**13.59 [.04, 27.13]**	**8.24**	**1.65**	**.099**
test vs. input × pre- vs. post- × matched vs. mismatched	7.86 [-6.15, 21.87]	8.52	.92	.356
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	9.22 [-4.52, 22.96]	8.35	1.10	.270
test vs. input × pre- vs. delayed- × matched vs. mismatched	-2.46 [-16.48, 11.57]	8.53	-.29	.773

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$

Second critical word

Table Appx. 117 Fixed effects of model analysis in the SPR test for the SRC-A structure at the second critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	**-8.62 [-13.19, -4.05]**	**2.78**	**-3.10**	**.002****
test vs. parsing	**10.19 [3.94, 16.44]**	**3.80**	**2.68**	**.007****
test vs. input	5.55 [-0.83, 11.93]	3.88	1.43	.152
pre- vs. post-	**11.74 [5.34, 18.14]**	**3.89**	**3.02**	**.003****
pre- vs. delayed-	**7.7 [1.30, 14.1]**	**3.89**	**1.98**	**.048***
matched vs. mismatched	**8.07 [1.55, 14.58]**	**3.96**	**2.04**	**.042***
test vs. parsing × pre- vs. post-	**-14.09 [-22.97, -5.21]**	**5.40**	**-2.61**	**.009****
test vs. input × pre- vs. post-	-5.85 [-14.98, 3.29]	5.56	-1.05	.293
test vs. parsing × pre- vs. delayed-	-8.31 [-17.28, .67]	5.46	-1.52	.128
test vs. input × pre- vs. delayed-	-8.53 [-17.68, .62]	5.56	-1.53	.125
test vs. parsing × matched vs. mismatched	-8.85 [-17.73, .02]	5.40	-1.64	.101
test vs. input × matched vs. mismatched	-6.58 [-15.65, 2.49]	5.52	-1.19	.233
pre- vs. post- × matched vs. mismatched	**-10.00 [-19.07, -.92]**	**5.52**	**-1.81**	**.070**
pre- vs. delayed- × matched vs. mismatched	-4.51 [-13.65, 4.64]	5.56	-.81	.418
test vs. parsing × pre- vs. post- × matched vs. mismatched	**14.64 [2.06, 27.23]**	**7.65**	**1.91**	**.056**
test vs. input × pre- vs. post- × matched vs. mismatched	8.40 [-4.59, 21.38]	7.89	1.06	.288
test vs. parsing × pre- vs. delayed post- × matched vs. mismatched	8.91 [-3.86, 21.68]	7.76	1.15	.251
test vs. input × pre- vs. delayed post- × matched vs. mismatched	8.44 [-4.61, 21.48]	7.93	1.06	.287

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$

Third critical word

Table Appx. 118 Fixed effects of model analysis in the SPR test for the SRC-A structure at the third critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*-	*p*
intercept	**-8.98 [-14.65, -3.31]**	**3.45**	**-2.60**	**.009****
test vs. parsing	**8.00 [.58, 15.42]**	**4.51**	**1.77**	**.076**
test vs. input	**11.15 [3.6, 18.7]**	**4.59**	**2.43**	**.015***
pre- vs. post-	**7.98 [.38, 15.57]**	**4.62**	**1.73**	**.084**
pre- vs. delayed-	**7.98 [.38, 15.57]**	**4.62**	**1.73**	**.084**
matched vs. mismatched	**18.15 [10.09, 26.20]**	**4.90**	**3.71**	**<.001*****
test vs. parsing × pre- vs. post-	**-11.29 [-21.8, -.79]**	**6.39**	**-1.77**	**.077**
test vs. input × pre- vs. post-	**-14.27 [-25.15, -3.39]**	**6.62**	**-2.16**	**.031***
test vs. parsing × pre- vs. delayed-	-9.25 [-19.9, 1.40]	6.47	-1.43	.153
test vs. input × pre- vs. delayed-	**-16.28 [-27.08, -5.48]**	**6.57**	**-2.48**	**.013***
test vs. parsing × matched vs. mismatched	**-14.40 [-24.9, -3.90]**	**6.38**	**-2.26**	**.024***
test vs. input × matched vs. mismatched	**-13.72 [-24.44, -2.99]**	**6.52**	**-2.10**	**.035***
pre- vs. post- × matched vs. mismatched	**-15.17 [-25.97, -4.38]**	**6.56**	**-2.31**	**.021***
pre- vs. delayed- × matched vs. mismatche	**-15.46 [-26.27, -4.65]**	**6.57**	**-2.35**	**.019***
test vs. parsing × pre- vs. post- × matched vs. mismatched	**20.43 [5.50, 35.36]**	**9.08**	**2.25**	**.025***
test vs. input × pre- vs. post- × matched vs. mismatched	**16.13 [0.67, 31.59]**	**9.40**	**1.72**	**.086**
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	12.14 [-2.95, 27.23]	9.18	1.32	.186
test vs. input × pre- vs. delayed- × matched vs. mismatched	14.67 [-.70, 30.03]	9.34	1.57	.116

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$; *** significantly different from zero when $\alpha<.001$

Whole sentence

Table Appx. 119 Fixed effects of model analysis in the SPR test for the SRC-A structure for the whole sentence (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	**-6.93 [-11.12, -2.73]**	**2.55**	**-2.72**	**.007****
test vs. parsing	5.40 [-0.26, 11.07]	3.44	1.57	.117
test vs. input	1.53 [-4.27, 7.32]	3.52	.43	.664
pre- vs. post-	**6.08 [0.27, 11.9]**	**3.53**	**1.72**	**.085**
pre- vs. delayed-	2.25 [-3.53, 8.02]	3.51	0.64	.522
matched vs. mismatched	**15.79 [9.82, 21.77]**	**3.63**	**4.35**	**<.001*****
test vs. parsing × pre- vs. post-	-6.20 [-14.29, 1.88]	4.91	-1.26	.207
test vs. input × pre- vs. post-	-3.32 [-11.62, 4.99]	5.05	-.66	.511
test vs. parsing × pre- vs. delayed-	.68 [-7.43, 8.8]	4.93	.14	.890
test vs. input × pre- vs. delayed-	-.02 [-8.31, 8.27]	5.04	-.01	.996
test vs. parsing × matched vs. mismatched	**-11.93 [-20, -3.87]**	**4.90**	**-2.44**	**.015***
test vs. input × matched vs. mismatched	-5.44 [-13.66, 2.78]	5.00	-1.09	.276
pre- vs. post- × matched vs. mismatched	**-11.52 [-19.75, -3.28]**	**5.01**	**-2.30**	**.022***
pre- vs. delayed- × matched vs. mismatched	**-10.03 [-18.22, -1.84]**	**4.98**	**-2.01**	**.044***
test vs. parsing × pre- vs. post- × matched vs. mismatched	**13.92 [2.47, 25.37]**	**6.96**	**2.00**	**.046***
test vs. input × pre- vs. post- × matched vs. mismatched	6.26 [-5.51, 18.04]	7.16	.88	.382
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	2.83 [-8.7, 14.36]	7.01	.40	.687
test vs. input × pre- vs. delayed- × matched vs. mismatched	4.27 [-7.46, 16.00]	7.13	.60	.550

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$; *** significantly different from zero when $\alpha < .001$

SRC-I

First critical word

Table Appx. 120 Fixed effects of model analysis in the SPR test for the SRC-I structure at the first critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	**6.24 [.86, 11.63]**	**3.28**	**1.91**	**.057***
test vs. parsing	-6.13 [-12.89, .62]	4.11	-1.49	.136
test vs. input	-4.68 [-11.58, 2.22]	4.20	-1.12	.265
pre- vs. post-	-5.73 [-12.69, 1.23]	4.23	-1.35	.176
pre- vs. delayed-	**-12.73 [-19.7, -5.76]**	**4.24**	**-3.00**	**.003****
matched vs. mismatched	-1.60 [-9.16, 5.96]	4.60	-.35	.728
test vs. parsing × pre- vs. post-	5.78 [-3.82, 15.38]	5.84	.99	.322
test vs. input × pre- vs. post-	5.52 [-4.45, 15.49]	6.06	.91	.363
test vs. parsing × pre- vs. delayed-	**14.19 [4.45, 23.92]**	**5.92**	**2.40**	**.017***
test vs. input × pre- vs. delayed-	**11.42 [1.44, 21.41]**	**6.07**	**1.88**	**.060**
test vs. parsing × matched vs. mismatched	2.38 [-7.14, 11.89]	5.79	.41	.681
test vs. input × matched vs. mismatched	5.7 [-4.04, 15.44]	5.92	.96	.336
pre- vs. post- × matched vs. mismatched	2.64 [-7.16, 12.45]	5.96	.44	.658
pre- vs. delayed- × matched vs. mismatched	6.12 [-3.75, 15.98]	6.00	1.02	.308
test vs. parsing × pre- vs. post- × matched vs. mismatched	1.54 [-12.03, 15.11]	8.25	.19	.852
test vs. input × pre- vs. post- × matched vs. mismatched	-7.34 [-21.4, 6.71]	8.55	-.86	.390
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	-7.87 [-21.67, 5.93]	8.39	-.94	.349
test vs. input × pre- vs. delayed- × matched vs. mismatched	-9.49 [-23.58, 4.60]	8.57	-1.11	.268

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$

Second critical word

Table Appx. 121 Fixed effects of model analysis in the SPR test for the SRC-I structure at the second critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	-2.43 [-7.21, 2.35]	2.91	-.84	.403
test vs. parsing	5.01 [-1.27, 11.3]	3.82	1.31	.190
test vs. input	5.9 [-0.52, 12.33]	3.90	1.51	.131
pre- vs. post-	-.16 [-6.65, 6.32]	3.94	-.04	.967
pre- vs. delayed-	-1.91 [-8.39, 4.57]	3.94	-.48	.628
matched vs. mismatched	5.93 [-0.83, 12.7]	4.11	1.44	.149
test vs. parsing × pre- vs. post-	-.80 [-9.72, 8.13]	5.43	-.15	.883
test vs. input × pre- vs. post-	1.31 [-7.93, 10.56]	5.62	.23	.815
test vs. parsing × pre- vs. delayed-	-.04 [-9.09, 9.00]	5.50	-.01	.994
test vs. input × pre- vs. delayed-	-.25 [-9.52, 9.02]	5.64	-.04	.965
test vs. parsing × matched vs. mismatched	**-10.78 [-19.69, -1.86]**	**5.42**	**-1.99**	**.047***
test vs. input × matched vs. mismatched	-3.59 [-12.67, 5.48]	5.52	-.65	.515
pre- vs. post- × matched vs. mismatched	-1.82 [-10.99, 7.35]	5.58	-.33	.744
pre- vs. delayed- × matched vs. mismatched	-2.18 [-11.35, 6.99]	5.57	-.39	.696
test vs. parsing × pre- vs. post- × matched vs. mismatched	12.14 [-0.55, 24.83]	7.71	1.57	.116
test vs. input × pre- vs. post- × matched vs. mismatched	-8.03 [-21.09, 5.02]	7.94	-1.01	.312
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	9.72 [-3.10, 22.54]	7.79	1.25	.213
test vs. input × pre- vs. delayed- × matched vs. mismatched	-1.28 [-14.36, 11.79]	7.95	-.16	.872

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$

Third critical word

Table Appx. 122 Fixed effects of model analysis in the SPR test for the SRC-I structure at the third critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	-1.78 [-7.74, 4.18]	3.63	-.49	.624
test vs. parsing	-1.28 [-8.83, 6.28]	4.59	-.28	.781
test vs. input	3.13 [-4.58, 10.85]	4.69	.67	.504
pre- vs. post-	-3.5 [-11.05, 4.06]	4.59	-.76	.446
pre- vs. delayed-	.71 [-6.81, 8.23]	4.57	.16	.876
matched vs. mismatched	**13.71 [5.45, 21.97]**	**5.02**	**2.73**	**.007****
test vs. parsing × pre- vs. post-	4.22 [-6.19, 14.62]	6.32	.67	.505
test vs. input × pre- vs. post-	.74 [-10.04, 11.51]	6.55	.11	.910
test vs. parsing × pre- vs. delayed-	.86 [-9.72, 11.43]	6.43	.13	.894
test vs. input × pre- vs. delayed-	-2.29 [-13.00, 8.43]	6.51	-.35	.726
test vs. parsing × matched vs. mismatched	-9.71 [-20.13, .72]	6.34	-1.53	.126
test vs. input × matched vs. mismatched	-10.49 [-21.24, .26]	6.53	-1.61	.109
pre- vs. post- × matched vs. mismatched	-5.92 [-16.66, 4.82]	6.53	-.91	.365
pre- vs. delayed- × matched vs. mismatched	**-12.77 [-23.5, -2.04]**	**6.52**	**-1.96**	**.050***
test vs. parsing × pre- vs. post- × matched vs. mismatched	11.43 [-3.39, 26.24]	9.01	1.27	.205
test vs. input × pre- vs. post- × matched vs. mismatched	5.64 [-9.73, 21.01]	9.34	.60	.546
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	12.53 [-2.45, 27.5]	9.10	1.38	.169
test vs. input × pre- vs. delayed- × matched vs. mismatched	10.17 [-5.15, 25.48]	9.31	1.09	.275

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$

Whole sentence

Table Appx. 123 Fixed effects of model analysis in the SPR test for the SRC-I structure for the whole sentence (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	-.45 [-5.12, 4.21]	2.84	-.16	.873
test vs. parsing	.85 [-4.76, 6.45]	3.40	.25	.804
test vs. input	4.58 [-1.16, 10.31]	3.49	1.31	.190
pre- vs. post-	-1.40 [-7.02, 4.21]	3.42	-.41	.681
pre- vs. delayed	-1.40 [-7.02, 4.21]	3.41	-.41	.682
matched vs. mismatched	**10.6 [4.06, 17.14]**	**3.98**	**2.67**	**.008****
test vs. parsing × pre- vs. post-	2.46 [-5.29, 10.21]	4.71	.52	.602
test vs. input × pre- vs. post-	-6.34 [-14.38, 1.70]	4.89	-1.30	.195
test vs. parsing × pre- vs. delayed	.26 [-7.6, 8.12]	4.78	.06	.956
test vs. input × pre- vs. delayed	-2.46 [-10.46, 5.54]	4.86	-.51	.613
test vs. parsing × matched vs. mismatched	**-9.14 [-16.91, -1.37]**	**4.72**	**-1.94**	**.053**
test vs. input × matched vs. mismatched	-5.02 [-13.02, 2.99]	4.87	-1.03	.303
pre- vs. post- × matched vs. mismatched	-4.05 [-12.01, 3.92]	4.84	-.84	.403
pre- vs. delayed × matched vs. mismatched	-6.02 [-14.00, 1.96]	4.85	-1.24	.215
test vs. parsing × pre- vs. post- × matched vs. mismatched	10.49 [-0.52, 21.49]	6.69	1.57	.117
test vs. input × pre- vs. post- × matched vs. mismatched	4.58 [-6.81, 15.97]	6.93	.66	.508
test vs. parsing × pre- vs. delayed × matched vs. mismatched	9.81 [-1.3, 20.91]	6.75	1.45	.147
test vs. input × pre- vs. delayed × matched vs. mismatched	1.97 [-9.42, 13.36]	6.92	.29	.776

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; ** significantly different from zero when $\alpha < .01$

ORC-A

First critical word

Table Appx. 124 Fixed effects of model analysis in the SPR test for the ORC-A structure at the first critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	-2.71 [-7.05, 1.63]	2.64	-1.03	.304
test vs. parsing	-3.29 [-9.21, 2.64]	3.60	-.91	.362
test vs. input	-2.77 [-8.84, 3.30]	3.69	-.75	.453
pre- vs. post-	-3.14 [-9.1, 2.82]	3.62	-.87	.386
pre- vs. delayed-	-3.04 [-9.03, 2.96]	3.65	-.83	.405
matched vs. mismatched	**-7.05 [-13.12, -.99]**	**3.69**	**-1.92**	**.056**
test vs. parsing × pre- vs. post-	**8.87 [.60, 17.14]**	**5.03**	**1.77**	**.078**
test vs. input × pre- vs. post-	6.62 [-1.95, 15.18]	5.21	1.27	.204
test vs. parsing × pre- vs. delayed-	6.17 [-2.2, 14.54]	5.09	1.21	.225
test vs. input × pre- vs. delayed-	3.11 [-5.53, 11.75]	5.25	.59	.554
test vs. parsing × matched vs. mismatched	**9.94 [1.69, 18.18]**	**5.01**	**1.98**	**.048***
test vs. input × matched vs. mismatched	3.84 [-4.59, 12.28]	5.13	.75	.454
pre- vs. post- × matched vs. mismatched	**10.82 [2.39, 19.25]**	**5.13**	**2.11**	**.035***
pre- vs. delayed- × matched vs. mismatched	**8.48 [.16. 96]**	**5.16**	**1.64**	**.100**
test vs. parsing × pre- vs. post- × matched vs. mismatched	**-14.9 [-26.57, -3.22]**	**7.10**	**-2.10**	**.036***
test vs. input × pre- vs. post- × matched vs. mismatched	-6.75 [-18.89, 5.39]	7.38	-.92	.361
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	**-13.97 [-25.83, -2.10]**	**7.21**	**-1.94**	**.053**
test vs. input × pre- vs. delayed- × matched vs. mismatched	-3.29 [-15.45, 8.88]	7.39	-.44	.657

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$

Second critical word

Table Appx. 125 Fixed effects of model analysis in the SPR test for the ORC-A structure at the second critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.27 [-3.58, 6.12]	2.95	.43	.666
test vs. parsing	-1.79 [-8.36, 4.78]	3.99	-.45	.654
test vs. input	-.71 [-7.47, 6.05]	4.11	-.17	.862
pre- vs. post-	-1.05 [-7.72, 5.62]	4.05	-.26	.796
pre- vs. delayed-	-5.79 [-12.46, .88]	4.06	-1.43	.153
matched vs. mismatched	-2.66 [-9.37, 4.04]	4.07	-.65	.513
test vs. parsing × pre- vs. post-	-3.88 [-13.05, 5.30]	5.58	-.70	.487
test vs. input × pre- vs. post-	-3.86 [-13.41, 5.69]	5.81	-.67	.506
test vs. parsing × pre- vs. delayed-	.77 [-8.56, 10.10]	5.67	.14	.892
test vs. input × pre- vs. delayed-	.71 [-8.83, 10.24]	5.80	.12	.903
test vs. parsing × matched vs. mismatched	1.59 [-7.52, 10.71]	5.54	.29	.774
test vs. input × matched vs. mismatched	-2.89 [-12.25, 6.46]	5.69	-.51	.611
pre- vs. post- × matched vs. mismatched	1.02 [-8.46, 10.50]	5.76	.18	.860
pre- vs. delayed- × matched vs. mismatched	5.5 [-3.92, 14.92]	5.73	.96	.337
test vs. parsing × pre- vs. post- × matched vs. mismatched	1.61 [-11.41, 14.63]	7.92	.20	.839
test vs. input × pre- vs. post- × matched vs. mismatched	7.55 [-5.98, 21.09]	8.23	.92	.359
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	-2.16 [-15.31, 10.99]	7.99	-.27	.787
test vs. input × pre- vs. delayed- × matched vs. mismatched	-1.72 [-15.19, 11.76]	8.19	-.21	.834

Note: parsing = parsing group; input = input flood group; test = test-only group

Third critical word

Table Appx. 126 Fixed effects of model analysis in the SPR test for the ORC-A structure at the third critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	-1.43 [-7.37, 4.52]	3.61	-.40	.693
test vs. parsing	-.32 [-8.25, 7.62]	4.82	-.07	.947
test vs. input	2.97 [-5.12, 11.05]	4.91	.60	.546
pre- vs. post-	1.57 [-6.54, 9.67]	4.93	.32	.751
pre- vs. delayed-	-.98 [-9.12, 7.17]	4.95	-.20	.843
matched vs. mismatched	.94 [-7.53, 9.42]	5.15	.18	.855
test vs. parsing × pre- vs. post-	-2.08 [-13.31, 9.14]	6.83	-.31	.760
test vs. input × pre- vs. post-	-3.83 [-15.39, 7.72]	7.03	-.55	.585
test vs. parsing × pre- vs. delayed-	.95 [-10.44, 12.33]	6.92	.14	.891
test vs. input × pre- vs. delayed-	-2.22 [-13.82, 9.38]	7.05	-.31	.753
test vs. parsing × matched vs. mismatched	4.40 [-6.95, 15.76]	6.90	.64	.524
test vs. input × matched vs. mismatched	2.26 [-9.28, 13.79]	7.01	.32	.747
pre- vs. post- × matched vs. mismatched	.98 [-10.70, 12.66]	7.10	.14	.890
pre- vs. delayed- × matched vs. mismatched	-.62 [-12.25, 11.02]	7.07	-.09	.930
test vs. parsing × pre- vs. post- × matched vs. mismatched	-.50 [-16.63, 15.63]	9.80	-.05	.959
test vs. input × pre- vs. post- × matched vs. mismatched	-.67 [-17.32, 15.98]	10.12	-.07	.947
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	-1.84 [-18.11, 14.43]	9.89	-.19	.853
test vs. input × pre- vs. delayed- × matched vs. mismatched	-7.38 [-23.88, 9.11]	10.03	-.74	.462

Note: parsing = parsing group; input = input flood group; test = test-only group

Whole sentence

Table Appx. 127 Fixed effects of model analysis in the SPR test for the ORC-A structure for the whole sentence (baseline of the test-only group)

fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.85 [-2.54, 6.23]	2.66	.69	.489
test vs. parsing	-2.33 [-8.1, 3.44]	3.51	-.67	.506
test vs. input	1.86 [-3.99, 7.70]	3.56	.52	.602
pre- vs. post-	-1.85 [-7.74, 4.05]	3.58	-.52	.606
pre- vs. delayed-	-4.78 [-10.7, 1.14]	3.60	-1.33	.184
matched vs. mismatched	1.39 [-4.87, 7.64]	3.80	.36	.716
test vs. parsing × pre- vs. post-	1.16 [-7.03, 9.36]	4.98	.23	.815
test vs. input × pre- vs. post-	-2.81 [-11.24, 5.62]	5.13	-.55	.584
test vs. parsing × pre- vs. delayed-	4.10 [-4.24, 12.44]	5.07	.81	.419
test vs. input × pre- vs. delayed-	-3.46 [-11.89, 4.96]	5.12	-.68	.499
test vs. parsing × matched vs. mismatched	6.12 [-2.11, 14.36]	5.01	1.22	.222
test vs. input × matched vs. mismatched	-2.06 [-10.44, 6.32]	5.09	-.40	.686
pre- vs. post- × matched vs. mismatched	4.38 [-4.05, 12.82]	5.13	.86	.393
pre- vs. delayed- × matched vs. mismatched	4.27 [-4.17, 12.72]	5.13	.83	.405
test vs. parsing × pre- vs. post- × matched vs. mismatched	-7.64 [-19.31, 4.03]	7.10	-1.08	.282
test vs. input × pre- vs. post- × matched vs. mismatched	.49 [-11.59, 12.57]	7.34	.07	.947
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	-8.45 [-20.32, 3.41]	7.21	-1.17	.241
test vs. input × pre- vs. delayed- × matched vs. mismatched	2.45 [-9.54, 14.45]	7.29	.34	.737

Note: parsing = parsing group; input = input flood group; test = test-only group

ORC-I

First critical word

Table Appx. 128 Fixed effects of model analysis in the SPR test for the ORC-I structure at the first critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	-1.26 [-5.68, 3.17]	2.69	-.47	.640
test vs. parsing	2.49 [-3.36, 8.34]	3.56	.70	.484
test vs. input	1.26 [-4.69, 7.20]	3.61	.35	.728
pre- vs. post-	-1.29 [-7.13, 4.55]	3.55	-.36	.717
pre- vs. delayed-	-1.08 [-6.88, 4.73]	3.53	-.31	.760
matched vs. mismatched	-5.83 [-11.87, .22]	3.67	-1.59	.113
test vs. parsing × pre- vs. post-	.76 [-7.32, 8.83]	4.91	.16	.877
test vs. input × pre- vs. post-	-.53 [-8.85, 7.79]	5.06	-.10	.917
test vs. parsing × pre- vs. delayed-	-3.57 [-11.74, 4.60]	4.97	-.72	.473
test vs. input × pre- vs. delayed-	-.50 [-8.78, 7.79]	5.04	-.10	.921
test vs. parsing × matched vs. mismatched	1.75 [-6.29, 9.79]	4.89	.36	.721
test vs. input × matched vs. mismatched	1.58 [-6.58, 9.75]	4.96	.32	.750
pre- vs. post- × matched vs. mismatched	5.24 [-2.97, 13.45]	4.99	1.05	.294
pre- vs. delayed- × matched vs. mismatched	7.82 [-.42, 16.05]	5.01	1.56	.119
test vs. parsing × pre- vs. post- × matched vs. mismatched	-3.83 [-15.22, 7.56]	6.92	-.55	.580
test vs. input × pre- vs. post- × matched vs. mismatched	-1.82 [-13.57, 9.93]	7.15	-.26	.799
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	0.13 [-11.43, 11.68]	7.02	.02	.986
test vs. input × pre- vs. delayed- × matched vs. mismatched	1.56 [-10.16, 13.29]	7.13	.22	.827

Note: parsing = parsing group; input = input flood group; test = test-only group

Second critical word

Table Appx. 129 Fixed effects of model analysis in the SPR test for the ORC-I structure at the second critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	5.85 [0.71, 10.99]	3.12	1.87	.062
test vs. parsing	-1.56 [-8.35, 5.24]	4.13	-.38	.707
test vs. input	-3.46 [-10.34, 3.42]	4.18	-.83	.409
pre- vs. post-	-5.64 [-12.51, 1.23]	4.18	-1.35	.177
pre- vs. delayed-	-3.38 [-10.26, 3.51]	4.19	-.81	.420
matched vs. mismatched	**-8.18 [-15.31, -1.04]**	**4.34**	**-1.89**	**.060**
test vs. parsing × pre- vs. post-	5.84 [-3.66, 15.34]	5.77	1.01	.312
test vs. input × pre- vs. post-	1.81 [-8.00, 11.62]	5.96	.30	.761
test vs. parsing × pre- vs. delayed-	-5.64 [-15.34, 4.06]	5.89	-.96	.339
test vs. input × pre- vs. delayed-	.73 [-9.03, 10.5]	5.94	.12	.902
test vs. parsing × matched vs. mismatched	2.14 [-7.34, 11.63]	5.77	.37	.710
test vs. input × matched vs. mismatched	1.05 [-8.61, 10.72]	5.87	.18	.858
pre- vs. post- × matched vs. mismatched	7.48 [-2.17, 17.14]	5.87	1.28	.203
pre- vs. delayed- × matched vs. mismatched	8.4 [-1.34, 18.13]	5.92	1.42	.156
test vs. parsing × pre- vs. post- × matched vs. mismatched	-6.84 [-20.23, 6.54]	8.14	-.84	.401
test vs. input × pre- vs. post- × matched vs. mismatched	.62 [-13.22, 14.45]	8.41	.07	.942
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	4.91 [-8.71, 18.52]	8.28	.59	.553
test vs. input × pre- vs. delayed- × matched vs. mismatched	5.44 [-8.39, 19.26]	8.40	.65	.518

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect

Third critical word

Table Appx. 130 Fixed effects of model analysis in the SPR test for the ORC-I structure at the third critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	.08 [-6.21, 6.37]	3.83	.02	.983
test vs. parsing	-.80 [-8.62, 7.02]	4.75	-.17	.867
test vs. input	3.92 [-4.06, 11.91]	4.85	.81	.419
pre- vs. post-	3.83 [-4.16, 11.82]	4.86	.79	.431
pre- vs. delayed-	-1.98 [-9.93, 5.98]	4.84	-.41	.683
matched vs. mismatched	5.85 [-3.11, 14.80]	5.45	1.07	.284
test vs. parsing × pre- vs. post-	-.83 [-11.86, 10.2]	6.70	-.12	.902
test vs. input × pre- vs. post-	**-11.69 [-23.18, -.20]**	**6.99**	**-1.67**	**.095**
test vs. parsing × pre- vs. delayed-	1.67 [-9.50, 12.85]	6.79	.25	.806
test vs. input × pre- vs. delayed-	-1.93 [-13.34, 9.49]	6.94	-.28	.782
test vs. parsing × matched vs. mismatched	-2.08 [-13.25, 9.10]	6.80	-.31	.760
test vs. input × matched vs. mismatched	-10.57 [-21.96, .83]	6.93	-1.53	.127
pre- vs. post- × matched vs. mismatched	-10.24 [-21.6, 1.12]	6.90	-1.48	.138
pre- vs. delayed- × matched vs. mismatched	-1.89 [-13.24, 9.46]	6.90	-.27	.784
test vs. parsing × pre- vs. post- × matched vs. mismatched	6.65 [-9.07, 22.36]	9.55	.70	.487
test vs. input × pre- vs. post- × matched vs. mismatched	**19.2 [2.91, 35.49]**	**9.90**	**1.94**	**.053**
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	1.05 [-14.86, 16.96]	9.67	.11	.914
test vs. input × pre- vs. delayed- × matched vs. mismatched	13.11 [-3.08, 29.31]	9.85	1.33	.183

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect

Whole sentence

Table Appx. 131 Fixed effects of model analysis in the SPR test for the ORC-I structure for the whole sentence (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.31 [-3.43, 6.04]	2.88	.45	.650
test vs. parsing	1.45 [-4.35, 7.25]	3.53	.41	.681
test vs. input	3.4 [-2.49, 9.30]	3.58	.95	.342
pre- vs. post-	1.59 [-4.19, 7.37]	3.51	.45	.650
pre- vs. delayed-	-0.89 [-6.62, 4.84]	3.48	-.26	.799
matched vs. mismatched	2.54 [-4.00, 9.09]	3.98	.64	.523
test vs. parsing × pre- vs. post-	-3.36 [-11.33, 4.62]	4.85	-.69	.489
test vs. input × pre- vs. post-	-6.85 [-15.14, 1.43]	5.04	-1.36	.174
test vs. parsing × pre- vs. delayed-	-1.71 [-9.73, 6.31]	4.88	-.35	.726
test vs. input × pre- vs. delayed-	-1.67 [-9.82, 6.48]	4.96	-.34	.736
test vs. parsing × matched vs. mismatched	-1.45 [-9.42, 6.52]	4.84	-.30	.764
test vs. input × matched vs. mismatched	-2.68 [-10.81, 5.45]	4.94	-.54	.588
pre- vs. post- × matched vs. mismatched	-2.41 [-10.57, 5.74]	4.96	-.49	.626
pre- vs. delayed- × matched vs. mismatched	1.81 [-6.37, 9.98]	4.97	.36	.716
test vs. parsing × pre- vs. post- × matched vs. mismatched	6.97 [-4.29, 18.23]	6.84	1.02	.309
test vs. input × pre- vs. post- × matched vs. mismatched	8.86 [-2.82, 20.55]	7.11	1.25	.212
test vs. parsing × pre- vs. delayed- × matched vs. mismatched	3.78 [-7.63, 15.19]	6.94	.55	.586
test vs. input × pre- vs. delayed- × matched vs. mismatched	1.28 [-10.32, 12.89]	7.06	.18	.856

Note: parsing = parsing group; input = input flood group; test = test-only group

Appendix 28 RQ2: Fixed effects of model analysis in the SPR test (baseline of the input flood group)

SRC-A

First critical word

Table Appx. 132 Fixed effects of model analysis in the SPR test for the SRC-A structure at the first critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	**-8.74 [-13.66, -3.81]**	**2.99**	**-2.92**	**.004****
input vs. parsing	**8.49 [1.72, 15.27]**	**4.12**	**2.06**	**.039***
input vs. test	4.78 [-2.14, 11.70]	4.21	1.14	.256
pre- vs. post-	3.1 [-4.00, 10.19]	4.31	.72	.473
pre- vs. delayed-	4.47 [-2.61, 11.54]	4.30	1.04	.300
match vs. mismatch	**9.76 [2.85, 16.66]**	**4.20**	**2.33**	**.020***
input vs. parsing × pre- vs. post-	-4.82 [-14.52, 4.88]	5.90	-.82	.414
input vs. test × pre- vs. post-	.05 [-9.84, 9.94]	6.01	.01	.993
input vs. parsing × pre- vs. delayed-	-3.95 [-13.77, 5.86]	5.96	-.66	.507
input vs. test × pre- vs. delayed-	-3.78 [-13.65, 6.10]	6.00	-.63	.529
input vs. parsing × match vs. mismatch	**-11.93 [-21.49, -2.36]**	**5.81**	**-2.05**	**.040***
input vs. test × match vs. mismatch	-.73 [-1.53, 9.06]	5.95	-.12	.902
pre- vs. post- × match vs. mismatch	-1.26 [-11.28, 8.75]	6.09	-.21	.835
pre- vs. delayed- × match vs. mismatch	**-1.94 [-2.95, -.94]**	**6.08**	**-1.80**	**.072**
input vs. parsing × pre- vs. post- × match vs. mismatch	5.73 [-7.98, 19.44]	8.33	.69	.492
input vs. test × pre- vs. post- × match vs. mismatch	-7.86 [-21.87, 6.15]	8.52	-.92	.356
input vs. parsing × pre- vs. delayed- × match vs. mismatch	11.68 [-2.19, 25.55]	8.43	1.39	.166
input vs. test × pre- vs. delayed- × match vs. mismatch	2.46 [-11.57, 16.48]	8.53	.29	.773

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; *significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$

Second critical word

Table Appx. 133 Fixed effects of model analysis in the SPR test for the SRC-A structure at the second critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	-3.06 [-7.63, 1.51]	2.78	-1.10	.270
input vs. parsing	4.64 [-1.61, 1.89]	3.80	1.22	.223
input vs. test	-5.55 [-11.93, .83]	3.88	-1.43	.152
pre- vs. post-	5.89 [-.63, 12.42]	3.97	1.49	.138
pre- vs. delayed-	-.83 [-7.38, 5.72]	3.98	-.21	.835
match vs. mismatch	1.49 [-4.99, 7.96]	3.94	.38	.706
input vs. parsing × pre- vs. post-	-8.25 [-17.22, .73]	5.46	-1.51	.131
input vs. test × pre- vs. post-	5.85 [-3.29, 14.98]	5.56	1.05	.293
input vs. parsing × pre- vs. delayed-	.22 [-8.86, 9.31]	5.52	.04	.968
input vs. test × pre- vs. delayed-	8.53 [-.62, 17.68]	5.56	1.53	.125
input vs. parsing × match vs. mismatch	-2.27 [-11.12, 6.57]	5.38	-.42	.672
input vs. test × match vs. mismatch	6.58 [-2.49, 15.65]	5.52	1.19	.233
pre- vs. post- × match vs. mismatch	-1.60 [-1.89, 7.69]	5.65	-.28	.777
pre- vs. delayed- × match vs. mismatch	3.93 [-5.37, 13.23]	5.66	.70	.487
input vs. parsing × pre- vs. post- × match vs. mismatch	6.25 [-6.5, 18.99]	7.75	.81	.420
input vs. test × pre- vs. post- × match vs. mismatch	-8.4 [-21.38, 4.59]	7.89	-1.06	.288
input vs. parsing × pre- vs. delayed- × match vs. mismatch	.47 [-12.4, 13.35]	7.83	.06	.952
input vs. test × pre- vs. delayed- × match vs. mismatch	-8.44 [-21.48, 4.61]	7.93	-1.06	.287

Note: parsing = parsing group; input = input flood group; test = test-only group

Third critical word

Table Appx. 134 Fixed effects of model analysis in the SPR test for the SRC-A structure at the third critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	2.17 [-3.47, 7.82]	3.43	.63	.527
input vs. parsing	-3.15 [-1.55, 4.25]	4.50	-.70	.484
input vs. test	**-11.15 [-18.7, -3.60]**	**4.59**	**-2.43**	**.015***
pre- vs. post-	-6.29 [-14.09, 1.50]	4.74	-1.33	.184
pre- vs. delayed-	**-8.3 [-15.99, -.62]**	**4.67**	**-1.78**	**.076**
match vs. mismatch	4.43 [-3.59, 12.44]	4.87	.91	.364
input vs. parsing × pre- vs. post-	2.98 [-7.68, 13.63]	6.48	.46	.646
input vs. test × pre- vs. post-	**14.27 [3.39, 25.15]**	**6.62**	**2.16**	**.031***
input vs. parsing × pre- vs. delayed-	7.03 [-3.67, 17.73]	6.51	1.08	.280
input vs. test × pre- vs. delayed-	**16.28 [5.48, 27.08]**	**6.57**	**2.48**	**.013***
input vs. parsing × match vs. mismatch	-.68 [-11.15, 9.79]	6.37	-.11	.915
input vs. test × match vs. mismatch	**13.72 [2.99, 24.44]**	**6.52**	**2.10**	**.036***
pre- vs. post- × match vs. mismatch	.96 [-1.11, 12.03]	6.73	.14	.887
pre- vs. delayed- × match vs. mismatch	-.79 [-11.72, 1.15]	6.65	-.12	.906
input vs. parsing × pre- vs. post- × match vs. mismatch	4.3 [-1.83, 19.43]	9.20	.47	.640
input vs. test × pre- vs. post- × match vs. mismatch	**-16.13 [-31.59, -.67]**	**9.40**	**-1.72**	**.086**
input vs. parsing × pre- vs. delayed- × match vs. mismatch	-2.53 [-17.7, 12.64]	9.22	-.27	.784
input vs. test × pre- vs. delayed- × match vs. mismatch	-14.67 [-3.03, .7]	9.34	-1.57	.117

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; *significantly differently from zero when $\alpha \leq .05$

Whole sentence

Table Appx. 135 Fixed effects of model analysis in the SPR test for the SRC-A structure for the whole sentence (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	-5.40 [-9.67, -1.12]	2.60	-2.08	.038*
input vs. parsing	3.87 [-1.85, 9.60]	3.48	1.11	.266
input vs. test	-1.53 [-7.32, 4.27]	3.52	-.43	.664
pre- vs. post-	2.77 [-3.17, 8.7]	3.61	.77	.444
pre- vs. delayed-	2.22 [-3.73, 8.18]	3.62	.62	.539
match vs. mismatch	1.35 [4.32, 16.38]	3.67	2.82	.005**
input vs. parsing × pre- vs. post-	-2.89 [-11.06, 5.29]	4.97	-.58	.561
input vs. test × pre- vs. post-	3.32 [-4.99, 11.62]	5.05	.66	.511
input vs. parsing × pre- vs. delayed-	.71 [-7.53, 8.95]	5.01	.14	.888
input vs. test × pre- vs. delayed-	.02 [-8.27, 8.31]	5.04	.01	.996
input vs. parsing × match vs. mismatch	-6.49 [-14.6, 1.61]	4.93	-1.32	.188
input vs. test × match vs. mismatch	5.44 [-2.78, 13.66]	5.00	1.09	.276
pre- vs. post- × match vs. mismatch	-5.26 [-13.67, 3.16]	5.12	-1.03	.304
pre- vs. delayed- × match vs. mismatch	-5.76 [-14.17, 2.64]	5.11	-1.13	.259
input vs. parsing × pre- vs. post- × match vs. mismatch	7.66 [-3.92, 19.24]	7.04	1.09	.277
input vs. test × pre- vs. post- × match vs. mismatch	-6.26 [-18.04, 5.51]	7.16	-.88	.382
input vs. parsing × pre- vs. delayed- × match vs. mismatch	-1.44 [-13.11, 1.24]	7.10	-.20	.840
input vs. test × pre- vs. delayed- × match vs. mismatch	-4.27 [-16.00, 7.46]	7.13	-.60	.550

Note: parsing = parsing group; input = input flood group; test = test-only group;

*significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$

SRC-I

First critical word

Table Appx. 136 Fixed effects of model analysis in the SPR test for the SRC-I structure at the first critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.56 [-3.76, 6.89]	3.24	.48	.630
parsing vs. parsing	-1.45 [-8.16, 5.26]	4.08	-.36	.722
parsing vs. test	4.68 [-2.22, 11.58]	4.20	1.12	.265
pre- vs. post-	-.21 [-7.36, 6.94]	4.35	-.05	.961
pre- vs. delayed-	-1.31 [-8.46, 5.85]	4.35	-.30	.764
match vs. mismatch	4.1 [-3.45, 11.66]	4.59	.89	.372
parsing vs. parsing × pre- vs. post-	.27 [-9.48, 1.01]	5.93	.05	.964
parsing vs. test × pre- vs. post-	-5.52 [-15.49, 4.45]	6.06	-.91	.363
parsing vs. parsing × pre- vs. delayed-	2.76 [-7.1, 12.62]	5.99	.46	.645
parsing vs. test × pre- vs. delayed-	**-11.42 [-21.41, -1.44]**	**6.07**	**-1.88**	**.060**
parsing vs. parsing × match vs. mismatch	-3.33 [-12.84, 6.19]	5.78	-.58	.565
parsing vs. test × match vs. mismatch	-5.7 [-15.44, 4.04]	5.92	-.96	.336
pre- vs. post- × match vs. mismatch	-4.7 [-14.78, 5.38]	6.13	-.77	.443
pre- vs. delayed- × match vs. mismatch	-3.38 [-13.45, 6.7]	6.12	-.55	.581
parsing vs. parsing × pre- vs. post- × match vs. mismatch	8.88 [-4.89, 22.66]	8.38	1.06	.289
parsing vs. test × pre- vs. post- × match vs. mismatch	7.34 [-6.71, 21.4]	8.55	.86	.390
parsing vs. parsing × pre- vs. delayed- × match vs. mismatch	1.63 [-12.31, 15.56]	8.47	.19	.848
parsing vs. test × pre- vs. delayed- × match vs. mismatch	9.49 [-4.6, 23.58]	8.57	1.11	.268

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect

Second critical word

Table Appx. 137 Fixed effects of model analysis in the SPR test for the SRC-I structure at the second critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	3.47 [-1.23, 8.18]	2.86	1.21	.225
input vs. parsing	-.89 [-7.12, 5.34]	3.79	-.24	.814
input vs. test	-5.9 [-12.33, .52]	3.91	-1.51	.131
pre- vs. post-	1.15 [-5.45, 7.75]	4.01	.29	.775
pre- vs. delayed-	-2.16 [-8.8, 4.48]	4.04	-.54	.593
match vs. mismatch	2.34 [-4.3, 8.98]	4.04	.58	.562
input vs. parsing × pre- vs. post-	-2.11 [-11.12, 6.9]	5.48	-.39	.700
input vs. test × pre- vs. post-	-1.31 [-1.56, 7.93]	5.62	-.23	.815
input vs. parsing × pre- vs. delayed-	.21 [-8.94, 9.36]	5.56	.04	.970
input vs. test × pre- vs. delayed-	.25 [-9.02, 9.52]	5.64	.04	.965
input vs. parsing × match vs. mismatch	-7.18 [-16, 1.64]	5.36	-1.34	.181
input vs. test × match vs. mismatch	3.59 [-5.48, 12.67]	5.52	.65	.515
pre- vs. post- × match vs. mismatch	**-9.86 [-19.16, -.56]**	**5.65**	**-1.74**	**.081**
pre- vs. delayed- × match vs. mismatch	-3.46 [-12.79, 5.87]	5.67	-.61	.542
input vs. parsing × pre- vs. post- × match vs. mismatch	**2.17 [7.39, 32.96]**	**7.77**	**2.60**	**.009****
input vs. test × pre- vs. post- × match vs. mismatch	8.03 [-5.02, 21.09]	7.94	1.01	.312
input vs. parsing × pre- vs. delayed- × match vs. mismatch	11.00 [-1.93, 23.93]	7.86	1.40	.162
input vs. test × pre- vs. delayed- × match vs. mismatch	1.28 [-11.79, 14.36]	7.95	.16	.872

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; ** significantly different from zero when α < .01

Third critical word

Table Appx. 138 Fixed effects of model analysis in the SPR test for the SRC-I structure at the third critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.35 [-4.52, 7.23]	3.57	.38	.705
input vs. parsing	-4.41 [-11.89, 3.08]	4.55	-.97	.333
input vs. test	-3.13 [-1.85, 4.58]	4.69	-.67	.504
pre- vs. post-	-2.76 [-1.45, 4.92]	4.67	-.59	.555
pre- vs. delayed-	-1.58 [-9.22, 6.06]	4.65	-.34	.734
match vs. mismatch	3.22 [-5.04, 11.48]	5.02	.64	.522
input vs. parsing × pre- vs. post-	3.48 [-7.02, 13.97]	6.38	.55	.586
input vs. test × pre- vs. post-	-.74 [-11.51, 1.04]	6.55	-.11	.910
input vs. parsing × pre- vs. delayed-	3.14 [-7.51, 13.8]	6.48	.49	.627
input vs. test × pre- vs. delayed-	2.29 [-8.43, 13]	6.51	.35	.726
input vs. parsing × match vs. mismatch	.78 [-9.64, 11.21]	6.34	.12	.901
input vs. test × match vs. mismatch	1.49 [-.26, 21.24]	6.53	1.61	.109
pre- vs. post- × match vs. mismatch	-.28 [-11.28, 1.71]	6.68	-.04	.966
pre- vs. delayed- × match vs. mismatch	-2.6 [-13.54, 8.33]	6.65	-.39	.695
input vs. parsing × pre- vs. post- × match vs. mismatch	5.79 [-9.21, 2.79]	9.12	.64	.526
input vs. test × pre- vs. post- × match vs. mismatch	-5.64 [-21.01, 9.73]	9.34	-.60	.546
input vs. parsing × pre- vs. delayed- × match vs. mismatch	2.36 [-12.75, 17.47]	9.19	.26	.797
input vs. test × pre- vs. delayed- × match vs. mismatch	-1.17 [-25.48, 5.15]	9.31	-1.09	.275

Note: parsing = parsing group; input = input flood group; test = test-only group

Whole sentence

Table Appx. 139 Fixed effects of model analysis in the SPR test for the SRC-I structure for the whole sentence (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	4.12 [-.53, 8.77]	2.83	1.46	.146
input vs. parsing	-3.73 [-9.32, 1.86]	3.40	-1.10	.272
input vs. test	-4.58 [-1.31, 1.16]	3.49	-1.31	.190
pre- vs. post-	**-7.74 [-13.5, -1.98]**	**3.50**	**-2.21**	**.027***
pre- vs. delayed-	-3.86 [-9.57, 1.85]	3.47	-1.11	.266
match vs. mismatch	5.58 [-.93, 12.09]	3.96	1.41	.160
input vs. parsing × pre- vs. post-	**8.79 [.94, 16.65]**	**4.78**	**1.84**	**.066**
input vs. test × pre- vs. post-	6.34 [-1.7, 14.38]	4.89	1.30	.195
input vs. parsing × pre- vs. delayed-	2.72 [-5.2, 1.65]	4.82	.57	.572
input vs. test × pre- vs. delayed-	2.46 [-5.54, 1.46]	4.86	.51	.613
input vs. parsing × match vs. mismatch	-4.12 [-11.87, 3.63]	4.71	-.88	.382
input vs. test × match vs. mismatch	5.02 [-2.99, 13.02]	4.87	1.03	.303
pre- vs. post- × match vs. mismatch	.53 [-7.62, 8.68]	4.95	.11	.914
pre- vs. delayed- × match vs. mismatch	-4.05 [-12.18, 4.09]	4.95	-.82	.414
input vs. parsing × pre- vs. post- × match vs. mismatch	5.91 [-5.23, 17.05]	6.77	.87	.383
input vs. test × pre- vs. post- × match vs. mismatch	-4.58 [-15.97, 6.81]	6.93	-.66	.509
input vs. parsing × pre- vs. delayed- × match vs. mismatch	7.83 [-3.38, 19.05]	6.82	1.15	.251
input vs. test × pre- vs. delayed- × match vs. mismatch	-1.97 [-13.36, 9.42]	6.92	-.29	.776

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly different from zero when $\alpha < .05$

ORC-A

First critical word

Table Appx. 140 Fixed effects of model analysis in the SPR test for the ORC-A structure at the first critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	-5.48 [-9.87, -1.10]	2.67	-2.06	.040*
input vs. parsing group	-.52 [-6.48, 5.45]	3.63	-.14	.887
input vs. test	2.77 [-3.3, 8.84]	3.69	.75	.453
pre- vs. post-	3.48 [-2.68, 9.63]	3.74	.93	.353
pre- vs. delayed post-	.07 [-6.15, 6.30]	3.78	.02	.985
match vs. mismatch	-3.21 [-9.28, 2.86]	3.69	-.87	.385
input vs. parsing group × pre- vs. post-	2.26 [-6.15, 1.67]	5.11	.44	.659
input vs. test × pre- vs. post-	-6.62 [-15.18, 1.95]	5.21	-1.27	.204
input vs. parsing group × pre- vs. delayed post-	3.07 [-5.47, 11.60]	5.19	.59	.555
input vs. test × pre- vs. delayed post-	-3.11 [-11.75, 5.53]	5.25	-.59	.554
input vs. parsing group × match vs. mismatch	6.09 [-2.16, 14.35]	5.02	1.22	.225
input vs. test × match vs. mismatch	-3.84 [-12.28, 4.59]	5.13	-.75	.454
pre- vs. post- × match vs. mismatch	4.07 [-4.66, 12.81]	5.31	.77	.443
pre- vs. delayed post- × match vs. mismatch	5.19 [-3.54, 13.92]	5.31	.98	.328
input vs. parsing group × pre- vs. post- × match vs. mismatch	-8.15 [-2.05, 3.75]	7.23	-1.13	.260
input vs. test × pre- vs. post- × match vs. mismatch	6.75 [-5.39, 18.89]	7.38	.92	.361
input vs. parsing group × pre- vs. delayed post- × match vs. mismatch	-1.68 [-22.71, 1.36]	7.32	-1.46	.145
input vs. test × pre- vs. delayed post- × match vs. mismatch	3.29 [-8.88, 15.45]	7.39	.44	.657

Note: parsing = parsing group; input = input flood group; test = test-only group; *

significantly different from zero when α < .05

Second critical word

Table Appx. 141 Fixed effects of model analysis in the SPR test for the ORC-A structure at the second critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	.56 [-4.27, 5.39]	2.94	.19	.849
input vs. parsing	-1.08 [-7.63, 5.48]	3.98	-.27	.787
input vs. test	.71 [-6.05, 7.47]	4.11	.17	.862
pre- vs. post-	-4.91 [-11.75, 1.92]	4.16	-1.18	.237
pre- vs. delayed-	-5.09 [-11.91, 1.74]	4.15	-1.23	.220
match vs. mismatch	-5.55 [-12.25, 1.14]	4.07	-1.36	.173
input vs. parsing × pre- vs. post-	-.02 [-9.31, 9.28]	5.65	.00	.998
input vs. test × pre- vs. post-	3.86 [-5.69, 13.41]	5.81	.67	.506
input vs. parsing × pre- vs. delayed-	.07 [-9.37, 9.51]	5.74	.01	.991
input vs. test × pre- vs. delayed-	-.71 [-1.24, 8.83]	5.80	-.12	.903
input vs. parsing × match vs. mismatch	4.48 [-4.63, 13.6]	5.54	.81	.419
input vs. test × match vs. mismatch	2.89 [-6.46, 12.25]	5.69	.51	.611
pre- vs. post- × match vs. mismatch	8.57 [-1.09, 18.23]	5.87	1.46	.145
pre- vs. delayed- × match vs. mismatch	3.78 [-5.86, 13.42]	5.86	.65	.519
input vs. parsing × pre- vs. post- × match vs. mismatch	-5.94 [-19.1, 7.21]	8.00	-.74	.458
input vs. test × pre- vs. post- × match vs. mismatch	-7.55 [-21.09, 5.98]	8.23	-.92	.359
input vs. parsing × pre- vs. delayed- × match vs. mismatch	-.44 [-13.75, 12.86]	8.09	-.06	.957
input vs. test × pre- vs. delayed- × match vs. mismatch	1.72 [-11.76, 15.19]	8.19	.21	.834

Note: parsing = parsing group; input = input flood group; test = test-only group

Third critical word

Table Appx. 142 Fixed effects of model analysis in the SPR test for the ORC-A structure at the third critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.54 [-4.30, 7.38]	3.55	.43	.665
input vs. parsing	-3.29 [-11.14, 4.57]	4.78	-.69	.492
input vs. test	-2.97 [-11.05, 5.12]	4.91	-.60	.546
pre- vs. post-	-2.27 [-1.51, 5.98]	5.01	-.45	.651
pre- vs. delayed-	-3.19 [-11.46, 5.07]	5.03	-.64	.525
match vs. mismatch	3.2 [-5.13, 11.53]	5.06	.63	.527
input vs. parsing × pre- vs. post-	1.75 [-9.58, 13.08]	6.89	.25	.800
input vs. test × pre- vs. post-	3.83 [-7.72, 15.39]	7.03	.55	.585
input vs. parsing × pre- vs. delayed-	3.16 [-8.31, 14.63]	6.97	.45	.650
input vs. test × pre- vs. delayed-	2.22 [-9.38, 13.82]	7.05	.31	.753
input vs. parsing × match vs. mismatch	2.15 [-9.1, 13.39]	6.84	.31	.754
input vs. test × match vs. mismatch	-2.26 [-13.79, 9.28]	7.01	-.32	.747
pre- vs. post- × match vs. mismatch	.31 [-11.56, 12.19]	7.22	.04	.965
pre- vs. delayed- × match vs. mismatch	-8.00 [-19.7, 3.7]	7.11	-1.13	.261
input vs. parsing × pre- vs. post- × match vs. mismatch	.17.00 [-16.1, 16.44]	9.89	.02	.986
input vs. test × pre- vs. post- × match vs. mismatch	.67 [-15.98, 17.32]	1.12	.07	.947
input vs. parsing × pre- vs. delayed- × match vs. mismatch	5.55 [-1.76, 21.85]	9.91	.56	.576
input vs. test × pre- vs. delayed- × match vs. mismatch	7.38 [-9.11, 23.88]	1.03	.74	.462

Note: parsing = parsing group; input = input flood group; test = test-only group

Whole sentence

Table Appx. 143 Fixed effects of model analysis in the SPR test for the ORC-A structure for the whole sentence (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	3.70 [-.63, 8.03]	2.63	1.41	.160
input vs. parsing	-4.19 [-9.92, 1.54]	3.48	-1.20	.229
input vs. test	-1.86 [-7.7, 3.99]	3.56	-.52	.602
pre- vs. post-	-4.66 [-1.69, 1.38]	3.67	-1.27	.204
pre- vs. delayed-	-8.24 [-14.25, -2.24]	3.65	-2.26	.024
match vs. mismatch	-.67 [-6.88, 5.53]	3.77	-.18	.858
input vs. parsing × pre- vs. post-	3.97 [-4.32, 12.27]	5.04	.79	.431
input vs. test × pre- vs. post-	2.81 [-5.62, 11.24]	5.13	.55	.584
input vs. parsing × pre- vs. delayed-	7.56 [-.84, 15.96]	5.11	1.48	.139
input vs. test × pre- vs. delayed-	3.46 [-4.96, 11.89]	5.12	.68	.499
input vs. parsing × match vs. mismatch	8.18 [-.02, 16.38]	4.98	1.64	.101
input vs. test × match vs. mismatch	2.06 [-6.32, 1.44]	5.09	.40	.686
pre- vs. post- × match vs. mismatch	4.88 [-3.77, 13.52]	5.26	.93	.354
pre- vs. delayed- × match vs. mismatch	6.73 [-1.81, 15.26]	5.19	1.30	.195
input vs. parsing × pre- vs. post- × match vs. mismatch	-8.13 [-19.96, 3.69]	7.19	-1.13	.258
input vs. test × pre- vs. post- × match vs. mismatch	-.49 [-12.57, 11.59]	7.34	-.07	.947
input vs. parsing × pre- vs. delayed- × match vs. mismatch	-1.91 [-22.83, 1.02]	7.25	-1.51	.133
input vs. test × pre- vs. delayed- × match vs. mismatch	-2.45 [-14.45, 9.54]	7.29	-.34	.737

Note: parsing = parsing group; input = input flood group; test = test-only group

ORC-I

First critical word

Table Appx. 144 Fixed effects of model analysis in the SPR test for the ORC-I structure at the first critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	.00 [-4.36, 4.35]	2.65	.00	.999
input vs. parsing	1.23 [-4.57, 7.03]	3.53	.35	.727
input vs. test	-1.26 [-7.2, 4.69]	3.61	-.35	.728
pre- vs. post-	-1.82 [-7.74, 4.11]	3.60	-.50	.614
pre- vs. delayed-	-1.57 [-7.49, 4.35]	3.60	-.44	.662
match vs. mismatch	-4.24 [-1.28, 1.8]	3.67	-1.16	.248
input vs. parsing × pre- vs. post-	1.29 [-6.85, 9.43]	4.95	.26	.795
input vs. test × pre- vs. post-	.53 [-7.79, 8.85]	5.06	.10	.917
input vs. parsing × pre- vs. delayed-	-3.07 [-11.32, 5.18]	5.02	-.61	.541
input vs. test × pre- vs. delayed-	.5 [-7.79, 8.78]	5.04	.10	.922
input vs. parsing × match vs. mismatch	.16 [-7.87, 8.2]	4.89	.03	.973
input vs. test × match vs. mismatch	-1.58 [-9.75, 6.58]	4.96	-.32	.750
pre- vs. post- × match vs. mismatch	3.42 [-5, 11.84]	5.12	.67	.504
pre- vs. delayed- × match vs. mismatch	**9.38 [1.03, 17.73]**	**5.08**	**1.85**	**.065**
input vs. parsing × pre- vs. post- × match vs. mismatch	-2.01 [-13.55, 9.53]	7.02	-.29	.775
input vs. test × pre- vs. post- × match vs. mismatch	1.82 [-9.93, 13.57]	7.15	.26	.799
input vs. parsing × pre- vs. delayed- × match vs. mismatch	-1.44 [-13.06, 1.19]	7.07	-.20	.839
input vs. test × pre- vs. delayed- × match vs. mismatch	-1.56 [-13.29, 1.16]	7.13	-.22	.827

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect

Second critical word

Table Appx. 145 Fixed effects of model analysis in the SPR test for the ORC-I structure at the second critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	2.39 [-2.57, 7.35]	3.02	.79	.428
input vs. parsing	1.90 [-4.77, 8.57]	4.05	.47	.639
input vs. test	3.46 [-3.42, 1.34]	4.18	.83	.409
pre- vs. post-	-3.83 [-1.83, 3.17]	4.26	-.90	.369
pre- vs. delayed-	-2.65 [-9.58, 4.29]	4.21	-.63	.530
match vs. mismatch	-7.12 [-14.18, -.06]	4.29	-1.66	.098
input vs. parsing × pre- vs. post-	4.03 [-5.56, 13.62]	5.83	.69	.490
input vs. test × pre- vs. post-	-1.81 [-11.62, 8]	5.96	-.30	.761
input vs. parsing × pre- vs. delayed-	-6.37 [-16.1, 3.36]	5.92	-1.08	.282
input vs. test × pre- vs. delayed-	-.73 [-1.5, 9.03]	5.94	-.12	.902
input vs. parsing × match vs. mismatch	1.09 [-8.35, 1.52]	5.74	.19	.850
input vs. test × match vs. mismatch	-1.05 [-1.72, 8.61]	5.87	-.18	.858
pre- vs. post- × match vs. mismatch	8.10 [-1.81, 18.01]	6.02	1.35	.179
pre- vs. delayed- × match vs. mismatch	13.83 [4.01, 23.66]	5.97	2.32	.021*
input vs. parsing × pre- vs. post- × match vs. mismatch	-7.46 [-21.02, 6.11]	8.25	-.90	.366
input vs. test × pre- vs. post- × match vs. mismatch	-.62 [-14.45, 13.22]	8.41	-.07	.942
input vs. parsing × pre- vs. delayed- × match vs. mismatch	-.53 [-14.2, 13.15]	8.31	-.06	.950
input vs. test × pre- vs. delayed- × match vs. mismatch	-5.44 [-19.26, 8.39]	8.40	-.65	.518

Note: parsing = parsing group; input = input flood group; test = test-only group; *

significantly differently from zero when $\alpha \leq .05$

Third critical word

Table Appx. 146 Fixed effects of model analysis in the SPR test for the ORC-I structure at the third critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	4.01 [-2.26, 1.27]	3.81	1.05	.293
input vs. parsing	-4.72 [-12.52, 3.08]	4.74	-1.00	.319
input vs. test	-3.93 [-11.91, 4.06]	4.85	-.81	.419
pre- vs. post-	-7.87 [-16.13, .4]	5.03	-1.57	.118
pre- vs. delayed-	-3.9 [-12.1, 4.3]	4.99	-.78	.434
match vs. mismatch	-4.72 [-13.66, 4.21]	5.43	-.87	.385
input vs. parsing × pre- vs. post-	1.86 [-.38, 22.1]	6.83	1.59	.112
input vs. test × pre- vs. post-	**11.69 [.2, 23.18]**	**6.99**	**1.67**	**.095**
input vs. parsing × pre- vs. delayed-	3.6 [-7.75, 14.95]	6.90	.52	.602
input vs. test × pre- vs. delayed-	1.93 [-9.49, 13.34]	6.94	.28	.782
input vs. parsing × match vs. mismatch	8.49 [-2.67, 19.65]	6.79	1.25	.211
input vs. test × match vs. mismatch	1.57 [-.83, 21.96]	6.93	1.53	.127
pre- vs. post- × match vs. mismatch	8.96 [-2.73, 2.65]	7.11	1.26	.208
pre- vs. delayed- × match vs. mismatch	11.22 [-.36, 22.8]	7.04	1.59	.111
input vs. parsing × pre- vs. post- × match vs. mismatch	-12.56 [-28.52, 3.4]	9.70	-1.29	.196
input vs. test × pre- vs. post- × match vs. mismatch	**-19.2 [-35.49, -2.91]**	**9.91**	**-1.94**	**.053**
input vs. parsing × pre- vs. delayed- × match vs. mismatch	-12.07 [-28.13, 4]	9.77	-1.24	.217
input vs. test × pre- vs. delayed- × match vs. mismatch	-13.12 [-29.31, 3.08]	9.85	-1.33	.183

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect

Whole sentence

Table Appx. 147 Fixed effects of model analysis in the SPR test for the ORC-I structure for the whole sentence (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	**4.71 [.04, 9.38]**	**2.84**	**1.66**	**.098**
input vs. parsing	-1.95 [-7.7, 3.79]	3.49	-.56	.576
input vs. test	-3.4 [-9.3, 2.49]	3.58	-.95	.342
pre- vs. post-	-5.26 [-11.2, .68]	3.61	-1.46	.145
pre- vs. delayed-	-2.56 [-8.37, 3.25]	3.53	-.73	.469
match vs. mismatch	-.14 [-6.64, 6.37]	3.95	-.04	.972
input vs. parsing × pre- vs. post-	3.5 [-4.6, 11.59]	4.92	.71	.478
input vs. test × pre- vs. post-	6.85 [-1.43, 15.14]	5.04	1.36	.174
input vs. parsing × pre- vs. delayed-	-.04 [-8.11, 8.04]	4.91	-.01	.994
input vs. test × pre- vs. delayed-	1.67 [-6.48, 9.82]	4.96	.34	.736
input vs. parsing × match vs. mismatch	1.23 [-6.71, 9.17]	4.83	.26	.799
input vs. test × match vs. mismatch	2.68 [-5.45, 1.81]	4.94	.54	.588
pre- vs. post- × match vs. mismatch	6.45 [-1.92, 14.83]	5.09	1.27	.205
pre- vs. delayed- × match vs. mismatch	3.09 [-5.16, 11.35]	5.02	.62	.538
input vs. parsing × pre- vs. post- × match vs. mismatch	-1.89 [-13.31, 9.53]	6.94	-.27	.785
input vs. test × pre- vs. post- × match vs. mismatch	-8.86 [-2.55, 2.82]	7.11	-1.25	.212
input vs. parsing × pre- vs. delayed- × match vs. mismatch	2.49 [-8.96, 13.95]	6.97	.36	.720
input vs. test × pre- vs. delayed- × match vs. mismatch	-1.28 [-12.89, 1.32]	7.06	-.18	.856

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect

Appendix 29 RQ2: Fixed effects of model analysis in the eye-tracking test (baseline of the test-only group)

SRC-A: *First critical word*

Table Appx. 148 Fixed effects of model analysis in the eye-tracking test for the SRC-A structure for the first critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.78 [1.72, 1.84]	.04	50.52	<.001***
linear	-.11 [-.29, .07]	.11	-1.01	.311
quadratic	.11 [-.08, .29]	.11	.98	.330
cubic	.02 [-.17, .20]	.11	.14	.887
pre- vs. post-	-.01 [-.09, .06]	.05	-.32	.750
pre- vs. delayed-	.00 [-.07, .08]	.05	.09	.926
test vs. parsing	-.01 [-.09, .07]	.05	-.26	.796
test vs. input	-.06 [-.14, .02]	.05	-1.20	.233
linear × pre- vs. post-	.00 [-.25, .26]	.16	.03	.977
linear × pre- vs. delayed-	.15 [-.12, .42]	.16	.92	.356
quadratic × pre- vs. post-	-.02 [-.28, .24]	.16	-.15	.884
quadratic × pre- vs. delayed-	-.14 [-.41, .13]	.16	-.87	.384
cubic × pre- vs. post-	-.09 [-.35, .17]	.16	-.59	.556
cubic × pre- vs. delayed-	.06 [-.21, .33]	.16	.37	.714
linear × test vs. parsing	.05 [-.21, .31]	.16	.32	.747
linear × test vs. input	.00 [-.27, .26]	.16	-.02	.985
quadratic × test vs. parsing	-.11 [-.37, .15]	.16	-.70	.485
quadratic × test vs. input	.00 [-.27, .27]	.16	-.02	.987
cubic × test vs. parsing	-.19 [-.44, .07]	.16	-1.20	.230
cubic × test vs. input	-.18 [-.45, .09]	.16	-1.12	.263
pre- vs. post- × test vs. parsing	.05 [-.06, .16]	.06	.76	.449
pre- vs. delayed- × test vs. parsing	-.01 [-.12, .10]	.07	-.13	.894
pre- vs. post- × test vs. input	.03 [-.08, .14]	.07	.40	.691
pre- vs. delayed- × test vs. input	.03 [-.08, .14]	.07	.45	.653
linear × pre- vs. post- × test vs. parsing	.08 [-.28, .45]	.22	.36	.716
linear × pre- vs. delayed- × test vs. parsing	-.16 [-.53, .21]	.22	-.70	.484
linear × pre- vs. post- × test vs. input	-.06 [-.44, .32]	.23	-.27	.784
linear × pre- vs. delayed- × test vs. input	-.02 [-.41, .36]	.23	-.09	.931
quadratic × pre- vs. post- × test vs. parsing	.21 [-.16, .57]	.22	.94	.347
quadratic × pre- vs. delayed- × test vs. parsing	.04 [-.33, .41]	.22	.18	.856
quadratic × pre- vs. post- × test vs. input	-.01 [-.39, .37]	.23	-.06	.954
quadratic × pre- vs. delayed- × test vs. input	.03 [-.36, .42]	.24	.12	.904
cubic × pre- vs. post- × test vs. parsing	.24 [-.13, .60]	.22	1.07	.287
cubic × pre- vs. delayed- × test vs. parsing	.24 [-.13, .61]	.23	1.06	.287
cubic × pre- vs. post- × test vs. input	.31 [-.07, .69]	.23	1.33	.183
cubic × pre- vs. delayed- × test vs. input	.23 [-.15, .62]	.23	1.00	.318

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

Second critical word

Table Appx. 149 Fixed effects of model analysis in the eye-tracking test for the SRC-A structure for the second critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.77 [1.66, 1.87]	.06	27.83	< .001***
linear	-.15 [-.33, .03]	.11	-1.33	.184
pre- vs. post-	-.06 [-.21, .08]	.09	-.71	.476
pre- vs. delayed-	-.12 [-.27, .03]	.09	-1.28	.200
test vs. parsing	-.06 [-.20, .09]	.09	-.66	.510
test vs. input	**-.16 [-.31, -.02]**	**.09**	**-1.84**	**.066**
linear × pre- vs. post-	.11 [-.14, .37]	.15	.75	.456
linear × pre- vs. delayed-	.13 [-.14, .39]	.16	.80	.427
linear × test vs. parsing	.22 [-.03, .47]	.15	1.43	.152
linear × test vs. input	.23 [-.02, .49]	.16	1.51	.130
pre- vs. post- × test vs. parsing	.18 [-.03, .38]	.13	1.41	.160
pre- vs. delayed- × test vs. parsing	.12 [-.09, .33]	.13	.93	.355
pre- vs. post- × test vs. input	.11 [-.10, .33]	.13	.89	.374
pre- vs. delayed- × test vs. input	**.35 [.13, .57]**	**.13**	**2.66**	**.008****
linear × pre- vs. post- × test vs. parsing	-.07 [-.43, .28]	.22	-.34	.736
linear × pre- vs. delayed- × test vs. parsing	-.13 [-.49, .23]	.22	-.59	.557
linear × pre- vs. post- × test vs. input	-.25 [-.61, .12]	.22	-1.10	.271
linear × pre- vs. delayed- × test vs. input	.02 [-.36, .39]	.23	.09	.933

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; ** significantly different from zero when α < .01; *** significantly different from zero when α<.001

Third critical word

Table Appx. 150 Fixed effects of model analysis in the eye-tracking test for the SRC-A structure for the third critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.75 [1.69, 1.81]	.04	49.11	< .001***
linear	-.01 [-.18, .16]	.10	-.11	.916
quadratic	-.04 [-.21, .12]	.10	-.44	.659
cubic	.01 [-.16, .18]	.10	.09	.928
pre- vs. post-	-.02 [-.1, .05]	.05	-.52	.600
pre- vs. delayed-	.03 [-.05, .11]	.05	.65	.518
test vs. parsing	.02 [-.07, .10]	.05	.32	.751
test vs. input	-.02 [-.10, .07]	.05	-.35	.729
linear × pre- vs. post-	.09 [-.15, .33]	.14	.63	.529
linear × pre- vs. delayed-	.16 [-.08, .41]	.15	1.10	.273
quadratic × pre- vs. post-	.05 [-.19, .28]	.14	.32	.750
quadratic × pre- vs. delayed-	-.01 [-.26, .24]	.15	-.07	.942
cubic × pre- vs. post-	-.01 [-.25, .22]	.14	-.08	.940
cubic × pre- vs. delayed-	.17 [-.07, .42]	.15	1.16	.246
linear × test vs. parsing	.00 [-.24, .24]	.14	-.01	.990
linear × test vs. input	-.14 [-.37, .10]	.15	-.93	.352
quadratic × test vs. parsing	.11 [-.13, .34]	.14	.74	.457
quadratic × test vs. input	-.11 [-.34, .13]	.15	-.73	.468
cubic × test vs. parsing	.16 [-.08, .39]	.14	1.09	.277
cubic × test vs. input	.10 [-.14, .34]	.15	.71	.478
pre- vs. post- × test vs. parsing	.03 [-.08, .13]	.07	.39	.697
pre- vs. delayed- × test vs. parsing	-.07 [-.18, .04]	.07	-1.09	.274
pre- vs. post- × test vs. input	-.01 [-.12, .10]	.07	-.12	.904
pre- vs. delayed- × test vs. input	-.04 [-.15, .07]	.07	-.61	.545
linear × pre- vs. post- × test vs. parsing	-.09 [-.43, .24]	.21	-.45	.650
linear × pre- vs. delayed- × test vs. parsing	-.09 [-.44, .25]	.21	-.45	.653
linear × pre- vs. post- × test vs. input	.16 [-.19, .50]	.21	.75	.456
linear × pre- vs. delayed- × test vs. input	.09 [-.26, .45]	.22	.43	.668
quadratic × pre- vs. post- × test vs. parsing	-.12 [-.45, .22]	.21	-.57	.572
quadratic × pre- vs. delayed- × test vs. parsing	-.19 [-.53, .15]	.21	-.93	.352
quadratic × pre- vs. post- × test vs. input	.22 [-.12, .56]	.21	1.04	.297
quadratic × pre- vs. delayed- × test vs. input	.15 [-.20, .51]	.22	.71	.476
cubic × pre- vs. post- × test vs. parsing	-.26 [-.60, .08]	.21	-1.25	.212
cubic × pre- vs. delayed- × test vs. parsing	-.12 [-.46, .22]	.21	-.57	.571
cubic × pre- vs. post- × test vs. input	.00 [-.34, .34]	.21	.02	.985
cubic × pre- vs. delayed- × test vs. input	-.32 [-.67, .04]	.21	-1.47	.141

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

SRC-I

First critical word

Table Appx. 151 Fixed effects of model analysis in the eye-tracking test for the SRC-I structure for the first critical word (baseline of the test-only group)

fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.8 [1.74, 1.86]	.03	51.94	<.001***
linear	-.02 [-.22, .17]	.12	-.18	.860
quadratic	-.14 [-.33, .05]	.12	-1.18	.236
cubic	-.1 [-.29, .09]	.12	-.85	.394
pre- vs. post-	-.04 [-.12, .04]	.05	-.84	.401
pre- vs. delayed	-.09 [-.17, -.01]	.05	-1.80	.072
test vs. parsing	-.01 [-.09, .07]	.05	-.14	.891
test vs. input	-.10 [-.18, -.02]	.05	-2.00	.046*
linear × pre- vs. post-	.14 [-.12, .41]	.16	.88	.381
linear × pre- vs. delayed	-.20 [-.48, .07]	.17	-1.20	.229
quadratic × pre- vs. post-	.14 [-.12, .41]	.16	.89	.375
quadratic × pre- vs. delayed	.00 [-.27, .27]	.17	.01	.995
cubic × pre- vs. post-	-.07 [-.34, .2]	.16	-.44	.663
cubic × pre- vs. delayed	.08 [-.19, .36]	.17	.50	.617
linear × test vs. parsing	.12 [-.15, .38]	.16	.71	.476
linear × test vs. input	.06 [-.22, .33]	.17	.33	.740
quadratic × test vs. parsing	.21 [-.06, .47]	.16	1.28	.201
quadratic × test vs. input	.19 [-.08, .46]	.17	1.14	.254
cubic × test vs. parsing	.22 [-.05, .49]	.16	1.35	.176
cubic × test vs. input	.16 [-.11, .44]	.17	.97	.331
pre- vs. post- × test vs. parsing	.08 [-.03, .19]	.07	1.17	.244
pre- vs. delayed × test vs. parsing	.08 [-.03, .19]	.07	1.16	.248
pre- vs. post- × test vs. input	**.13 [.02, .24]**	**.07**	**1.93**	**.053**
pre- vs. delayed × test vs. input	**.12 [.01, .23]**	**.07**	**1.73**	**.084**
linear × pre- vs. post- × test vs. parsing	**-.43 [-.81, -.04]**	**.23**	**-1.84**	**.067**
linear × pre- vs. delayed × test vs. parsing	-.05 [-.43, .33]	.23	-.22	.826
linear × pre- vs. post- × test vs. input	-.15 [-.54, .23]	.23	-.65	.517
linear × pre- vs. delayed × test vs. input	.34 [-.05, .73]	.24	1.44	.149
quadratic × pre- vs. post- × test vs. parsing	-.18 [-.57, .2]	.23	-.78	.433
quadratic × pre- vs. delayed × test vs. parsing	-.18 [-.56, .2]	.23	-.78	.437
quadratic × pre- vs. post- × test vs. input	-.13 [-.52, .25]	.24	-.57	.571
quadratic × pre- vs. delayed × test vs. input	.13 [-.26, .51]	.24	.53	.595
cubic × pre- vs. post- × test vs. parsing	-.08 [-.46, .31]	.23	-.33	.741
cubic × pre- vs. delayed × test vs. parsing	-.27 [-.65, .11]	.23	-1.18	.240
cubic × pre- vs. post- × test vs. input	.05 [-.34, .43]	.23	.19	.847
cubic × pre- vs. delayed × test vs. input	-.08 [-.47, .3]	.24	-.36	.721

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; *** significantly different from zero when $\alpha < .001$

Second critical word

Table Appx. 152 Fixed effects of model analysis in the eye-tracking test for the SRC-I structure for the second critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*-value
intercept	1.72 [1.6, 1.83]	.07	23.98	< .001***
linear	**.23 [.03, .42]**	**.12**	**1.92**	**.056**
pre- vs. post-	.01 [-.14, .17]	.09	.15	.881
pre- vs. delayed-	.03 [-.14, .19]	.10	.26	.794
test vs. parsing	-.04 [-.20, .13]	.10	-.38	.708
test vs. input	-.02 [-.19, .15]	.10	-.19	.847
linear × pre- vs. post-	**-.37 [-.64, -.10]**	**.16**	**-2.24**	**.025***
linear × pre- vs. delayed-	-.25 [-.52, .03]	.17	-1.44	.149
linear × test vs. parsing	-.24 [-.51, .03]	.16	-1.47	.141
linear × test vs. input	-.12 [-.4, .16]	.17	-.68	.496
pre- vs. post- × test vs. parsing	.01 [-.21, .22]	.13	.05	.961
pre- vs. delayed- × test vs. parsing	.03 [-.19, .25]	.13	.23	.820
pre- vs. post- × test vs. input	.05 [-.17, .28]	.14	.40	.692
pre- vs. delayed- × test vs. input	-.01 [-.24, .22]	.14	-.06	.955
linear × pre- vs. post- × test vs. parsing	.33 [-.04, .70]	.23	1.48	.140
linear × pre- vs. delayed- × test vs. parsing	.00 [-.38, .38]	.23	.00	.998
linear × pre- vs. post- × test vs. input	.05 [-.34, .43]	.23	.21	.838
linear × pre- vs. delayed- × test vs. input	.26 [-.14, .65]	.24	1.07	.284

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; *** significantly different from zero when $\alpha<.001$

Third critical word

Table Appx. 153 Fixed effects of model analysis in the eye-tracking test for the SRC-I structure for the third critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.76 [1.7, 1.82]	.04	47.25	< .001***
linear	-.07 [-.26, .11]	.11	-.67	.501
pre- vs. post-	.02 [-.06, .10]	.05	.45	.654
pre- vs. delayed-	.04 [-.04, .12]	.05	.90	.367
test vs. parsing	.03 [-.05, .12]	.05	.61	.543
test vs. input	-.02 [-.11, .06]	.05	-.44	.662
linear × pre- vs. post-	.10 [-.15, .36]	.15	.68	.498
linear × pre- vs. delayed-	.22 [-.03, .47]	.15	1.43	.153
linear × test vs. parsing	.07 [-.18, .32]	.15	.44	.664
linear × test vs. input	.13 [-.13, .39]	.16	.84	.402
pre- vs. post- × test vs. parsing	-.01 [-.12, .11]	.07	-.09	.931
pre- vs. delayed- × test vs. parsing	.02 [-.09, .12]	.07	.23	.820
pre- vs. post- × test vs. input	-.01 [-.12, .11]	.07	-.12	.902
pre- vs. delayed- × test vs. input	.03 [-.09, .14]	.07	.39	.694
linear × pre- vs. post- × test vs. parsing	-.12 [-.47, .24]	.21	-.54	.587
linear × pre- vs. delayed- × test vs. parsing	-.15 [-.49, .20]	.21	-.71	.476
linear × pre- vs. post- × test vs. input	-.28 [-.64, .09]	.22	-1.25	.211
linear × pre- vs. delayed- × test vs. input	-.34 [-.71, .02]	.22	-1.56	.118

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

ORC-A

First critical word

Table Appx. 154 Fixed effects of model analysis in the eye-tracking test for the ORC-A structure for the first critical word (baseline of the test-only group)

fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.82 [1.71, 1.94]	.07	26.74	<.001***
linear	-.08 [-.27, .11]	.12	-.69	.492
quadratic	-.13 [-.32, .07]	.12	-1.08	.280
pre- vs. post-	-.04 [-.2, .12]	.10	-.44	.663
pre- vs. delayed-	-.04 [-.2, .12]	.10	-.37	.709
test vs. parsing	.02 [-.13, .18]	.09	.25	.800
test vs. input	-.06 [-.22, .1]	.10	-.61	.541
linear × pre- vs. post-	-.07 [-.35, .21]	.17	-.43	.670
linear × pre- vs. delayed-	.23 [-.05, .50]	.17	1.37	.171
quadratic × pre- vs. post-	.02 [-.26, .30]	.17	.13	.898
quadratic × pre- vs. delayed-	.16 [-.12, .44]	.17	.96	.339
linear × test vs. parsing	.09 [-.17, .36]	.16	.57	.566
linear × test vs. input	.14 [-.13, .41]	.17	.86	.392
quadratic × test vs. parsing	.03 [-.24, .3]	.16	.18	.856
quadratic × test vs. input	.08 [-.19, .36]	.17	.50	.616
pre- vs. post- × test vs. parsing	.09 [-.14, .32]	.14	.63	.527
pre- vs. delayed- × test vs. parsing	.01 [-.21, .23]	.13	.05	.961
pre- vs. post- × test vs. input	.16 [-.07, .39]	.14	1.14	.255
pre- vs. delayed- × test vs. input	.02 [-.21, .24]	.14	.12	.902
linear × pre- vs. post- × test vs. parsing	.30 [-.10, .70]	.24	1.24	.217
linear × pre- vs. delayed- × test vs. parsing	-.22 [-.60, .16]	.23	-.96	.339
linear × pre- vs. post- × test vs. input	.03 [-.36, .43]	.24	.14	.888
linear × pre- vs. delayed- × test vs. input	-.30 [-.68, .09]	.23	-1.27	.204
quadratic × pre- vs. post- × test vs. parsing	.04 [-.36, .44]	.24	.17	.868
quadratic × pre- vs. delayed- × test vs. parsing	-.17 [-.55, .21]	.23	-.73	.469
quadratic × pre- vs. post- × test vs. input	-.07 [-.46, .33]	.24	-.27	.785
quadratic × pre- vs. delayed- × test vs. input	.00 [-.38, .39]	.23	.00	.999

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

Second critical word

Table Appx. 155 Fixed effects of model analysis in the eye-tracking test for the ORC-A structure for the second critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.69 [1.62, 1.76]	.04	41.51	< .001***
linear	-.03 [-.23, .16]	.12	-.30	.766
quadratic	.05 [-.14, .24]	.12	.45	.657
cubic	-.12 [-.31, .07]	.12	-1.03	.303
pre- vs. post-	.00 [-.09, .09]	.06	-.04	.971
pre- vs. delayed-	.03 [-.07, .12]	.06	.48	.629
test vs. parsing	.08 [-.02, .17]	.06	1.33	.183
test vs. input	.07 [-.03, .16]	.06	1.13	.258
linear × pre- vs. post-	-.03 [-.30, .24]	.17	-.17	.862
linear × pre- vs. delayed-	.13 [-.15, .41]	.17	.75	.453
quadratic × pre- vs. post-	-.13 [-.40, .15]	.17	-.76	.448
quadratic × pre- vs. delayed-	.10 [-.18, .38]	.17	.59	.556
cubic × pre- vs. post-	.04 [-.24, .31]	.17	.22	.828
cubic × pre- vs. delayed-	.26 [-.02, .54]	.17	1.53	.126
linear × test vs. parsing	-.04 [-.31, .24]	.17	-.22	.829
linear × test vs. input	.16 [-.12, .43]	.17	.95	.345
quadratic × test vs. parsing	-.13 [-.4, .14]	.17	-.80	.423
quadratic × test vs. input	-.07 [-.34, .2]	.17	-.44	.663
cubic × test vs. parsing	.18 [-.09, .45]	.16	1.10	.271
cubic × test vs. input	.10 [-.17, .38]	.17	.62	.538
pre- vs. post- × test vs. parsing	-.05 [-.18, .09]	.08	-.58	.559
pre- vs. delayed- × test vs. parsing	-.04 [-.17, .09]	.08	-.48	.633
pre- vs. post- × test vs. input	.03 [-.10, .16]	.08	.36	.722
pre- vs. delayed- × test vs. input	-.05 [-.18, .08]	.08	-.60	.551
linear × pre- vs. post- × test vs. parsing	.17 [-.23, .56]	.24	.69	.491
linear × pre- vs. delayed- × test vs. parsing	-.08 [-.47, .30]	.23	-.36	.721
linear × pre- vs. post- × test vs. input	-.16 [-.55, .23]	.24	-.67	.501
linear × pre- vs. delayed- × test vs. input	-.39 [-.79, .00]	.24	-1.64	.101
quadratic × pre- vs. post- × test vs. parsing	.04 [-.35, .44]	.24	.18	.855
quadratic × pre- vs. delayed- × test vs. parsing	-.05 [-.44, .34]	.24	-.21	.832
quadratic × pre- vs. post- × test vs. input	.38 [-.01, .78]	.24	1.60	.109
quadratic × pre- vs. delayed- × test vs. input	-.14 [-.53, .25]	.24	-.60	.551
cubic × pre- vs. post- × test vs. parsing	-.02 [-.41, .38]	.24	-.08	.939
cubic × pre- vs. delayed- × test vs. parsing	-.31 [-.70, .08]	.23	-1.32	.186
cubic × pre- vs. post- × test vs. input	.08 [-.31, .48]	.24	.34	.733
cubic × pre- vs. delayed- × test vs. input	-.22 [-.61, .17]	.24	-.93	.354

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

Third critical word

Table Appx. 156 Fixed effects of model analysis in the eye-tracking test for the ORC-A structure for the third critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.76 [1.7, 1.81]	.03	55.23	< .001***
linear	.23 [.06, .41]	.11	2.19	.028*
quadratic	-.08 [-.25, .1]	.11	-.73	.468
cubic	-.11 [-.29, .06]	.11	-1.07	.285
pre- vs. post-	-.01 [-.08, .06]	.04	-.26	.796
pre- vs. delayed-	.04 [-.03, .11]	.04	.99	.323
test vs. parsing	-.03 [-.10, .05]	.04	-.59	.559
test vs. input	-.01 [-.08, .07]	.04	-.13	.898
linear × pre- vs. post-	-.08 [-.32, .16]	.15	-.57	.568
linear × pre- vs. delayed-	-.17 [-.42, .08]	.15	-1.13	.261
quadratic × pre- vs. post-	.08 [-.16, .31]	.14	.53	.599
quadratic × pre- vs. delayed-	-.01 [-.26, .23]	.15	-.10	.922
cubic × pre- vs. post-	.18 [-.06, .42]	.14	1.23	.218
cubic × pre- vs. delayed-	-.05 [-.30, .20]	.15	-.34	.734
linear × test vs. parsing	-.25 [-.49, .00]	.15	-1.63	.103
linear × test vs. input	-.31 [-.56, -.07]	.15	-2.08	.038*
quadratic × test vs. parsing	.08 [-.17, .33]	.15	.53	.597
quadratic × test vs. input	.07 [-.18, .32]	.15	.46	.648
cubic × test vs. parsing	.01 [-.23, .26]	.15	.08	.939
cubic × test vs. input	.12 [-.13, .36]	.15	.77	.442
pre- vs. post- × test vs. parsing	.06 [-.04, .15]	.06	1.00	.316
pre- vs. delayed- × test vs. parsing	-.01 [-.11, .08]	.06	-.24	.808
pre- vs. post- × test vs. input	.05 [-.04, .15]	.06	.91	.363
pre- vs. delayed- × test vs. input	-.08 [-.18, .02]	.06	-1.37	.170
linear × pre- vs. post- × test vs. parsing	.12 [-.22, .47]	.21	.58	.560
linear × pre- vs. delayed- × test vs. parsing	.18 [-.17, .52]	.21	.85	.395
linear × pre- vs. post- × test vs. input	.17 [-.18, .52]	.21	.82	.414
linear × pre- vs. delayed- × test vs. input	.13 [-.23, .48]	.22	.58	.561
quadratic × pre- vs. post- × test vs. parsing	-.02 [-.36, .33]	.21	-.07	.942
quadratic × pre- vs. delayed- × test vs. parsing	-.05 [-.39, .29]	.21	-.24	.808
quadratic × pre- vs. post- × test vs. input	-.03 [-.38, .31]	.21	-.16	.873
quadratic × pre- vs. delayed- × test vs. input	.11 [-.24, .46]	.22	.51	.613
cubic × pre- vs. post- × test vs. parsing	.03 [-.32, .37]	.21	.13	.896
cubic × pre- vs. delayed- × test vs. parsing	.05 [-.29, .39]	.21	.24	.813
cubic × pre- vs. post- × test vs. input	-.05 [-.4, .29]	.21	-.25	.800
cubic × pre- vs. delayed- × test vs. input	-.12 [-.47, .23]	.21	-.57	.567

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when $\alpha \leq .05$; *** significantly different from zero when $\alpha < .001$

ORC-I

First critical word

Table Appx. 157 Fixed effects of model analysis in the eye-tracking test for the ORC-I structure for the first critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.75 [1.63, 1.86]	.07	25.60	< .001***
linear	.12 [-.08, .31]	.12	.99	.323
pre- vs. post-	.06 [-.10, .23]	.10	.61	.541
pre- vs. delayed-	.00 [-.17, .16]	.10	-.03	.974
test vs. parsing	.04 [-.13, .2]	.10	.37	.715
test vs. input	.01 [-.16, .17]	.10	.09	.931
linear × pre- vs. post-	-.14 [-.42, .14]	.17	-.82	.415
linear × pre- vs. delayed-	-.06 [-.34, .23]	.18	-.32	.750
linear × test vs. parsing	**-.31 [-.59, -.03]**	**.17**	**-1.81**	**.070**
linear × test vs. input	.26 [-.02, .54]	.17	1.50	.133
pre- vs. post- × test vs. parsing	-.1 [-.33, .13]	.14	-.73	.464
pre- vs. delayed- × test vs. parsing	-.03 [-.26, .20]	.14	-.21	.837
pre- vs. post- × test vs. input	-.03 [-.27, .21]	.14	-.21	.835
pre- vs. delayed- × test vs. input	-.03 [-.27, .21]	.15	-.20	.845
linear × pre- vs. post- × test vs. parsing	.14 [-.26, .53]	.24	.56	.573
linear × pre- vs. delayed- × test vs. parsing	.36 [-.05, .76]	.25	1.45	.149
linear × pre- vs. post- × test vs. input	-.40 [-.81, .01]	.25	-1.61	.108
linear × pre- vs. delayed- × test vs. input	-.14 [-.56, .28]	.26	-.56	.574

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; *** significantly different from zero when α<.001

Second critical word

Table Appx. 158 Fixed effects of model analysis in the eye-tracking test for the ORC-I structure for the second critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*-value	*p*-value
intercept	1.74 [1.67, 1.81]	.04	4.89	< .001***
linear	.07 [-.12, .26]	.12	.62	.535
pre- vs. post-	.02 [-.07, .12]	.06	.38	.704
pre- vs. delayed-	.05 [-.04, .14]	.06	.89	.373
test vs. parsing	.06 [-.04, .16]	.06	.95	.343
test vs. input	.03 [-.07, .13]	.06	.45	.651
linear × pre- vs. post-	-.05 [-.32, .22]	.17	-.30	.764
linear × pre- vs. delayed-	.03 [-.24, .31]	.17	.19	.851
linear × test vs. parsing	-.02 [-.29, .25]	.16	-.11	.915
linear × test vs. input	.01 [-.26, .28]	.17	.07	.947
pre- vs. post- × test vs. parsing	-.05 [-.19, .08]	.08	-.69	.493
pre- vs. delayed- × test vs. parsing	-.07 [-.2, .06]	.08	-.84	.404
pre- vs. post- × test vs. input	-.08 [-.22, .05]	.08	-1.00	.316
pre- vs. delayed- × test vs. input	-.04 [-.18, .09]	.08	-.53	.599
linear × pre- vs. post- × test vs. parsing	.2 [-.18, .58]	.23	.85	.393
linear × pre- vs. delayed- × test vs. parsing	-.14 [-.52, .25]	.23	-.60	.552
linear × pre- vs. post- × test vs. input	-.11 [-.5, .29]	.24	-.44	.660
linear × pre- vs. delayed- × test vs. input	-.14 [-.54, .26]	.24	-.59	.558

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

Third critical word

Table Appx. 159 Fixed effects of model analysis in the eye-tracking test for the ORC-I structure for the third critical word (baseline of the test-only group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.76 [1.7, 1.81]	.03	51.54	<.001***
linear	.06 [-.12, .24]	.11	.57	.572
quadratic	.12 [-.06, .31]	.11	1.10	.271
pre- vs. post-	.00 [-.07, .07]	.04	-.04	.971
pre- vs. delayed-	.06 [-.01, .13]	.04	1.35	.177
test vs. parsing	.04 [-.03, .12]	.05	.93	.355
test vs. input	.02 [-.06, .10]	.05	.34	.732
linear × pre- vs. post-	.03 [-.22, .28]	.15	.18	.855
linear × pre- vs. delayed-	.02 [-.24, .28]	.16	.15	.881
quadratic × pre- vs. post-	-.18 [-.43, .07]	.15	-1.19	.235
quadratic × pre- vs. delayed-	-.12 [-.38, .14]	.16	-.78	.435
linear × test vs. parsing	.03 [-.23, .29]	.16	.19	.849
linear × test vs. input	-.36 [-.63, -.09]	.16	-2.23	.026*
quadratic × test vs. parsing	.03 [-.23, .29]	.16	.18	.855
quadratic × test vs. input	-.18 [-.45, .08]	.16	-1.13	.259
pre- vs. post- × test vs. parsing	-.03 [-.13, .07]	.06	-.52	.606
pre- vs. delayed- × test vs. parsing	-.06 [-.16, .04]	.06	-1.00	.315
pre- vs. post- × test vs. input	-.05 [-.16, .05]	.06	-.88	.379
pre- vs. delayed- × test vs. input	-.07 [-.18, .03]	.06	-1.15	.249
linear × pre- vs. post- × test vs. parsing	-.18 [-.54, .18]	.22	-.83	.407
linear × pre- vs. delayed- × test vs. parsing	-.01 [-.37, .35]	.22	-.04	.965
linear × pre- vs. post- × test vs. input	.29 [-.08, .66]	.23	1.30	.192
linear × pre- vs. delayed- × test vs. input	.22 [-.16, .59]	.23	.95	.343
quadratic × pre- vs. post- × test vs. parsing	.16 [-.2, .52]	.22	.75	.452
quadratic × pre- vs. delayed- × test vs. parsing	.08 [-.28, .44]	.22	.37	.715
quadratic × pre- vs. post- × test vs. input	.03 [-.34, .39]	.22	.12	.903
quadratic × pre- vs. delayed- × test vs. input	.12 [-.25, .49]	.23	.52	.600

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when $\alpha \leq .05$; *** significantly different from zero when $\alpha < .001$

Appendix 30 RQ2: Fixed effects of model analysis in the eye-tracking test (baseline of the input flood group)

SRC-A: *First critical word*

Table Appx. 160 Fixed effects of model analysis in the eye-tracking test for the SRC-A structure for the first critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.72 [1.66, 1.78]	.04	46.76	<.001***
linear	-.12 [-.31, .08]	.12	-.98	.328
quadratic	.11 [-.09, .30]	.12	.91	.363
cubic	-.17 [-.36, .03]	.12	-1.41	.159
pre- vs. post-	.01 [-.07, .09]	.05	.25	.805
pre- vs. delayed-	.04 [-.05, .12]	.05	.72	.470
input vs. parsing	.05 [-.03, .13]	.05	.96	.336
input vs. test	.06 [-.02, .14]	.05	1.20	.233
linear × pre- vs. post-	-.06 [-.34, .22]	.17	-.35	.730
linear × pre- vs. delayed-	.13 [-.15, .41]	.17	.77	.441
quadratic × pre- vs. post-	-.04 [-.32, .24]	.17	-.21	.831
quadratic × pre- vs. delayed-	-.11 [-.39, .16]	.17	-.68	.498
cubic × pre- vs. post-	.22 [-.06, .5]	.17	1.27	.205
cubic × pre- vs. delayed-	**.29 [.02, .57]**	**.17**	**1.75**	**.080**
linear × input vs. parsing	.05 [-.21, .32]	.16	.33	.741
linear × input vs. test	0.00 [-.26, .27]	.16	.02	.985
quadratic × input vs. parsing	-.11 [-.37, .16]	.16	-.67	.506
quadratic × input vs. test	.00 [-.27, .27]	.16	.02	.987
cubic × input vs. parsing	-.01 [-.27, .26]	.16	-.03	.974
cubic × input vs. test	.18 [-.09, .45]	.16	1.12	.263
pre- vs. post- × input vs. parsing	.02 [-.09, .13]	.07	.33	.742
pre- vs. delayed- × input vs. parsing	-.04 [-.15, .07]	.07	-.60	.552
pre- vs. post- × input vs. test	-.03 [-.14, .08]	.07	-.40	.691
pre- vs. delayed- × input vs. test	-.03 [-.14, .08]	.07	-.45	.653
linear × pre- vs. post- × input vs. parsing	.14 [-.24, .53]	.23	.62	.535
linear × pre- vs. delayed- × input vs. parsing	-.14 [-.51, .24]	.23	-.60	.550
linear × pre- vs. post- × input vs. test	.06 [-.32, .44]	.23	.27	.784
linear × pre- vs. delayed- × input vs. test	.02 [-.36, .41]	.23	.09	.931
quadratic × pre- vs. post- × input vs. parsing	.22 [-.16, .6]	.23	.96	.337
quadratic × pre- vs. delayed- × input vs. parsing	.01 [-.36, .39]	.23	.06	.956
quadratic × pre- vs. post- × input vs. test	.01 [-.37, .39]	.23	.06	.954
quadratic × pre- vs. delayed- × input vs. test	-.03 [-.42, .36]	.24	-.12	.904
cubic × pre- vs. post- × input vs. parsing	-.07 [-.45, .31]	.23	-.31	.758
cubic × pre- vs. delayed- × input vs. parsing	.01 [-.37, .38]	.23	.03	.980
cubic × pre- vs. post- × input vs. test	-.31 [-.69, .07]	.23	-1.33	.183
cubic × pre- vs. delayed- × input vs. test	-.23 [-.62, .15]	.23	-1.00	.318

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; *** significantly different from zero when α<.001

Second critical word

Table Appx. 161 Fixed effects of model analysis in the eye-tracking test for the SRC-A structure for the second critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.6 [1.5, 1.71]	.06	25.38	<.001***
linear	.09 [-.09, .27]	.11	.81	.419
pre- vs. post-	.05 [-.10, .21]	.09	.55	.581
pre- vs. delayed-	.23 [.08, .39]	.09	2.48	.013*
input vs. parsing	.11 [-.04, .25]	.09	1.23	.220
input vs. test	**.16 [.02, .31]**	**.09**	**1.84**	**.066**
linear × pre- vs. post-	-.13 [-.40, .14]	.16	-.81	.418
linear × pre- vs. delayed-	.15 [-.12, .41]	.16	.91	.363
linear × input vs. parsing	-.02 [-.27, .23]	.15	-.11	.909
linear × input vs. test	-.23 [-.49, .02]	.16	-1.51	.130
pre- vs. post- × input vs. parsing	.06 [-.15, .27]	.13	.47	.636
pre- vs. delayed- × input vs. parsing	**-.23 [-.44, -.02]**	**.13**	**-1.83**	**.067**
pre- vs. post- × input vs. test	-.11 [-.33, .10]	.13	-.89	.374
pre- vs. delayed- × input vs. test	-.35 [-.57, -.13]	.13	-2.66	.008**
linear × pre- vs. post- × input vs. parsing	.17 [-.20, .54]	.23	.77	.443
linear × pre- vs. delayed- × input vs. parsing	-.15 [-.51, .21]	.22	-.67	.501
linear × pre- vs. post- × input vs. test	.25 [-.12, .61]	.22	1.10	.271
linear × pre- vs. delayed- × input vs. test	-.02 [-.39, .36]	.23	-.09	.933

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$; *** significantly different from zero when $\alpha<.001$

Third critical word

Table Appx. 162 Fixed effects of model analysis in the eye-tracking test for the SRC-A structure for the third critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.73 [1.67, 1.79]	.04	47.28	<.001***
linear	-.15 [-.32, .03]	.10	-1.40	.162
quadratic	-.15 [-.32, .02]	.10	-1.44	.150
cubic	.11 [-.06, .28]	.10	1.08	.281
pre- vs. post-	-.03 [-.11, .05]	.05	-.67	.506
pre- vs. delayed-	-.01 [-.09, .07]	.05	-.22	.829
input vs. parsing	.03 [-.05, .12]	.05	.66	.511
input vs. test	.02 [-.07, .1]	.05	.35	.730
linear × pre- vs. post-	.25 [.00, .49]	.15	1.62	.105
linear × pre- vs. delayed-	**.26 [.00, .51]**	**.16**	**1.65**	**.099**
quadratic × pre- vs. post-	**.26 [.01, .51]**	**.15**	**1.74**	**.082**
quadratic × pre- vs. delayed-	.14 [-.11, .4]	.15	.92	.356
cubic × pre- vs. post-	-.01 [-.25, .24]	.15	-.05	.963
cubic × pre- vs. delayed-	-.14 [-.40, .11]	.15	-.93	.354
linear × input vs. parsing	.13 [-.11, .37]	.15	.91	.362
linear × input vs. test	.14 [-.10, .37]	.15	.93	.352
quadratic × input vs. parsing	.21 [-.03, .45]	.15	1.46	.146
quadratic × input vs. test	.11 [-.13, .34]	.15	.73	.468
cubic × input vs. parsing	.05 [-.19, .29]	.15	.37	.714
cubic × input vs. test	-.1 [-.34, .14]	.15	-.71	.478
pre- vs. post- × input vs. parsing	.03 [-.08, .14]	.07	.50	.619
pre- vs. delayed- × input vs. parsing	-.03 [-.14, .08]	.07	-.46	.645
pre- vs. post- × input vs. test	.01 [-.10, .12]	.07	.12	.904
pre- vs. delayed- × input vs. test	.04 [-.07, .15]	.07	.61	.545
linear × pre- vs. post- × input vs. parsing	-.25 [-.6, .10]	.21	-1.18	.239
linear × pre- vs. delayed- × input vs. parsing	-.19 [-.54, .16]	.21	-.88	.380
linear × pre- vs. post- × input vs. test	-.16 [-.5, .19]	.21	-.75	.456
linear × pre- vs. delayed- × input vs. test	-.09 [-.45, .26]	.22	-.43	.668
quadratic × pre- vs. post- × input vs. parsing	-.33 [-.68, .01]	.21	-1.58	.115
quadratic × pre- vs. delayed- × input vs. parsing	-.35 [-.69, .00]	.21	-1.64	.100
quadratic × pre- vs. post- × input vs. test	-.22 [-.56, .12]	.21	-1.04	.297
quadratic × pre- vs. delayed- × input vs. test	-.15 [-.51, .2]	.22	-.71	.476
cubic × pre- vs. post- × input vs. parsing	-.26 [-.61, .09]	.21	-1.24	.216
cubic × pre- vs. delayed- × input vs. parsing	.20 [-.15, .55]	.21	.95	.345
cubic × pre- vs. post- × input vs. test	.00 [-.34, .34]	.21	-.02	.985
cubic × pre- vs. delayed- × input vs. test	.32 [-.04, .67]	.21	1.47	.141

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; *** significantly different from zero when α<.001

SRC-I

First critical word

Table Appx. 163 Fixed effects of model analysis in the eye-tracking test for the SRC-I structure for the first critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.7 [1.64, 1.76]	.04	48.33	<.001***
linear	.03 [-.16, .23]	.12	.29	.769
quadratic	.05 [-.14, .24]	.12	.43	.667
cubic	.06 [-.13, .26]	.12	.52	.600
pre- vs. post-	**.09 [.01, .17]**	**.05**	**1.88**	**.061**
pre- vs. delayed-	.03 [-.05, .11]	.05	.65	.516
input vs. parsing	**.09 [.01, .17]**	**.05**	**1.90**	**.058**
input vs. test	.10 [.02, .18]	.05	2.00	.046*
linear × pre- vs. post-	-.01 [-.29, .27]	.17	-.06	.955
linear × pre- vs. delayed-	.14 [-.13, .41]	.17	.84	.404
quadratic × pre- vs. post-	.01 [-.27, .29]	.17	.07	.948
quadratic × pre- vs. delayed-	.13 [-.15, .4]	.17	.75	.451
cubic × pre- vs. post-	-.03 [-.3, .25]	.17	-.15	.881
cubic × pre- vs. delayed-	.00 [-.28, .27]	.17	-.01	.995
linear × input vs. parsing	.06 [-.21, .33]	.16	.38	.708
linear × input vs. test	-.06 [-.33, .22]	.17	-.33	.740
quadratic × input vs. parsing	.02 [-.25, .29]	.16	.11	.912
quadratic × input vs. test	-.19 [-.46, .08]	.17	-1.14	.254
cubic × input vs. parsing	.06 [-.21, .33]	.16	.36	.723
cubic × input vs. test	-.16 [-.44, .11]	.17	-.97	.331
pre- vs. post- × input vs. parsing	-.05 [-.17, .06]	.07	-.77	.443
pre- vs. delayed- × input vs. parsing	-.04 [-.15, .07]	.07	-.62	.535
pre- vs. post- × input vs. test	**-.13 [-.24, -.02]**	**.07**	**-1.93**	**.053**
pre- vs. delayed- × input vs. test	**-.12 [-.23, -.01]**	**.07**	**-1.73**	**.084**
linear × pre- vs. post- × input vs. parsing	-.27 [-.66, .12]	.24	-1.16	.247
linear × pre- vs. delayed- × input vs. parsing	**-.39 [-.77, -.01]**	**.23**	**-1.70**	**.089**
linear × pre- vs. post- × input vs. test	.15 [-.23, .54]	.23	.65	.517
linear × pre- vs. delayed- × input vs. test	-.34 [-.73, .05]	.24	-1.44	.149
quadratic × pre- vs. post- × input vs. parsing	-.05 [-.44, .34]	.24	-.21	.836
quadratic × pre- vs. delayed- × input vs. parsing	-.30 [-.68, .08]	.23	-1.32	.188
quadratic × pre- vs. post- × input vs. test	.13 [-.25, .52]	.24	.57	.571
quadratic × pre- vs. delayed- × input vs. test	-.13 [-.51, .26]	.24	-.53	.595
cubic × pre- vs. post- × input vs. parsing	-.12 [-.51, .27]	.24	-.52	.606
cubic × pre- vs. delayed- × input vs. parsing	-.19 [-.56, .19]	.23	-.81	.421
cubic × pre- vs. post- × input vs. test	-.05 [-.43, .34]	.23	-.19	.847
cubic × pre- vs. delayed- × input vs. test	.08 [-.3, .47]	.24	.36	.721

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; *** significantly different from zero when $\alpha < .001$

Second critical word

Table Appx. 164 Fixed effects of model analysis in the eye-tracking test for the SRC-I structure for the second critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.7 [1.58, 1.82]	.07	22.88	<.001***
linear	.11 [-.09, .31]	.12	.90	.367
pre- vs. post-	.07 [-.09, .23]	.10	.70	.485
pre- vs. delayed-	.02 [-.14, .18]	.10	.18	.855
input vs. parsing	-.02 [-.18, .15]	.10	-.17	.864
input vs. test	.02 [-.15, .19]	.10	.19	.847
linear × pre- vs. post-	**-.32 [-.60, -.04]**	**.17**	**-1.90**	**.058**
linear × pre- vs. delayed-	.01 [-.27, .29]	.17	.07	.944
linear × input vs. parsing	-.12 [-.4, .15]	.17	-.75	.456
linear × input vs. test	.12 [-.16, .40]	.17	.68	.496
pre- vs. post- × input vs. parsing	-.05 [-.27, .17]	.13	-.36	.722
pre- vs. delayed- × input vs. parsing	.04 [-.18, .26]	.13	.29	.774
pre- vs. post- × input vs. test	-.05 [-.28, .17]	.14	-.40	.692
pre- vs. delayed- × input vs. test	.01 [-.22, .24]	.14	.06	.955
linear × pre- vs. post- × input vs. parsing	.29 [-.09, .66]	.23	1.25	.213
linear × pre- vs. delayed- × input vs. parsing	-.26 [-.64, .12]	.23	-1.12	.263
linear × pre- vs. post- × input vs. test	-.05 [-.43, .34]	.23	-.21	.838
linear × pre- vs. delayed- × input vs. test	-.26 [-.65, .14]	.24	-1.07	.284

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; *** significantly different from zero when α<.001

Third critical word

Table Appx. 165 Fixed effects of model analysis in the eye-tracking test for the SRC-I structure for the third critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.74 [1.67, 1.8]	.04	45.56	< .001***
linear	.06 [-.13, .24]	.11	.51	.607
pre- vs. post-	.01 [-.07, .1]	.05	.26	.792
pre- vs. delayed-	.07 [-.01, .15]	.05	1.42	.155
input vs. parsing	.05 [-.03, .14]	.05	1.05	.294
input vs. test	.02 [-.06, .11]	.05	.44	.662
linear × pre- vs. post-	-.17 [-.43, .09]	.16	-1.09	.278
linear × pre- vs. delayed-	-.12 [-.38, .13]	.16	-.79	.429
linear × input vs. parsing	-.07 [-.32, .19]	.15	-.43	.669
linear × input vs. test	-.13 [-.39, .13]	.16	-.84	.402
pre- vs. post- × input vs. parsing	.00 [-.11, .12]	.07	.04	.968
pre- vs. delayed- × input vs. parsing	-.01 [-.12, .1]	.07	-.18	.855
pre- vs. post- × input vs. test	.01 [-.11, .12]	.07	.12	.902
pre- vs. delayed- × input vs. test	-.03 [-.14, .09]	.07	-.39	.694
linear × pre- vs. post- × input vs. parsing	.16 [-.2, .51]	.22	.73	.464
linear × pre- vs. delayed- × input vs. parsing	.19 [-.16, .54]	.21	.91	.363
linear × pre- vs. post- × input vs. test	.28 [-.09, .64]	.22	1.25	.211
linear × pre- vs. delayed- × input vs. test	.34 [-.02, .71]	.22	1.56	.118

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

ORC-A

First critical word

Table Appx. 166 Fixed effects of model analysis in the eye-tracking test for the ORC-A structure for the first critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.76 [1.65, 1.88]	.07	25.85	<.001***
linear	.06 [-.13, .25]	.12	.52	.601
quadratic	-.04 [-.24, .15]	.12	-.37	.710
pre- vs. post-	.12 [-.05, .28]	.10	1.17	.243
pre- vs. delayed-	-.02 [-.17, .14]	.09	-.21	.837
input vs. parsing	.08 [-.07, .24]	.09	.88	.381
input vs. test	.06 [-.10, .22]	.10	.61	.541
linear × pre- vs. post-	-.04 [-.32, .24]	.17	-.22	.824
linear × pre- vs. delayed-	-.07 [-.34, .20]	.16	-.42	.678
quadratic × pre- vs. post-	-.04 [-.33, .24]	.17	-.26	.797
quadratic × pre- vs. delayed-	.16 [-.11, .43]	.16	.99	.324
linear × input vs. parsing	-.05 [-.32, .22]	.16	-.30	.767
linear × input vs. test	-.14 [-.41, .13]	.17	-.86	.392
quadratic × input vs. parsing	-.05 [-.32, .21]	.16	-.33	.740
quadratic × input vs. test	-.08 [-.36, .19]	.17	-.50	.616
pre- vs. post- × input vs. parsing	-.07 [-.3, .16]	.14	-.49	.622
pre- vs. delayed- × input vs. parsing	-.01 [-.23, .21]	.13	-.08	.938
pre- vs. post- × input vs. test	-.16 [-.39, .07]	.14	-1.14	.255
pre- vs. delayed- × input vs. test	-.02 [-.24, .21]	.14	-.12	.902
linear × pre- vs. post- × input vs. parsing	.27 [-.14, .67]	.24	1.09	.277
linear × pre- vs. delayed- × input vs. parsing	.08 [-.3, .45]	.23	.34	.736
linear × pre- vs. post- × input vs. test	-.03 [-.43, .36]	.24	-.14	.888
linear × pre- vs. delayed- × input vs. test	.30 [-.09, .68]	.23	1.27	.204
quadratic × pre- vs. post- × input vs. parsing	.11 [-.30, .51]	.24	.44	.664
quadratic × pre- vs. delayed- × input vs. parsing	-.17 [-.54, .21]	.23	-.74	.461
quadratic × pre- vs. post- × input vs. test	.07 [-.33, .46]	.24	.27	.785
quadratic × pre- vs. delayed- × input vs. test	.00 [-.39, .38]	.23	.00	.999

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

Second critical word

Table Appx. 167 Fixed effects of model analysis in the eye-tracking test for the ORC-A structure for the second critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.76 [1.69, 1.82]	.04	43.13	<.001***
linear	.12 [-.07, .32]	.12	1.04	.300
quadratic	-.02 [-.21, .17]	.12	-.17	.864
cubic	-.02 [-.21, .18]	.12	-.16	.875
pre- vs. post-	.03 [-.07, .12]	.06	.46	.647
pre- vs. delayed-	-.02 [-.11, .07]	.06	-.36	.719
input vs. parsing	.01 [-.08, .10]	.06	.19	.851
input vs. test	-.07 [-.16, .03]	.06	-1.13	.258
linear × pre- vs. post-	-.19 [-.47, .09]	.17	-1.10	.270
linear × pre- vs. delayed-	-.26 [-.54, .01]	.17	-1.57	.116
quadratic × pre- vs. post-	.26 [-.03, .54]	.17	1.49	.136
quadratic × pre- vs. delayed-	-.04 [-.32, .23]	.17	-.25	.802
cubic × pre- vs. post-	.12 [-.16, .4]	.17	.69	.493
cubic × pre- vs. delayed-	.04 [-.24, .31]	.17	.23	.816
linear × input vs. parsing	-.19 [-.47, .08]	.17	-1.17	.244
linear × input vs. test	-.16 [-.43, .12]	.17	-.95	.345
quadratic × input vs. parsing	-.06 [-.33, .21]	.16	-.37	.714
quadratic × input vs. test	.07 [-.20, .34]	.17	.44	.663
cubic × input vs. parsing	.08 [-.19, .35]	.16	.48	.633
cubic × input vs. test	-.10 [-.38, .17]	.17	-.62	.538
pre- vs. post- × input vs. parsing	-.08 [-.21, .06]	.08	-.92	.356
pre- vs. delayed- × input vs. parsing	.01 [-.12, .14]	.08	.13	.896
pre- vs. post- × input vs. test	-.03 [-.16, .1]	.08	-.36	.722
pre- vs. delayed- × input vs. test	.05 [-.08, .18]	.08	.60	.551
linear × pre- vs. post- × input vs. parsing	.33 [-.07, .73]	.24	1.34	.181
linear × pre- vs. delayed- × input vs. parsing	.31 [-.08, .69]	.23	1.32	.187
linear × pre- vs. post- × input vs. test	.16 [-.23, .55]	.24	.67	.501
linear × pre- vs. delayed- × input vs. test	.39 [0, .79]	.24	1.64	.101
quadratic × pre- vs. post- × input vs. parsing	-.34 [-.74, .06]	.24	-1.39	.166
quadratic × pre- vs. delayed- × input vs. parsing	.09 [-.29, .48]	.23	.40	.693
quadratic × pre- vs. post- × input vs. test	-.38 [-.78, .01]	.24	-1.60	.109
quadratic × pre- vs. delayed- × input vs. test	.14 [-.25, .53]	.24	.60	.551
cubic × pre- vs. post- × input vs. parsing	-.10 [-.50, .30]	.24	-.41	.682
cubic × pre- vs. delayed- × input vs. parsing	-.09 [-.47, .29]	.23	-.38	.703
cubic × pre- vs. post- × input vs. test	-.08 [-.48, .31]	.24	-.34	.733
cubic × pre- vs. delayed- × input vs. test	.22 [-.17, .61]	.24	.93	.354

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

Third critical word

Table Appx. 168 Fixed effects of model analysis in the eye-tracking test for the ORC-A structure for the third critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.75 [1.7, 1.8]	.03	55.11	<.001***
linear	-.08 [-.25, .1]	.11	-.74	.458
quadratic	-.01 [-.18, .17]	.11	-.08	.937
cubic	.00 [-.17, .18]	.11	.02	.987
pre- vs. post-	.04 [-.03, .11]	.04	1.01	.313
pre- vs. delayed-	-.04 [-.11, .03]	.04	-.95	.341
input vs. parsing	-.02 [-.09, .05]	.04	-.46	.648
input vs. test	.01 [-.07, .08]	.04	.13	.898
linear × pre- vs. post-	.09 [-.16, .34]	.15	.59	.558
linear × pre- vs. delayed-	-.04 [-.3, .21]	.16	-.29	.776
quadratic × pre- vs. post-	.04 [-.21, .29]	.15	.28	.780
quadratic × pre- vs. delayed-	.09 [-.16, .35]	.15	.61	.542
cubic × pre- vs. post-	.13 [-.12, .37]	.15	.83	.409
cubic × pre- vs. delayed-	-.17 [-.43, .08]	.15	-1.13	.258
linear × input vs. parsing	.07 [-.18, .32]	.15	.45	.654
linear × input vs. test	.31 [.07, .56]	.15	2.08	.038*
quadratic × input vs. parsing	.01 [-.24, .26]	.15	.07	.942
quadratic × input vs. test	-.07 [-.32, .18]	.15	-.46	.648
cubic × input vs. parsing	-.1 [-.35, .14]	.15	-.70	.487
cubic × input vs. test	-.12 [-.36, .13]	.15	-.77	.442
pre- vs. post- × input vs. parsing	.01 [-.09, .1]	.06	.09	.933
pre- vs. delayed- × input vs. parsing	.07 [-.03, .16]	.06	1.16	.247
pre- vs. post- × input vs. test	-.05 [-.15, .04]	.06	-.91	.363
pre- vs. delayed- × input vs. test	.08 [-.02, .18]	.06	1.37	.170
linear × pre- vs. post- × input vs. parsing	-.05 [-.4, .3]	.22	-.23	.817
linear × pre- vs. delayed- × input vs. parsing	.05 [-.3, .4]	.21	.25	.806
linear × pre- vs. post- × input vs. test	-.17 [-.52, .18]	.21	-.82	.414
linear × pre- vs. delayed- × input vs. test	-.13 [-.48, .23]	.22	-.58	.561
quadratic × pre- vs. post- × input vs. parsing	.02 [-.33, .37]	.21	.09	.932
quadratic × pre- vs. delayed- × input vs. parsing	-.16 [-.51, .19]	.21	-.75	.452
quadratic × pre- vs. post- × input vs. test	.03 [-.31, .38]	.21	.16	.873
quadratic × pre- vs. delayed- × input vs. test	-.11 [-.46, .24]	.22	-.51	.613
cubic × pre- vs. post- × input vs. parsing	.08 [-.27, .43]	.21	.38	.706
cubic × pre- vs. delayed- × input vs. parsing	.17 [-.17, .52]	.21	.82	.415
cubic × pre- vs. post- × input vs. test	.05 [-.29, .4]	.21	.25	.800
cubic × pre- vs. delayed- × input vs. test	.12 [-.23, .47]	.21	.57	.567

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when $\alpha \leq .05$; *** significantly different from zero when $\alpha<.00$

ORC-I

First critical word

Table Appx. 169 Fixed effects of model analysis in the eye-tracking test for the ORC-I structure for the first critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.75 [1.63, 1.87]	.07	23.86	< .001***
linear	.37 [.16, .59]	.13	2.93	.003**
pre- vs. post-	.03 [-.14, .2]	.10	.30	.765
pre- vs. delayed-	-.03 [-.21, .14]	.11	-.30	.763
input vs. parsing	.03 [-.14, .2]	.10	.27	.789
input vs. test	-.01 [-.17, .16]	.10	-.09	.931
linear × pre- vs. post-	-.54 [-.84, -.24]	.18	-2.99	.003**
linear × pre- vs. delayed-	-.20 [-.5, .11]	.19	-1.08	.282
linear × input vs. parsing	-.57 [-.86, -.28]	.18	-3.19	.001**
linear × input vs. test	-.26 [-.54, .02]	.17	-1.50	.133
pre- vs. post- × input vs. parsing	-.07 [-.31, .16]	.14	-.51	.612
pre- vs. delayed- × input vs. parsing	.00 [-.24, .24]	.15	.00	.998
pre- vs. post- × input vs. test	.03 [-.21, .27]	.14	.21	.835
pre- vs. delayed- × input vs. test	.03 [-.21, .27]	.15	.20	.845
linear × pre- vs. post- × input vs. parsing	.54 [.13, .95]	.25	2.17	.030*
linear × pre- vs. delayed- × input vs. parsing	.50 [.08, .91]	.25	1.97	.049*
linear × pre- vs. post- × input vs. test	.40 [-.01, .81]	.25	1.61	.108
linear × pre- vs. delayed- × input vs. test	.14 [-.28, .56]	.26	.56	.574

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$; *** significantly different from zero when $\alpha<.001$

Second critical word

Table Appx. 170 Fixed effects of model analysis in the eye-tracking test for the ORC-I structure for the second critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.77 [1.7, 1.84]	.04	41.14	< .001***
linear	.08 [-.11, .28]	.12	.70	.485
pre- vs. post-	-.06 [-.16, .04]	.06	-1.03	.301
pre- vs. delayed-	.01 [-.09, .1]	.06	.13	.898
input vs. parsing	.03 [-.07, .13]	.06	.49	.624
input vs. test	-.03 [-.13, .07]	.06	-.45	.651
linear × pre- vs. post-	-.15 [-.44, .13]	.17	-.90	.369
linear × pre- vs. delayed-	-.11 [-.4, .18]	.18	-.63	.529
linear × input vs. parsing	-.03 [-.3, .25]	.17	-.17	.864
linear × input vs. test	-.01 [-.28, .26]	.17	-.07	.947
pre- vs. post- × input vs. parsing	.03 [-.1, .16]	.08	.34	.736
pre- vs. delayed- × input vs. parsing	-.02 [-.16, .11]	.08	-.29	.775
pre- vs. post- × input vs. test	.08 [-.05, .22]	.08	1.00	.316
pre- vs. delayed- × input vs. test	.04 [-.09, .18]	.08	.53	.599
linear × pre- vs. post- × input vs. parsing	.3 [-.09, .7]	.24	1.28	.201
linear × pre- vs. delayed- × input vs. parsing	.00 [-.39, .4]	.24	.01	.990
linear × pre- vs. post- × input vs. test	.11 [-.29, .5]	.24	.44	.660
linear × pre- vs. delayed- × input vs. test	.14 [-.26, .54]	.24	.59	.558

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

Third critical word

Table Appx. 171 Fixed effects of model analysis in the eye-tracking test for the ORC-I structure for the third critical word (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*t*	*p*
intercept	1.78 [1.72, 1.83]	.04	5.39	<.001***
linear	-.30 [-.49, -.1]	.12	-2.52	.012*
quadratic	-.06 [-.25, .13]	.12	-.51	.612
pre- vs. post-	-.06 [-.13, .02]	.05	-1.25	.213
pre- vs. delayed-	-.01 [-.09, .06]	.04	-.29	.773
input vs. parsing	.03 [-.05, .11]	.05	.57	.573
input vs. test	-.02 [-.1, .06]	.05	-.34	.732
linear × pre- vs. post-	**.32 [.05, .59]**	**.17**	**1.94**	**.052**
linear × pre- vs. delayed-	.24 [-.03, .51]	.16	1.46	.144
quadratic × pre- vs. post-	-.15 [-.42, .11]	.16	-.95	.344
quadratic × pre- vs. delayed-	-.01 [-.27, .26]	.16	-.03	.973
linear × input vs. parsing	.39 [.12, .66]	.16	2.41	.016*
linear × input vs. test	.36 [.09, .63]	.16	2.23	.026*
quadratic × input vs. parsing	.21 [-.05, .47]	.16	1.32	.186
quadratic × input vs. test	.18 [-.08, .45]	.16	1.13	.259
pre- vs. post- × input vs. parsing	.02 [-.08, .13]	.06	.37	.712
pre- vs. delayed- × input vs. parsing	.01 [-.09, .11]	.06	.19	.852
pre- vs. post- × input vs. test	.05 [-.05, .16]	.06	.88	.379
pre- vs. delayed- × input vs. test	.07 [-.03, .18]	.06	1.15	.249
linear × pre- vs. post- × input vs. parsing	-.48 [-.85, -.10]	.23	-2.08	.037*
linear × pre- vs. delayed- × input vs. parsing	-.23 [-.59, .14]	.22	-1.01	.312
linear × pre- vs. post- × input vs. test	-.29 [-.66, .08]	.23	-1.30	.192
linear × pre- vs. delayed- × input vs. test	-.22 [-.59, .16]	.23	-.95	.343
quadratic × pre- vs. post- × input vs. parsing	.14 [-.23, .51]	.23	.61	.543
quadratic × pre- vs. delayed- × input vs. parsing	-.04 [-.4, .32]	.22	-.18	.861
quadratic × pre- vs. post- × input vs. test	-.03 [-.39, .34]	.22	-.12	.903
quadratic × pre- vs. delayed- × input vs. test	-.12 [-.49, .25]	.23	-.52	.600

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; *** significantly different from zero when $\alpha<.001$

Appendix 31 RQ2: Fixed effects of model analysis in the oral production test (baseline of the input flood group)

SRC-A

Table Appx. 172 Fixed effects of model analysis in the oral production test for SRC-A structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-.25 [-1.08, .58]	.50	-.50	.618	.78 [.34, 1.78]
input vs. parsing	.50 [-.64, 1.65]	.69	.73	.467	1.66 [.53, 5.19]
input vs. test	.39 [-.78, 1.56]	.71	.55	.581	1.48 [.46, 4.77]
pre- vs. post-	**1.10 [.08, 2.12]**	**.62**	**1.77**	**.077**	**2.99 [1.08, 8.29]**
pre- vs. delayed-	2.10 [.86, 3.34]	.75	2.78	.005**	8.13 [2.35, 28.09]
input vs. parsing × pre- vs. post-	.46 [-.91, 1.84]	.84	.55	.580	1.59 [.40, 6.27]
input vs. test × pre- vs. post-	-1.00 [-2.44, .43]	.87	-1.15	.250	.37 [.09, 1.54]
input vs. parsing × pre- vs. delayed-	.35 [-1.24, 1.94]	.97	.36	.719	1.42 [.29, 6.94]
input vs. test × pre- vs. delayed-	-1.13 [-2.80, .54]	1.01	-1.11	.266	.32 [.06, 1.72]

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; ** significantly different from zero when $\alpha < .01$

SRC-I

Table Appx. 173 Fixed effects of model analysis in the oral production test for SRC-I structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	.20 [-.93, 1.32]	.68	.29	.772	1.22 [.40, 3.76]
input vs. parsing	.04 [-1.49, 1.56]	.93	.04	.970	1.04 [.22, 4.78]
input vs. test	-.36 [-1.93, 1.21]	.95	-.38	.705	.70 [.15, 3.34]
pre- vs. post-	.83 [-.54, 2.20]	.83	1.00	.320	2.29 [.58, 9.00]
pre- vs. delayed-	2.05 [.39, 3.72]	1.01	2.03	.042*	7.80 [1.48, 41.12]
input vs. parsing × pre- vs. post-	2.53 [.59, 4.46]	1.18	2.15	.032*	12.52 [1.81, 86.58]
input vs. test × pre- vs. post-	-.28 [-2.15, 1.59]	1.14	-.25	.803	.75 [.12, 4.89]
input vs. parsing × pre- vs. delayed-	.81 [-1.30, 2.93]	1.28	.64	.526	2.26 [.27, 18.65]
input vs. test × pre- vs. delayed-	-.23 [-2.40, 1.94]	1.32	-.17	.864	.80 [.09, 6.99]

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when $\alpha \leq .05$

ORC-A

Table Appx. 174 Fixed effects of model analysis in the oral production test for ORC-A structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-1.61 [-3.67, .45]	1.25	-1.29	.198	.20 [.03, 1.56]
input vs. parsing	-.69 [-3.55, 2.18]	1.74	-.39	.693	.50 [.03, 8.81]
input vs. test	**-3.52 [-6.60, -.44]**	**1.87**	**-1.88**	**.060**	**.03 [.00, .64]**
pre- vs. post-	1.67 [.79, 2.55]	.54	3.11	.002**	5.31 [2.20, 12.87]
pre- vs. delayed-	4.88 [3.63, 6.12]	.76	6.43	<.001***	131.32 [37.74, 456.95]
input vs. parsing × pre- vs. post-	3.97 [2.33, 5.62]	1.00	3.97	<.001***	53.21 [1.26, 275.91]
input vs. test × pre- vs. post-	1.71 [.28, 3.15]	.87	1.96	.049*	5.54 [1.32, 23.22]
input vs. parsing × pre- vs. delayed-	2.79 [.57, 5.01]	1.35	2.06	.039*	16.28 [1.76, 15.38]
input vs. test × pre- vs. delayed-	-.51 [-2.22, 1.19]	1.04	-.49	.621	.60 [.11, 3.30]

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$; *** significantly different from zero when $\alpha < .001$

ORC-I

Table Appx. 175 Fixed effects of model analysis in the oral production test for ORC-I structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-1.13 [-2.88, .63]	1.06	-1.06	.291	.32 [.06, 1.87]
input vs. parsing	-.82 [-3.25, 1.61]	1.48	-.55	.580	.44 [.04, 5.01]
input vs. test	**-2.80 [-5.34, -.26]**	**1.55**	**-1.81**	**.070**	**.06 [.00, .77]**
pre- vs. post-	3.31 [.69, 5.92]	1.59	2.08	.037*	27.27 [2.00, 372.08]
pre- vs. delayed-test	5.09 [2.85, 7.33]	1.36	3.74	<.001***	162.05 [17.25, 1522.40]
input vs. parsing × pre- vs. post-	4.58 [.80, 8.36]	2.30	1.99	.046*	97.51 [2.23, 4269.33]
input vs. test × pre- vs. post-	-2.39 [-6.30, 1.52]	2.38	-1.01	.314	.09 [.00, 4.56]
input vs. parsing × pre- vs. delayed-	2.78 [-.16, 5.72]	1.79	1.55	.120	16.11 [.85, 305.31]
input vs. test × pre- vs. delayed-	-1.26 [-4.27, 1.75]	1.83	-.69	.490	.28 [.01, 5.73]

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; *** significantly different from zero when $\alpha < .001$

Appendix 32 RQ2: Fixed effects of model analysis of the metalinguistic knowledge test (baseline of the input flood group)

Task of deciding match or mismatch

Matched items

ORC-A

Table Appx. 176 Fixed effects of model analysis for the task of deciding match or mismatch (matched item) in the metalinguistic knowledge test for ORC-A structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	2.88 [1.84, 3.91]	.63	4.56	<.001***	17.73 [6.28, 5.03]
input vs. parsing	-.27 [-1.56, 1.03]	.79	-.34	.733	.76 [.21, 2.79]
input vs. test	1.15 [-.76, 3.07]	1.16	.99	.323	3.16 [.47, 21.46]
pre- vs. post-	.01 [-1.37, 1.39]	.84	.02	.987	1.01 [.26, 4.03]
pre- vs. delayed-	1.16 [-.75, 3.08]	1.17	1.00	.318	3.20 [.47, 21.77]
input vs. parsing × pre- vs. post-	1.45 [-.87, 3.76]	1.41	1.03	.303	4.25 [.42, 42.87]
input vs. test × pre- vs. post-	16.46 [-516.16, 549.11]	183.92	.09	.929	1.41E+7 [0, 3.4E+138]
input vs. parsing × pre- vs. delayed-	-.84 [-3.15, 1.47]	1.41	-.60	.550	.43 [.04, 4.36]
input vs. test × pre- vs. delayed-	-1.86 [-4.65, .93]	1.70	-1.10	.272	.16 [.01, 2.52]

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

ORC-I

Table Appx. 177 Fixed effects of model analysis for the task of deciding match or mismatch (matched item) in the metalinguistic knowledge test for ORC-I structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	3.59 [2.07, 5.12]	.93	3.88	<.001***	36.38 [7.93, 166.93]
input vs. parsing	1.26 [-.76, 3.27]	1.23	1.02	.306	3.51 [.47, 26.36]
input vs. test	-.04 [-1.53, 1.45]	.91	-.05	.963	.96 [.22, 4.26]
pre- vs. post-	.47 [-1.15, 2.08]	.98	.48	.633	1.6 [.32, 8.04]
pre- vs. delayed-	-.06 [-1.53, 1.41]	.89	-.06	.950	.95 [.22, 4.12]
input vs. parsing × pre- vs. post-	-1.65 [-4.21, .91]	1.56	-1.06	.290	.19 [.01, 2.49]
input vs. test × pre- vs. post-	.92 [-1.64, 3.47]	1.55	.59	.555	2.5 [.19, 32.27]
input vs. parsing × pre- vs. delayed-	.00 [-2.82, 2.82]	1.72	.00	.998	1.00 [.06, 16.86]
input vs. test × pre- vs. delayed-	1.26 [-1.22, 3.74]	1.51	.84	.402	3.53 [.30, 42.11]

Note: parsing = parsing group; input = input flood group; test = test-only group; *** significantly different from zero when α<.001

Mismatched items

SRC-A

Table Appx. 178 Fixed effects of model analysis for the task of deciding match or mismatch (mismatched item) in the metalinguistic knowledge test for SRC-A structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	2.39 [1.20, 3.57]	.72	3.31	.001**	1.86 [3.32, 35.58]
input vs. parsing	.00 [-1.23, 1.23]	.75	-.01	.996	1.00 [.29, 3.41]
input vs. test	.09 [-1.16, 1.34]	.76	.12	.907	1.09 [.31, 3.83]
pre- vs. post-	.61 [-.90, 2.12]	.92	.67	.504	1.85 [.41, 8.34]
pre- vs. delayed-	.71 [-.79, 2.21]	.91	.78	.435	2.04 [.45, 9.12]
input vs. parsing × pre- vs. post-	.93 [-.82, 2.67]	1.06	.87	.382	2.51 [.44, 14.50]
input vs. test × pre- vs. post-	-.36 [-1.95, 1.23]	.97	-.37	.710	.70 [.14, 3.42]
input vs. parsing × pre- vs. delayed-	1.24 [-.66, 3.13]	1.15	1.08	.282	3.45 [.52, 22.95]
input vs. test × pre- vs. delayed-	-.49 [-2.10, 1.13]	.98	-.50	.621	.61 [.12, 3.10]

Note: parsing = parsing group; input = input flood group; test = test-only group; ** significantly different from zero when α<.01

SRC-I

Table Appx. 179 Fixed effects of model analysis for the task of deciding match or mismatch (mismatched item) in the metalinguistic knowledge test for SRC-I structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	2.64 [1.11, 4.17]	.93	2.84	.004**	14.05 [3.05, 64.79]
input vs. parsing	.35 [-1.14, 1.84]	.91	.39	.697	1.42 [.32, 6.32]
input vs. test	-1.69 [-3.10,-2.80]	.86	-1.97	.049	.18 [.05, .76]
pre- vs. post-	.30 [-1.32, 1.92]	.99	.31	.758	1.36 [.27, 6.85]
pre- vs. delayed-	1.47 [-.57, 3.51]	1.24	1.19	.235	4.36 [.57, 33.43]
input vs. parsing × pre- vs. post-	**2.39 [.08, 4.70]**	**1.40**	**1.70**	**.089**	**1.90 [1.08, 109.55]**
input vs. test × pre- vs. post-	**1.74 [.16, 3.31]**	**.96**	**1.81**	**.070**	**5.69 [1.18, 27.48]**
input vs. parsing × pre- vs. delayed-	.17 [-1.68. 2.03]	1.13	.16	.877	1.19 [.19, 7.62]
input vs. test × pre- vs. delayed-	.85 [-.87, 2.58]	1.05	.81	.417	2.34 [.42, 13.16]

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; ** significantly different from zero when α<.01

ORC-A

Table Appx. 180 Fixed effects of model analysis for the task of deciding match or mismatch (mismatched item) in the metalinguistic knowledge test for ORC-A structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	1.73 [.76, 2.69]	.59	2.95	.003**	5.63 [2.15, 14.77]
input vs. parsing	.82 [-.45, 2.10]	.78	1.06	.289	2.28 [.63, 8.19]
input vs. test	.33 [-.91, 1.57]	.75	.44	.661	1.39 [.40, 4.81]
pre- vs. post-	.53 [-.47, 1.53]	.61	.87	.383	1.70 [.63, 4.62]
pre- vs. delayed-	.76 [-.28, 1.80]	.63	1.21	.228	2.14 [.76, 6.04]
input vs. parsing × pre- vs. post-	.77 [-.84, 2.39]	.98	.79	.432	2.16 [.43, 1.89]
input vs. test × pre- vs. post-	.51 [-.97, 1.99]	.90	.57	.571	1.66 [.38, 7.28]
input vs. parsing × pre- vs. delayed-	1.03 [-.75, 2.81]	1.08	.95	.342	2.80 [.47, 16.64]
input vs. test × pre- vs. delayed-	.95 [-.68, 2.57]	.99	.96	.339	2.58 [.51, 13.12]

Note: parsing = parsing group; input = input flood group; test = test-only group; ** significantly different from zero when $\alpha < .01$

ORC-I

Table Appx. 181 Fixed effects of model analysis for the task of deciding match or mismatch (mismatched item) in the metalinguistic knowledge test for ORC-I structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	1.41 [.44, 2.38]	.59	2.39	.017*	4.09 [1.55, 1.75]
input vs. parsing	-.54 [-1.70, .61]	.70	-.77	.439	.58 [.18, 1.84]
input vs. test	-1.46 [-2.62, -.29]	.71	-2.06	.039*	.23 [.07, .74]
pre- vs. post-	.26 [-.68, 1.20]	.57	.46	.648	1.30 [.51, 3.31]
pre- vs. delayed-	.34 [-.59, 1.27]	.57	.60	.548	1.41 [.55, 3.56]
input vs. parsing × pre- vs. post-	1.28 [-.03, 2.60]	.80	1.61	.108	3.61 [.97, 13.44]
input vs. test × pre- vs. post-	1.66 [.35, 2.97]	.80	2.09	.037*	5.26 [1.42, 19.48]
input vs. parsing × pre- vs. delayed-	1.73 [.35, 3.10]	.84	2.07	.039*	5.63 [1.42, 22.28]
input vs. test × pre- vs. delayed-	2.64 [1.21, 4.07]	.87	3.03	.002**	13.96 [3.34, 58.41]

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$

Task of sentence correction

SRC-A

Table Appx. 182 Fixed effects of model analysis for the task of sentence correction in the metalinguistic knowledge test for SRC-A structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	**1.49 [.35, 2.63]**	**.70**	**2.15**	**.032***	**4.44 [1.42, 13.92]**
input vs. parsing	.38 [-1.12, 1.88]	.91	.42	.674	1.47 [.33, 6.58]
input vs. test	**-1.68 [-3.17, -.19]**	**.91**	**-1.85**	**.064**	**.19 [.04, .83]**
pre- vs. post-	.55 [-.43, 1.52]	.59	.92	.356	1.73 [.65, 4.59]
pre- vs. delayed-	1.03 [-.03, 2.09]	.64	1.60	.110	2.80 [.97, 8.08]
input vs. parsing × pre- vs. post-	**2.23 [.54, 3.91]**	**1.02**	**2.18**	**.030***	**9.29 [1.72, 5.08]**
input vs. test × pre- vs. post-	**1.50 [.07, 2.93]**	**.87**	**1.73**	**.084**	**4.48 [1.08, 18.67]**
input vs. parsing × pre- vs. delayed-	1.60 [-.12, 3.32]	1.05	1.53	.127	4.94 [.88, 27.59]
input vs. test × pre- vs. delayed-	1.32 [-.23, 2.88]	.94	1.40	.161	3.76 [.79, 17.75]

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$

SRC-I

Table Appx. 183 Fixed effects of model analysis for the task of sentence correction in the metalinguistic knowledge test for SRC-I structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	**1.43 [.18, 2.67]**	**.76**	**1.88**	**.060**	**4.17 [1.20, 14.49]**
input vs. parsing	-.06 [-1.53, 1.41]	.89	-.07	.946	.94 [.22, 4.08]
input vs. test	-2.15 [-3.65, -.65]	.91	-2.36	.018	.12 [.03, .52]
pre- vs. post-	1.14 [-.21, 2.49]	.82	1.39	.166	3.11[.81, 12.01]
pre- vs. delayed-	**1.69 [.01, 3.38]**	**1.02**	**1.65**	**.098**	**5.44 [1.01, 29.37]**
input vs. parsing × pre- vs. post-	**2.13 [.35, 3.91]**	**1.08**	**1.97**	**.049***	**8.42 [1.42, 49.87]**
input vs. test × pre- vs. post-	**2.04 [.46, 3.61]**	**.96**	**2.13**	**.033***	**7.66 [1.59, 36.92]**
input vs. parsing × pre- vs. delayed-	1.12 [-.54, 2.79]	1.01	1.11	.268	3.07 [.58, 16.22]
input vs. test × pre- vs. delayed-	1.54 [-.14, 3.21]	1.02	1.51	.132	4.64 [.87, 24.79]

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$

ORC-A

Table Appx. 184 Fixed effects of model analysis for the task of sentence correction in the metalinguistic knowledge test for ORC-A structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	1.41 [.51, 2.31]	.55	2.59	.010*	4.09 [1.67, 1.03]
input vs. parsing	.51 [-.72, 1.75]	.75	.68	.496	1.67 [.48, 5.75]
input vs. test	-.45 [-1.66, .75]	.73	-.62	.536	.63 [.19, 2.12]
pre- vs. post-	.65 [-.30, 1.61]	.58	1.12	.262	1.92 [.74, 5.01]
pre- vs. delayed-	.16 [-.76, 1.07]	.56	.28	.776	1.17 [.47, 2.93]
input vs. parsing × pre- vs. post-	1.28 [-.29, 2.86]	.96	1.34	.180	3.61 [.75, 17.41]
input vs. test × pre- vs. post-	1.13 [-.28. 2.54]	.86	1.32	.187	3.09 [.76, 12.65]
input vs. parsing × pre- vs. delayed-	2.22 [.52, 3.92]	1.03	2.15	.032*	9.20 [1.68, 5.38]
input vs. test × pre- vs. delayed-	2.30 [.81, 3.79]	.91	2.54	.011*	9.95 [2.24, 44.18]

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when $\alpha \leq .05$

ORC-I

Table Appx. 185 Fixed effects of model analysis for the task of sentence correction in the metalinguistic knowledge test for ORC-I structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	**1.06 [.06, 2.06]**	**.61**	**1.74**	**.082**	**2.88 [1.06, 7.81]**
input vs. parsing	-.45 [-1.67, .78]	.75	-.60	.549	.64 [.19, 2.18]
input vs. test	-1.71 [-2.96, -.46]	.76	-2.26	.024*	.18 [.05, .63]
pre- vs. post-	.73 [-.23, 1.69]	.58	1.26	.209	2.08 [.80, 5.42]
pre- vs. delayed-	.12 [-.79, 1.03]	.55	.22	.830	1.13 [.45, 2.81]
input vs. parsing × pre- vs. post-	1.24 [-.11, 2.60]	.82	1.51	.131	3.46 [.90, 13.40]
input vs. test × pre- vs. post-	1.90 [.52, 3.27]	.84	2.27	.023*	6.66 [1.69, 26.29]
input vs. parsing × pre- vs. delayed-	2.36 [.95, 3.76]	.85	2.77	.006**	1.54 [2.60, 42.80]
input vs. test × pre- vs. delayed-	3.53 [2.01, 5.04	.92	3.83	<.001***	34.01 [7.48, 154.61]

Note: parsing = parsing group; input = input flood group; test = test-only group; bold typeface indicates a reliable effect; * significantly differently from zero when $\alpha \leq .05$; ** significantly different from zero when $\alpha < .01$; *** significantly different from zero when $\alpha<.001$

Task of reason explanation

SRC-A

Table Appx. 186 Fixed effects of model analysis for the task of reason explanation in the metalinguistic knowledge test for SRC-A structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-5.85 [-8.48,-3.21]	1.60	-3.65	<.001***	.00 [.00,.04]
input vs. parising	.04 [-2.6,2.68]	1.61	.03	.980	1.04 [.07,14.61]
input vs. test	-2.65 [-5.74,.45]	1.88	-1.41	.159	.07 [0,1.56]
pre- vs. post-	.22 [-1.35,1.78]	.95	.23	.820	1.24 [.26,5.92]
pre- vs. delayed-	.15 [-1.4,1.7]	.94	.16	.876	1.16 [.25,5.46]
input vs. parising × pre- vs. post-	3.94 [1.78,6.1]	1.31	3.00	.003**	51.51 [5.93,447.8]
input vs. test × pre- vs. post-	1.00 [-1.88,3.87]	1.75	.57	.568	2.71 [.15,48.04]
input vs. parising × pre- vs. delayed-	2.86 [.73,4.98]	1.29	2.21	.027*	17.41 [2.08,145.84]
input vs. test × pre- vs. delayed-	2.67 [-.14,5.48]	1.71	1.57	.117	14.51 [.87,24.98]

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when α ≤ .05; ** significantly different from zero when α < .01; *** significantly different from zero when α<.001

SRC-I

Table Appx. 187 Fixed effects of model analysis for the task of reason explanation in the metalinguistic knowledge test for SRC-I structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-1.14 [-13.48,-6.8]	2.03	-4.99	<.001***	.00 [.00,.00]
input flood vs. parsing	-7.57 [-13.4,-1.74]	3.54	-2.14	.033*	.00 [.00,.18]
input flood vs. test	-.07 [-4.31,4.17]	2.58	-.03	.979	.93 [.01,64.7]
pre- vs. post-	.09 [-2.01,2.19]	1.28	.07	.944	1.09 [.13,8.94]
pre- vs. delayed-	1.34 [-.63,3.32]	1.20	1.12	.263	3.83 [.53,27.65]
input flood vs. parsing × pre- vs. post-	1.8 [5.6,16.01]	3.16	3.42	.001**	4.91E+04 [27.16,8.93E+06]
input flood vs. test × pre- vs. post-	-.99 [-4.21,2.23]	1.96	-.51	.612	.37 [.01,9.27]
input flood vs. parsing × pre- vs. delayed-	8.2 [3.31,13.1]	2.98	2.76	.006**	3658.86 [27.39,4.89E+05]
input flood vs. test × pre- vs. delayed-	-1.23 [-4.15,1.68]	1.77	-.70	.486	.29 [.02,5.36]

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when α ≤ .05; ** significantly different from zero when α < .01; *** significantly different from zero when α<.001

ORC-A

Table Appx. 188 Fixed effects of model analysis for the task of reason explanation in the metalinguistic knowledge test for ORC-A structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*	*p*	OR[CI]
Intercept	-7.78[-11.27,-4.29]	2.12	-3.66	<.001***	.00 [.00,.01]
input vs. parsing	.01 [-3.74,3.77]	2.28	.01	.995	1.02 [.02,43.21]
input vs. test	-.45 [-3.45,2.55]	1.82	-.25	.804	.64 [.03,12.79]
pre- vs. post-	.40 [-1.39,2.18]	1.08	.37	.715	1.49 [.25,8.83]
pre- vs. delayed-	.55 [-1.18,2.28]	1.05	.52	.604	1.73 [.31,9.74]
input vs. parsing × pre- vs. post-	3.33 [.85,5.82]	1.51	2.21	.027*	28.1 [2.34,336]
input vs. test × pre- vs. post-	-.64 [-3.35,2.07]	1.65	-.39	.698	.53 [.03,7.94]
input vs. parsing × pre- vs. delayed-	2.87 [.49,5.26]	1.45	1.99	.047*	17.72 [1.64,192]
input vs. test × pre- vs. delayed-	.49 [-2.06,3.04]	1.55	.32	.751	1.64 [.13,2.92]

Note: parsing = parsing group; input = input flood group; test = test-only group; * significantly differently from zero when $\alpha \leq .05$; *** significantly different from zero when $\alpha<.001$

ORC-I

Table Appx. 189 Fixed effects of model analysis for the task of reason explanation in the metalinguistic knowledge test for ORC-I structure (baseline of the input flood group)

Fixed effects	Estimate [CI]	SE	*z*-value	*p*-value	OR[CI]
Intercept	-5.42 [-7.73,-3.1]	1.41	-3.85	<.001***	.00 [0,.04]
input vs. parsing	-1.6 [-4.35,1.15]	1.67	-.96	.338	.20 [.01,3.14]
input vs. test	-1.4 [-3.93,1.13]	1.54	-.91	.362	.25 [.02,3.09]
delayed- vs. post-	-1.01 [-2.73,.7]	1.04	-.97	.331	.36 [.07,2.02]
delayed- vs. delayed-	-.46 [-2.05,1.13]	.97	-.48	.634	.63 [.13,3.1]
input vs. parsing × delayed- vs. post-	5.8 [3.14,8.45]	1.61	3.59	<.001***	329.37[23.17,4682.74]
input vs. test × delayed- vs. post-	1.01 [-1.55,3.58]	1.56	.65	.516	2.76 [.21,35.87]
input vs. parsing × delayed- vs. delayed-	4.08 [1.61,6.55]	1.50	2.72	.007**	.59 [4.99,697.11]
input vs. test × delayed- vs. delayed-	2.22 [-.12,4.56]	1.42	1.56	.118	9.21 [.89,95.45]

Note: parsing = parsing group; input = input flood group; test = test-only group; ** significantly different from zero when $\alpha < .01$; *** significantly different from zero when $\alpha<.001$

References

Alanen, R. (1995). Input enhancement and rule presentation in second language acquisition. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning and teaching* (Tech. Rep. No. 9; pp. 259–302) Honolulu: University of Hawai'i, Second Language Teaching & Curriculum Center.

Alderson, J. C., Clapham, C., & Steel, D. (1997). Metalinguistic knowledge, language aptitude and language proficiency. *Language teaching research*, 1(2), 93-121.

Allen, L. Q. (2000). Form-meaning connections and the French causative: An experiment in processing instruction. *Studies in Second Language Acquisition*, 69-84.

Altmann, G. T. (2011). Language can mediate eye movement control within 100 milliseconds, regardless of whether there is anything to move the eyes to. *Acta psychologica*, *137*(2), 190-200.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247-264.

Altmann, G. T., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of memory and language*, *57*(4), 502-518.

Andringa, S., & Curcic, M. (2015). How explicit knowledge affects online L2 processing: Evidence from differential object marking acquisition. *Studies in Second Language Acquisition*, *37*(2), 237-268.

Andringa, S., de Glopper, K., & Hacquebord, H. (2011). Effect of explicit and implicit instruction on free written response task performance. *Language Learning*, *61*(3), 868-903.

Audacity Team (2018). Audacity(R): Free Audio Editor and Recorder [Computer application]. Version 2.3.0 retrieved September 29th 2018 from https://audacityteam.org/.

Avery, N., & Marsden, E. J. (2019). A meta-analysis of sensitivity to grammatical information during self-paced reading: Towards a framework of reference for reading time effect sizes. *Studies in Second Language Acquisition*, 1055-1087.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of memory and language*, *59*(4), 457-474.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255-278.

Bartoń, K. (2020). MuMIn: Multi-Model Inference. R package version 1.43.17, https://CRAN.R-project.org/package=MuMIn

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-8, http://CRAN.R-project.org/package=lme4.

Benati, A. (2004). The effects of processing instruction and its components on the acquisition of gender agreement in Italian. *Language Awareness*, *13*(2), 67-80.

Benati, A. (2005). The effects of processing instruction, traditional instruction and meaning—output instruction on the acquisition of the English past simple tense. *Language Teaching Research*, *9*(1), 67-93.

Ben-Shachar, MS., Lüdecke, D., Makowski, D. (2020). Effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, 5(56), 2815. doi: 10.21105/joss.02815, https://doi.org/10.21105/joss.02815.

Brouwer, S., Sprenger, S., & Unsworth, S. (2017). Processing grammatical gender in Dutch: Evidence from eye movements. *Journal of Experimental Child Psychology*, *159*, 50-65.

Brown, J. D. (2014). Classical theory reliability. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1165 – 1181). Oxford, UK: Wiley – Blackwell.

Cadierno, T. (1995). Formal instruction from a processing perspective: An investigation into the Spanish past tense. *The Modern Language Journal*, *79*(2), 179-193.

Chambers, C. G., Tanenhaus, M. K., Eberhard, K. M., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains during real-time language comprehension. *Journal of memory and language*, *47*(1), 30-49.

Chen, B., Ning, A., Bi, H., & Dunlap, S. (2008). Chinese subject-relative clauses are more difficult to process than the object-relative clauses. *Acta Psychologica*, *129*(1), 61-65.

Chen, J., Bowerman, M., Huettig, F., & Majid, A. (2010). Do language-specific categories shape conceptual processing? Mandarin classifier distinctions influence eye gaze behavior, but only during linguistic processing. *Journal of Cognition and*

Culture, *10*(1-2), 39-58.

Clahsen, H., & Felser, C. (2006a). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*(1), 3-42.

Clahsen, H., & Felser, C. (2006b). Continuity and shallow structures in language processing. *Applied Psycholinguistics*, *27*(1), 107-126.

Clahsen, H., & Felser, C. (2018). Some notes on the shallow structure hypothesis. *Studies in Second Language Acquisition*, 40(3), 693-706.

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

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*.

Coumel, M., Ushioda, E., & Messenger, K. (2020). Between- and within-group variation in first and second language syntactic priming. https://doi.org/10.31234/osf.io/wzj7g

Culman, H., Henry, N., & VanPatten, B. (2009). The role of explicit information in instructed SLA: an on-line study with processing instruction and German accusative case inflections. *Die Unterrichtspraxis/Teaching German*, *42*(1), 19-31.

Cunnings, I., Batterham, C., Felser, C., & Clahsen, H. (2010). Constraints on L2 learners' processing of wh-dependencies. *Research in second language processing and parsing*, 87-112.

Cunnings, I., Fotiadou, G., & Tsimpli, I. (2017). Anaphora resolution and reanalysis during L2 sentence processing: Evidence from the visual world paradigm. *Studies in second language acquisition*, *39*(4), 621-652.

Curcic, M., Andringa, S., & Kuiken, F. (2019). The role of awareness and cognitive aptitudes in L2 predictive language processing. *Language Learning*, *69*, 42-71.

Dallas, A., & Kaan, E. (2008). Second language processing of filler-gap dependencies by late learners. *Language and Linguistics Compass*, *2*(3), 372-388.

DeKeyser, R. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge Handbook of Instructed Second Language Acquisition* (1st ed., pp. 15-32). Routledge.

DeKeyser, R. M. (1997). Beyond explicit rule learning: Automatizing second language

morphosyntax. *Studies in second language acquisition*, 195-221.

DeKeyser, R. M., & Sokalski, K. J. (1996). The differential role of comprehension and production practice. *Language Learning*, *46*(4), 613-642.

DeKeyser, R., & Prieto Botana, G. (2015). The effectiveness of processing instruction in L2 grammar acquisition: A narrative review. *Applied Linguistics*, *36*(3), 290-305.

DeKeyser, R.M. (2015). Skill acquisition theory. In: B. VanPatten, & J. Williams (Eds.), Theories in second language acquisition: An introduction (pp. 94–112). 2nd edition. New York: Routledge.

Diessel, H. (2004). *The acquisition of complex sentences* (Vol. 105). Cambridge University Press.

Diessel, H., & Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, *81*(4), 882-906.

Dijkgraaf, A., Hartsuiker, R. J., & Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism: Language and Cognition*, *20*(5), 917-930.

Doughty, C. (1991). Second language instruction does make a difference: Evidence from an empirical study of SL relativization. *Studies in second language acquisition*, *13*(4), 431-469.

Doughty, C. J., & Williams, J. (1998) Pedagogical choices in focus on form. In C. J. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 197–261) New York: Cambridge University Press.

Dracos, M., & Henry, N. (2021). The Role of Task-Essential Training and Working Memory in Offline and Online Morphological Processing. *Languages*, *6*(1), 24.

Dussias, P. E., & Scaltz, T. R. C. (2008). Spanish–English L2 speakers' use of subcategorization bias information in the resolution of temporary ambiguity during second language reading. *Acta psychologica*, *128*(3), 501-513.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied linguistics*, *27*(2), 164-194.

Ellis, R. (2009). Implicit and explicit learning, knowledge and instruction. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, and H. Reinders (Eds): *Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching* (pp. 3-26).

Multilingual Matters.

Felser, C., & Roberts, L. (2007). Processing wh-dependencies in a second language: A cross-modal priming study. *Second Language Research*, *23*(1), 9-36.

Fernández, C. (2008). Reexamining the role of explicit information in processing instruction. *Studies in Second Language Acquisition*, 277-305.

Foucart, A., & Frenck-Mestre, C. (2011). Grammatical gender processing in L2: Electrophysiological evidence of the effect of L1–L2 syntactic similarity. *Bilingualism: Language and Cognition*, *14*(3), 379-399.

Fox, B. A., & Thompson, S. A. (1990). A discourse explanation of the grammar of relative clauses in English conversation. *Language*, 297-316.

Friedmann, N., & Novogrodsky, R. (2004). The acquisition of relative clause comprehension in Hebrew: A study of SLI and normal development. *Journal of Child language*, *31*(3), 661-681.

Friedmann, N., Belletti, A., & Rizzi, L. (2009). Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua*, *119*(1), 67-88.

Frizelle, P., Thompson, P., Duta, M., & Bishop, D. V. (2019). Assessing Children's Understanding of Complex Syntax: A Comparison of Two Methods. *Language Learning*, *69*(2), 255-291.

Fujita, H., & Cunnings, I. (2021). Reanalysis processes in non-native sentence comprehension. *Bilingualism: Language and Cognition*, 1-14.

Gass, S. (1982). From theory to practice. In M. Hynes & W. Rutherford (Eds.), *On TESOL' 81: Selected papers from the Fifteenth Annual Conference of Teachers of English to Speakers of Other Languages* (pp.129-139). Washington, DC: TESOL.

Gennari, S. P., Mirkovic, J., & MacDonald, M. C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive psychology*, *65*(2), 141-176.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1-76.

Godfroid, A. (2019). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide.* Routledge.

Green, P. S., & Hecht, K. (1992). Implicit and explicit grammar: An empirical study. *Applied Linguistics*, *13*(2), 168-184.

Grüter, T., & Rohde, H. (2013). L2 processing is affected by RAGE: Evidence from reference resolution. In *the 12th conference on Generative Approaches to Second Language Acquisition (GASLA)*.

Grüter, T., Lau, E., & Ling, W. (2020). How classifiers facilitate predictive processing in L1 and L2 Chinese: The role of semantic and grammatical cues. *Language, Cognition and Neuroscience*, *35*(2), 221-234.

Grüter, T., Lew-Williams, C., & Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem?. *Second Language Research*, *28*(2), 191-215.

Grüter, T., Zhu, Y. A., & Jackson, C. N (2021). Forcing prediction increases priming and adaptation in second language production. In E. Kaan& T. Grüter (Eds.), *Prediction in second language processing and learning*(pp. 208–231). John Benjamins.

Hamburger, H., Crain, S., (1982). Relative acquisition. In S.A. Kuczaj. (Ed.), *Language Development: Syntax and Semantics* (Vol. II) (pp. 245-274). Erlbaum, Hillsdale, NJ.

Hawkins, J. A. (1999). Processing complexity and filler-gap dependencies across grammars. *Language*, 244-285.

He, W., Xu, N., & Ji, R. (2017). Effects of age and location in Chinese relative clauses processing. *Journal of psycholinguistic research*, *46*(5), 1067-1086.

Henry, N., Culman, H., & VanPatten, B. (2009). More on the effects of explicit information in instructed SLA: A partial replication and a response to Fernández (2008). *Studies in Second Language Acquisition*, 559-575.

Henry, N., Jackson, C. N., & Dimidio, J. (2017). The role of prosody and explicit instruction in processing instruction. *The Modern Language Journal*, *101*(2), 294-314.

Henry, N., Jackson, C. N., & Hopp, H. (2020). Cue coalitions and additivity in predictive processing: The interaction between case and prosody in L2 German. *Second Language Research*.

Henshaw, F. (2012). How effective are affective activities? Relative benefits of two types of structured input activities as part of a computer-delivered lesson on the Spanish subjunctive. *Language Teaching Research*, *16*(3), 393-414.

Hernández, T. (2008). The effect of explicit instruction and input flood on students' use of Spanish discourse markers on a simulated oral proficiency interview. *Hispania*,

665-675.

Hernández, T. A. (2011). Re-examining the role of explicit instruction and input flood on the acquisition of Spanish discourse markers. *Language Teaching Research*, *15*(2), 159-182.

Hernández, T. A. (2018). Input flooding. *The TESOL Encyclopedia of English Language Teaching*, 1-7.

Hopp, H. (2015). Semantics and morphosyntax in predictive L2 sentence processing. *International Review of Applied Linguistics in Language Teaching*, *53*(3), 277-306.

Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Language Research*, *32*(2), 277-307.

Hopp, H., & Lemmerth, N. (2018). Lexical and syntactic congruency in L2 predictive gender processing. *Studies in Second Language Acquisition*, *40*(1), 171-199.

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*(1), 3-27.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, *137*(2), 151-171.

Hulstijn, J. H. (2012). Incidental learning in second language acquisition. *The encyclopedia of applied linguistics*.

Inhoff, A. W., & Radach, R. (1998). Definition and computation of oculomotor measures in the study of cognitive processes. *Eye guidance in reading and scene perception*, 29-53.

Ionin, T., & Montrul, S. (2010). The role of L1 transfer in the interpretation of articles with definite plurals in L2 English. *Language learning*, *60*(4), 877-925.

Issa, B. I., & Morgan-Short, K. (2019). Effects of external and internal attentional manipulations on second language grammar development: An eye-tracking study. *Studies in Second Language Acquisition*, *41*(2), 389-417.

Izumi, S. (2002). Output, input enhancement, and the noticing hypothesis: An experimental study on ESL relativization. *Studies in second language acquisition*, 541-577.

Izumi, S. (2003). Processing difficulty in comprehension and production of relative

clauses by learners of English as a second language. *Language learning*, *53*(2), 285-323.

Jackson, C. N. (2007). The use and non-use of semantic information, word order, and case markings during comprehension by L2 learners of German. *The Modern Language Journal*, *91*(3), 418-432.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57-83.

Jiang, N. (2018). *Second Language Processing: An Introduction*. New York & London: Routledge.

Johnson, M. A., Turk-Browne, N. B., & Goldberg, A. E. (2013). Prediction plays a key role in language development as well as processing. *Behavioral and Brain Sciences*, *36*(4), 360.

Juffs, A. (1998). Main verb versus reduced relative clause ambiguity resolution in L2 sentence processing. *Language Learning*, *48*(1), 107-147.

Juffs, A., & Harrington, M. (1995). Parsing effects in second language sentence processing: Subject and Object Asymmetries in wh-Exaraction. *Studies in second language acquisition*, 483-516.

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different?. *Linguistic Approaches to Bilingualism*, *4*(2), 257-282.

Kamide, Y., Scheepers, C., & Altmann, G. T. (2003). Integration of syntactic and semantic information in predictive processing: Cross-linguistic evidence from German and English. *Journal of psycholinguistic research*, *32*(1), 37-55.

Kasprowicz, R. E., Marsden, E., & Sephton, N. (2019). Investigating distribution of practice effects for the learning of foreign language verb morphology in the young learner classroom. *The Modern Language Journal*, *103*(3), 580-606.

Kasprowicz, R., & Marsden, E. (2018). Towards ecological validity in research into input-based practice: Form spotting can be as beneficial as form-meaning practice. *Applied Linguistics*, *39*(6), 886-911.

Keating, G. D. (2009). Sensitivity to violations of gender agreement in native and nonnative Spanish: An eye-movement investigation. *Language Learning*, *59*(3), 503-535.

Keating, G. D., & Jegerski, J. (2015). Experimental designs in sentence processing research: A methodological review and user's guide. *Studies in Second Language Acquisition*, *37*(1), 1-32.

Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic inquiry*, *8*(1), 63-99.

Kempe, V., & MacWhinney, B. (1998). The acquisition of case marking by adult learners of Russian and German. *Studies in second language acquisition*, 543-587.

Kidd, E., Brandt, S., Lieven, E., & Tomasello, M. (2007). Object relatives made easy: A cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Language and cognitive processes*, *22*(6), 860-897.

Kim, C. E., & O'Grady, W. (2016). Asymmetries in Children's Production of Relative Clauses: Data from English and Korean.

Kim, E., Montrul, S., & Yoon, J. (2015). The on-line processing of binding principles in second language acquisition: Evidence from eye tracking. *Applied Psycholinguistics*, *36*(6), 1317.

Kim, Y., Skalicky, S., & Jung, Y. (2020). The Role of Linguistic Alignment on Question Development in Face-to-Face and Synchronous Computer-Mediated Communication Contexts: A Conceptual Replication Study. *Language Learning*, *70*(3), 643-684.

Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Oxford: Pergamon Press.

Labelle, M. (1990). Predication, wh-movement, and the development of relative clauses. *Language acquisition*, *1*(1), 95-119.

Lee, E. K., Lu, D. H. Y., & Garnsey, S. M. (2013). L1 word order and sensitivity to verb bias in L2 processing. *Bilingualism*, *16*(4), 761.

Lee, S. K., & Huang, H. T. (2008). Visual input enhancement and grammar learning: A meta-analytic review. *Studies in Second Language Acquisition*, 307-331.

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*(3), 193-198.

Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of memory and*

language, *63*(4), 447-464.

Lin, C. J. C., & Bever, T. G. (2006). Subject preference in the processing of relative clauses in Chinese. In *Proceedings of the 25th west coast conference on formal linguistics* (Vol. 25, pp. 254-260). Somerville, MA: Cascadilla Proceedings Project.

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning*, *65*(S1), 185-207.

LoCoco, V. (1987). Learner comprehension of oral and written sentences in German and Spanish: The importance of word order. In B. VanPatten, T. Dvorak & J. F. Lee (Eds.)*, Foreign language learning: A research perspective* (pp. 119-129). Rowley, MA: Newbury House.

Long, M. H. (1983). Does second language instruction make a difference? A review of research. *TESOL quarterly*, *17*(3), 359-382.

Long, M. H. (1983). Does second language instruction make a difference? A review of research. *TESOL quarterly*, *17*(3), 359-382.

Macdonald, R., Brandt, S., Theakston, A., Lieven, E., & Serratrice, L. (2020). The Role of Animacy in Children's Interpretation of Relative Clauses in English: Evidence From Sentence–Picture Matching and Eye Movements. *Cognitive science*, *44*(8), e12874.

MacWhinney, B., Bates, E., & Kliegl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *Journal of verbal learning and verbal behavior*, *23*(2), 127-150.

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within-and between-language competition. *Bilingualism*, *6*(2), 97.

Marinis, T., Roberts, L., Felser, C., & Clahsen, H. (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition*, 53-78.

Marsden, E. (2006). Exploring input processing in the classroom: An experimental comparison of processing instruction and enriched input. *Language Learning*, *56*(3), 507-566.

Marsden, E., & Chen, H. Y. (2011). The roles of structured input activities in processing instruction and the kinds of knowledge they promote. *Language Learning*, *61*(4), 1058-1098.

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*(5),

861-904.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314-324. doi:10.3758/s13428-011-0168-7

Mazerolle, M.J. (2020). AICcmodavg: Model selection and multimodel inference based on (Q)AIC(c). R package version 2.3-1, https://cran.r-project.org/package=AICcmodavg.

McManus, K., & Marsden, E. (2017). L1 explicit instruction can improve L2 online and offline performance. *Studies in Second Language Acquisition*, *39*(3), 459-492.

McManus, K., & Marsden, E. (2018). Online and offline effects of L1 practice in L2 grammar learning: A partial replication. *Studies in Second Language Acquisition*, 459-475.

McManus, K., & Marsden, E. (2019). Signatures of automaticity during practice: Explicit instruction about L1 processing routines can improve L2 grammatical processing. *Applied Psycholinguistics*, *40*(1), 205-234.

Meyer, A. M., Mack, J. E., & Thompson, C. K. (2012). Tracking passive sentence comprehension in agrammatic aphasia. *Journal of Neurolinguistics*, *25*(1), 31-43.

Microsoft Corporation. (2018). *Microsoft Excel*. Retrieved from https://office.microsoft.com/excel

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of memory and language*, *59*(4), 475-494.

Mitsugi, S., & Macwhinney, B. (2016). The use of case marking for predictive processing in second language Japanese. *Bilingualism*, *19*(1), 19.

Morgan-Short, K., Steinhauer, K., Sanz, C., & Ullman, M. T. (2012). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *Journal of cognitive neuroscience*, *24*(4), 933-947.

Nicklin, C., & Plonsky, L. (2020). Outliers in L2 Research in Applied Linguistics: A Synthesis and Data Re-Analysis. *Annual Review of Applied Linguistics*, *40*, 26-55.

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language learning*, *50*(3), 417-528.

O'Grady, W., Lee, M., & Choo, M. (2003). A subject-object asymmetry in the acquisition

of relative clauses in Korean as a second language. *Studies in Second Language Acquisition*, 433-448.

Omaki, A., & Schulz, B. (2011). Filler-gap dependencies and island constraints in second-language sentence processing. *Studies in Second Language Acquisition*, 563-588.

Papadopoulou, D., & Clahsen, H. (2003). Parsing strategies in L1 and L2 sentence processing: A study of relative clause attachment in Greek. *Studies in Second Language Acquisition*, 501-528.

Phillips, C., & Ehrenhofer, L. (2015). The role of language processing in language acquisition. *Linguistic approaches to bilingualism*, *5*(4), 409-453.

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal*, *100*(2), 538-553.

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, *102*(4), 713-731.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912.

Pu, M. M. (2007). The distribution of relative clauses in Chinese discourse. *Discourse Processes*, *43*(1), 25-53.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you!. *Educational and psychological measurement*, 79(1), 200-210.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. *Journal of memory and language*, *57*(1), 1-23.

Renou, J. M. (2000). Learner accuracy and learner performance: The quest for a link. *Foreign Language Annals*, *33*(2), 168-180.

Roberts, L. (2016). Self-paced reading and L2 grammatical processing. In A. Mackey & E. Marsden (Eds.), *Advancing Methodology and Practice* (pp. 70-84). Routledge.

Roberts, L., & Felser, C. (2011). Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics*, 299-331.

Roberts, L., & Liszka, S. A. (2013). Processing tense/aspect-agreement violations on-line in the second language: A self-paced reading study with French and German L2 learners of English. *Second Language Research*, *29*(4), 413-439.
Roberts, L., & Liszka, S. A. (2013). Processing tense/aspect-agreement violations on-line in the second language: A self-paced reading study with French and German L2 learners of English. *Second Language Research*, *29*(4), 413-439.
Robinson, P. (1995). Aptitude, awareness, and the fundamental similarity of implicit and explicit second language learning. In R. Schmidt (Ed.) *Attention and awareness in foreign language learning* (pp.303-358). Honolulu: University of Hawai’i at Manoa.

Roehr, K. (2006). Metalinguistic knowledge in L2 task performance: A verbal protocol analysis. Language Awareness, 15(3), 180-198.

Roehr, K. (2008). Metalinguistic knowledge and language ability in university-level L2 learners. *Applied Linguistics*, *29*(2), 173-199.

Roehr-Brackin, K. (2014). Explicit knowledge and processes from a usage-based perspective: The developmental trajectory of an instructed L2 learner. *Language Learning*, *64*(4), 771-808.

Sanz, C., & Morgan-Short, K. (2004). Positive evidence versus explicit rule presentation and explicit negative feedback: A computer-assisted study. *Language learning*, *54*(1), 35-78.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied linguistics*, *11*(2), 129-158.

Schwartz, B. D. (1998). The second language instinct. *Lingua*, *106*(1-4), 133-160.

Sekerina, I. A., & Sauermann, A. (2015). Visual attention and quantifier-spreading in heritage Russian bilinguals. *Second language research*, *31*(1), 75-104.

Sekerina, I. A., & Trueswell, J. C. (2011). Processing of contrastiveness by heritage Russian bilinguals. *Bilingualism: Language and cognition*, *14*(3), 280-300.

Sheldon, A. (1974). The role of parallel function in the acquisition of relative clauses in English. *Journal of verbal learning and verbal behavior*, *13*(3), 272-281.

Staub, A., & Clifton Jr, C. (2006). Syntactic prediction in language comprehension: evidence from either... or. *Journal of experimental psychology: Learning, memory, and cognition*, *32*(2), 425.

Su, Y. C., Lee, S. E., & Chung, Y. M. (2007). Asyntactic thematic role assignment by Mandarin aphasics: A test of the Trace-Deletion Hypothesis and the Double Dependency Hypothesis. *Brain and Language*, *101*(1), 1-18.

Tallerman, M. (2014). *Understanding syntax*. Routledge.

Tellier, A., & Roehr-Brackin, K. (2013). The development of language learning aptitude and metalinguistic awareness in primary-school children: A classroom study. UNSPECIFIED. Essex Research Reports in Linguistics, University of Essex, Colchester, UK.

Thompson-Lee, S. (2021). *Teaching learners to process morphosyntactic cues: The passive voice in second language English*. (Unpublished doctoral dissertation). The University of York.

Tolentino, L. C., & Tokowicz, N. (2014). Cross-language similarity modulates effectiveness of second language grammar instruction. *Language Learning*, *64*(2), 279-309.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of memory and language*, *47*(1), 69-90.

Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language*, *53*(2), 204-224

Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(3), 559-599.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443.

VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition*. Greenwood Publishing Group.

VanPatten, B. (2002). Processing instruction: An update. *Language learning*, *52*(4), 755-803.

VanPatten, B. (2004). Input processing in SLA. In B. VanPatten (Ed.), *Processing*

instruction: Theory, research, and commentary (pp. 5-32). New Jersey: Lawrence Erlbaum Associates, Inc.

VanPatten, B. (2005). Processing instruction. In C. Sanz (Ed.), *Mind and context in adult second language acquisition* (pp. 267–81). Washington, DC: Georgetown University Press.

VanPatten, B. (2015). Foundations of processing instruction. *International Review of Applied Linguistics in Language Teaching*, *53*(2), 91-109.

VanPatten, B. (2020). Input processing in adult second language acquisition. In B. VanPatten, G. D. Keating & S. Wulff (Eds.), *Theories in second language acquisition: An introduction* (3rd. ed.) (pp. 105-127). Milton: Routledge.

VanPatten, B., & Borst, S. (2012). The roles of explicit information and grammatical sensitivity in processing instruction: Nominative-accusative case marking and word order in German L2. *Foreign Language Annals*, *45*(1), 92-109.

VanPatten, B., & Cadierno, T. (1993a). Explicit instruction and input processing. *Studies in second language acquisition*, 225-243.

VanPatten, B., & Cadierno, T. (1993b). Input processing and second language acquisition: A role for instruction. *The Modern Language Journal*, *77*(1), 45-57.

VanPatten, B., & Jegerski, Jill. (2010). *Research in second language processing and parsing*. Amsterdam: John Benjamins Pub.

VanPatten, B., & Oikkenon, S. (1996). Explanation versus structured input in processing instruction. *Studies in Second Language Acquisition*, 495-510.

VanPatten, B., & Smith, M. (2019). Word-order typology and the acquisition of case marking: A self-paced reading study in Latin as a second language. *Second Language Research*, *35*(3), 397-420.

VanPatten, B., & Wong, W. (2004). Processing instruction and the French causative: Another replication. In B. VanPatten (Ed.), *Processing instruction: Theory, research, and commentary* (pp. 97-118). New Jersey: Lawrence Erlbaum Associates, Inc.

Weber, A., & Paris, G. (2004). The origin of the linguistic gender effect in spoken-word recognition: Evidence from non-native listening. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 26, No. 26).

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Williams, J. N. (2006). Incremental interpretation in second language sentence processing. *BILINGUALISM LANGUAGE AND COGNITION*, *9*(1), 71.

Williams, J. N., Mobius, P., & Kim, C. (2001). Native and non-native processing of English wh-questions: Parsing strategies and plausibility constraints. *Applied Psycholinguistics*, *22*(4), 509-540.

Winke, P. M., Godfroid, A., & Gass, S. M. (2013). Introduction to the special issue: Eye-movement recordings in second language research. *Studies in Second Language Acquisition*, *35*(2), 205-212.

Witzel, J., Witzel, N., & Nicol, J. (2012). Deeper than shallow: Evidence for structure-based parsing biases in second-language sentence processing. *Applied Psycholinguistics*, *33*(2), 419-456.

Wong, W., & Ito, K. (2018). The effects of processing instruction and traditional instruction on L2 online processing of the causative construction in French: An eye-tracking study. *Studies in Second Language Acquisition*, *40*(2), 241-268.

Zeileis, A., Hothorn, T. (2002). "Diagnostic Checking in Regression Relationships." *R News*, **2**(3), 7–10. https://CRAN.R-project.org/doc/Rnews/.

Zukowski, A. (2009). Elicited production of relative clauses in children with Williams syndrome. *Language and Cognitive Processes*, *24*(1), 1-43.