2

SELF-PACED READING

Jill Jegerski

History of the Method

The self-paced reading (SPR) method was invented by psycholinguists in the 1970s (Aaronson & Scarborough, 1976; Mitchell & Green, 1978). SPR is so simple in design that it would be easy to assume today that it predates modern eye-tracking's appearance in reading research, but in reality the two methods in their more primitive forms appeared at around the same time, when a newfound access to computers fostered some of the most significant advances in mental chronometry since its development just over a century prior (Donders, 1868/1969, as cited in Baayen & Milin, 2010). These game-changing developments in methods for psycholinguistic research arose out of a desire in cognitive psychology to measure language comprehension processes in real time and with "tasks that are as similar as possible to normal reading" (Mitchell & Green, 1978, p. 610). Self-paced reading was the simplest way to meet these goals using modern technology and for this reason it flourished in popularity and has persisted over time—unlike many of its predecessors, including click migration (Fodor & Bever, 1965), the phoneme-monitoring task (Foss, 1970), and the sentence classification task (Forster & Olbrei, 1973). Nearly forty years after its development, SPR is still the most fundamental experimental measure employed by psycholinguists interested in processing at or above the level of the sentence. SPR was also the first on-line (i.e., real-time) method to be applied in non-native sentence processing research.

The first published investigation to apply the SPR method in the study of second language acquisition (SLA) was Juffs and Harrington (1995). The immediate theoretical motivation for the study was the debate among generative linguists in SLA as to whether observed differences between native speakers and adult second

language (L2) learners were true differences in underlying grammatical competence, perhaps due to a lack of access to Universal Grammar after a critical period, or more superficial differences that arose due to the time constraints of real-time processing and were thus limited to performance. Proponents of the performance position had proposed, based on previous evidence collected via grammaticality judgments and global reaction times, that divergent behavior among non-natives could be due to processing difficulty rather than to the acquisition of a nontarget grammar (see, e.g., White & Juffs, 1998, for a more detailed discussion of the competence versus performance debate). Thus, the initial motivation for employing the SPR method in second language research was to measure linguistic performance in a way that complemented the grammaticality judgment as a measure of linguistic competence.

A decade later, when Clahsen and Felser (2006) reviewed the literature on second language processing in their development of the Shallow Structure Hypothesis, interest in psycholinguistic methods in SLA research had begun to take hold and over a dozen published studies using the self-paced reading method were available. Several of these, like Juffs and Harrington (1995), followed the generative linguistics tradition in SLA and focused on *wh-* movement (Juffs, 2005; Marinis, Roberts, Felser, & Clahsen, 2005; Williams, Möbius, & Kim, 2001) or clitics and causatives (Hoover & Dwivedi, 1998). Other researchers began to incorporate and adapt ambiguity and anomaly paradigms from the psycholinguistics tradition, such as relative clause attachment (Dussias, 2003; Felser, Roberts, Gross, & Marinis, 2003; Papadopoulou & Clahsen, 2003), subject-object ambiguity (Juffs, 1998a, 2004; Juffs & Harrington, 1996), verbal ambiguity (Juffs, 1998b), and broken agreement (Jiang, 2004). By 2009, the study of L2 processing had flourished and SPR was the single most popular on-line method among researchers at the first Conference on Second Language Processing and Parsing held at Texas Tech University, accounting for 37% of all research papers presented at the conference (see VanPatten & Jegerski, 2010, for select examples). At the same time, dozens of additional journal articles have reported self-paced reading studies of SLA, to the extent that they have become almost too numerous to be covered comprehensively and current literature reviews tend to need to be much more narrow in focus.

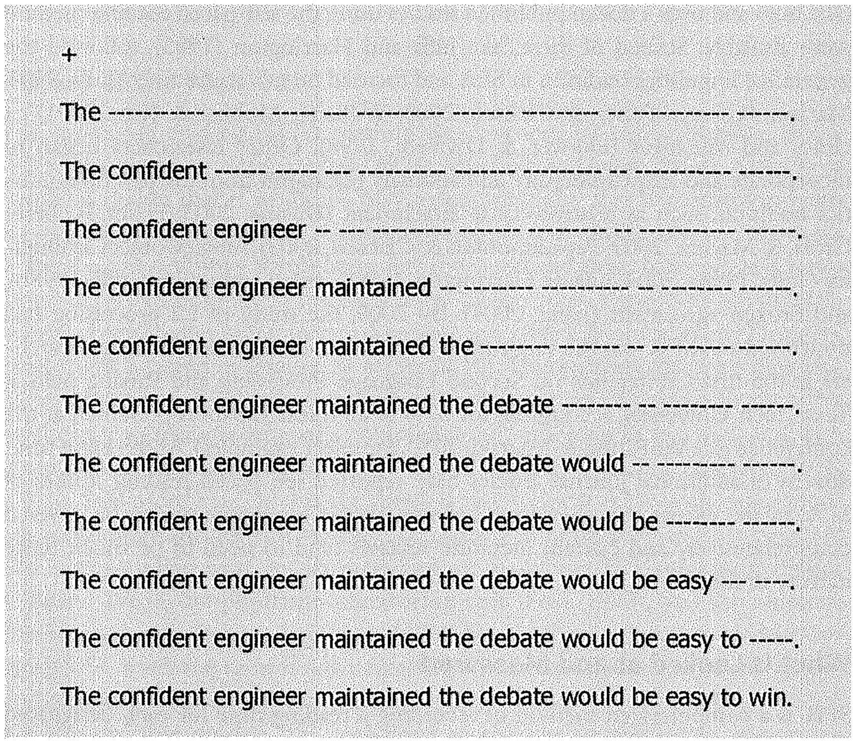
What is Looked at and Measured

SPR is a computerized method of recording a reading time for each designated segment (i.e., a word or phrase) of a sentence or series of sentences that is presented as an experimental stimulus. It is commonly referred to as *self-paced* and has also been called *subject-paced* because the research participant determines how long to spend reading each segment, which contrasts with fixed-pace methods like rapid serial visual presentation, or RSVP, where reading times are predetermined by the researcher. Specifically, in SPR, a button press causes the first segment of a sentence to appear together with a series of dashes masking the remainder of the stimulus, then when the participant is ready to continue a second button press reveals the next segment, then the next, and so on until the entire sentence has been read. Historically, self-paced reading has been a general term that includes several different formats. First, the display can be *cumulative,* meaning once a stimulus segment is revealed it remains visible to the participant as the next segment is revealed and the next and so on, until the entire sentence is finally displayed all together (as illustrated in Figure 2.1), or *noncumulative,* meaning only one segment is visible at a time and every time a new segment is revealed the previous one is remasked (as illustrated in Figure 2.2). Additionally, the display can be *centered,* meaning that every segment appears in the center of the display screen and overwrites the previous segment (as seen in Figure 2.3), or *linear,* meaning that segments appear in linear succession from left to right with no spatial overlap, much as they would in normal reading. However, the cumulative display is problematic because most participants develop a reading strategy in which they reveal several segments of a stimulus at a time before reading them all at once (Ferreira & Henderson, 1990; Just, Carpenter, & Wooley, 1982) and the centered display is avoided with SPR because it is less like normal reading—though it can be necessary with some other methods, like ERPs (for further information regarding the ERP method, see Morgan-Short & Tanner, Chapter 6, this volume). For these reasons, virtually all SPR studies now elect for a noncumulative linear display, which is also referred to as the *moving window(s)*[1] technique because successive button presses cause the unmasked segment of text to proceed like a moving window across the computer screen.

kumulativní
nekumulativní
×
centrovaný
×
lineární

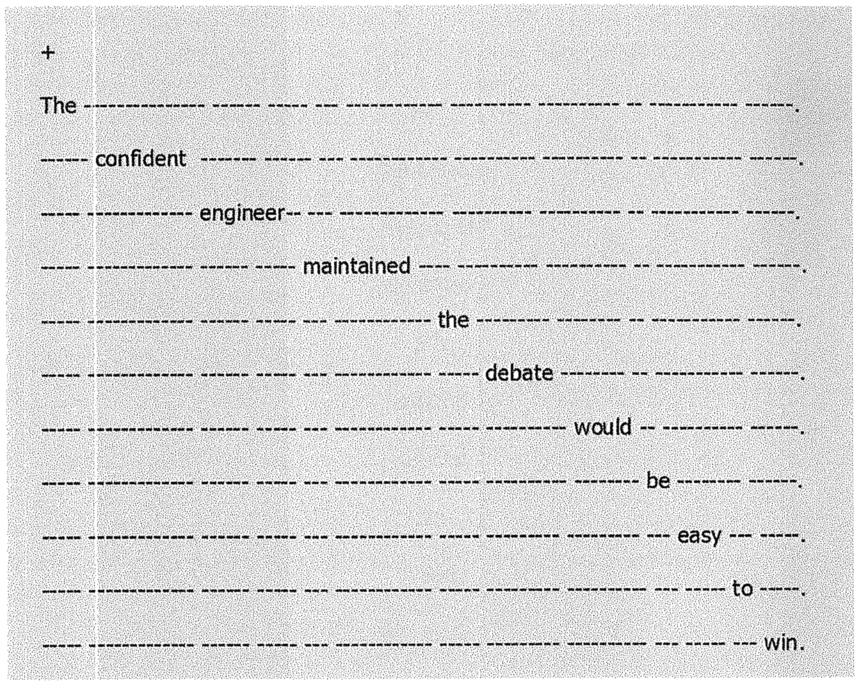
proti kumulativní a centrovaný

moving windows



```
+
The ---------- ------ ----- --- ---------- ------- --------- -- ----------- ------.
The confident ------ ----- --- ---------- ------- --------- -- ----------- ------.
The confident engineer -- --- --------- ------- --------- -- ----------- ------.
The confident engineer maintained -- -------- ---------- -- -------- ------.
The confident engineer maintained the ------- ------- -- --------- ------.
The confident engineer maintained the debate -------- -- --------- ------.
The confident engineer maintained the debate would -- --------- ------.
The confident engineer maintained the debate would be -------- ------.
The confident engineer maintained the debate would be easy --- ----.
The confident engineer maintained the debate would be easy to -----.
The confident engineer maintained the debate would be easy to win.
```

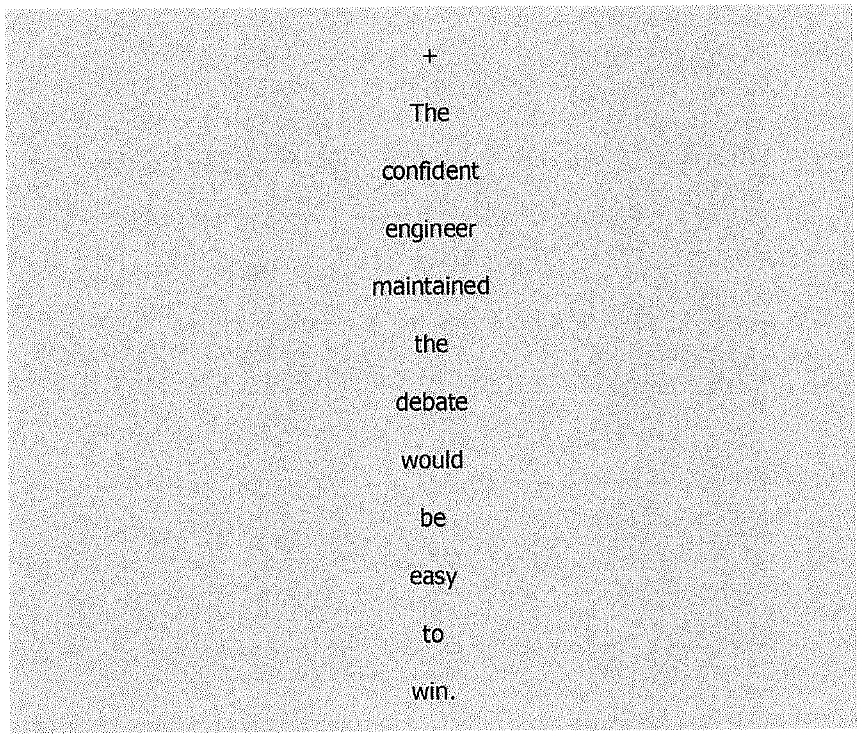
FIGURE 2.1 Illustration of self-paced reading, cumulative linear format with word-by-word segmentation. Each line of text shown above would be vertically centered in a separate display on a computer monitor. The participant presses a button to move through successive displays and computer software records the time between button presses, which is the primary dependent variable. A gray background is used because it is easier on the eyes than a white background.



```
+
The ------------- ------ ----- --- ----------- --------- --------- -- ----------- ------.
----- confident ------ ----- --- ----------- --------- --------- -- ----------- ------.
----- ------------- engineer-- --- ----------- --------- --------- -- ----------- ------.
----- ------------- ------ ----- maintained --- --------- --------- -- ----------- ------.
----- ------------- ------ ----- --- ----------- the -------- -------- -- ----------- ------.
----- ------------- ------ ----- --- ----------- ----- debate -------- -- ----------- ------.
----- ------------- ------ ----- --- ----------- ----- --------- would -- ----------- ------.
----- ------------- ------ ----- --- ----------- ----- --------- -------- be -------- ------.
----- ------------- ------ ----- --- ----------- ----- --------- --------- --- easy --- -----.
----- ------------- ------ ----- --- ----------- ----- --------- --------- --- ----- to -----.
----- ------------- ------ ----- --- ----------- ----- --------- --------- --- ----- --- win.
```

FIGURE 2.2 Illustration of self-paced reading, non-cumulative linear format with word-by-word segmentation. Also known as the moving window format, this is the most common type of self-paced reading.

The basic premise behind self-paced reading is that the eyes can be a window on cognition. Just and Carpenter (1980) proposed the eye-mind assumption, which states that the amount of time taken to read a word reflects the amount of time needed to process the word. While subsequent research has revealed that the connection between reading times and processing is in reality more complex, the basic assumption still holds in the broad sense and reading time data, as a specific class of reaction times (i.e., response times or response latencies), are interpreted with



+

The

confident

engineer

maintained

the

debate

would

be

easy

to

win.

FIGURE 2.3 Illustration of self-paced reading, centered format with word-by-word segmentation.

RTs

the goal of drawing inferences about the cognitive processing of language. Specifically, relatively longer reading times are taken as indications of processing difficulty, while faster reading times are interpreted as a sign that facilitation occurred.

tracing diffic[illegible]

Most SPR paradigms examine processing difficulties that arise during the reading of sentences that contain what could be classified as an ambiguity, an anomaly, or a distance dependency. Ambiguities arise where the grammar permits two or more distinct syntactic interpretations of a word or phrase in the sentence and observable processing strategy often occurs when the (native) parser tends towards one interpretation over the other. Such structural ambiguity can be either local, meaning it occurs temporarily during reading but is resolved within the same sentence, or global, meaning that even after the whole sentence has been read the ambiguity remains. Examples of local syntactic ambiguities include subject-object ambiguity (L1: Trueswell & Kim, 1998; L2: Juffs & Harrington, 1996) and reduced relative clause ambiguity (L1: MacDonald, 1994; L2: Juffs, 1998a). Local or temporary ambiguities are also referred to as *garden path* phenomena because such sentences are designed to initially lead the reader in the wrong direction with regard to the structure of the sentence. Garden path effects are evident in increased SPR times at or after the point in the sentence where it becomes evident to the reader that the initial interpretation was incorrect. In the example (1) below, taken from Trueswell and Kim (1998), longer reading times were observed on the embedded verb *would be* in the ambiguous version in (1a) versus in the unambiguous version in (1b).

(1) *Subject-Object Ambiguity*

a. The confident engineer maintained the debate would be easy to win.
b. The confident engineer maintained that the debate would be easy to win.

With global ambiguities, such as the attachment of ambiguous relative clauses (L1: Cuetos & Mitchell, 1988; L2: Dussias, 2003) or prepositional phrases (L1: Taraban & McClelland, 1988; L2: Pan & Felser, 2011), on the other hand, SPR effects are usually evident around the point where disambiguated versions of the stimuli become inconsistent with participants' preferred interpretation. For instance, in the example (2) below, a stimulus item from Dussias (2003), native speakers of Spanish exhibited longer reading times on the sentence-final phrase *con su esposo* "with her husband" in the forced low-attachment version in (2a) versus the forced high-attachment version in (2b), because Spanish in general tends toward high attachment for ambiguous relative clauses. Such disambiguation can be accomplished using pragmatic-contextual information, as in this example, or with grammatical dependencies like gender or number agreement, though each of these adds a layer of complexity in processing and thus has the potential to obscure reading time effects caused by the experimental manipulation, especially among non-native readers. Stimuli with local or global syntactic ambiguities are intended to present processing difficulty in the form of forced syntactic reanalysis at the point of disambiguation, but they remain grammatical in all versions.

(2) *Relative Clause Ambiguity (Disambiguated)*

a. El perro mordió al cuñado de la maestra que vivió en Chile con su esposo.

"The dog bit the brother-in-law of the (female) teacher who lived in Chile with her husband."

b. El perro mordió a la cuñada del maestro que vivió en Chile con su esposo.

"The dog bit the sister-in-law of the (male) teacher who lived in Chile with her husband."

anomalies

A second class of processing phenomena targeted with SPR are anomalies, which include specific violations of grammar (i.e., error recognition or grammaticality paradigms) as well as inconsistent or noncanonical permutations of word order, semantics, discourse, and other syntactic and extrasyntactic factors that are presented in the experimental stimuli. Numerous different anomaly paradigms are available to researchers in psycholinguistics, but some examples of specific phenomena that have been examined among non-natives using SPR include gender agreement (Sagarra & Herschensohn, 2011), number agreement (Foote, 2011), tense/aspect agreement (Roberts, 2009), and case marking (Jegerski, 2012a). Stimuli with violations of the linguistic principles guiding such phenomena commonly induce longer reading times at or after the point of the violation, presumably because the parser has difficulty incorporating a word that does not fit with the existing representation of the sentence. For instance, in sentences like (3) below, a stimulus from Foote (2011), reading times on the postverbal word *de* "from" were longer in the ungrammatical stimulus condition in (3b) versus the grammatical version in (3a), as can be seen in the reading times in Table 2.1, which are also graphed in Figure 2.4 (along with hypothetical data for the other sentence regions not in Table 2.1).

(3) *Number Agreement Violation*

a. Veo que tu padre es de Texas.
b. *Veo que tu padre son de Texas.

"I see that your father is/*are from Texas."

A third type of sentence-level phenomenon that can be exploited to study processing behavior with SPR is the distance dependency. Dependency paradigms examine the computation or recognition of a syntactic relationship between two elements in the stimulus that are usually nonadjacent in the lin ear word order, which presents a particular challenge in processing. Examples include *wh-* movement (L1: Crain & Fodor, 1985; L2: Williams, Möbius, & Kim, 2001) and broken agreement (L1: Nicol, Forster, & Veres, 1997; L2: Jiang, 2004).

TABLE 2.1 Reading times in milliseconds from an SPR experiment using a grammar violation paradigm (Adapted from Foote, 2011)

Veo que tu "I see that your	*padre* father		*es/son* is/are		*de . . .* from . . ."	
	Mean	SD	Mean	SD	Mean	SD
L1 Group						
Ungrammatical	389	66	452	94	488	109
Grammatical	414	86	409	52	403	75
Difference	−25		43		85	
L2 Group						
Ungrammatical	360	29	408	43	539	83
Grammatical	373	43	394	47	379	26
Difference	−13		14		160	

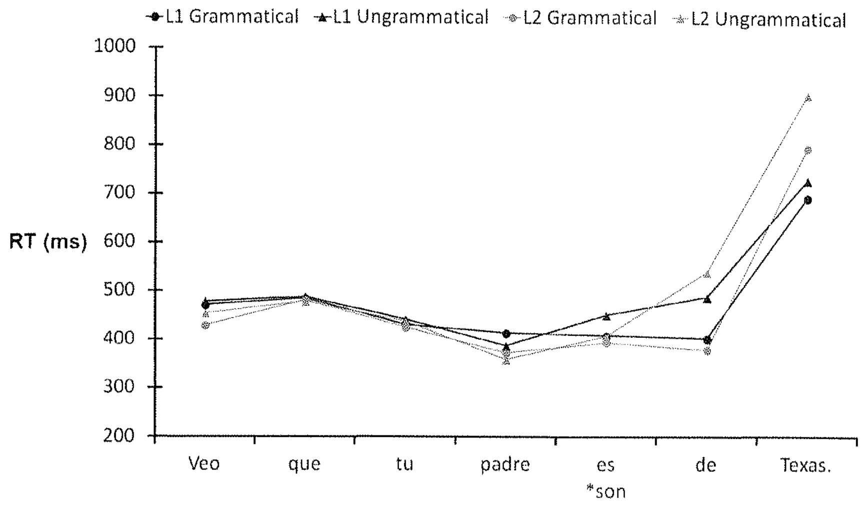


FIGURE 2.4 Example line graph of hypothetical reading time data from a SPR experiment using a grammar violation paradigm (based on a stimulus and partial data from Foote, 2011).

To illustrate, the examples of broken agreement in (4) below, taken from Jiang (2004), require the parser to compute subject-verb number agreement across several intervening words, the number feature of which can interfere with the computation. Thus, relatively longer reading times were observed on the verb in the Singular-Plural-Singular condition in (4a) versus in the Singular-Singular-Singular condition in (4b), presumably because of interference from the plural noun *cabinets* in (4a). As with syntactic ambiguities, distance dependencies can induce SPR effects without the introduction of grammar violations or other anomalies, manipulating instead the words and phrases that intervene between the two dependent elements.

(4) *Broken Agreement*

a. The key to the cabinets was kept in the main office.
b. The key to the cabinet was kept in the main office.

Over a period of nearly thirty-five years, all three types of SPR paradigms have been employed to study fundamental questions in native language sentence processing such as whether the parser considers multiple plausible analyses simultaneously or sequentially, whether all types or modules of linguistic information are immediately available or only syntax is active at first, what heuristics motivate different processing preferences, and to what extent these basic principles vary cross-linguistically, among others. Non-native sentence processing research is a relatively newer area of study that can be uniquely informative with regard to these pre-existing broad questions in psycholinguistics, and which has also begun to articulate its own research agenda within the field of second language study. SPR investigations have focused on the issue of learnability and age effects in processing, on the closely related debate as to whether divergence in adult SLA is rooted in competence or performance, and on the question of L1 transfer in processing, so far with relatively less attention dedicated to other L2 questions like mapping the developmental trajectory of non-native processing behavior. Thus, in most cases the SPR method has been employed to measure linguistic skill and knowledge for the purpose of making comparisons, either between native and non-native processing in the L2, between native processing in the L1 and non-native processing in the L2, or between the L2 processing behaviors of participant groups with different native languages.

Comparison on the basis of SPR data can be designed and interpreted from at least two different perspectives. First, because grammatical processing relies on existing knowledge of grammar that is stored in memory, the SPR method in L2 research was first viewed as complementary to previously established measures like grammaticality and acceptability judgments. From this angle, SPR data can be seen as an indirect measure of grammatical competence and are often regarded as a relatively more direct or more implicit measure of grammar than off-line judgments because the time constraints of on-line processing presumably allow less room for the application of explicit grammar rules. The most common SPR paradigms employed in this vein of research target grammar violations or anomalies and distance dependencies, both of which can be linked to the formal linguistics traditions of grammaticality judgments with relative ease. Sensitivity to an experimental manipulation of grammar, in the form of increased reading times at or near the site of a violation, is interpreted as evidence that the relevant underlying grammatical competence has been acquired. This is of course assuming that such sensitivity is also evident among a comparison group of native speakers and can therefore be reasonably expected, given that even violation-based reading time effects—which tend to be more robust and more reliable than those that occur with dependencies or ambiguities—can sometimes be inconsistent among native speakers.

Second, the SPR method can be used as a measure of performance or processing behavior itself, a perspective that is becoming dominant as the study of L2 processing expands and the body of existing published research grows. A variety of reading time effects are targeted in this line of investigation, which includes ambiguities as well as distance dependencies and anomalies. The interpretation of data can be considerably less straightforward than when SPR is employed as an indirect measure of grammatical competence, especially when the method is used to compare native and non-native processing. To illustrate, data interpretation is fairly straightforward when a group of native readers exhibits a reading time effect that is not at all evident among a group of non-native participants, as most researchers would agree that such an outcome indicates a difference between native and non-native processing. There are occasions, however, where a group of native readers shows an SPR effect that is even more pronounced among the non-native readers, meaning that the effect is sustained over more than one region of interest or it surfaces again during sentence wrap-up or while answering a poststimulus distractor question. Particularly if the SPR effect in question is presumed to signal syntactic reanalysis, a more pronounced effect among non-native readers could be interpreted as a sign of additional processing difficulty rather than native-like processing skill. Another experimental outcome that can be subject to multiple interpretations is when non-native participants display a reading time effect that occurs a region or two later than that exhibited by the native readers, or perhaps does not surface until wrap-up occurs at the last region of the stimulus. In both of these scenarios, there is some room for debate as to whether the observed differences between native and non-native processing are critical, meaning whether they represent qualitative or merely quantitative differences.

In general, reading time data from SPR experiments are more nuanced and thus tend to demand more complex interpretation than data from off-line measures like grammaticality judgments. In addition, the use of SPR to measure target L2 behavior is also by nature paradoxical, because most known L1 reading time effects are assumed to indicate some type of processing difficulty. In other words, we are investigating whether non-native readers have learned to have the same problems that native readers have while processing a given type of stimulus. In some cases, the interpretation of L2 SPR data can be relatively straightforward, but it is not always clear whether increased reading times among L2 learners reflect the target native-like processing difficulty induced by experimental manipulation of the stimuli or a different type of difficulty that has to do with the limitations of L2 processing. In the former scenario, increased SPR reading times would be interpreted as evidence of native-like processing *strategy*, whereas in the latter they would be taken as evidence of an L2-specific processing *struggle*.

Issues in the Development and Presentation of Stimuli

The creation of an SPR experiment entails the embedding of each stimulus within a trial, or experimental series of related events, which usually consists of three components: a cue, a stimulus, and a distractor. The cue phase is fairly straightforward and is the same for all trials; a "+"or similar symbol is displayed in isolation in the same screen location where the first letter of the first word in the stimulus will subsequently appear. Hence, the cue appears towards the left side of the display for a language that is read left-to-right and towards the right side for a language that is read right-to-left. The purpose of the cue is to encourage participants to direct their gaze at the location of the first word of the stimulus before it appears. Otherwise, the reading time for the first region of interest of a given stimulus may or may not include time spent initially dwelling on another screen location plus the time spent to bring the gaze to the location of the beginning of the stimulus. This would make SPR less precise as an experimental measure because, at a minimum, reading times for the sentence-initial region would be artificially but uniformly inflated, while a more likely scenario is that reading times would be affected inconsistently and therefore display unnecessarily high levels of variance.

After the cue comes the stimulus. SPR stimuli are usually one sentence in length, given that the method is most often used to measure sentence-level comprehension behavior, although discourse-level phenomena can be studied through the addition of one or more sentences that establish a context prior to the appearance of the target sentence. In either case, the development of stimuli entails the creation of a list of experimental sentences or *items,* which are directed at the research questions guiding the investigation. Within each experimental item, there are multiple versions (usually two to four) referred to as *conditions,* which correspond to the researcher's manipulation of independent linguistic variables and are thus determined by the experimental design. In order to maintain control between stimulus conditions, the corresponding regions to be directly compared in the statistical analysis should be as near to identical as possible, given the constraints of the particular experiment, as in (4) above. If the experimental manipulation of linguistic variables necessitates the use of different lexical items in different stimulus conditions (e.g., if the manipulation involves lexical specifications like verb transitivity or semantic properties), then these should be counterbalanced across stimulus conditions, minimally for objective variables like length in characters and syllables, frequency, and similarity to the equivalent L1 lexical item (including cognate status), and then verified via statistical comparisons, because such variables are known to affect reading speed. Depending on the type of stimuli and the objectives of the experiment, it may also be necessary to counterbalance the different lexical items (or phrases) to be directly compared with regard to subjective variables such as imageability, comprehensibility, concreteness, or semantic properties. As such variables are not easily measured via objective means, they are measured instead via a norming procedure, in which a group of research subjects (native speakers) are asked to rate a list of words or phrases with regard to one or more characteristics. The ratings are then compiled and analyzed and used to create counterbalanced experimental stimuli (see Sunderman, Chapter 8, this volume, for a more detailed discussion of norming procedures).

items
conditions

use
of
different
lexical
items

Once the experimental sentences have been created, they are broken down into regions of interest, which participants will read one-by-one and will each correspond to a separate data point in the form of a reading time in milliseconds (ms). The researcher chooses whether to use word-by-word or phrase-by-phrase segmentation, both of which are illustrated in (5) and (6) below with examples from Juffs (1998b) and Pliatsikas and Marinis (2013), respectively. The decision between the two types of segmentation is typically a compromise between the conflicting goals of maximizing the level of detail in the reading time data and maximizing the ecological validity of the experimental task. In other words, a word-by-word segmentation yields more precise data because more data points are collected per stimulus, whereas a phrase-by-phrase segmentation is closer to normal reading and may therefore eliminate some unnatural effects induced by the SPR task itself, such as a tendency towards highly incremental processing. On the other hand, data collected in a word-by-word fashion can be easily converted to phrase-by-phrase mode by summing reading times across multiple words/regions, but once an experiment has been run in the phrase-by-phrase mode, there is no way to break the data down into word-by-word reading times without rerunning the experiment. The phrase-by-phrase mode also has the added complication of potentially influencing processing behavior through the particular grouping of words into phrases and especially the length of the phrases (Gilboy & Sopena, 1996). One way to determine the effect of stimulus segmentation on a given set of materials would be to run a pilot experiment testing two or more segmentations. Both the word-by-word and the phrase-by-phrase segmentation are common in the literature and with both modes the number of total regions per stimulus depends on the length of the stimuli, but should be exactly the same for all items in a given experiment.

region
of
interest
word-
word
x
phrase

yield
a
might

pilot

(5) *Word-by-word segmentation*

Before / Mary / ate / the / pizza / arrived / from / the / local / restaurant.

(6) *Phrase-by-phrase segmentation*

The manager who / the secretary claimed / that / the new salesman / had pleased / will raise company salaries.

Within regions of interest, be they words or phrases, it is important that there be grammatical equivalency across experimental items, such that if Region 1 is a subject noun phrase (NP) in Item 1, then it should also be a subject NP for Item 2, Item 3, and all the other stimulus items. Each region of interest should also be roughly equivalent in length across different stimulus items (this is not to be confused with the control of regions across conditions of a single item, which should be identical except for the critical region).

grammatical
equivalence

The number of stimuli per condition is usually eight to twelve, which means that the number of sentences created for an experiment can range from 16 (2 conditions × 8 stimuli per condition) to 48 (4 conditions × 12 stimuli per condition). Because these target stimuli represent only 25 to 35% of the total experiment and even non-natives at a very high level of L2 proficiency are not commonly asked to read more than 150 to 200 sentences maximum per research session, individual SPR experiments rarely include more than 48 target stimuli. If the research questions driving a given investigation necessitate the inclusion of so many variables that the total number of target stimuli would be higher, then the variables can be broken down into separate, smaller experiments. Such simplicity has the added

stimuli
~
condition

benefits of avoiding complexity in the statistical analyses that can render them largely uninterpretable and also keeping the required number of participants to a reasonable number, as will be explained further below.

filler

In addition to the experimental stimuli created through the manipulation of linguistic variables, the other 65 to 75% of the sentences in an SPR experiment are not related to the research questions. There is no single consensus in the literature with regard to the ideal ratio of target versus total nontarget stimuli for an SLA experiment, but there is some evidence that a very low proportion of these might affect reading behavior during SPR and that 50% nontarget sentences is the minimum acceptable amount (Havik, Roberts, van Hout, Schreuder, & Haverkort, 2009). Up to half of the nontarget sentences may be distractors, which are created through the manipulation of entirely different linguistic variables and may serve either to balance the target stimuli (e.g., unambiguous distractors to complement similar but ambiguous target sentences) or as target stimuli for another experiment. Whether or not the experiment includes distractors, the remaining sentences are fillers, or unrelated sentences with no specific linguistic target. All target stimuli, distractors, and fillers are created to be comparable with regard to length and other superficial characteristics so that participants cannot easily identify the target sentences. For the same reason, manipulated variables that could potentially be recognized by research participants, such as grammar violations, should also be balanced across target and nontarget stimuli. In other words, if half of the target stimuli are ungrammatical, then half of the distractors and/or fillers should also be ungrammatical, again to avoid making target stimuli stand out. Most experiments present written instructions and eight to ten additional fillers as practice items so that participants can become familiar with the SPR procedure and the distractor questions that usually follow each of the stimuli before any potentially meaningful data are recorded.

!

practice

presentation lists!

counter-balance

Once all of the stimuli have been finalized, the presentation lists are created. In an ideal world, there would be only one list because all participants would read all stimulus items in all conditions, but in reality this could cause any number of undesirable presentation effects, including priming and ordering effects, as well as increasing the likelihood that participants would become consciously aware of the linguistic target of the experiment. To prevent such complications, each participant reads each stimulus item only once in one of its conditions, but still reads an equal number of target stimuli in each of the conditions. In order to have each stimulus read in all of its conditions, multiple counterbalanced presentation lists are created so that one subgroup of participants reads a stimulus item in the first condition, another group reads it in the second condition, and so on. Counterbalancing here means that each participant contributes an equal number of data points to each level of a variable (e.g., by reading eight items in each of four conditions), because there may be individual differences in reading speed or other characteristics among participants. To illustrate, if there are four stimulus conditions called a, b, c, and d, and there are 32 stimulus items numbered 1 to 32 (note that the total number of target stimuli is always a multiple of the number of conditions), then there would be four presentations lists, as illustrated in Table 2.2.

TABLE 2.2 Illustration of counterbalancing the first 8 of 32 stimuli in an experiment with four conditions across four presentation lists

	Item 1	*Item 2*	*Item 3*	*Item 4*	*Item 5*	*Item 6*	*Item 7*	*Item 8*	. . .
List I	1a	2b	3c	4d	5a	6b	7c	8d	. . .
List II	1b	2c	3d	4a	5b	6c	7d	8a	. . .
List III	1c	2d	3a	4b	5c	6d	7a	8b	. . .
List IV	1d	2a	3b	4c	5d	6a	7b	8c	. . .

X

Looking at just the first eight items in Table 2.2 and ignoring for now the other 24, if the first four participants each read one of the four lists, then they would each contribute two data points in each stimulus condition and the variable levels would be counterbalanced with regard to any differences in the individual subjects. That is, that participant reading List I would yield two reading time data points in Condition a, two in Condition b, two in Condition c, and two in Condition d, so individual variations such as reading speed should not affect the outcome of the experiment. There can be individual differences among stimulus items as well, so the same type of counterbalancing applies to the stimuli. In the above example, the variable levels would be counterbalanced with regard to items as long as the number of participants is a multiple of four. If there are four participants, as above, then Item 1 contributes one data point to each condition. If there are eight participants, two would read each of the four presentation lists and Item 1 would contribute two data points to each condition, and so on. The ideal experiment would have both types of counterbalancing, both by subjects and by items. In reality, the number of items is always a multiple of the number of stimulus conditions, but at the present time it is actually quite common in the literature to see numbers of participants (within each group) that are not exact multiples of the number of stimulus conditions because planning the number of items is usually more feasible than controlling the exact number of subjects that yield usable data, especially with L2 participants. As long as the presentation lists are rotated continually through each participant group (as opposed to using List I for the first ten participants, then List II for the next ten, etc.), then the final number of data points contributed by each item to each condition will not vary by more than one.

Of course, the stimuli are not presented in numerical order and are not presented in the same order to all participants. For the ordering of stimuli within each presentation list, pseudorandomization is the preferred technique. Total randomization would be ideal as far as minimizing any effects of presentation order, especially because the order could be unique for each participant instead of only

pseudo-randomi-zace

X totál randomi

for each list, but a problem with this method is that several experimental sentences would sometimes appear one right after the other (and in theory they could even all appear together), which can lead to priming or can draw participants' attention. Unfortunately, most experimental software can be set to automatically randomize stimuli, but not to pseudorandomize with specified limitations. Thus, to prevent target stimuli from being clumped together, a limited number of randomizations are usually created and then corrected so that no two similar sentences appear in succession. Randomized lists of numbers can be generated using, for example, the RANDBETWEEN function in Excel or the online Research Randomizer (Urbaniak & Plous, 2011).

distraktor
úloha

SPR studies typically involve some type of distractor task, in the form of questions that follow some or all of the stimuli, which serves in the first place to ensure that cognitive processes are engaged throughout and the participant does not end up pressing buttons without paying attention to the experimental stimuli on display. In addition, a well-designed distractor task can also prevent the participant from consciously reflecting on the primary SPR task and altering their behavior accordingly. The selection of a distractor task for an experiment should be intentional, as the type of task has been shown to affect sentence processing behavior (Aaronson & Scarborough, 1976; Havik et al., 2009; Leeser, Brandl, & Weissglass, 2011; but cf. Jackson & Bobb, 2009). Most SPR distractor tasks fall into one of two categories: acceptability judgments or comprehension questions, but given the critical distinction in SLA between acquired implicit linguistic knowledge and learned explicit information (Ellis, 2007; Krashen, 1981) and the ease with which most adult second language learners are influenced by explicit rules, it seems highly desirable in most cases to opt for meaning-based comprehension questions. After all, SPR behavior is only informative and generalizable to the extent that it reflects the same cognitive processes that are engaged during normal reading and during language comprehension in general. Furthermore, the use of metalinguistic distractor tasks can affect processing strategy, even causing certain reading time effects. For example, Leeser et al. (2011) found that intermediate L2 learners of Spanish showed on-line sensitivity to gender agreement violations during self-paced reading when the distractor task was a grammaticality judgment, but not when it was a meaningful comprehension question. Regardless of the type of distractor task, the questions or judgments tend to be binary choice items that are counterbalanced with regard to the number of correct "a" versus "b" or "acceptable" versus "unacceptable" answers.

[illegible]
1
[illegible]
[illegible]

If the researcher confirms that meaningful comprehension questions are indeed the best distractor task, but is also interested in collecting complementary data via an off-line measure like an acceptability judgment, then this task should be administered independently and not prior to the self-paced reading task, in order to avoid unnatural cross-contamination effects like priming on the reading time results. The two tasks should be constructed using different items and should be separated by additional activities, as priming is also known to affect the results of acceptability judgments (Luka & Barsalou, 2005).

An additional point of variation in the literature is with regard to how often distractor questions appear, meaning after every stimulus or randomly after only a fraction of the stimuli, such as one in four. Either method would be sufficient where the only purpose of the distractor question is that mentioned first above: to engage participants in the SPR task while simultaneously diverting their conscious attention from it. However, the comprehension questions that follow experimental stimuli are receiving increasingly more attention in L2 sentence processing research as potential loci of delayed processing effects. For instance, Roberts and Felser (2011) and Jegerski (2012b) reported and analyzed the accuracy rates and reaction times for responses to comprehension questions that appeared after every stimulus in their investigations of subject-object ambiguities. Like other researchers, they interpreted lower accuracy rates and longer reaction times as indications of processing difficulty that was delayed or spilled over from the stimulus that immediately preceded a comprehension question. Given the potential for comprehension question response and reaction time data to provide additional insight into the time-course of sentence processing, it seems that in most cases it would be advantageous to present distractor questions after every single stimulus rather than only a fraction.

[illegible]
[illegible]

Scoring, Data Analysis, and Reporting Results

raw
data

Raw data collected via the self-paced reading method include reaction times in ms as well as qualitative responses for every event in the experiment that allowed input from the participant. For example, each region of the experimental sentences yields a numerical reading time plus a categorical record of the button that was pressed to advance to the next display (usually a space bar on the keyboard or a green key on a button box), so a single sentential stimulus will easily have ten or more data points associated with it, depending on how long it is and what type of segmentation was used. A distractor question also has a corresponding reaction time plus a record of which button was pressed to answer (typically one of two buttons designated on the keyboard or a button box for responding to binary choice questions). These reaction time and button press data are compiled and stored as one output file per participant by the experimental software, so there is no true scoring or coding to be conducted manually with SPR. It is usually necessary, however, to compile and sort the data before performing statistical analyses.

Data output files group data by trial (i.e., stimulus), with each data point corresponding to one row in a list or table, and all trials are listed in the order in which they were presented during the experiment. Thus, in a raw data file target items are intermingled with practice items and fillers, numerical reaction times are mixed with categorical distractor task responses, and all the different events or steps within the trial—the initial cue symbol "+" or similar, Region 1, Region 2,

TABLE 2.3 Excerpt of an unsorted SPR data output file

List	*Subject*	*Trial*	*Event*	*Response*	*RT*
List B	9121	Instructions	Instructions	GREEN KEY	29258
List B	9121	Practice 1	[SubEvent 1]	GREEN KEY	710
List B	9121	Practice 1	[SubEvent 2]	GREEN KEY	804
List B	9121	Practice 1	[SubEvent 3]	GREEN KEY	930
List B	9121	Practice 1	[SubEvent 4]	GREEN KEY	905
List B	9121	Practice 1	Practice Ques 1	A KEY	3624
List B	9121	Practice 2	[SubEvent 1]	GREEN KEY	761
List B	9121	Practice 2	[SubEvent 2]	GREEN KEY	1700
List B	9121	Practice 2	[SubEvent 3]	GREEN KEY	1357
List B	9121	Practice 2	[SubEvent 4]	GREEN KEY	1218
List B	9121	Practice 2	Practice Ques 2	B KEY	6465
List B	9121	11a	[SubEvent 1]	GREEN KEY	1175
List B	9121	11a	[SubEvent 2]	GREEN KEY	882
List B	9121	11a	[SubEvent 3]	GREEN KEY	604
List B	9121	11a	[SubEvent 4]	GREEN KEY	642
List B	9121	11a	Comp Ques	A KEY	4780
List B	9121	58filler	[SubEvent 1]	GREEN KEY	729
List B	9121	58filler	[SubEvent 2]	GREEN KEY	970
List B	9121	58filler	[SubEvent 3]	GREEN KEY	625
List B	9121	58filler	[SubEvent 4]	GREEN KEY	614
List B	9121	58filler	Comp Ques	B KEY	4007
List B	9121	25b	[SubEvent 1]	GREEN KEY	681
List B	9121	25b	[SubEvent 2]	GREEN KEY	636
List B	9121	25b	[SubEvent 3]	GREEN KEY	1103
List B	9121	25b	[SubEvent 4]	GREEN KEY	929
List B	9121	25b	Comp Ques	A KEY	4646
List B	9121	45filler	[SubEvent 1]	GREEN KEY	851
List B	9121	45filler	[SubEvent 2]	GREEN KEY	730
List B	9121	45filler	[SubEvent 3]	GREEN KEY	986
List B	9121	45filler	[SubEvent 4]	GREEN KEY	1409
List B	9121	45filler	Comp Ques	B KEY	4140

Note: This part of the data file represents only the instructions presented at the beginning of the experiment, two practice items, two experimental items, and two fillers, so a complete data file would be much longer. Each "SubEvent" is a region of interest for the stimulus.

Region 3, any subsequent stimulus regions, and the distractor question—are listed together and also in order of appearance, as seen in Table 2.3. Because each stimulus presentation list includes a different randomization of the stimuli as well as different conditions of each stimulus, the data files will initially look different for each of the presentation lists. The raw data output files are commonly in the .txt format and can easily be opened in Excel or a similar spreadsheet program, where the sorting, linking, and macros features greatly facilitate the task of preparing data for analysis. The statistical software packages SPSS and R (R Development Core Team, 2005) both have additional features and capacity for large data sets. The initial steps in preparing data for analysis typically include compiling all of the data files into a single master file that identifies individual participants by number and specifies values for any grouping variables like nativeness or L2 proficiency level, and separating out the experimental items from the practice, distractor , and filler items, which in some cases may be analyzed separately or even included in measures of individual participant characteristics like average reading speed or overall comprehension accuracy (see Jegerski, 2012b, and Roberts & Felser, 2011, for examples of how all items have been included in posthoc measures of individual reading speed). A single master data file can be modified in order to conduct different types of statistical analyses, though it can be helpful to create a separate file or spreadsheet within a workbook for each part of the data that will undergo independent statistical analyses: one for each stimulus region with reading times, one for the distractor question with reading times, and one for the distractor question with categorical response data.

[illegible]

Each of these separate data sets would contain, for example, 3200 rows of data in an experiment with 40 target stimuli and 80 participants, and will be treated independently from this point forward for the purposes of statistical analyses. If the statistical analyses are to be linear mixed-effects models in R (R Development Core Team, 2005; see Baayen, Davidson, & Bates, 2008, for a motivation for the recent trend in psycholinguistics towards mixed-effects models and away from traditional parametric statistics), then there is no need to compute aggregate means or to conduct separate items analysis, and data trimming is either very minimal or entirely unnecessary (see Baayen & Milin, 2010, for further discussion of data trimming with mixed-effects models). If the more traditional ANOVAs and *t*-tests are to be conducted, then the data first need to be trimmed and converted to aggregate means, each step by both subject as well as item (see Clark, 1973, for the motivation behind items analysis in psycholinguistics). With both types of analyses, linear mixed-effects models or ANOVAs, it is possible to clean the data by removing data points from trials that ended in inaccurate distractor question responses and to transform the data, both of which are described in greater detail below.

Regardless of whether the statistical analyses are traditional parametric statistics or the newer mixed-effects models, it is currently common in both L1 and L2 SPR studies to eliminate and ignore reading time data that correspond to incorrect distractor question responses, under the assumption that inaccuracy is an indication that the participant may not have been paying attention while reading the experimental sentence. This convention is most clearly justified in the study of native language processing, where errors tend to be infrequent and the process by which readers arrive at inaccurate responses to comprehension questions is usually not of interest. It may also make sense to follow suit in research on very advanced near-natives, because their error rates are usually low as well. But particularly in those investigations where participants are not at the highest levels of proficiency, data from trials with inaccurate distractor question responses can be quite numerous (10 to 20% or more of trials) and may prove to be enlightening with regard to the mechanisms of (and obstacles to) development of L2 processing strategy, especially when analyzed independently. For example, Juffs and Harrington (1996) included data from incorrect grammaticality judgments in their report, conducting separate analyses of the data corresponding to accurate and inaccurate responses. Based on the two separate analyses, the researchers concluded that the ESL learners in their study were more likely to make an accurate grammaticality judgment for those trials in which they had dwelled longer on a critical region while reading the sentence, an interesting observation that could not have been deduced solely based on the data from accurate responses. It is also common to examine and analyze data from trials with incorrect responses in the study of aphasia, at least when such data are numerous enough to be informative (see Caplan & Waters, 2003, for an example from a self-paced listening study, or Dickey, Choy, & Thompson, 2007, for an example with visual world eye-tracking). Data from inaccurate trials have thus far been mostly ignored in SLA research, however, so at the present time there is very limited empirical evidence of how this type of data cleansing may affect experimental outcomes.

parametric tests

After the data from inaccurate trials are separated out and either discarded or reserved for independent analysis, the remaining steps in preparing the data for parametric tests like ANOVAs and *t*-tests are to trim the reaction time data by subject and by item and to compute aggregate means, also by subject and by item, for both numerical reaction time data and categorical response data from the distractor task. Both steps can be conducted without changing the format of the master data spreadsheet files, but it may be helpful for those new to items analysis to at least imagine the data in two different spreadsheet layouts, one by subject and one by item, in order to better understand the two types of analysis. For organization by subject, illustrated with hypothetical reading time data in Table 2.4, the data are arranged such that participants are listed down the left most column and each one corresponds to a row of data from all the different stimuli read by that subject. The stimuli are listed across the top, sorted by condition and then by item (e.g., 1a, 2a, 3a, etc., 1b, 2b, 3b, etc.). Thus, in the subjects layout, each row corresponds to a subject and each column to an item. This layout is useful for understanding both how to trim the data by subject and how to calculate aggregate means by subject. For the items layout, the same data from the subjects layout is transposed so that each row represents an item and each column represents a subject, as illustrated with hypothetical data in Table 2.5. The item layout can be helpful for visualizing how both data trimming and the computation of aggregate means produce different results when conducted by item rather than by subject. It is also interesting to note that any one of the group means (i.e., a mean of the individual means, calculated for each participant group within each stimulus condition), if calculated on untrimmed data, comes out the same by subject or by item, but the standard deviation differs. This is why data trimming is conducted separately by subject and by item and why it can result in (usually very small) differences in the group means by subject versus by item.

TABLE 2.4 Reading time data from SPR organized by subject

Subject	*Item 1a*	*Item 1b*	*Item 2a*	*Item 2b*	*Item 3a*	*Item 3b*	*Item 4a*	*Item 4b*
1	763			797	660			848
2		1029	883			959	823	
3	374			498	384			530
4		623	687			1102	655	

Note: This is a hypothetical partial data set from only four participants and four items in two conditions (these data are partial because of space limitations; real data would represent at least 16 stimuli and 16 participants). For ANOVAs and *t*-tests, a mean score for each subject for each stimulus condition (a, b) would be calculated on the complete data set, which in this example would yield two mean scores per row/subject. Stimulus Type or Condition would be a repeated measure or within-subjects variable because the same subject contributes to both levels of the variable.

TABLE 2.5 Reading time data from SPR organized by item

Item	*Sub 1/a*	*Sub 1/b*	*Sub 2/a*	*Sub 2/b*	*Sub 3/a*	*Sub 3/b*	*Sub 4/a*	*Sub 4/b*
1	763			1029	374			623
2		797	883			498	687	
3	660			959	384			1102
4		848	823			530	655	

Note: This is the same hypothetical partial data set from Table 2.4. For ANOVAs and *t*-tests, a mean score for each item for each stimulus condition (a, b) would be calculated on the complete data set, which in this example would yield two mean scores per row/item. Stimulus Type/Condition would be a repeated measure because the same item contributes to both levels of the variable.

outliers

Data trimming is the process of cleaning reaction time data to minimize the effects of those data points which appear to have been influenced by external factors unrelated to language processing, such as minor distractions and disruptions during the SPR experiment, which can obscure real reading time effects through the addition of extraneous variance and the resulting reduction in experimental power. Trimming involves the identification and removal or replacement of extreme data points, known as outliers. Outliers are presumed to reflect measurement error rather than authentic processing behavior and their elimination is most important to maximizing the accuracy and power of parametric tests that are conducted on aggregate means, like *t*-tests and ANOVAs, and is less critical with linear mixed-effects models, although a few very extreme outliers may be removed for mixed-effects models as well. In the case of the former, the purpose of identifying outliers is to ensure that aggregate means are as accurate as possible and minimally affected by extreme values, because once the means are calculated, the range of values they represent is no longer part of the data. Given that mixed-effects models do not rely on aggregate means, the full range of values remains in the data on which the statistical tests are conducted and the presence of outliers is thus not such a concern.

There is no single accepted method for dealing with outliers in SPR data, and early L2 studies tended to leave reading time data untrimmed, but outliers can obscure reading time effects that would otherwise prove significant, so it is preferable to minimize their influence. Two common procedures for identifying outliers are the designation of an absolute cutoff and the calculation of a cutoff based on standard deviations, while two ways to mitigate the effects of outliers, once they have been identified, are deletion and substitution with a more moderate value. For the absolute cutoff method, real response times of less than 100 ms are generally not possible (Luce, 1986) and with SPR in particular reading times of less than 200 ms likely reflect unintentional button presses, so lower cutoffs usually fall within the range of those two values. Higher cutoffs vary and have been set, for instance, at 3000 ms for all participants (Roberts & Felser, 2011), 3000 ms for native readers and 4000 ms for non-native readers (Havik et al., 2009), or 6000 ms for all participants (Jackson, 2010). The absolute cutoff method has the advantage of being conducted only once on the data, uniformly across reading time data from all stimulus regions, with no need to differentiate at this point by subject and by item. A disadvantage of using absolute cutoffs is that they are uniform across participant groups and experimental conditions and thus can potentially neutralize subtle differences.

abs/cf

Especially where there is considerable variability between subjects or between items, it may be preferable to use the more conservative standard deviation approach to identifying outliers rather than setting absolute values. With the standard deviation method, reading times that fall greater than two to three standard deviations away from a mean, calculated either for each individual or for each participant group in each stimulus condition, are judged to be outliers. Individual calculations should be conducted where there is notable variability between subjects. A disadvantage of the standard deviation approach is that it is considerably more time consuming, especially because it is conducted twice for each stimulus region, once by subjects and once by items. Some researchers prefer to use a combination of both of the above methods, identifying the most dramatic outliers first using absolute cutoffs and then using a standard deviation method to identify additional outliers.

SD

After the outliers have been identified using one of the above methods, they are either deleted from the data set or replaced with a more moderate value, such as the absolute value or standard deviation value used as the cutoff, or treated with a combination of both deletion and substitution. For example, extremely low reading time values of 100 to 200 ms are presumably erroneous and uninformative, and therefore can reasonably be deleted. Extremely high values, on the other hand, might reflect real processing difficulty, so it makes sense to replace them with more moderate values that are still relatively high, such as the cutoff value used to identify outliers. Spreadsheet and statistical software can greatly facilitate the identification of outliers, their deletion, and their substitution with more moderate values. In the end, by the time the reading time data are submitted to statistical tests, the trimming of outliers has rarely affected more than 5% of the total data set, although it is acceptable to go as high as 10% (Ratcliff, 1993).

del.
high
values

Once the data have been trimmed, aggregate means can be calculated and ANOVAs and *t*-tests conducted on the means. For subjects analysis, one mean is calculated for each participant within each stimulus condition. For items analysis, one mean is calculated for each item in each stimulus condition. ANOVAs and *t*-tests, which are the most common statistical analyses conducted on SPR data in SLA research to date, both determine whether any observed differences are likely to reflect real differences in the stimuli conditions or participant groups. These tests can be performed using a commercial statistical program like SPSS or free software like R (R Development Core Team, 2005); basic *t*-tests can also be conducted in Excel using the TTEST function, which can be useful for preliminary analyses. A typical ANOVA for a simple SPR experiment is a 2 (groups) × 2 (stimulus conditions), which is run as a mixed design because group is a between-subjects factor and stimulus condition is a within-subjects factor, or repeated measure, in which the same subject contributes to both levels of the variable. In the case of an interaction, posthoc *t*-tests can be conducted to make the comparisons within each group. Statistics from these ANOVAs and *t*-tests are reported as F_1 and t_1 in order to distinguish them from those generated in the analyses by item, which are referred to as F_2 and t_2. For the analyses by item, the ANOVA for a simple experiment would again be a 2 × 2, but here both group and stimulus condition are within-subjects factors because the item is held constant across both levels of both variables. Similar sets of analyses by subject and then by item are conducted for each region of interest in the stimuli and for the distractor questions. One of the basic assumptions of parametric statistics like ANOVAs and *t*-tests is that the data are in a normal distribution, but reading time data are positively skewed and proportional data from binary choice distractor questions are negatively skewed, so the transformation of both types of data before submitting them to such analyses can also improve statistical power.

The complete results of the statistical analyses, be they parametric tests or mixed-effects models, are reported when presenting or publishing the results of an SPR experiment or series of experiments. Analyses for stimulus regions prior to the region in which the conditions differ because of the experimental manipulation do not typically yield any significant effects or interactions—barring any failure to counterbalance extraneous variables in the experimental materials—and are often not reported. In addition, descriptive statistics in the form of group means are also reported as line graphs (often with error bars) or as a table of reading times, as in Figure 2.4 and in Table 2.1, respectively. Accuracy data from distractor questions are typically displayed as separate tables or bar graphs.

An Exemplary Study

A good example of how SPR can be employed to address SLA issues can be found in Jackson (2010), a research article which provides a good level of detail on the methodology as well as the complete set of stimuli used for the experiment. This investigation was designed to address the questions of whether L1 processing strategy is transferred to L2 processing and whether L2 readers pay more attention to lexical-semantic information while processing than do L1 readers. To answer these questions, the study exploited a temporary subject-object ambiguity that can arise in German due to flexible word order, case marking that is sometimes optional, and the occurrence of lexical verbs in either verb-second or verb-final positions, depending on whether the verb tense is simple or complex, respectively. A sample of a stimulus item in its four conditions is given in (7), where slashes indicate the phrase-by-phrase segmentation used for the SPR task. Psycholinguistic research has shown that native speakers of German take longer to read the object-first version of similarly disambiguated sentences, even when a compound verb tense means that the lexical verb does not appear until the end of the sentence (Schriefers, Friederici, & Kühn, 1995). This shows not only that they prefer subject-first word order, but also that they begin to assign argument roles before encountering the lexical verb, purely on the basis of syntactic cues like case markers.

(7)

a. *Subject-first, Simple past*

Welche Ingenieurin / traf / den Chemiker / gestern Nachmittag / im Café / bevor / der Arbeitstag / anfing?
Which NOM$_{/ACC}$ engineer / met / the$_{ACC}$ chemist / yesterday afternoon / in-the café / before / the work-day / began?
"Which engineer met the chemist yesterday afternoon in the café, before the workday began?"

b. *Object-first, Simple past*

Welche Ingenieurin / traf / der Chemiker . . .
Which$_{NOM/ACC}$ engineer / met / the$_{NOM}$ chemist . . .
"Which engineer did the chemist meet . . . "

c. *Subject-first, Present perfect*

Welche Ingenieurin / hat / den Chemiker / gestern Nachmittag / getroffen . . .
Which$_{NOM/ACC}$ engineer / has / the$_{ACC}$ chemist / yesterday afternoon / met . . .
"Which engineer has met the chemist yesterday afternoon . . . "

d. *Object-first, Present perfect*

Welche Ingenieurin / hat / der Chemiker / gestern Nachmittag / getroffen . . .
Which$_{NOM/ACC}$ engineer / has / the$_{NOM}$ chemist / yesterday afternoon / met . . .
"Which engineer has the chemist met yesterday afternoon . . ."

The participants in Jackson's (2010) study were native and advanced nonnative speakers of German, with 22 L1 English-L2 German, 20 L1 Dutch-L2 German, and 24 German native speakers. Dutch is similar to German with regard to the differential positioning of lexical verbs in simple versus complex verb tenses described above, while English is not, so the purpose of including the two L2 participant groups was to better address the issue of L1 transfer. Participants each read 32 target items, with eight in each of the four conditions illustrated above, along with 64 filler items in a linear, noncumulative (i.e., moving window) SPR procedure with phrase-by-phrase segmentation. The distractor task was meaning-based and asked participants to indicate whether a statement on the screen was consistent with the meaning of the stimulus sentence that preceded it.

Mixed-design ANOVAs performed on the reading times from the critical region underlined in (7) revealed that all three participant groups had similar difficulty in processing object-first word order, regardless of whether the verb tense was simple past or present perfect. Analyses of reading times for the clause-final region, however, showed that only the L1 English participants had greater difficulty with the stimuli in the present perfect conditions after reaching the lexical verb at the end of the clause. Jackson (2010) concluded that the L1 English readers may have transferred a lexical verb-driven strategy from English while processing sentences in L2 German, but that such relatively higher sensitivity to lexically-based information is not an inherent characteristic of L2 processing because the L1 Dutch readers exhibited SPR effects that were similar to those of the native German readers.

Pros and Cons in Using the Method

Pros

- SPR is inexpensive. It is probably the most economical on-line method for sentence processing research and is thus accessible to a wide range of researchers, including those conducting pilot studies and student research projects. Experiments can be built and run on a basic desktop or laptop using either free software that runs entirely script-based experiments (e.g., DMDX by Forster & Forster, 1999; Linger by Rhode, 2001; PsyScope by Cohen, MacWhinney, Flatt, & Provost, 1993) or relatively inexpensive commercial software that adds the convenience of a graphical user interface (e.g., E-Prime by Schneider, Eschmann, & Zuccolotto, 2002; SuperLab, 1992). [illegible]
- SPR is highly portable. Because there is no special equipment outside the computer that runs the software and perhaps a small response device, SPR experiments can be conducted virtually anywhere. [illegible]
- SPR is efficient. The researcher does not have to supervise participants as closely as with some other methods, where it can be necessary to monitor and make adjustments to equipment while the experiment is in progress. This convenience, combined with the low cost of start-up, makes it feasible [illegible]

for a single researcher to run as many as six or seven subjects at a time if the laboratory facilities are sufficient.

- SPR is an exceptionally covert measure of sentence processing. Participants' conscious attention can easily be diverted away from language to a distractor task such as answering comprehension questions, which is also more familiar to them as an assessment than SPR. Additionally, participants do not need to know beforehand that the software program is recording their reading times, they are not likely to have previous assumptions regarding SPR because it is one of few methods used exclusively for psycholinguistic research, and they do not come into contact with any highly specialized technical equipment that could further invite conscious, task-specific strategy.
- SPR materials can also be relatively covert with regard to their linguistic targets. While some SPR paradigms do employ stimuli with grammatical violations that may invite explicit judgments or the activation of metalinguistic knowledge, particularly among participants formally trained in the L2, it is also possible to obtain significant reading time effects with more subtle paradigms in which all stimuli are grammatical.
- SPR can detect spillover and sentence wrap-up effects. Increased reading times on the stimulus region immediately following the site of an immediate effect or at the end of a sentence are assumed to reflect later phases of comprehension and can be indicators of processing difficulty that is either persistent or delayed, which is especially useful in L2 research.

Cons

- Text must be segmented word-by-word or phrase-by-phrase in order to generate reading time data with any level of detail. This presentation format is different from that typically seen outside of the laboratory environment and may present an additional processing load, probably because in "normal" college-level reading (as measured with eye-tracking), regressions account for about 15% of eye movements. This can raise questions of ecological validity. During debriefing in the laboratory, a few of my research participants (less than 2%) have even commented that they find the word-by-word reading mode difficult and have speculated that the SPR experiments in which they participated were tests of memory.
- Research subjects must repeatedly press a button while reading in the self-paced mode, which was historically regarded as a potentially unnatural distraction because it was markedly different from reading a book or other printed text. However, this difference between what occurs inside versus outside the laboratory is no longer as dramatic, as most participants are now accustomed to using input devices while reading because of contemporary technology like text messaging, browsing the Internet via a smart phone, or using computer software with a graphical user interface and pointing device like a mouse or touchpad.

- SPR requires that participants be relatively fluent readers. Thus, great care must be taken when using the method to study populations with relatively less reading experience in the target language, such as beginning L2 learners, heritage bilinguals, and even some native speakers with low levels of literacy. Also, the study of languages without developed and easily digitized writing systems is of course not possible with SPR.
- SPR can generate task-specific effects, especially when stimuli are segmented word-by-word. These effects include reading times that are generally slower than normal and delayed processing effects that can spill over into the next region.

Discussion and Practice

Questions

1) Give two examples of participant populations that could easily be studied using the SPR method, giving specific details regarding their L1, L2, proficiency level, and demographic characteristics. Then, give two more examples of populations for which groups it might be more desirable to select another experimental measure.
2) Define the terms *item, condition,* and *region of interest* as they relate to self-paced reading stimuli, being clear about the differences between them. Be sure to use concrete examples to support your explanations.
3) Select an example of a published research study of L2 processing using the SPR method, either from among those mentioned in this chapter or from recent SLA and psycholinguistics journals, and analyze its research design. This analysis should include a) the specific research questions guiding the study, b) the independent variables that were manipulated through the choice of participants and/or stimuli, c) the type of statistical analyses run on the data, and d) a critique of the design, with particular attention to the relationship between the research questions, the independent variables, and the statistical analyses.
4) Consider the hypothetical incomplete data set in Tables 2.4 and 2.5. Calculate the mean scores by subject and by item (you should have a total of 4 × 2 = 8 means for each table). Do there appear to be any differences in reading times between the two stimulus conditions (a, b), either by subject and/or by item? Give an example of what type of stimuli might yield this data set and which region of those experimental sentences these data might represent.
5) List the steps in preparing SPR data from one experiment for statistical analyses via ANOVAs and *t*-tests. How does the list change if the analyses are to be via linear mixed-effects models?

Research Project Option A

The off-line distractor task is an integral part of sentence comprehension research, yet not much is known about how it might influence on-line processing behavior.

One project that would serve as a good introduction to the SPR method and its potential effects on participant behavior would be to replicate a published SPR study using two different distractor task conditions: acceptability judgment and meaningful comprehension question. (If the complete set of stimuli is not part of the published article, materials, and permission to use them, can usually be obtained by writing the author.) Such an investigation could be informative with just a single L2 participant group, though ideally a comparison group of native speakers would also be included.

Research Project Option B

The majority of existing L2 sentence processing research includes participants of only one proficiency level, so our knowledge of the developmental trajectory of non-native sentence comprehension behavior is quite limited. Another good introduction to the SPR method would be to replicate a published study and include at least two participant groups with different levels of proficiency in the L2. With both this option and Research Project Option A above, the target language could be the same as in the original study, or the stimuli could be translated into another language—assuming that there is sufficient grammatical similarity between the two languages.

Note

1. In order to avoid confusion, it is worth noting that the term *moving window* is also sometimes used to refer to a gaze-contingent eye-tracking paradigm in which parafoveal vision is masked or blurred in order to examine its role in reading comprehension and in visual perception in general. In this type of eye-tracking, the "moving window" is a small, round window of clear vision that corresponds to the reader's foveal region of vision, which moves around the display screen as the reader's gaze shifts. Thus, in both eye-tracking and self-paced reading the moving window is a spatially limited view of the stimulus that is surrounded by a larger area that is masked.

Suggested Readings

Dussias, P. E., & Piñar, P. (2010). Effects of reading span and plausibility in the reanalysis of wh- gaps by Chinese-English L2 speakers. *Second Language Research, 26*, 443–472.

Jegerski, J. (2012b). The processing of temporary subject-object ambiguities in native and near-native Mexican Spanish. *Bilingualism: Language and Cognition, 15*, 721–735.

Roberts, L., & Felser, C. (2011). Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics, 32*, 299–331.

Sagarra, N., & Herschensohn, J. (2011). Proficiency and animacy effects on L2 gender agreement processes during comprehension. *Language Learning, 61*, 80–116.

References

Aaronson, D., & Scarborough, H. S. (1976). Performance theories for sentence coding: Some quantitative evidence. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 56–70.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12–28.

Caplan, D., & Waters, G. S. (2003). On-line syntactic processing in aphasia: studies with auditory moving windows presentation. *Brain and Language, 84*(2), 222–249.

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

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359.

Cohen, J. D., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope [Computer software]. Retrieved from http://psy.ck.sissa.it/.

Crain, S., & Fodor, J. D. (1985). How can grammars help parsers? In D. R. Dowty, L. Karttunen, & A.M. Zwicky (Eds.), *Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives* (pp. 94–128). Cambridge, UK: Cambridge University Press.

Cuetos, F., & Mitchell, D.C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition, 30*, 73–105.

Dickey, M. W., Choy, J. J., & Thompson, C. K. (2007). Real-time comprehension of wh-movement in aphasia: Evidence from eyetracking while listening. *Brain and Language, 100*(1), 1–22.

Donders, F. (1868/1969). On the speed of mental processes. *Acta Psychologica, 30*, 412–431. (Translated by W. G. Koster).

Dussias, P. (2003). Syntactic ambiguity resolution in second language learners: Some effects of bilinguality on L1 and L2 processing strategies. *Studies in Second Language Acquisition, 25*, 529–557.

Ellis, N. C. (2007). Implicit and explicit knowledge about language. In J. Cenoz & N. H. Hornberger (Eds.) *Encyclopedia of Language and Education, Second Edition, Volume 6: Knowledge about Language* (pp. 119–132). New York: Springer.

Felser, C., Roberts, L., Gross, R., & Marinis, T. (2003). The processing of ambiguous sentences by first and second language learners of English. *Applied Psycholinguistics, 24*, 453–489.

Ferreira, F., & Henderson, J. (1990). Use of verb information during syntactic parsing: Evidence from eye tracking and word by word self-paced reading. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16*, 555–568.

Fodor, J. A., & Bever, T. G. (1965). The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior, 4*, 414–420.

Foote, R. (2011). Integrated knowledge of agreement in early and late English-Spanish bilinguals. *Applied Psycholinguistics, 21*, 187–220.

Forster, K. I., & Forster, J. C. (1999). DMDX [Computer software]. Tucson, AZ: University of Arizona/Department of Psychology.

Forster, K. I., & Olbrei, I. (1973). Semantic heuristics and syntactic analysis. *Cognition, 2*, 319–347.

Foss, D. J. (1970). Some effects of ambiguity upon sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 9*, 699–706.

Gilboy, E., & Sopena, J. M. (1996). Segmentation effects in the processing of complex NPs with relative clauses. In M. Carreiras, J. E. García-Albea, & N. Sebastián-Gallés (Eds.), *Language Processing in Spanish* (pp. 191–206). Mahwah, New Jersey: Erlbaum Associates.

Havik, E., Roberts, L., van Hout, R., Schreuder, R., & Haverkort, M. (2009). Processing subject-object ambiguities in the L2: A self-paced reading study with German L2 learners of Dutch. *Language Learning, 59*, 73–112.

Hoover, M., & Dwivedi, V. (1998). Syntactic processing by skilled bilinguals. *Language Learning, 48,* 1–29.

Jackson, C. N. (2010). The processing of subject-object ambiguities by English and Dutch L2 learners of German. In B. VanPatten & J. Jegerski (Eds.), *Research in second language processing and parsing* (pp. 207–230). Amsterdam: John Benjamins.

Jackson, C. N., & Bobb, S. C. (2009). The processing and comprehension of *wh-* questions among second language speakers of German. *Applied Psycholinguistics, 30*(4), 603–636.

Jegerski, J. (2012a). The processing of case markers in near-native Mexican Spanish. Poster presented at the 25th Annual CUNY Conference on Human Sentence Processing, New York, NY.

Jegerski, J. (2012b). The processing of temporary subject-object ambiguities in native and near-native Mexican Spanish. *Bilingualism: Language and Cognition, 15*(4), 721–735.

Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics, 25*(4), 603–634.

Juffs, A. (1998a). Main verb versus reduced relative clause ambiguity resolution in second language sentence processing. *Language Learning, 48,* 107–147.

Juffs, A. (1998b). Some effects of first language argument structure and syntax on second language processing. *Second Language Research, 14,* 406–424.

Juffs, A. (2004). Representation, processing, and working memory in a second language. *Transactions of the Philological Society, 102,* 199–225.

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

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

Juffs, A., & Harrington, M. (1996). Garden path sentences and error data in second language processing research. *Language Learning, 46,* 286–324.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixation to comprehension. *Psychological Review, 87,* 329–354.

Just, M. A., Carpenter, P. A., & Wooley, J. D. (1982). Paradigms and Processes in Reading Comprehension. *Journal of Experimental Psychology: General, 111*(2), 228–238.

Krashen, S. (1981). *Second language acquisition and second language learning.* Oxford, UK: Pergamon.

Leeser, M., Brandl, A., & Weissglass, C. (2011). Task effects in second language sentence processing research. In P. Trofimovich & K. McDonough (Eds.), *Applying priming methods to L2 learning, teaching, and research: Insights from psycholinguistics* (pp. 179–198). Amsterdam: John Benjamins.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization.* New York: Oxford University Press.

Luka, B. J., & Barsalou, L. W. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language, 52,* 436–459.

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes, 9,* 121–136.

Marinis, T., Roberts, L., Felser, C., & Clahsen, H. (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition, 27,* 53–78.

Mitchell, D.C., & Green, D. W. (1978). The effects of context and content on immediate processing in reading. *Quarterly Journal of Experimental Psychology, 30*(4), 609–636.

Nicol, J. L., Forster, K. I., & Veres, C. (1997). Subject-verb agreement processes in comprehension. *Journal of Memory and Language, 36,* 569–587.

Pan, H.-Y., & Felser, C. (2011). Referential context effects in L2 ambiguity resolution: Evidence from self-paced reading. *Lingua, 121,* 221–236.

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, 24,* 501–528.

Pliatsikas, C., & Marinis, T. (2013). Processing empty categories in a second language: When naturalistic exposure fills the (intermediate) gap. *Bilingualism: Language and Cognition, 16*(1), 167–182.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114,* 510–532.

R Development Core Team. (2005). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org.

Rhode, D. L. T. (2001). Linger [Computer software]. Cambridge, MA: MIT TedLab. Retrieved from http://tedlab.mit.edu/~dr.

Roberts, L. (2009). The L1 influences the on-line comprehension of tense/aspect in L2 English. Paper presented at the L2 Processing and Parsing Conference, Lubbock, TX.

Roberts, L., & Felser, C. (2011). Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics, 32*(2), 299–331.

Sagarra, N., & Herschensohn, J. (2011). Proficiency and animacy effects on L2 gender agreement processes during comprehension. *Language Learning, 61,* 80–116.

Schneider, W., Eschmann, A., & Zuccolotto, A. (2002). E-Prime [Computer software]. Pittsburgh, PA: Psychology Software Tools Inc.

Schriefers, H., Friederici, A.D., & Kühn, K. (1995). The processing of locally ambiguous relative clauses in German. *Journal of Memory and Language, 34,* 499–520.

SuperLab (Version 4.0.5) [Computer software]. (1992). San Pedro, CA: Cedrus Corporation.

Taraban, R., & McClelland, J. L. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of Memory and Language, 27,* 1–36.

Trueswell, J. C., & Kim, A. E. (1998). How to prune a garden path by nipping it in the bud: Fast priming of verb argument structure. *Journal of Memory and Language, 39,* 102–123.

Urbaniak, G. C., & Plous, S. (2011). Research Randomizer (Version 3.0) [Computer software]. Retrieved from http://www.randomizer.org/

VanPatten, B., & Jegerski, J. (Eds.) (2010). *Research in second language processing and parsing.* Amsterdam: John Benjamins.

White, L., & Juffs, A. (1998). Constraints on wh- movement in two different contexts of non-native language acquisition: Competence and processing. In S. Flynn, G. Martohardjono, & W. O'Neil (Eds.), *The generative study of second language acquisition* (pp. 111–129). Mahwah, NJ: Lawrence Erlbaum.

Williams, J., Möbius, P., & Kim, C. (2001). Native and non-native processing of English wh-questions: Parsing strategies and plausibility constraints. *Applied Psycholinguistics, 22,* 509–540.