

A methodological synthesis of self-paced reading in second language research

EMMA MARSDEN and SOPHIE THOMPSON
University of York

LUKE PLONSKY
Northern Arizona University

Received: February 2, 2017 Accepted for publication: December 12, 2017

ADDRESS FOR CORRESPONDENCE

Emma Marsden, University of York, Centre for Research into Language Learning and Use,
Department of Education, York, United Kingdom YO10 5DD. E-mail: emma.marsden@york.ac.uk

ABSTRACT

Self-paced reading tests (SPRs) are being increasingly adopted by second language (L2) researchers. Using SPR with L2 populations presents specific challenges, and its use is still evolving in L2 research (as well as in first language research, in many respects). Although the topic of several narrative overviews (Keating & Jegerski, 2015; Roberts, 2016), we do not have a comprehensive picture of its usage in L2 research. Building on the growing body of systematic reviews of research practices in applied linguistics (e.g., Liu & Brown, 2015; Plonsky, 2013), we report a methodological synthesis of the rationales, study contexts, and methodological decision making in L2 SPR research. Our comprehensive search yielded 74 SPRs used in L2 research. Each instrument was coded along 121 parameters, including: reported rationales and study characteristics, indicating the scope and nature of L2 SPR research agendas; design and analysis features and reporting practices, determining instrument validity and reliability; and materials transparency, affecting reproducibility and systematicity of agendas. Our findings indicate an urgent need to standardize the use and reporting of this technique, requiring empirical investigation to inform methodological decision making. We also identify several areas (e.g., study design, sample demographics, instrument construction, data analysis, and transparency) where SPR research could be improved to enrich our understanding of L2 processing, reading, and learning.

Keywords: foreign language learning; moving window; open science; research design; research methodology; second language learning; self-paced reading; synthesis; systematic review

Self-paced reading (SPR) is an online computer-assisted research technique in which participants read sentences, broken into words or segments, at a pace they control by pressing a key. The time elapsed (reaction time; RT) between each keypress is recorded. Underlying this technique is an assumption that participant RTs indicate their knowledge of and/or sensitivity to linguistic phenomena relative to other linguistic phenomena. The technique was originally used to investigate first language (L1) reading mechanisms (Aronson & Scarborough, 1976), including word recognition in sentential contexts, meaning representation, and real-time

parsing (building syntactic structures), among native speakers, often highly literate, monolingual adults. The method has been increasingly adopted by researchers interested in L2 phenomena, yet special challenges are presented when using SPR for L2 research. As the applicability and rigor of usage of this technique in L2 research has not been scoped systematically, one of the main purposes of the current study is to identify why and how L2 researchers have used this method. For example, although SPR is thought to offer a window into processes that are largely automatic (i.e., fast and without awareness), L2 learners are often of varying proficiencies, experiences, and ages. They also have a wide range of L2 reading skills and are more likely than many adult L1 participants to have explicit knowledge of the language due to formal L2 instruction. Thus, how the nature of knowledge and mechanisms elicited by SPRs, and the instruments used alongside them, are discussed and operationalized by L2 researchers is worthy of empirical and systematic investigation.

L2 learners are also unique in that they bring to the task of reading a complex set of phenomena due to their highly entrenched L1 representations and processing routines, along with varying degrees of L1 reading expertise. To illustrate, in L2 research an inverse relationship is generally expected between proficiency and the time needed to process words or segments in an SPR test (higher proficiency = *faster*). However, we might also expect more advanced users to process anomalies or disambiguations more *slowly* than less proficient users who may be less sensitive to the target structure. In addition, we might anticipate certain effects to obtain as a function of different L1s, depending on the particular theory about the role of the L1 in real-time processing, representation, and learning of an L2. Other questions specific to L2 research are whether L2 online processing is fundamentally different to L1 processing (e.g., more superficial or “shallower”; see below) and the extent to which it is different to *offline* knowledge in the L2 compared to the L1; investigations into these questions inform our understanding of differences between L1 and L2 learning. The extent to which all these issues and relevant participant characteristics have been investigated, operationalized, and reported is, therefore, of high importance, and can provide the field of L2 research with data about the purpose and nature of its own practices and on relations between data elicitation, analysis, and theorizing.

Our focus on SPR in L2 research had several motivations. First, as noted above, L2 populations present specific areas of interest and, therefore, entail particular methodological decisions and reporting requirements. Second, SPR is increasingly popular. Of course, other methods for investigating L2 knowledge and reading exist (e.g., rapid serial visual presentation; RSVP; Boo & Conklin, 2015; whereby the researcher, rather than the participant, controls the pace), but SPR is often thought to provide certain advantages, no doubt reflected by its increasing popularity. RSVP is much less commonly used perhaps because SPR, unlike RSVP, leaves control over exposure time to the participant (as in natural reading) and, as such, can concurrently *measure* processing time, thus reflecting *online* cognitive mechanisms (see Just, Carpenter, & Woolley, 1982). Similarly, other methods exist for investigating online processing, such as eye tracking and event-related potentials, but, again, SPR presents some advantages, including: its relative ease of administration and cost; its elicitation of behavioral data rather than neurological data where links between constructs and their signatures are still relatively nascent and

debated (Morgan-Short, 2014); and its comparability to eye tracking in its capacity to tap into cognitive processes (Just et al., 1982; see Keating & Jegerski, 2015, for a narrative review of these three online techniques). Third, our focus on SPR allows us to drill down with a high level of detail, in the space available, into substantive and methodological issues that pertain to this particular technique and are specific to L2 research. These issues include comprehension measures, participant sampling and reporting practices, segmentation decisions in different languages, and the extent and nature of investigations such as patterns of L1-L2 combinations and different processing phenomena. Fourth, a systematic methodological review is already available for another online processing technique (see Lai et al., 2013, for eye tracking). Thus, given our interest in the context of L2 research, we limited the scope of inquiry for our synthesis accordingly, i.e., to studies employing SPR with L2 users as participants. Future studies could compare L1 to L2 SPR research, or SPR to RSVP. (See Bowles, 2010; Lai et al., 2013; Yan, Maeda, Lv, & Ginther, 2016, for similar rationales underpinning systematic methodological reviews of think-alouds, eye-tracking, and elicited imitation, respectively).

With many of the issues and challenges described thus far in mind, Keating and Jegerski (2015) provide particularly useful guidance on SPR, addressing design, administration, data preparation, and analysis procedures (see also Jegerski & Van-Patten, 2013; Roberts, 2016, for methodological guidance and commentary on key studies). The present study complements these and other relevant discussions (e.g., Clahsen & Felser, 2006; Jiang, 2012; Juffs & Rodríguez, 2015) to systematically examine the purpose and use of SPR in L2 research. More specifically, we apply a synthetic/meta-analytic technique, namely, systematic methodological synthesis, to understand the following:

1. The extent to which SPR has been used in L2 research and the research areas that such studies have addressed
2. The contexts, demographics, design features, and instrumentation used in L2 SPR research
3. The features of L2 SPR tests and corresponding analyses
4. The methodological transparency of L2 SPR research

By investigating these characteristics within a comprehensive body of research using SPR, we sought to better understand why and how this technique has been used in L2 research. It is not our intention to criticize the efforts of previous researchers but, rather, to highlight issues and practices that relate to construct validity, reliability, and reproducibility. We use our results to indicate where empirically grounded standard practices might be useful and also to indicate specific study and participant characteristics that would extend the agendas thus far investigated using SPRs. Our study thereby complements and builds on foundational discussions put forward by others (e.g., Keating & Jegerski, 2015).

RESEARCH AIMS AND RATIONALES FOR USING SPR IN L2 RESEARCH

Methodological syntheses cover a wide range of issues that cannot all be justified in a background section of a journal article. As noted above, we refer the reader

to several existing narrative reviews and guides, which do an excellent job of laying out the substantive and methodological considerations in the use of SPR. Those works were also highly influential in motivating the current study and in the development of our scheme for coding our studies. The majority of the background section that follows is, therefore, limited to issues that require further explanation, particularly those where greater inferencing was needed to code for features in our sample of primary studies. These issues are as follows: the reported rationales for using SPR; the broad research aims of studies; the processing phenomena and linguistic features investigated; the sentence regions analyzed; and the nature of processing/knowledge elicited.

Overarching rationales and research aims of SPR research

Two broad questions, both central to much of L2 research, have driven the use of SPR: the extent and nature of differences between native and nonnative language acquisition and knowledge, and the role of the L1 in L2 development (i.e., cross-linguistic influence). In the former, SPRs have been used to investigate the extent to which L1 (native) and L2 (nonnative) processing draw on fundamentally different mechanisms, such as access to and nature of linguistic representations. For example, there is evidence that L2 adult learners access superficial linguistic syntactic information as compared to adults processing their L1 (Clahsen & Felser, 2006; Marinis, Roberts, Felser, & Clahsen, 2005). SPR data have also been used to show, however, that nativelike syntactically based processing can occur (Dussias, 2003; Juffs, 1998; Williams, Mobius, & Kim, 2001), and that this can depend on, for example, proficiency (Dekydtspotter & Outcalt, 2005; Hopp, 2006), the complexity of syntactic structures, the nature of the task (Havik, Roberts, van Hout, Schreuder, & Haverkort, 2009), and the type of learning experience (immersion vs. more form-focused, Pliatsikas & Marinis, 2013a).

A closely related agenda investigates the extent to which the L1 influences L2 processing, learning, or representations. A key principle motivating this line of research is as follows: if the speed of processing is affected on words or structures that share some similarity with the L1 compared to others that do not, we might assume that L1 representations are activated, at some level, during reading (Koda, 2005). Such findings have been used to suggest that the L1 influenced or continues to influence L2 learning or use, via the lexicon (Bultena, Dijkstra, & van Hell, 2014; Ibáñez, Macizo, & Bajo, 2010) or morphosyntax (Dussias, 2003; Hopp, 2009; Jiang, Novokshanova, Masuda, & Wang, 2011; Marull, 2015).

In the current study, we systematically review the designs of SPRs that have addressed these broad questions and the processing phenomena they target (e.g., ambiguity resolution or anomaly detection). We also systematically review the features that have served as the linguistic targets in this line of inquiry.

Sentence processing phenomena and the linguistic “critical regions”

A large part of L2 sentence processing research is based on the idea that initial parses can be erroneous and reanalysis is required (this reanalysis has been theorized in various ways; see Van-Gompel, 2013, for a detailed overview). For example, in (1) the reader could interpret “the mistake” (an ambiguity) as a direct

object that completes a subject–verb–object parse, and not as a reduced relative clause (i.e., “The ticket agent admitted that the mistake ...”), until the disambiguation point “might” is reached. This could then result in a reinterpretation of the sentence by parsing the ellipsed “that,” observable in slower processing during or after the disambiguation point.

- (1) “The ticket agent admitted the mistake might not have been caught” (Dussias & Cramer Scaltz, 2008, p. 505).

Some studies have manipulated the plausibility of the noun following the first verb to investigate temporary ambiguity resolution (known as “garden-pathing”). For example, in the version of example (2) with “milk,” an initial parse of “milk” as a direct object, rather than as the subject of a new clause, would require reanalysis on encountering “disappeared” (the disambiguation point) to reach the correct interpretation.

- (2) “As the girl drank the milk /dog disappeared from the kitchen” (Roberts & Felser, 2011, p. 328).

In this example, the parser encounters an optionally transitive verb (drink) and so can expect an object. However, in the version with “dog,” the parser might slow down because this object does not fit with the semantics of “drink” (especially in the absence of punctuation and prosody). However, because it is an implausible direct object of “drink,” it is more likely than “milk” to receive a correct initial parse, that is, as the subject of an upcoming coordinating clause. Thus, sentences containing nouns that are implausible as objects may result in quicker recovery in the disambiguating region (“disappeared”) compared to nouns that initially seemed plausible objects.¹ That is, patterns of RTs indicate sensitivity to verb semantics and arguments, and this sensitivity may vary as a function of similarity/difference between features in the L1 versus L2, or native/nonnativeness.

The relevant point for the current *methodological* review is that decisions about *which* words or segments to manipulate and analyze should be reported explicitly and be broadly systematic across studies investigating related phenomena. Thus, the choice and reporting of which region to analyze is critical to construct validity in SPR research and, as such, is one of the features we examine in our review.

Underlying constructs: Processing and knowledge types

Another more common technique used to elicit sensitivity to morphosyntax in the input is the grammaticality (or acceptability) judgment test (JT). Compared to JTs, SPRs allow researchers to determine with more precision the moment where difficulty (or processing cost) or facilitation (processing ease) arises, without seeking an explicit and offline judgment. Researchers use this information to infer that a representation (of, e.g., morphosyntax or lexicon) is sufficiently well established in a participant’s mind for them to demonstrate sensitivity to it, without intentionally drawing on awareness or explicit knowledge (see Vafaei, Suzuki, & Kachisnke, 2016). Thus, one reason that researchers turn to SPRs is that they are thought to provide a window into implicit processing and, possibly, into learners’ implicit

underlying linguistic knowledge representations. However, many L2 researchers recognize a distinction between processing and knowledge. Within this position, investigating online processing per se does not predetermine a particular assumption about the type of knowledge or the nature of linguistic representations that processing mechanisms draw on. Consequently, SPRs can and have been used by researchers with a range of theoretical perspectives (e.g., generative and cognitive).

Related to the issue of the knowledge constructs being elicited is the fact that SPR, by definition, is both in the written modality (in contrast to self-paced listening; see Padapdopoulou, Tsimpli, & Amvrazis, 2013) and is not time constrained by the researcher (i.e., untimed, in contrast to RSVP). There is evidence that these test characteristics are more likely to allow access to awareness and even explicit knowledge (Ellis, 2005; Kim & Nam, 2016; Spada, 2017; Vafaei et al., 2016). In addition, the early stages of reading itself begin as conscious processes, and can be accounted for by skill acquisition theories (DeKeyser, 2015; Laberge & Samuels, 1974; Tunmer & Nicholson, 2010).

Although not a full account of these issues, we have touched upon them as they informed our decision to code certain features: (a) the rationales discussed for using an SPR and (b) the extent to which authors discussed the nature of knowledge and processing (e.g., implicit, explicit, or automatized). They also informed our decision to examine design features that can affect participants' attentional focus and awareness of the target of the test: (c) the use of other instruments (e.g., JTs) in the same study and (d) the focus of comprehension questions (if used) on particular words in the sentences in relation to the target feature.

METHODOLOGICAL SYNTHESIS IN SECOND LANGUAGE RESEARCH

A number of useful narrative discussions of different online data collection techniques exist. These include Frenck-Mestre (2005) and Roberts and Siyanova-Chanturia (2013) on eye-movement techniques; Kotz (2009) on event-related potentials and functional magnetic resonance imaging; and Bowles (2010) and Leow, Grey, Marijuan, and Moorman (2014) on concurrent think-alouds. Several publications also focus on online sentence processing techniques (Jegerski & VanPatten, 2013; Keating & Jegerski, 2015; Marinis, 2010; Roberts, 2012; Witzel, Witzel, & Forster, 2012), with one that focusses uniquely on SPR (Roberts, 2016). The present study differs from these in its exclusive focus on SPR and, critically, the comprehensive and systematic nature of our approach: *methodological synthesis*.

In methodological synthesis, unlike other types of synthetic and meta-analytic research, the focus is not so much on aggregating substantive findings but, rather, on the methods that have produced them. In doing so, this approach draws heavily on the synthetic ethic developing in applied linguistics (Norris & Ortega, 2006); it is also closely tied to the methodological reform movement taking place in the field and efforts to understand and investigate "study quality" (Plonsky, 2013).

Methodological synthesis involves collecting a representative or exhaustive sample of studies with a common interest, which are then coded systematically for different study features, research practices, and so forth. This procedure has been used to examine methodologies *within* large substantive domains, such as interaction in second language acquisition (SLA; Plonsky & Gass, 2011), written corrective

feedback (Liu & Brown, 2015), and task-based learner production (Plonsky & Kim, 2016). Methodological syntheses have also looked *across* domains focusing on a particular technique, procedure, or set of practices, such as: designs, analyses, and reporting practices in quantitative research (Plonsky, 2013); classroom experiment designs (Marsden & Torgerson, 2012); factor analysis (Plonsky & Gonulal, 2015); and instrument reporting practices (Derrick, 2016).

The methodological syntheses carried out to date in applied linguistics have provided a number of insights derived from describing and evaluating their domains of inquiry. Findings include underpowered samples, a lack of demographic diversity, and, in terms of analyses, an overreliance on techniques that are not always appropriate to the data or research questions (Plonsky & Oswald, 2017). Of additional concern is the lack of transparency about both instrumentation (e.g., Derrick, 2016; Marsden, Mackey, & Plonsky, 2016) and data and analysis reporting practices, which are critical to enable consumers and synthesists to capitalize on reports (Norris & Ortega, 2000).

To our knowledge, only two *systematic* reviews of individual data elicitation techniques in applied linguistics have been conducted to date: Bowles's (2010) meta-analysis of reactivity in think-alouds, and Yan et al.'s (2016) meta-analysis on the validity of elicited imitation tests (see also Lai et al.'s 2013 review of eye-tracking in the wider domain of education research). These studies provide comprehensive data on the methods they target. No reviews of this nature exist for SPR. The current synthesis aimed to produce such a review by providing a comprehensive examination of the amount, purpose, scope, nature of usage, reporting, and transparency of SPRs.

RESEARCH QUESTIONS

The ultimate goal of the study was to provide an empirical evidence base regarding the use of SPR in L2 research that could help to improve the rigor and scope of future research. Within the domain of L2 research that is reported in journal articles, the following research questions guided the study:

- RQ1: How much L2 research using SPR is there, and what are its stated aims and rationales?
- RQ2: What are the study and participant characteristics in L2 SPR research?
- RQ3: What are the L2 SPR instrument design characteristics?
- RQ4: How are L2 SPR data cleaning and statistical procedures carried out and reported?
- RQ5: What is the extent of L2 SPR instrument transparency?

METHOD

The present study adheres to best practices in research synthesis at all stages, including searching for studies, clarifying inclusion/exclusion criteria, piloting the coding scheme, and analyzing synthetic data.

Study selection

We aimed to find all peer-reviewed journal articles reporting the use of one or more SPRs in a study investigating second language, foreign language, or bilingual

(but not child bilingual) participants. Following Plonsky and Brown (2015), we searched a variety of sources: Linguistics and Language Behaviour Abstracts (LLBA), PsycInfo, IRIS (Marsden, et al., 2016), and the L2 Research Corpus (a collection of around 8,000 articles from 16 journals from 1980 until the present day held by Plonsky). There was no a priori start date; the search concluded in March 2016 (this did not include studies that were only in online format by this date as LLBA and PsycInfo do not index them). Any studies published in journals were eligible for inclusion. We recognize that this may render our syntheses susceptible to a certain type of publication bias. A number of book chapters (e.g., Bannai, 2011; Fernández & Souza, 2016; Suda, 2015; White & Juffs, 1998) and eight doctoral dissertations were excluded. However, we believe that our journal-based sample is representative of the population of research employing SPR. Further, this approach provides for enhanced systematicity and replicability (Plonsky & Derrick, 2016; Plonsky & Gass, 2011). In any case, publication bias was less of a concern in our study as we were not aggregating substantive effects as in a meta-analysis. In addition, we wished to focus on usage and reporting of SPRs that have been approved through the journal peer review system.

After various trials, our ultimate search terms for LLBA and PsycInfo were (self-paced reading OR subject-paced reading OR moving window) AND (learning OR acquisition OR biling* OR language OR multiling*), with “peer-reviewed” checked. This resulted in 384 hits in LLBA and 250 hits in PsycInfo. The L2 Research Corpus yielded an additional 14 studies that had not been found by LLBA or PsycInfo because those databases do not search the full texts (just the title, abstract, and keywords). After eliminating duplicates and excluding any studies that focused only on L1 acquisition or child bilinguals, the final sample consisted of 64 studies, reporting a total of 74 SPR tests. Included studies are marked in the references with*.

Coding

Our data collection instrument, a coding scheme, can be found in full in Appendix A (online supplementary material) and on the IRIS database (iris-database.org). Most items were categorical (e.g., absent/present; English/French/Spanish etc.; gender/tense etc.), although a few allowed for open-ended text (e.g., rationales for SPR use).

The scheme was developed through a process of rigorous piloting by the authors, involving 10 iterations, with additions and refinements of categories, definitions, and values within coding parameters at each stage. The initial scheme was informed by previous literature (Jegerski & VanPatten, 2013; Keating & Jegerski, 2015; Roberts, 2016) and was used by the three authors to independently code two randomly selected studies (Amato & MacDonald, 2010; Roberts & Liszka, 2013). Disagreements were resolved and unclear codes amended. The revised scheme was then used by the second author to code nine studies, and further refinements were discussed with the first author. Each study was then coded by the second author.

To check coding reliability, a second coder, who was not involved in the development of the coding scheme but who had considerable training and experience

in meta-analytic research, was trained to use the scheme. He then independently coded 15 (20%) of the 74 SPRs. These SPRs were chosen quasi-randomly, ensuring they came from different studies. For just 6 of the 121 coding categories agreement fell below 75%; for these, the first coder either amended her coding or requested the second coder reconsider. We then recalculated agreement, and all categories reached at least 80%. In terms of Cohen's kappa (κ), out of 121 coding parameters (113 of which allowed κ to be calculated) all $\kappa \geq 0.63$, with just 6 exceptions exhibiting κ of 0.48, 0.31, 0.52, 0.55, 0.44, and 0.36, but high agreement rates of 93%, 80%, 93%, 87%, 87%, and 80% respectively. We attribute these apparent discrepancies between percentage agreement and κ s to the very high consistency of values within those items (e.g., almost all "zeros"), which leads to overly conservative κ estimates. The final overall mean agreement was 94%, with a mean interrater reliability of $\kappa = 0.86$. See Appendix A (in online Supplementary Materials) and www.iris-database.org for % agreement and κ on each coding category. To benchmark this against other methodological syntheses, Plonsky (2013) reported an interrater reliability agreement rate of 82%, $\kappa = 0.56$, Plonsky and Derrick (2016) $\kappa = 0.74$, and Marsden, Morgan-Short, Thompson, and Abugaber (2018) 89% agreement and mean $\kappa = 0.80$.

FINDINGS AND DISCUSSION

We present our findings below organized according to our research questions. Given the wide range of issues covered by methodological syntheses and space constraints, we also include most of our discussion in this section. This approach, though somewhat nontraditional, allows us to present interpretations of our findings in closer proximity to their associated data. Given the number of unique quantitative results, we felt that this style of presentation would be helpful and more efficient than the standard approach.

RQ1: How much L2 research using SPR is there, and what are its stated aims and rationales?

Our search revealed a total of 74 SPRs in 64 individual articles (7 of which used multiple SPRs) used in L2 research. The majority of these studies ($k = 42$) were published since 2010, illustrating the increasing popularity of this technique (Figure 1). The earliest example of L2 SPR appears in Juffs and Harrington (1995), approximately 20 years after the early L1 SPR studies. Our sample spans 21 journals, with most published in *Applied Psycholinguistics* (14), *Bilingualism: Language and Cognition* (9), *Studies in Second Language Acquisition* (9), *Language Learning* (8), *Second Language Research* (8), and a small number in other journals.

Rationales given for using SPR: Knowledge and processing. A total of 52 studies included some rationale for using an SPR (beyond a general interest in examining online processing). We found a total of 129 individual (tokens of) rationales. These were first coded "bottom-up" to extract keywords, and we then searched for these keywords across all articles. This produced the seven main themes, shown in Table 1. Two broad types of rationales emerged: one relating to learner

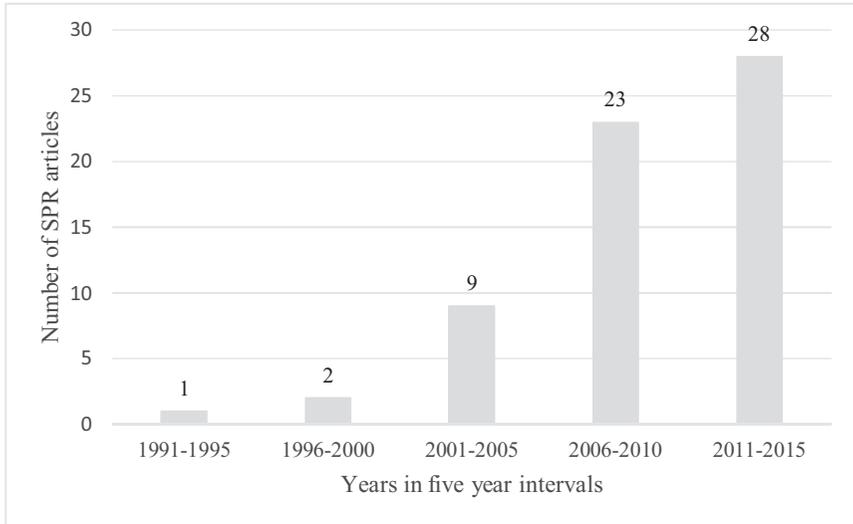


Figure 1. Number of journal articles reporting SPRs over time. The year 2016 is excluded because this review only went up to the first quarter of 2016 (three articles).

knowledge (40 tokens across 26 articles) and one relating to processing mechanisms or phenomena (89 tokens across 57 articles). Twenty-six articles referred to both knowledge and processing. Twenty-three articles used the word “processing” alone to explain their use of the technique.

Although many rationales were given related to implicit knowledge and processing, we found little in-depth discussion of the nature of knowledge or processing. We found no challenges to the notion that SPRs in L2 research are a measure of implicit knowledge, and no discussion of a potential role for awareness or attention. When explicit knowledge was mentioned, it was in relation to SPR reducing access to it or in relation to other measures being used in the same study to elicit a different type of knowledge to the SPR. This perhaps reflects a consensus that reactions in SPRs are deemed to operate below the level of consciousness, though empirical validation of this would be useful. For example, some have argued that conscious thought can occur 300 ms after registration of a stimulus (Dehaene, 2014). SPR has clearly been used to investigate relations between offline knowledge and online processing, as reflected in the 29 studies mentioning both in their rationales, and this was often manifested in studies that incorporated other measures alongside an SPR. However, we did not find any studies looking at the concurrent development of processing and L2 knowledge over time (discussed below).

The broad aims and processing phenomena investigated. The key aim for 34 out of the 64 studies was to investigate differences between native and nonnative online processing. The vast majority of these (30/34) used a native comparison group within the same study (the remaining 4 compared their findings to previous studies that used different SPRs or other measures).

Table 1. *Rationales provided for using an SPR*

Broad rationale type	Key words given in rationale	<i>k</i> number of studies	Example comments
Knowledge	Automatic knowledge/automaticity	18	“SPR ... offers a more objective way to determine whether certain linguistic knowledge is an integrated part of one’s automatic competence.” (Jiang, 2004, p. 610)
	Implicit knowledge	8	“... the study incorporated ... a SPRT ... which should draw on IK [implicit knowledge] to a greater extent [than JTs].” (Vafaei et al., 2016, p. 423)
	“Not” explicit knowledge (expressed directly, by stating SPR avoids explicit knowledge, or indirectly, by using another test to elicit explicit knowledge)	14	“To determine whether L2 learners had explicit knowledge of number agreement in object clitics, the participants also completed an acceptability judgment task.” (Coughlin & Tremblay, 2013, p. 627)
Processing mechanism	Implicit processing	5	“... an accepted psycholinguistic tool for getting at implicit processing of language (Mitchell, 2004).” (VanPatten, Keating, & Leaser, 2012, p. 118)
	Online (including online processing, time windows of processing, sentence processing, L2 processing, real-time processing)	49	“self-paced reading... to examine reading processes ... and to reflect different time windows of processing” (Bultena et al., 2014, p. 1220)
	Facilitation effects (including shorter RTs)	11	“... cognate should facilitate reading if L1 and L2 are activated simultaneously, if not (serial processing) then there shouldn’t be any difference in RTs.” (Mazico & Bajo 2006, p. 4)
	Processing difficulty	24	“The rationale for the SPR task is that increased reading times for a particular segment (relative to the same segment in a control condition) indicate a relatively higher processing difficulty at this point during the parse.” (Papadopoulou & Clahsen, 2003, p. 13)

Twenty-one studies investigated both cross-linguistic influence and also differences between native and nonnative online processing. Nineteen of these used a native-speaker group for comparison, and 11 used different L1 groups for comparison.

Five studies had the sole key aim of investigating cross-linguistic influence: one of these had more than one L1 group as a between-subject factor, whereas four addressed this question without an L1 comparison group (three used SPRs in the participants' L1 and L2, and one manipulated the similarity of L2 verbs to those in the L1). One (of these five) also had a native comparison group.

Four studies had other aims: one used an artificial language to investigate the early stages of acquisition, one used novel words to investigate vocabulary learning, one investigated the effect of translation and repetition, and another validated measures of implicit and explicit knowledge.

Of the 74 SPR tests, the majority (40) were used to investigate the processing and resolution of ambiguities (13 global and 27 local/temporary/garden path ambiguities). Global ambiguities remain after the reader has processed the entire sentence (e.g., "Peter fell in love with the daughter of the psychologist who studied in California"; Dussias, 2003, p. 541), whereas local/garden path ambiguities result in an initial syntactic misanalysis and are then disambiguated at later a point in the sentence (e.g., "After Bill drank the water proved to be poisoned"; Roberts, 2016, p. 59).

Twenty-two SPRs were used to investigate the processing of and sensitivity to anomalies, seven of which investigated multiple features (e.g., gender and number; Sagarra & Herschensohn, 2010, 2011). Of those investigating just one feature, the most common was gender ($k = 6$), then number ($k = 4$). Other commonly investigated features (some combined with other features) included tense (3), aspect (2), person (4), and number (7).

Twelve other SPR tests did not clearly fall into any of those three categories (global ambiguity, local ambiguity, or anomaly). Three investigated syntactic distance dependency (Amato & MacDonald, 2010; Coughlin & Tremblay, 2013; Marinis et al., 2005). Others investigated how cognates affect processing (Bultena et al., 2014; Ibáñez et al., 2010), the effect of text type on reading speed (Lazarte & Barry, 2008; Yamashita & Ichikawa, 2010), the plausibility of collocations (Lim & Christianson, 2013), and novel word learning (Bordag, Kirschenbaum, Tschirner, & Opitz, 2015).

Other instrumentation used alongside SPRs. Of the 74 SPRs, 57 (77%) were used in coordination with other instruments (Table 2). Almost half of the studies (31) used a JT, enabling researchers to investigate relationships between online processing and offline performance. Of the 31, 14 did so in an "integrated" fashion. That is, a JT item was provided after each SPR trial, prompting participants to indicate acceptability or plausibility of some morphosyntactic feature. This is a critical design decision, as orienting attention on particular features, in ways that participants might anticipate across trials, may affect response times and raise awareness of the target feature being tested. For example, Havik et al. (2009) found that when L2 learners (particularly those with a higher working memory

Table 2. *Number of SPRs used with another instrument*

Other instrument	<i>k</i>
Separate judgment task	17
Working memory/reading span test	15
Integrated judgment task	14
Other test	7
Lexical decision task	2
Eye tracking	1
Semantic priming task	1

capacity) made judgments after trials, they manifested similar patterns of RTs to native speakers.

JTs were also administered *after* SPR tests in 15 studies, and in just 2 studies a JT *preceded* an SPR. Study design in this respect seemed largely in line with Keating and Jegerski's (2015) observation that explicit JTs should be administered after SPRs. The rationale behind this is that JTs that precede or are integrated within SPRs may raise the participants' awareness about a study's target.

However, none of the studies used a measure to determine the nature or magnitude of awareness during the SPR tests, such as retrospective subjective measures or knowledge source judgments (Rebuschat, 2013). Thus, despite SPRs being untimed and written, and tapping into a process (reading) that is explicit in its early stages, it remains for future research to investigate the extent to which participants are aware during SPR of the linguistic focus of the study. Collecting such information will be especially important if SPRs are to be considered measures of implicit processing or knowledge (see Vafaei et al., 2016).

RQ2: What are the study and participant characteristics in L2 SPR research?

Contexts and languages. Thirty-four of the 74 SPRs were used in second language and 33 in foreign language contexts, and 6 were used in both contexts (i.e., in two or more sites). One study used an artificial language. The vast majority of participants were university students (59/74 SPRs). Fifty-four studies included instructed language learners and at least 2 were students from education or applied linguistics departments. Such participants likely possess a specific nature of language competence (Hulstijn, 2015) and above average meta-linguistic knowledge (Roehr, 2008), which may affect the nature and speed of reading processes (as suggested by Keating & Jegerski, 2015, p. 27).

Table 3 shows the range and frequency of L1 and target languages investigated. Fourteen of the 20 studies using learners with different L1s used the L1 as a between-subject variable (Table 4). The other 6 studies grouped learners into a single group, regardless of their L1.

Participant sample sizes. Whole-study sample sizes ranged from 12 (Macizo & Bajo, 2006) to 133 (Sagarra & Herschensohn, 2010), with a mean of 46.58

Table 3. *Languages used in SPRs*

Language	Number of SPRs (<i>k</i>) with participants with L1...	Number of SPRs (<i>k</i>) with participants with L2...
English	9	43
Chinese	9	2
Spanish	5	11
Greek	4	1
German	3	6
Korean	3	1
Dutch	3	2
Japanese	3	0
French	1	2
Russian	1	0
Arabic	1	0
Multiple Ls	20	0

Table 4. *Types and numbers of comparisons between different L1s*

Combination of L1s	Number of SPRs
English and Chinese	2
English and Spanish	1
English and Dutch	1
English + 2 others	2
German + 2 others	1
Japanese + 1 other	2
Japanese + 2 others	2
Japanese + 3 others	2
Spanish + 3 others	1

($SD = 26.12$, median = 43.5). Subgroup sample sizes ranged from 10 to 69, with a mean of 26.91 ($SD = 11.15$, median = 24). This is somewhat higher than Plonsky’s (2013) finding, from 606 primary studies, of a median subgroup sample size of 19, and yet lower than his median whole study sample size of 60.² This difference might be due in part to the fact that administering SPRs can be done relatively easily in groups in labs. None of the studies reported an a priori power analysis, and very few reported effect sizes, which, among other benefits, would facilitate subsequent power analyses.

Of the 36 SPRs used to compare multiple groups, 14 had the same sample size across groups. However the mean difference between group (subsample) sizes was 8.9 ($SD = 10.4$, range 1–50). Such sample size differences may require specific statistical techniques (e.g., Games Howell post hoc paired comparison tests for nonequal sample sizes, or nonparametric tests). We did not find any studies that explicitly addressed unequal sample sizes in their analyses.

Participant proficiencies and study design. Participant groups were labeled by the studies' authors as beginner (6), intermediate (18), advanced (53), near native (6), and bilingual (10).³ This is not typical of the general propensity for L2 research to oversample intermediate learners (see Plonsky, 2013) and shows the relative neglect of online processing research among lower proficiency levels. As well as ease of participant recruitment, this might be for several reasons. One might be the underpinning assumption that SPRs tap into comprehension processes, and successful comprehension is more likely among higher proficiencies. Another might be that one of the research aims that drove the use of SPR among L2 researchers (i.e., fundamental differences between native and nonnative processing) is thought to require high proficiency/high exposure to have given the SLA process maximum opportunity to reach an "end state." However, it is of course possible, and of theoretical and pedagogical interest, to investigate online processing among less proficient learners (e.g., manipulating the comprehensibility of the stimuli). In our study sample, a relatively low number (16/64 studies) used a cross-sectional design using proficiency as a between-group factor. The majority of these compared what the authors referred to as "intermediate" and "advanced" learners (12 studies).

We found no examples of longitudinal research using SPR, defined as within-subject comparisons over time on the same SPR.⁴ However, the number (16) of *cross-sectional* studies may reflect a growing interest in the developmental trajectory of online processing. Nevertheless, this number is surprisingly low, given the interest in the role of processing as a driver in the acquisition process (Chang, Dell, & Bock, 2006; O'Grady 2005; Philips & Ehrenhofer, 2015; Seidenberg & MacDonald, 1999). In our studies, we found little discussion of interfaces between processing/learning/knowledge, perhaps a reflection of SPRs being initially employed in L2 research under the premise that offline knowledge (such as access to a Universal Grammar, often elicited via JTs) was distinct from online behaviors. Thus, it remains to be explored the extent to which SPR has potential for investigating whether processing or anticipatory effects have a causal role in driving and constraining acquisition, or are more of a "symptom/product" of other acquisition mechanisms (e.g., Foucart, 2015; Huettig & Mani, 2016; Kaan, 2015).

We add a note of caution to our findings about proficiency levels. In terms of the measurement of proficiency, it was reassuring to find only three studies that just used educational level to *assume* proficiency and none that just used self-rating (i.e., the vast majority of studies used a measure to select or group participants). Thirty-four studies reported one proficiency indicator, detailed in Table 5, and the remaining studies used more than one.

A good number of studies used standardized proficiency tests, though there was a wide range, even within one language: for English: TOEFL (3), IELTS (3), The MELAB (3), and Cambridge Proficiency Test (2). Across all articles, 36% did not report using a standardized test. Thirty-three percent used a measure adapted or designed specifically for the study, though did not report native speaker scores or whether it was a single test or a battery of tests, which can be indicators of measurement validity and reliability (Hulstijn, 2010). Determining proficiency level remains an important endeavor to help comparability and replicability across primary studies (as noted by Bowden, 2016; Norris & Ortega, 2012).

Table 5. *Studies using a single indicator of proficiency (k = 34)*

Type of test used	Number of studies
Standardized proficiency test	21
Assumed from educational level	3
Test adapted from standardized proficiency test	4
Test specifically designed for the study	3
Other measure	2 (tests developed from cited past research)
Artificial language (so no test needed)	1
Self-rating	0

RQ3: What are the L2 SPR instrument design characteristics?

Development of stimuli. Twenty-five SPRs reported using or adapting materials that had been used in previous published studies, perhaps reflecting a relatively strong systematicity within research agendas using SPR and/or a healthy collaborative ethic within the SPR community (though see section on transparency below).

Thirty of the 74 SPRs were reported as having been checked for plausibility, acceptability, or grammaticality before the main study, as part of stimuli development. We found inconsistencies in nomenclature of this stage of stimuli design. For example, 3/30 referred to these procedures as “norming,” 2 as “piloting,” and 1 as “base-line” tests. Of the 22 studies investigating anomalies, just 3 reported checking stimuli prior to testing, 1 of these altering the stimuli to be “more natural” or “unambiguously grammatical or ungrammatical” after native speaker feedback (Vafae et al., 2016, p. 17). Checking perceived naturalness with native speakers could be particularly important if using this population as a comparison: native speaker sensitivity to unnatural (though grammatical) language may affect RTs.

Frequency of lexical items across conditions (such as grammatical/ungrammatical, ambiguous/unambiguous, plausible/implausible, high/low attachment) can also affect RTs (as discussed by Keating & Jegerski, 2015). In 26/64 studies, lexical frequency was addressed in some way, either by design, descriptively, or statistically. For example, 16 studies in the sample consulted corpora to select words from specific frequency bands.

Noncritical items. Keating and Jegerski (2015) define distractors as “intentionally designed to contain a specific linguistic form or structure [...] to counterbalance some characteristic of the critical stimuli that might otherwise make them stand out to the participant” (p. 16). Fillers are defined as “unrelated sentences that are not intended to elicit any specific type of processing effects” (p. 16). In our sample, nomenclature varied, with the terms “filler” and “distractor” being used interchangeably across studies and sometimes within studies. In Table 6 we report the frequency of terms as used by the authors.

Across all studies, the mean number of critical items was 43.26 ($SD = 32.26$), of fillers and/or distractors 50.55 ($SD = 43.45$), and practice items 5.01 ($SD = 5.20$).

Table 6. *Balance of critical and noncritical items*

Type of stimuli, as reported by authors	SPRs ($k = 74$)
Critical items	74
Only fillers	48
Only distractors	6
Fillers and distractors	8
Practice items	42
None reported (fillers or distractors)	12

Using too few noncritical compared to critical items may raise awareness of the experimental target. Fifty-seven out of the 62 SPRs that had noncritical items included 50% or more noncritical items (fillers and/or distractors) compared to critical items, which falls in line with Keating and Jegerski's suggestion of the minimum ratio of noncritical to critical items, though by no means always met their recommendation of 75% noncritical sentences. Furthermore, the mean difference between numbers of critical and noncritical items (fillers or distractors) was -0.30 ($SD = 56.92$), ranging from -224 to 148 . This range indicates a need for research to investigate the effects that this ratio has on results, with a view to providing an evidence base for more standardization in this design decision.

Length of conditions and stimuli. In addition to overall instrument length, which may cause participant fatigue and thereby threaten a study's internal validity, other design characteristics that can affect construct validity include the number of items and lists in relation to the number of conditions, and the length of sentences, segments, words, and critical regions.

The recommended ratio of 8 to 12 items per condition (Keating & Jegerski 2015) is thought to address the fact that too many items per condition can fatigue participants or make them accustomed to structures or features and thus show less sensitivity to the manipulations.⁵ In contrast, too few items per condition does not provide sufficient data for many statistical procedures. According to this recommendation, a study with four conditions requires 32–48 items; this was met by 15 of the 44 studies that used four conditions. The other studies with four conditions used either between 6 and 31 or between 49 and 114. The range of items per condition is presented in Table 7, and again demonstrates that an evidence base for standardization would be helpful.

Sentence length. Sentence length can affect processing ease as the start and end of sentences are thought to place the least burden on working memory (Pienemann & Kessler, 2011). However, sentence length was not reported in almost half the studies (30/64), so it was not always possible to ascertain whether the analyzed words occurred at the same point in each sentence, particularly problematic for 22 of these 30 studies that did not provide the full stimuli. In studies that reported sentence length, lengths were not always uniform across trials within a study (e.g., 1 study reported sentences ranging from 9 to 15 words).

Table 7. *Numbers of items, lists, and conditions*

No. of conditions (<i>k</i> of SPRs/71) ^a	Range of no. of lists (<i>k</i> where not reported)	Recommended no. of items per SPR	Actual range of no. of items across SPRs	No. of SPRs within recommended range
2 (11)	1–4 (3)	16–24	10–60	5/11
3 (6)	1–4 (1)	24–36	30–60	2/6
4 (44)	1–8 (3)	32–48	6–114	15/44
6 (4)	3–4 (2)	48–72	36–78	2/4
8 (5)	2–8 (1)	64–96	40–80	2/5
9 (1)	1 (0)	72–108	54	0/1

Note: ^aThree studies were excluded due to the design not requiring conditions or because the number of conditions was unclear.

Length of presented segments (including single words). The most common presentation length was word-by-word (46/74). Out of the 28 SPRs that used multi-word segments, 5 stated that individual segment length had been controlled, and 6 that the number of segments had been controlled. The other 17 did not report the length or number of segments, again particularly problematic when full stimuli are not provided.

Word length. Five out of 74 (7%) reported controlling for the number of syllables in each word. Four of these 5 gave a range of syllables per word, such as 2–4 (Sagarra & Herschensohn, 2010). One used a *t* test to compare syllable length (Macizo & Bajo, 2006).

Comprehension questions: Attentional focus during sentence processing. A central tenet of SPR tests is that participants try to comprehend what they are reading. This is perhaps particularly important if the intention is to elicit implicit processing and knowledge, so participants may be conscious of extracting meaning but not structure or form. The majority of the SPRs included comprehension questions (CQs; *k* = 57, 77%), with various rationales: 18 gave a rationale of ‘checking understanding,’ 17 of ‘ensuring that participants were paying attention/on task,’ 5 gave both reasons, and 17 gave no (clear) reason. Keating and Jegerski also suggest analysing RTs on responses to CQs. Two in our sample of studies analyzed RTs on all CQs and 3 on CQs following fillers only.

CQs can repeatedly focus participants’ attention on specific regions of sentences by repetitively, over many trials, asking about the meaning of the same region of the sentences. Thus, the region that CQs focus participants’ attention on is critical for construct validity (as raised awareness about specific regions can affect RTs and claims about implicitness or orientation to sentential meaning). A few researchers have intentionally aimed to focus participants’ attention on and check interpretation of the region that is analyzed for RTs. Others aim to do the opposite. That is, the CQ is not intended to draw (repeated) attention to the regions that are analyzed, so that slower RTs cannot be ascribed to paying special attention for the purpose of

answering the CQ. To investigate these features, we set out to examine the CQs in relation to the analyzed regions. Twenty-five of the 50 studies using CQs provided no example of the CQs (7 of these described them as yes/no questions). Of the other 25 studies, 21 provided one example of a CQ that followed a critical trial. However, one isolated example (or even two) does not enable the reader to determine the nature of the CQs across the critical trials or whole test. Four studies provided multiple examples. These are given in Appendix B, alongside the SPR trial that the CQ followed, as well as the critical region analyzed (CRA), and a commentary is provided on the relation between the focus of the CQ and the CRA. This small subset of studies showed a mixed picture of design choices. One set of CQs focused on words within the CRA, as intended in the study, because interpretation of the CRA was central to the research questions; another set of CQs also focused on words within the CRA though it was not clear whether this was intentional; one set sometimes focused on words in the analyzed region and sometimes not; and for one set, the focus of the CQ was not discernible.

Despite the lack of attention and clarity on this issue, it is central to the construct validity of SPR tests, affecting claims about whether the critical region was understood, and whether participants became aware of the target feature or (de)sensitized to a particular anomaly. The nature of the CQ also determines decisions about which trials to analyze, such as only those where the CQ was answered correctly (see below). Of most relevance here is that we were only able to discern relevant details from those studies that provided sufficient examples of their stimuli.

To sum up findings related to RQ3, it seems that having full sets of stimuli and CQs available would allow researchers, reviewers, editors, and would-be replicators to better evaluate study and instrument quality, compare across studies, and design future SPRs.

RQ4: How are L2 SPR data cleaning and statistical procedures carried out and reported?

Data cleaning. SPR data must be examined for statistical outliers as outliers can heavily influence subsequent analyses, especially for null hypothesis significance tests, such as analyses of variance (ANOVAs), commonly applied to SPR data. We therefore coded for any discernible patterns or norms of practice. Of the 48 studies that reported removing outliers, in order to identify those outliers 20 studies used participant RTs, 11 used item RTs, and the other 17 used both participant and item RTs. Twenty of these 48 studies reported using *SDs* above or below the mean to identify outliers, with a mode 2.5 *SD* ($k = 8/20$), 17 studies used predetermined millisecond cutoff ranges, and 11 studies used both *SD* and cutoffs. The smallest lower cutoff was for RTs <100 ms and the largest upper limit was RTs >25000 ms (for total trials). The modal range was 200 ms to 2000 ms (4 studies), and the modal lower cutoff was <200 ms (3 studies). The mean reading speed of a native speaker has been found to be around 250 ms/word (Milton & Fitzpatrick, 2013), and such information might inform future empirical investigations into principled elimination of unnaturally fast key presses (as suggested by Conroy & Cupples, 2010). The upper cutoff ranged from >2000 ms to 20000 ms, with wide variability across studies, with a mode of >2000 ms (7 studies). As proficiency affects reading

speed, it might be that cutoffs vary between studies using participants of different proficiencies, although no such discernible pattern emerged from our study sample: 5 of the 7 studies using RTs >2000 ms tested advanced learners; in the 6 studies with intermediate participants that used RTs to trim data, the upper cutoff ranged from >2000 ms to >20000 ms; in the 3 studies testing beginners, the upper cutoff ranged from >2000 to >3500.

In sum, research with SPRs would benefit from empirically based norms for the identification of outliers, as variation in this respect could affect the comparability of results between studies.

Incorrect responses to CQs were also used to remove trials or participants. Twenty-two out of the 50 studies using CQs analyzed only the items with correct responses, whereas 14 studies gave a specific accuracy rate (e.g., over 80%) for a participant's data to be included. Nine studies analyzed all data regardless of the correctness of responses. Some studies that *included* data for trials followed by incorrect responses provided reasons such as (a) the L2 participants' responses did not differ significantly from NSs (Hopp, 2016); (b) to avoid a high number of missing values (Jegerski, 2016); and (c) not all items were followed by CQs (Rah & Adone, 2010).

Although two studies have investigated the relationship between CQ error rate and RTs (Keating, Jegerski, & VanPatten, 2016; Xu, 2014), in general, the practice of eliminating trials with incorrect responses probably relates to the tendency, observed earlier, to investigate online processing where comprehension is high. Investigating processing where there are comprehension difficulties (a frequent phenomenon for L2 learners) remains a relatively neglected area of research.

Statistics used. RT data from 58 SPRs were analyzed using ANOVA to identify within-subject effects (e.g., low vs. high attachment, anomalous vs. correct, or L1 vs. L2) and between-group effects (e.g., proficiencies, L1s, or learning contexts). This practice aligns with findings that point to the widespread dominance of ANOVA and its variants when other choices might be more appropriate (Plonsky & Oswald, 2017). Other analyses of RT data included general linear mixed effects models ($k = 7$ SPRs); t tests ($k = 7$), and correlations with confirmatory factor analysis ($k = 1$). Note that if using mixed effects models, the routine removal of outliers is not always necessary (Baayen & Milin, 2010).

Statistical reporting for the majority of SPRs (57/74) did not include effect sizes. Eleven provided eta squared (η^2) or partial η^2 (the often flawed default provided by SPSS) following the ANOVA, and just two studies reported Cohen's d (Kato, 2009; Vafaei et al., 2016). Although η^2 and partial η^2 provide information about the amount of variance accounted for by the omnibus test (e.g., ANOVA; how much group membership explains the dependent variable), there are complications surrounding the use and interpretation of η^2 (Larson-Hall, 2016, p. 149; see also Norouzian & Plonsky, 2018). Most important, the d family of effect sizes provides information about the paired comparisons that are usually of most theoretical interest (and omnibus effects are usually broken down into paired comparisons anyway). While not currently standard practice in SPR research, effect sizes of mean differences between groups or conditions determines the magnitude of difference. This is especially informative where comparisons should receive a nuanced rather

than a dichotomous interpretation, given the multitude of factors we know to affect SLA. Rather than an “absence versus presence” of L1 influence on L2 processing, or “difference versus no difference” between native and nonnative processing, effect sizes such as *d* enable us to interpret the relative size of differences in one study (with one set of learners, on one linguistic feature) compared to another study. Thus far, very few meta-analyses have been done on studies using RTs (and those that have included effect sizes for RT data needed to extract information from primary studies in order to calculate them; e.g., Adesope, Lanvin, Thompson, & Ungerleider, 2010).

Providing effect sizes in future studies based on SPR data would greatly facilitate meta-analyses, power analyses, cross-study comparisons, and more nuanced interpretations. Another concern is that we do not yet have a feel for interpreting effect sizes in studies using RTs from SPRs, for example, whether they are “small,” “medium,” or “large” (relative to the general tendencies presented by Cohen, 1988, or L2 field-specific ones by Plonsky & Oswald, 2014). As we know that different types of instrument tend to yield different effect sizes, this is an important consideration for future research.

Segments analyzed. Decisions about which parts of a sentence are predicted to reveal effects directly relate to the construct validity of the elicitation technique. Analyses are carried out on different segments depending on which regions are deemed critical and whether researchers consider effects may be observed pre-, during, or postcritical regions (e.g., spillover effects). We documented the nature of these choices, as a function of the processing and linguistic phenomena under investigation, to determine the level of consistency across similar studies. Nomenclature was not always consistent. For instance, the term “spillover” was used to refer to a “critical region,” as by Omaki and Schulz (2011, p. 577) and to a “postcritical region,” as by Coughlin and Tremblay (2013, p. 629).

In order to better understand the use of such terms and the phenomena they represent, we extracted the examples of regions analyzed from a subset of articles. We selected studies that investigated the same processing phenomena with the same (or similar) linguistic feature.⁶ Again, we emphasize we do not aim to criticize any individual study, but rather to draw together different analysis choices with a view to illustrating potential benefits of methodological transparency and replication.

Local ambiguity. Twenty-seven SPRs investigated local ambiguity resolution (or garden path), in which the stimuli have an ambiguous region followed by a disambiguating region.⁷ Out of these, we found three groups of comparable studies: four focused on subject/object ambiguity, four on antecedent attachment preferences in relative clauses, and two on reduced relative clauses. See Appendix C for the segments that were presented and analyzed in each of these groups, with detailed commentary comparing presentation and analysis decisions within each group of related studies. A few studies (e.g., Felser, Roberts, Marinis, & Gross 2003; Papadopoulou & Clahsen, 2003) reported having carried out analyses on all regions and, finding no statistical significance (as predicted) prior to the ambiguous region, reported the inferential statistics only from the ambiguous region onward. Some studies (e.g., Roberts & Felser, 2011; Marinis et al., 2005) provided descriptive

statistics (numerically or graphically, respectively) for all regions and carried out inferential statistics only for particular regions, for example, the ambiguity onward. Others presented data and analyses only for the regions that were either predetermined or selected after data collection on the basis of descriptive statistics.

Our close examination of groups of comparable studies (Appendix C) revealed some key similarities but also a number of important differences in analysis regions: five differences between the four studies focusing on subject/object ambiguity; four differences between the three studies on attachment preferences; and two differences between the two studies on reduced relative clauses.

Global ambiguity. Out of 13 SPRs used to investigate global ambiguity resolution, we were able to compare three pairs of studies: one pair focusing on subject–object assignment in German, one pair on subject–object *wh*-questions in English, and one pair on subject–object *wh*-questions in German. See Appendix D for the segments presented and analyzed, with detailed commentary. The pair of studies focussing on *wh*-questions in English analyzed directly comparable regions, whereas the other two pairs each had two differences in their presentation and analysis decisions.

While analysis decisions will inevitably vary to some extent between studies, more similarities might be hoped for so as to allow better cross-study comparisons and future meta-analyses. We found comparability to be threatened for a number of reasons. For example, when the presentation format varied (word by word vs. multiword segments, or different multiword segments), then in one study RTs were the sum (or mean) for one group of words whereas in another study RTs were for different or individual words. These are critical design decisions that can affect parsing behaviors (for discussion, see De Vincenzi & Job, 1995; Gilboy & Sopena, 1996; MacDonald, 1994). Another problem is that where one study found effects in one region, another study did not analyze the equivalent region. One possible way forward to both enhance comparability and not stifle exploratory analyses is, when reporting results, to clearly separate confirmatory analyses, which allow comparison with previous studies, from exploratory analyses, which present new analyses (see Chambers, 2013; Marsden, Morgan-Short, Trofimovich, & Ellis, 2018).

RQ5: What is the extent of L2 SPR instrument transparency?

One of several aspects of transparency that we coded for is the provision of stimuli. The majority of studies (49/64) had only a brief example of stimuli available (e.g., one or two items). Between 2000 and 2009 27% of SPRs were available in full in the article, and the remaining 73% gave just examples in the article (i.e., accessible with journal subscription). Since 2010 the proportion of articles providing full stimuli rose to 46%, though for 54% of articles only example items were available. Table 8 illustrates the transparency of materials.

As yet, no clear relationship between publication outlet and instrument availability is observable. However, this may change as more major journals⁸ begin to recognize authors for fully open methodological transparency, by, for example, adopting the Centre for Open Science badge scheme (Blohowiak et al., 2016), which has been shown to increase the long-term availability of materials and data (Kidwell et al., 2016).

Table 8. *Availability of SPR stimuli*

Location	No. of SPR tests ($k = 74$) available at time of coding [currently]
In full ^a on IRIS	2 [46]
In full in article ^b and IRIS	1 [40]
In full in article	16
Just example(s) in article and IRIS	4 [15]
Just example in article	50
Author's website	1

Note: ^a“in full” means all the critical SPR items, but not necessarily distractors, fillers, comprehension questions, etc. ^bThis is including journal supplementary materials, i.e., behind the journal paywall.

As can be seen in Table 8, as a follow-up to the current study, we sought to establish a “special collection” of SPR materials on IRIS (Marsden et al., 2016) in order to improve materials transparency in this domain. The positive response we had is testimony to the willingness of researchers to engage in collaborative effort. We hope that this collection will serve as a reference corpus for future syntheses and substantive meta-analyses and as a research methods training tool, as well as serve to stimulate and facilitate replication.

Another important feature of transparent reporting about instrumentation is that of measurement reliability. We found two studies that reported reliability coefficients for the RTs, using Cronbach α . Improved reporting of reliability would help our understanding of measurements taken with SPRs, the error in the data, the psychometric properties of SPRs, and future instrument development.

FURTHER DISCUSSION AND FUTURE DIRECTIONS

Our review identified a good deal of coherence in terms of research aims and systematicity of agendas (across languages, processing phenomena, participant proficiencies and ages, and linguistic features). By contrast, this review also found massive variability in the SPRs used to investigate and advance those agendas. To name just a few: theoretical positions and assumptions to motivate the use of SPRs were occasionally, but certainly not uniformly, detailed explicitly. Reporting of some participant characteristics could also be patchy. SPRs were found to be used both with and without JTs and with or without CQs, not always with a clear or consistent rationale to justify these choices. Features of the instruments employed (e.g., number of items, sentence length, segment length, and item:condition ratios) were highly variable or regularly omitted from the report, as were critical data such as measures of internal consistency (i.e., reliability) and effect sizes. Data cleaning procedures varied widely, and regions of analysis in some related studies were also disparate; both of these issues can directly affect the outcomes of an analysis. Equally concerning as the inconsistency and opaqueness that we observed is our poor understanding of how these and many other aspects of SPR design might

actually impact study results. We note that Keating and Jegerski (2015) had warned of a number of these issues. The current study goes beyond those comments, providing quantitative data based on a systematic synthesis of published empirical work to illustrate their pervasiveness and severity.

In concluding the paper, we indicate several directions for future use of this technique that, we believe, will lead to more informative SPR-based findings and interpretations. In doing so, we hasten to note again that we are building on some of the work of Keating and Jegerski but with the empirical support of the current review to motivate our comments.

Enhancing the scope of research agendas using SPR

Our data on study design and participant characteristics suggest several avenues that are currently largely neglected.

Sample demographics of SPR studies are skewed in line with L2 research in general, with a propensity to investigate English as an L1 or L2 (Norris & Ortega, 2000), and we found no evidence to suggest that trends in this respect are changing over time. Similarly, participants tended to be university students, often from language, linguistics, psychology, or education departments, thus limiting our understanding of L2 online reading processes from SPR data to the more highly educated and possibly meta-linguistically aware sections of society.

Perhaps due to the fact that SPRs were initially used in adult L1 research, successful comprehension of every sentence has been assumed to be necessary or at least desirable in most L2 SPR studies, with many researchers removing trials with incorrect responses to CQs. The extent to which sensitivity to morphosyntax changes with comprehension difficulty (e.g., less familiar lexical items) or individual differences (e.g., working memory capacity) seems worthy of future empirical effort (Hopp, 2016; Sagarra, 2008; VanPatten, 2015). One consequence of this is that we found insufficient numbers of studies that would enable a meta-analysis of the relationship between proficiency and processing phenomena. This was partly because there were only 17 studies that compared different proficiencies, with a limited range of proficiencies (advanced/near native vs. native) and with little homogeneity of measures (as noted by Norris & Ortega, 2003; Wu & Ortega, 2013); for example, only 5 of the studies that investigated proficiency used a standardized proficiency test to provide a reliable benchmark for comparisons.

Nevertheless, a bright spot in our findings was relatively high consistency in terms of the two main research agendas addressed using SPRs to date: we found 52 studies that investigated differences between native and nonnative online reading; and 26 studies that investigated cross-linguistic influence. This body of research may be ripe, at least in the not too distant future, for meta-analyses of these two major questions. Despite the challenges that we have raised (e.g., of comparability and transparency), such a meta-analysis would have a very important advantage: it would draw on data from a single elicitation technique, thus avoiding the oft-cited problem of meta-analyses collapsing data from different outcome measures that may tap into different phenomena (i.e., the “apples and oranges” problem).

Though requiring additional effort, effect sizes could be extracted from data in the primary studies, as most provided means, standard deviations, and *n*.

Methodological rigor

It was a positive indication of collaboration that 25 SPRs reported drawing on previously used stimuli. However, full scrutiny of the design of most stimuli was not possible in most cases due to the lack of availability (either in appendices or elsewhere), and reporting did not compensate for this. For example, a comprehensive synthesis of how lexical items are selected was not possible, an important consideration for future research as there is evidence that word length and lexical and collocational frequency (both L1 and L2) can affect reading times (Bultena et al., 2014; Hopp, 2016; Ibáñez et al., 2010). Two ways of addressing such issues are by using letter-length corrected residual reading times (Ferreira & Clifton, 1986; Lee, Lu, & Garnsey, 2013) and/or mixed effects models with item as a random factor (Barr, Levy, Scheeper, & Tily, 2016; Cunnings, 2012). We found 11/64 of L2 studies to date reported using residual reading times and 7 using mixed effects models, indicating there is some way to go to integrate these into our methodological toolkit.

Reporting and transparency

Other issues we observed relate to the reporting and consistency of data cleaning procedures, nomenclature (e.g., piloting and norming), and analysis. We hope to have illustrated the inseparability of methodological transparency and construct validity and reliability.

With respect to the data resulting from SPR tests, reporting the means, standard deviations, and results of all statistical analyses carried out, ideally on segments that are comparable across studies and on posttrial CQs, would facilitate comparisons and future meta-analyses. Reporting of effect sizes in primary studies and comparing these to others (Plonsky & Oswald, 2014) will provide a more accurate and informative depiction of the magnitude of the relationships being investigated.

In addition to more comprehensive reporting, providing the field with access to materials including stimuli (critical, distractors, and fillers), CQs, software scripts, and data cleaning and analysis procedures, would inspire more confidence among reviewers and readers. Improved reporting alone would rarely, if ever, capture all aspects of instrument design (Derrick, 2016), partly because conceptual and methodological innovation usually occur before reporting conventions become established. Greater materials transparency also reduces reinvention of the wheel and, in many cases, helps to build on previous efforts (Marsden et al., 2016). We are a community of researchers, and we owe it to each other to behave like one. Provision of materials also facilitates replications with different sample demographics, target features, contexts, and so forth. While not complete, the special collection of SPRs on the IRIS database has now increased the open availability of full SPR stimuli from 2 to 46.⁹ We hope this will stimulate an expansion of the scope and practice of replicating SPR research.

In terms of analysis, in some cases it was unclear whether the choice of segments for which to present analyses was made a priori (hypothesis testing), or postanalysis (exploratory). Of course, both approaches have their merits. We hope that this review and greater transparency of stimuli and analyses will inform and improve the consistency of future decisions about word/segment presentation and analysis.

Conclusion

One of the most basic findings of this study concerns, very simply, the extent to which L2 researchers have used SPRs. Although not as frequent as, for instance, JTs or cloze tests, SPR is part of the methodological repertoire of a growing number of L2 scholars. The motivation behind this project was to inform these efforts, and although providing a largely retrospective account, we hope to have highlighted some of the many choices inherent in utilizing SPRs. Perhaps most critical, we also hope to have stimulated future empirical examination of the impact of these methodological choices on findings and, consequently, on our ability to account for the findings. Finally, our approach of subjecting the research process to empirical scrutiny at the primary and synthetic levels can certainly be applied to other procedures. Doing so can only serve to promote a greater understanding of and confidence in our methods and findings.

APPENDIX B

Examples of comprehension questions (CQs) with analyzed segments, in studies providing more than one example of a comprehension question on critical trials

Table B.1. *Examples of CQs with analyzed segments, in studies providing more than one example of a CQ on critical trials*

Study	Target feature	Example stimuli critical regions/targets underlined	Example CQ (total <i>k</i> of CQs available)	Commentary: CQ in relation to critical region analyzed (CRA)
Study A ^a	Verb incongruence– plurality (i & ii) and subcategorization (iii & iv)	i. “The child was watching some of the rabbit(s) in the room.” ii. “The professor noticed a few of his friend(s) in the picture.” iii. “The teacher wanted/insisted the student to start all over again.” iv. “They had to teach/train the employees Chinese before sending them to China.”	i. “Was the child chasing some chickens?” ii. “Were the professor’s friends in the picture?” iii. “Was the teacher satisfied with the student’s work?” iv. “Were the employees to be sent to China?” (64, all in appendix, pp. 964–967)	All CQ focus on (all or part of) CRA. Pattern across all CQs.
Study B	Subject object ambiguities– weak garden paths (i & ii) and strong garden paths (iii & iv)	i. “The spokesman confirmed the story/doctor had surprised the president yesterday.” ii. “The man believed the girl/book had upset very many people.” iii. “As the men drank the beer/song pleased everybody very much.” iv. “While the neighbor visited the boy/car passed by the house.”	i. “Had the story surprised the president?” ii. “Had the girl/book upset many people?” iii. “Did the beer/song make everybody unhappy?” iv. “Did the neighbor pass by the house?” (40, all in appendix, pp. 327–329)	All focus on CRA as intended, to measure offline interpretation. These two types of orientation of attention pattern across all CQs.

Table B.1 (cont.)

Study	Target feature	Example stimuli critical regions/targets underlined	Example CQ (total k of CQs available)	Commentary: CQ in relation to critical region analyzed (CRA)
Study C	Gender agreement– animate versus inanimate	<p>i. “El abuelo adora <u>al nieto nuevo/*nueva</u> de la familia. The grandfather adores the family’s new grandson.”</p> <p>ii. “El abuelo adora <u>el auto nuevo/*nueva</u> de la familia. The grandfather adores the family’s new car.” (p. 94)</p>	<p>i. “El abuelo adora el auto/al nieto nuevo? Does the grandfather adore the new car/grandson?”</p> <p>ii. “¿El hombre detesta el auto/al nieto nuevo? Does the man detest the new car/grandson?” (2, p. 95)</p>	Both examples require some focus on CRA, but a pattern not discernible from these two examples and it could be pre-CRA, on the verb.
Study D	Plural inflection of the object– simple quantifier phrases (i & ii) and partitive quantifier phrases (iii & iv)	<p>i. “Haley ordered <u>four brown leather chair(s)</u> for her new condo.”</p> <p>ii. “Erin fixed <u>several old broken desk(s)</u> with some old tools.”</p> <p>iii. “Frank promoted <u>three of the engineer(s)</u> to high management positions.”</p> <p>iv. “Alex gave <u>several of his</u> toy(s) to his baby cousins.”</p>	<p>i. “Was Haley planning to put the leather chairs in her new condo?”</p> <p>ii. “Did Erin repair the desk without any tools?”</p> <p>iii “Did any engineers get promoted?”</p> <p>iv. “Did Alex give his toys only to his classmates?” (all online supplementary materials)</p>	In i and ii, the focus is on a region after the CRA. iii focuses on the CRA and pre-CRA (“promoted”). iv focuses on both the CRA (“several of” is needed for interpretation) and also the post-CRA region.

Note: ^aWe reiterate, the purpose of this synthesis is not to critique individual studies but to highlight potential issues regarding consistency and transparency for further research. Study A in this table is not the same as Study A in subsequent tables.

APPENDIX C

Segments analyzed in studies investigating temporary (local) ambiguity

Table C.1. Segments analyzed in studies investigating temporary (local) ambiguity^z

Linguistic feature	Study	Segments analyzed	Example stimuli (as presented in the article)	Commentary on group of related studies
Subject–object ambiguities in Spanish	Study A	The postverbal NP, the region with the main clause verb, and the sentence final region.	<p>“Cuando el escultor acabó/volvió <u>la obra</u> <u>tenía tres metros</u> <u>de altura</u> When the sculptor finished/came back the piece <u>was ten feet</u> <u>in height</u>” (p. 726)</p> <p>“Después de que comieron/hablaron el pollo <u>se enfrió</u> <u>de una vez</u>. After they ate/talked <u>the chicken</u> got cold <u>right away</u>.” (p. 733)</p>	<p><i>Similarity.</i> All three studies analyzed some or all of the ambiguous noun phrase and the disambiguating verb region.</p> <p><i>Differences.</i></p> <p>(a) Study A presented multiword segments; Studies B, C, and D presented word by word.</p> <p>(b) Studies A, C, and D analyzed words or segments <i>after</i> the disambiguating verbs; Study B did not.</p> <p>(c) Studies A, B, and D analyzed the article with the ambiguous noun, each in different ways: A as the total RT on the multiword segment, B as the mean RT on the article and noun, and D as the sum of the RT on article and noun and also included the verb; study C did not analyze the RT on the article.</p>
Subject–object ambiguities in English (verb bias)	Study B	Analysis on temporarily ambiguous NP (mean of article + noun), and disambiguating verb region (mean of modal + be).	<p>“The club members understood (that) <u>the</u> <u>bylaws</u> <u>would</u> <u>be</u> applied to everyone.</p> <p>The ticket agent admitted (that) <u>the</u> <u>mistake</u> <u>might</u> <u>be</u> hard to correct.” (p. 765)</p>	<p>(c) Study A analyzed the segments as presented, that is, sums of words in segment; Study B analyzed the mean RTs of the two words in each region analyzed; Study C provided analysis of the segments as presented that is, individual words; and Study D summed the RTs on the individual words in the two segments analyzed.</p>
Subject–object ambiguities in English (strong–weak garden paths)	Study C	Graphically presented mean RTs on all 10 segments. ANOVAs presented for segments 6–10 for strong GP, 5–10 for weak GP (i.e., ambiguous noun onward).	<p>“While the neighbor visited the boy/car <u>passed</u> by <u>the</u> <u>house</u>. (strong GP)</p> <p>The man confessed the <u>truth/king</u> was <u>not</u> <u>believed</u> by <u>anyone</u>” (weak GP) (p. 328)</p>	<p>(d) Study A analyzed the segments as presented, that is, sums of words in segment; Study B analyzed the mean RTs of the two words in each region analyzed; Study C provided analysis of the segments as presented that is, individual words; and Study D summed the RTs on the individual words in the two segments analyzed.</p>

Table C.1 (*cont.*)

Linguistic feature	Study	Segments analyzed	Example stimuli (as presented in the article)	Commentary on group of related studies
Subject–object ambiguities in English (strong–weak garden paths)	Study D	Sum of the article + noun + V and the sum of the following three words.	“Before Mary ate the pizza arrived from the local restaurant. After Mary died her husband married a woman from Texas.” (p. 415)	(e) Study C examined the RTs on the comprehension questions; Studies A, B, and D did not.
Relative clauses high/low attachment in Dutch (disambiguation by S–V number agreement)	Study A	Analyses presented for the disambiguating region, defined as the disambiguating auxiliary, the past participle, the preposition/conjunction, and the determiner.	“Daar is de machinist die de conducteurs heeft/hebben bevrijd uit het brandende Treinstel” (p. 82) That is the engine-driver who the guards has/have saved from the burning train-carriage.”	<i>Similarity.</i> All four studies presented analyses of the disambiguation region: aux verb and following verb plus complement. <i>Differences</i>
Relative clauses high/low attachment in English (disambiguation by S–V number agreement)	Study B (expt. 2 & 4)	Mean and <i>SD</i> given for all regions. ANOVA carried out on all segments, and reported for disambiguating auxiliary verb and the sentence final region.	“ <u>The dean liked</u> the secretary of the professors <u>who</u> was/were reading a letter <u>The dean liked</u> the professors with the secretary <u>who</u> was/were reading a letter.” (p. 464) “ <u>The clerk asked for</u> the consultants of the economist <u>who</u> was/were reading the reports. <u>The clerk asked for</u> the economist with the consultants <u>who</u> was/were reading the reports” (p. 483)	(a) Study A was word-by-word presentation; Studies B, C, and D was presentation by segments. (b) Studies B, C, and D reported analysis of all regions after the disambiguation point; Study A did not. (c) Study A analyzed the three words following the disambiguating word on each separate word; Studies B, C, and D analyzed them as one segment (d) Study D carried out analyses on the RTs to the CQ; Studies A, B, and C did not.

Table C.1 (*cont.*)

Linguistic feature	Study	Segments analyzed	Example stimuli (as presented in the article)	Commentary on group of related studies
	Study C	Means calculated for all segments; inferential statistics reported for disambiguating region and the final region.	“The journalist interviewed the assistant of the inspectors who was/were looking very serious” (p. 303)	
Relative clauses high/low attachment in Greek (disambiguation by S-ADJ gender agreement)	Study D (expt. 2)	Descriptive data for all segments; all segments analyzed, analysis of 4th and 5th segments reported	“ <u>Enas kirios fonakse ton fititi tis kathighitrias pu fenotan apoghoitevmenos/eni apo to neoekpedheftiko sistima</u> A man called the student [MASC] of/with the teacher [FEM] who seemed disappointed [MASC/FEM] by the new educational system.” (p. 513)	

Table C.1 (cont.)

Linguistic feature	Study	Segments analyzed	Example stimuli (as presented in the article)	Commentary on group of related studies
Reduced relative clauses in English	Study A	Ambiguous verb, cue, and disambiguating verb	<p>“The brown sparrow <u>seen</u> by the hungry cat <u>pecked</u> at an insect. The brown sparrow <u>noticed</u> on an upper branch <u>pecked</u> at an insect. The brown sparrow <u>noticed</u> almost every day <u>pecked</u> at an insect.” (p. 90)</p>	<p><i>Similarity.</i> Both studies analyzed three segments including the ambiguous verb and two subsequent regions.</p> <p><i>Differences.</i> (a) Study B presented and analyzed “by” in the same segment as the ambiguous verb; Study A presented and analyzed the ambiguous verb alone. (b) Study B included the predicate following the disambiguating verb, that is, the whole sentence final region; Study A analyzed the disambiguating verb only.</p>
	Study B		“Segments 3, 4, and 5” p. 1121	

APPENDIX D

Segments analyzed in studies investigating global ambiguity

Table D.1. *Studies investigating global ambiguity and the segments analyzed*

Linguistic feature	Studies	Segments analyzed	Example stimuli (as presented in the article)	
Subject–object assignment in German	Study A	Average RT for whole item used for analysis	<p>“Peter kann sehen, dass das Spiel den Trainer ärgert. (SO/–ANIM) <u>Peter kann sehen, dass den Trainer das Spiel ärgert. (OS/ –ANIM)</u> (Peter can see that the game angers the coach.)”</p> <p>Peter kann sehen, dass das Kind den Trainer ärgert. (SO/ + ANIM) <u>Peter kann sehen, dass den Trainer das Kind ärgert. (OS/ + ANIM)</u> (Peter can see that the child angers the coach).” (p. 401).</p>	<p><i>Differences.</i></p> <p>(a) Study A analyzed whole sentence reading time and Study B analyzed the sum of SO + adverb + OS in the subordinate clause.</p> <p>(b) Study A used whole-sentence presentation; Study B used presentation in multiword segments.</p>
Subject–object assignment in German	Study B	Sum of segments 3, 4, and 5	<p>“Ich glaube, dass den Arbeiter am Dienstag der Lehrling abgelenkt hat I think that the worker-OBJ on Tuesday the apprentice-SUBJ distracted Ich glaube, dass der Chefarzt am Morgen den Oberarzt überredet hat</p>	

Table D.1 (cont.)

Linguistic feature	Studies	Segments analyzed	Example stimuli (as presented in the article)	
<i>Wh</i> -questions in English, subject/object	Study A	The three words after the main verb. Average RTs of region presented and analyzed.	<p>I think that <u>the chief physician</u>-SUBJ in the morning <u>the senior physician</u>-OBJ persuaded.” (p. 479)</p> <p>“Who did the police declare killed the pedestrian? Who did the police declare the pedestrian <u>killed</u>? Who did the police know <u>killed</u> <u>the</u> pedestrian? Who did the police know <u>the</u> pedestrian <u>killed</u>?” (p. 452)</p>	<i>Similarity</i> . Both studies presented analyses of the same word-by-word regions.
<i>Wh</i> -questions in English, subject/object	Study B	“The critical region was defined as the verb and noun phrase in the complement clause” (p. 208). Average RTs of region presented and analyzed.	<p>“Who do you think <u>met</u> <u>the</u> <u>tourists</u> in front of the museum? Who do you think <u>the</u> <u>tourists</u> <u>met</u> in front of the museum?” (p. 204)</p>	

Table D.1 (cont.)

Linguistic feature	Studies	Segments analyzed	Example stimuli (as presented in the article)	
<i>Wh</i> -questions in German (nominative and accusative case marking)	Study A	Means for all 5 segments then, based on differences in means, ANOVAs on segments 3 and 4.	<p>“Welche Ingenieurin traf den/der Chemiker <u>gestern Nachmittag</u> <u>im Café?</u>”</p> <p>Which [NOM/ACC] engineer met <u>the</u> [ACC/NOM] chemist <u>yesterday afternoon</u> in the cafe?” (p. 887)</p>	<p><i>Similarity.</i> Both studies analyzed the 2nd NP (article + noun ACC or NOM) and the next multiword segment</p> <p><i>Differences.</i></p> <p>(a) Following the ACC/NOM NP, Study B’s adverbial segment was sentence final; Study A included an additional preposition wrap up region.</p> <p>(b) Study B provided ANOVAs on all segments prior to the NP; Study A did not.</p>
<i>Wh</i> -questions in German (nominative and accusative case marking)	Study B	“The <i>wh</i> -element, the matrix verb, the matrix subject, the past participle, the complement verb, and the complement noun phrase...and sentence final prepositional phase” (p. 616) ANOVA for each segment.	<p>“Wer/wen denkst du, vermisste den/der Lehrer <u>in den Ferien?</u>”</p> <p>Who [NOM/ACC] <u>think</u> you, <u>missed</u> the [ACC/NOM] <u>teacher</u> <u>during the vacation?</u>” (p. 613)</p>	

ACKNOWLEDGMENTS

An earlier version of part of this study contributed to the second author's dissertation for a masters in applied linguistics and ESRC funded PhD at the University of York. The project was also partially funded by the British Academy (Academy Research Project AN110002 and AN110003) and the Economic and Social Research Council (RES-062-23-2946). We are grateful to Yasser Teimouri (Georgetown University) for his help in second coding some of the data, and to Leah Roberts (University of York) for her useful comments on an earlier version of this manuscript. The coding scheme for the synthesis is available on IRIS (www.iris-database.org). We are very grateful to the researchers who have made their self-paced reading tests openly available in a special collection on IRIS. The current systematic review was presented at the *Second Language Research Forum* (Columbus, Ohio, 2017), and we thank the attendees for their helpful questions. We are grateful to three anonymous reviewers for their insightful feedback.

NOTES

1. Several sentence processing models exist. For example, two-stage (universal or garden-path) models assume an initial minimal interpretation based on syntactic information followed by reanalysis (e.g., Frazier & Fodor, 1978). Integrated (interactional or constraint-based) models predict multiple sources of information are used (syntactic, pragmatic, lexical, and frequency) for simultaneous and competing interpretations (e.g., MacDonald, 1994). Although L2 research has not had as a *primary* aim to test these models, constraint-based models predict language-specific, and therefore L1, influence on L2 processing. Thus, some L2 research findings align with one or other model for both L1 and L2, and some with syntactic models for L1 but constraint-based models for L2 (see Rah & Adone, 2010; Yang & Shih, 2013).
2. Drawing on the 2013 data, though not reported in that article.
3. These numbers exceed the number of studies, 64, because some studies compared different proficiency groups. "Beginner" covered all author labels that included the word "beginner" (e.g., post-beginner).
4. McManus & Marsden (2017 & in press), using SPR as a pre-, post- and delayed posttest, fell outside the time scope of the review.
5. The extent to which RTs were affected by the point at which the trial occurs within a test was not discussed in the majority of our studies, and more research is needed to assess how RTs are affected by the length of test.
6. We were unable to select studies investigating sensitivity to anomalies, as there could be no expectation of consistency in the regions analyzed given the variety of linguistic features investigated (e.g., number agreement on verbs, nouns and pronouns, and gender agreement on pronouns and adjectives).
7. The terminology for the region following the ambiguous segment varied, sometimes referred to as the *disambiguating region*, others as *prefinal*, *final*, or *sentence final region*.
8. Three currently do this in the field of L2 research: *Language Learning*, *The Modern Language Journal*, *Studies in Second Language Acquisition*.
9. The special collection can be accessed by clicking on the "special collection" button on the Search and Download page at www.iris-database.org.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit <https://doi.org/10.1017/S0142716418000036>

REFERENCES

* Indicates an article included in our study.

- 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.
- Adesope, O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research*, 80, 207–245. doi:10.3102/0034654310368803
- * Amato, M., & MacDonald, M. (2010). Sentence processing in an artificial language: Learning and using combinatorial constraints. *Cognition*, 116, 143–148. doi:10.1016/j.cognition.2010.04.001
- Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3, 12–28. doi:10.21500/20112084.807
- Bannai, M. (2011). The nature of variable sensitivity to agreement violations in L2 English. *EUROSLA Yearbook*, 11, 115–137.
- Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68, 255–278. doi:10.1016/j.jml.2012.11.001
- Blohowiak, B., Cohoon, J., de-Wit, L., Farach, F., Hasselman, F., & DeHaven, A. (2016). Badges to acknowledge open practices. *Centre for Open Science*. Retrieved from <https://osf.io/tyvxz/>
- Boo, Z., & Conklin, K. (2015). The impact of rapid serial visual presentation (RSVP) on reading by nonnative speakers. *Journal of Second Language Teaching and Research*, 4, 111–129.
- * Bordag, D., Kirschenbaum, A., Tschirmer, E., & Opitz, A. (2015). Incidental acquisition of new words during reading in L2: Inference of meaning and its integration in the L2 mental lexicon. *Bilingualism: Language and Cognition*, 18, 372–390. doi:10.1017/S1366728914000078
- Bowden, H. W. (2016). Assessing second-language oral proficiency for research. *Studies in Second Language Acquisition*, 38, 1–29. doi:10.1017/S0272263115000443
- Bowles, M. (2010). *The think-aloud controversy in language acquisition research*. New York: Routledge.
- * Bultena, S., Dijkstra, T., & van Hell, J. G. (2014). Cognate effects in sentence context depend on word class, L2 proficiency, and task. *Quarterly Journal of Experimental Psychology*, 67, 1214–1241. doi:10.1080/17470218.2013.853090
- Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex [Editorial]. *Cortex*, 49, 609–610. doi:10.1016/j.cortex.2012.12.016
- Chang, F., Dell, G. S., & Bock, J. K. (2006). Becoming syntactic. *Psychological Review*, 113, 234–272. doi:10.1037/0033-295X.113.2.234
- * Clahsen, H., & Felser, C. (2006). Continuity and shallow structures in language processing. *Applied Psycholinguistics*, 27, 107–126. doi:10.1017/S0142716406060206
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale, NJ: Erlbaum.
- * Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than non-formulaic language by native and non-native speakers? *Applied Linguistics*, 29, 72–89. doi:10.1093/applin/amm022
- * Conroy, M., & Cupples, L. (2010). We could have loved and lost, or we could never have loved at all. *Studies in Second Language Acquisition*, 32, 523–552. doi:10.1017/S0272263110000252

- * Coughlin, C., & Tremblay, A. (2013). Proficiency and working memory based explanations for non-native speakers' sensitivity to agreement in sentence processing. *Applied Psycholinguistics*, 34, 615–646. doi:[10.1017/S0142716411000890](https://doi.org/10.1017/S0142716411000890)
- * Cui, Y. (2013). L2 processing of relative clauses in Mandarin. *Arizona Working Papers in SLA & Teaching*, 20, 20–39.
- Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, 28, 369–382. doi:[10.1177/0267658312443651](https://doi.org/10.1177/0267658312443651)
- Dehaene, S. (2014). The signatures of a conscious thought. In *Consciousness and the brain: Deciphering how the brain codes our thoughts* (pp. 115–160). New York: Penguin.
- DeKeyser, R. (2015). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition* (pp. 94–112). New York: Routledge.
- * Dekydtspotter, L., & Outcalt, S. (2005). A syntactic bias in scope ambiguity resolution in the processing of English-French cardinality interrogatives: Evidence for informational encapsulation. *Language Learning*, 55, 1–36. doi:[10.1111/j.0023-8333.2005.00288.x](https://doi.org/10.1111/j.0023-8333.2005.00288.x)
- Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly*, 50, 132–153. doi:[10.1002/tesq.217](https://doi.org/10.1002/tesq.217)
- De Vincenzi, M., & Job, R. (1995). An investigation of late-closure: The role of syntax, thematic structure, and pragmatics in initial and final interpretation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21, 1303–1321. doi:[10.1037/0278-7393.21.5.1303](https://doi.org/10.1037/0278-7393.21.5.1303)
- * Dong, Y., Wen, Y., Zeng, X., & Ji, Y. (2015). Exploring the cause of English pronoun gender errors by Chinese learners of English: Evidence from the self-paced reading paradigm. *Journal of Psycholinguistic Research*, 44, 733–747. doi:[10.1007/s10936-014-9314-6](https://doi.org/10.1007/s10936-014-9314-6)
- * Dussias, P. E. (2003). Syntactic ambiguity resolution in L2 learners: Some effects of bilinguality on L1 and L2 processing strategies. *Studies in Second Language Acquisition*, 25, 529–557. doi:[10.1017.S0272263103000238](https://doi.org/10.1017/S0272263103000238)
- * Dussias, P. E., & Cramer Scaltz, T. R. (2008). Spanish-English L2 speakers' use of subcategorization bias information in the resolution of temporary ambiguity during second language reading. *Acta Psychologica*, 128, 501–513. doi:[10.1016/j.actpsy.2007.09.004](https://doi.org/10.1016/j.actpsy.2007.09.004)
- * Dussias, P. E., & Piñar, P. (2010). Effects of reading span and plausibility in the reanalysis of wh-gaps by Chinese-English second language speakers. *Second Language Research*, 26, 443–472. doi:[10.1177/0267658310373326](https://doi.org/10.1177/0267658310373326)
- Ellis, N. C. (2005). At the interface: Dynamic interaction of explicit and implicit language knowledge. *Studies in Second Language Acquisition*, 27, 305–352. doi:[10.1017/S027226310505014X](https://doi.org/10.1017/S027226310505014X)
- * Felser, C., & Roberts, L. (2007). Processing wh-dependencies in a second language: A cross-modal priming study. *Second Language Research*, 23, 9–36. doi:[10.1177/0267658307071600](https://doi.org/10.1177/0267658307071600)
- * Felser, C., Roberts, L., Marinis, T., & Gross, R. (2003). The processing of ambiguous sentences by first and second language learners of English. *Applied Psycholinguistics*, 24, 453–489. doi:[10.1017.S0142716403000237](https://doi.org/10.1017.S0142716403000237)
- * Fender, M. (2003). English word recognition and word integration skills of native Arabic- and Japanese-speaking learners of English as a second language. *Applied Psycholinguistics*, 24, 289–315. doi:[10.1017.S014271640300016X](https://doi.org/10.1017.S014271640300016X)
- Fernández, E. M., & Souza, R. A. (2016). Walking bilinguals across language boundaries: On-line and off-line techniques. In R. R. Heredia, J. Altarriba, & B. A. Cieślicka (Eds.), *Methods in bilingual reading comprehension research* (pp. 33–60). New York: Springer.
- Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, 25, 348–368.
- Foucart, A. (2015). Prediction is a question of experience. *Linguistic Approaches to Bilingualism*, 5, 465–469. doi:[10.1075/lab.5.4.04fou](https://doi.org/10.1075/lab.5.4.04fou)
- Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, 6, 291–325.

- Frenc-Mestre, C. (2005). Eye-movement recording as a tool for studying syntactic processing in a second language: A review of methodologies and experimental findings. *Second Language Research*, 21, 175–198. doi:[10.1191/0267658305sr257oa](https://doi.org/10.1191/0267658305sr257oa)
- Gilboy, E., & Sopena, J. (1996). Segmentation effects in the processing of complex NPs with RCs. In M. Carreiras, J. García-Albea, & N. Sebastian-Gallés (Eds.), *Language processing in Spanish* (pp. 191–206). Mahwah, NJ: Erlbaum.
- * 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. doi:[10.1111/j.1467-9922.2009.00501.x](https://doi.org/10.1111/j.1467-9922.2009.00501.x)
- * Hopp, H. (2006). Syntactic features and reanalysis in near-native processing. *Second Language Research*, 22, 369–397. doi:[10.1191/0267658306sr272oa](https://doi.org/10.1191/0267658306sr272oa)
- * Hopp, H. (2009). The syntax-discourse interface in near-native L2 acquisition: Off-line and on-line performance. *Bilingualism: Language and Cognition*, 12, 463–483. doi:[10.1017/S1366728909990253](https://doi.org/10.1017/S1366728909990253)
- * Hopp, H. (2010). Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. *Lingua*, 120, 901–931. doi:[10.1016/j.lingua.2009.06.004](https://doi.org/10.1016/j.lingua.2009.06.004)
- * Hopp, H. (2016). The timing of lexical and syntactic processes in second language sentence comprehension. *Applied Psycholinguistics*, 37, 1253–1280. doi:[10.1017/S0142716415000569](https://doi.org/10.1017/S0142716415000569)
- Huettig, F., & Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Language, Cognition and Neuroscience*, 31, 19–31. doi:[10.1080/23273798.2015.1072223](https://doi.org/10.1080/23273798.2015.1072223)
- Hulstijn, J. (2010). In S. Unsworth & E. Blom (Eds.), *Experimental methods in language acquisition research* (pp. 185–201). Amsterdam: John Benjamins.
- Hulstijn, J. (2015). *Language proficiency in native and non-native speakers: Theory and research*. Amsterdam: John Benjamins.
- * Ibáñez, A. J., Macizo, P., & Bajo, M. T. (2010). Language access and language selection in professional translators. *Acta Psychologica*, 135, 257–266. doi:[10.1016/j.actpsy.2010.07.009](https://doi.org/10.1016/j.actpsy.2010.07.009)
- * Jackson, C. N. (2008a). Processing strategies and the comprehension of sentence-level input by L2 learners of German. *System*, 36, 388–406. doi:[10.1016/j.system.2008.02.003](https://doi.org/10.1016/j.system.2008.02.003)
- * Jackson, C. N. (2008b). Proficiency level and the interaction of lexical and morphosyntactic information during L2 sentence processing. *Language Learning*, 58, 875–909. doi:[10.1111/j.1467-9922.2008.00481.x](https://doi.org/10.1111/j.1467-9922.2008.00481.x)
- * Jackson, C., & Bobb, S. (2009). The processing and comprehension of wh-questions among second language speakers of German. *Applied Psycholinguistics*, 30, 603–636. doi:[10.1017/S014271640999004X](https://doi.org/10.1017/S014271640999004X)
- * Jackson, C., & Dussias, P. (2009). Cross-linguistic differences and their impact on L2 sentence processing. *Bilingualism: Language and Cognition*, 12, 65–82. doi:[10.1017/S1366728908003908](https://doi.org/10.1017/S1366728908003908)
- * Jackson, C., & Roberts, L. (2010). Animacy affects the processing of subject-object ambiguities in the second language: Evidence from self-paced reading with German second language learners of Dutch. *Applied Psycholinguistics*, 31, 671–691. doi:[10.1017/S0142716410000196](https://doi.org/10.1017/S0142716410000196)
- * Jackson, C., & van Hell, J. (2011). The effects of L2 proficiency level on the processing of wh-questions among Dutch second language speakers of English. *International Review of Applied Linguistics in Language Teaching*, 49, 195–219. doi:[10.1515/iral.2011.012](https://doi.org/10.1515/iral.2011.012)
- * Jegerski, J. (2012). The processing of subject-object ambiguities in native and near-native Mexican Spanish. *Bilingualism: Language and Cognition*, 15, 721–735. doi:[10.1017/S1366728911000654](https://doi.org/10.1017/S1366728911000654)
- * Jegerski, J. (2016). Number attraction effects in near-native Spanish sentence comprehension. *Studies in Second Language Acquisition*, 38, 5–33. doi:[10.1017/S027226311400059X](https://doi.org/10.1017/S027226311400059X)
- Jegerski, J., & VanPatten, B. (2013). *Research methods in second language psycholinguistics*. New York: Routledge.
- * Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics*, 25, 603–634. doi:[10.1017.S0142716404001298](https://doi.org/10.1017/S0142716404001298)

- * Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language learning. *Language Learning*, 57, 1–33. doi:[10.1111/j.1467-9922.2007.00397.x](https://doi.org/10.1111/j.1467-9922.2007.00397.x)
- Jiang, N. (2012). *Conducting reaction time research in second language studies*. New York: Routledge.
- * Jiang, N., Novokshanova, E., Masuda, K., & Wang, X. (2011). Morphological congruency and the acquisition of L2 morphemes. *Language Learning*, 61, 940–967. doi:[10.1111/j.1467-9922.2010.00627.x](https://doi.org/10.1111/j.1467-9922.2010.00627.x)
- * Juffs, A. (1998). Main verb versus reduced relative clause ambiguity resolution in L2 sentence processing. *Language Learning*, 48, 107–147. doi:[10.1111/1467-9922.00034](https://doi.org/10.1111/1467-9922.00034)
- * 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. doi:[10.1191/0267658305sr2550a](https://doi.org/10.1191/0267658305sr2550a)
- * Juffs, A., & Harrington, M. (1995). Parsing effects in second language sentence processing. *Studies in Second Language Acquisition*, 17, 483–516. doi:[10.1017/S027226310001442X](https://doi.org/10.1017/S027226310001442X)
- * Juffs, A., & Harrington, M. (1996). Garden path sentences and error data in second language sentence processing research. *Language Learning*, 46, 286–324. doi:[10.1111/j.1467-1770.1996.tb01237.x](https://doi.org/10.1111/j.1467-1770.1996.tb01237.x)
- Juffs, A., & Rodríguez, G. (2015). *Second language sentence processing*. Oxford: Routledge.
- Just, M., Carpenter, P., & Woolley, J. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111, 228–238.
- Kaan, E. (2015). Knowing without predicting, predicting without learning. *Linguistic Approaches to Bilingualism*, 5, 482–486. doi:[10.1075/lab.5.4.07kaa](https://doi.org/10.1075/lab.5.4.07kaa)
- * Kato, S. (2009). Interaction between processing and maintenance in online L2 sentence comprehension: Implications for linguistic threshold hypothesis. *Asian EFL Journal*, 11, 235–273.
- Keating, G., & Jegerski, J. (2015). Experimental designs in sentence processing research. *Studies in Second Language Acquisition*, 37, 1–32. doi:[10.1017/S0272263114000187](https://doi.org/10.1017/S0272263114000187)
- * Keating, G., Jegerski, J., & Vanpatten, B. (2016). Online processing of subject pronouns in monolingual and heritage bilingual speakers of Mexican Spanish. *Bilingualism: Language and Cognition*, 19, 36–49. doi:[10.1017/S1366728914000418](https://doi.org/10.1017/S1366728914000418)
- Kidwell, M., Lazarević, L., Baranski, E., Hardwicke, T., Piechowski, S., Falkenberg, L., -S., . . . Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLOS Biology*, 14, 1–15. doi:[10.1371/journal.pbio.1002456](https://doi.org/10.1371/journal.pbio.1002456)
- Kim, J.-E., & Nam, H. (2016). Measures of implicit knowledge revisited: Processing modes, time, pressure, and modality. *Studies in Second Language Acquisition*. Advance online publication. doi:[10.1017/S0272263115000510](https://doi.org/10.1017/S0272263115000510)
- * Kim, S. H., & Kim, J. H. (2012). Frequency effects in L2 multiword unit processing: Evidence from self-paced reading. *TESOL Quarterly*, 46, 831–841. doi:[10.1002/tesq.66](https://doi.org/10.1002/tesq.66)
- Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge: Cambridge University Press.
- Kotz, S. A. (2009). A critical review of ERP and fMRI evidence on L2 syntactic processing. *Brain and Language*, 109, 68–74. doi:[10.1016/j.bandl.2008.06.002](https://doi.org/10.1016/j.bandl.2008.06.002)
- Laberge, D., & Samuels, S. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293–323. doi:[10.1016/0010-0285\(74\)90015-2](https://doi.org/10.1016/0010-0285(74)90015-2)
- Lai, M.-L., Tsai, M.-J., Yang, F.-Y., Hsu, C.-Y., Liu, T.-C., Lee, S. W.-Y., & Tsai, C.-C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*, 10, 90–115. doi:[10.1016/j.edurev.2013.10.001](https://doi.org/10.1016/j.edurev.2013.10.001)
- Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. New York: Routledge.
- * Lazarte, A. A., & Barry, S. (2008). Syntactic complexity and L2 academic immersion effects on readers' recall and pausing strategies for English and Spanish texts. *Language Learning*, 58, 785–834. doi:[10.1111/j.1467-9922.2008.00479.x](https://doi.org/10.1111/j.1467-9922.2008.00479.x)
- * Lee, E.-K., Lu, D. H.-Y., & Garnsey, S. M. (2013). L1 word order and sensitivity to verb bias in L2 processing. *Bilingualism: Language and Cognition*, 16, 761–775. doi:[10.1017/S1366728912000776](https://doi.org/10.1017/S1366728912000776)

- Leow, R., Grey, S., Marijuan, S., & Moorman, C. (2014). Concurrent data elicitation procedures, processes, and the early stages of L2 learning: A critical overview. *Second Language Research*, 30, 111–127. doi:[10.1177/0267658313511979](https://doi.org/10.1177/0267658313511979)
- * Lim, J. H., & Christianson, K. (2013). Second language sentence processing in reading for comprehension and translation. *Bilingualism: Language and Cognition*, 16, 518–537. doi:[10.1017/S1366728912000351](https://doi.org/10.1017/S1366728912000351)
- Liu, Q., & Brown, D. (2015). Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing*, 30, 66–81. doi:[10.1016/j.jslw.2015.08.011](https://doi.org/10.1016/j.jslw.2015.08.011)
- MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, 9, 121–136. doi:[10.1080/01690969408402115](https://doi.org/10.1080/01690969408402115)
- * Macizo, P., & Bajo, M. T. (2006). Reading for repetition and reading for translation: Do they involve the same processes? *Cognition*, 99, 1–34. doi:[10.1016/j.cognition.2004.09.012](https://doi.org/10.1016/j.cognition.2004.09.012)
- Marinis, T. (2010). Using on-line processing methods in language acquisition research. In S. Unsworth & E. Blom (Eds.), *Experimental methods in language acquisition research* (pp. 139–162). Amsterdam: John Benjamins.
- * Marinis, T., Roberts, L., Felsler, C., & Clahsen, H. (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition*, 27, 53–78. doi:[10.1017/S0272263105050035](https://doi.org/10.1017/S0272263105050035)
- Marsden, E., Mackey, A., & Plonsky, L. (2016). The IRIS repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 1–21). New York: Routledge.
- Marsden, E. J., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews, and recommendations for the field. *Language Learning* 68. Advance online publication.
- Marsden, E., Morgan-Short, K., Trofimovich, P., & Ellis, N. (2018). Introducing Registered Reports at *Language Learning*: Promoting transparency, replication, and a synthetic ethic in the language sciences [Editorial]. *Language Learning*, 68. Advance online publication.
- Marsden, E., & Torgerson, C. (2012). Single group, pre- and post-research designs: Some methodological concerns. *Oxford Review of Education*, 38, 583–616. doi:[10.2307/41702779](https://doi.org/10.2307/41702779)
- Marull, C. H. (2015). Syntactic position constrains cross-linguistic activation. *Linguistic Approaches to Bilingualism*, 5, 153–179. doi:[10.1075/lab.5.2.01mar](https://doi.org/10.1075/lab.5.2.01mar)
- McManus, K., & Marsden, E. (2017). L1 explicit instruction can improve L2 online and offline performance. *Studies in Second Language Acquisition*, 39, 459–492. doi:[10.1017/S027226311600022X](https://doi.org/10.1017/S027226311600022X)
- McManus, K., & Marsden, E. (in press). Online and offline effects of L1 practice in L2 grammar learning: A partial replication. *Studies in Second Language Acquisition*. doi:[10.1017/S0272263117000171](https://doi.org/10.1017/S0272263117000171)
- * Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics*, 32, 129–148. doi:[10.1093/applin/amq035](https://doi.org/10.1093/applin/amq035)
- Milton, J., & Fitzpatrick, T. (Eds.). (2013). *Dimensions of vocabulary knowledge*. London: Palgrave Macmillan.
- Morgan-Short, K. (2014). Electrophysiological approaches to understanding second language acquisition: A field reaching its potential. *Annual Review of Applied Linguistics*, 34, 15–36. doi:[10.1017/S026719051400004X](https://doi.org/10.1017/S026719051400004X)
- * Mueller, J., & Jiang, N. (2013). The acquisition of the Korean honorific affix (uji) by advanced L2 learners. *Modern Language Journal*, 97, 318–339. doi:[10.1111/j.1540-4781.2013.12005.x](https://doi.org/10.1111/j.1540-4781.2013.12005.x)
- Norouzian, R., & Plonsky, L. (2018). Eta- and partial eta-squared in L2 research: A cautionary review and guide to more appropriate usage. *Second Language Research*, 34, 257–271. doi:[10.1177/0267658316684904](https://doi.org/10.1177/0267658316684904).

- Norris, J., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, 50, 417–528. doi:10.1111/0023-8333.00136
- Norris, J., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty & M. Long (Eds.), *Handbook of second language acquisition* (pp. 717–761). Malden, MA: Wiley-Blackwell.
- Norris, J., & Ortega, L. (Eds.). (2006). *Synthesizing research on language learning and teaching*. Amsterdam: John Benjamins.
- Norris, J. M., & Ortega, L. (2012). Assessing learner knowledge. In S. M. Gass & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 573–589). New York: Routledge.
- O'Grady, W. (2005). *Syntactic carpentry: An emergentist approach to syntax*. New York: Routledge.
- * Omaki, A., & Schulz, B. (2011). Filler-gap dependencies and island constraints in second-language sentence processing. *Studies in Second Language Acquisition*, 33, 563–588. doi:10.1017/S0272263111000313
- * Pan, H.-Y., Schimke, S., & Felsler, C. (2015). Referential context effects in non-native relative clause ambiguity resolution. *International Journal of Bilingualism*, 19, 298–313. doi:10.1177/1367006913515769
- * Papadopoulou, D., & Clahsen, H. (2003). Parsing strategies in L1 and L2 sentence processing. *Studies in Second Language Acquisition*, 25, 501–528. doi:10.1017/S0272263103000214
- Papadopoulou, D., Tsimpli, I. M., & Amvrazis, N. (2013). Self-paced listening. In J. Jegerski & B. VanPatten (Eds.), *Research methods in second language psycholinguistics* (pp. 50–68). New York: Routledge.
- Philips, C., & Ehrenhofer, L. (2015). The role of language processing in language acquisition. *Linguistic Approaches to Bilingualism*, 5, 409–453. doi:10.1075/lab.5.4.01phi
- Pienemann, M., & Kessler, J. U. (2011). *Studying processability theory: An introductory textbook*. Amsterdam: John Benjamins.
- * Pliatsikas, C., & Marinis, T. (2013a). Processing empty categories in a second language: When naturalistic exposure fills the (intermediate) gap. *Bilingualism: Language and Cognition*, 16, 167–182. doi:10.1017/S136672891200017X
- * Pliatsikas, C., & Marinis, T. (2013b). Processing of regular and irregular past tense morphology in highly proficient second language learners of English: A self-paced reading study. *Applied Psycholinguistics*, 34, 943–970. doi:10.1017/S0142716412000082
- Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35, 655–687. doi:10.1017/S0272263113000399
- Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (Or why 18 meta-analyses of feedback have different results). *Second Language Research*, 31, 267–278. doi:10.1177/0267658314536436
- Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *Modern Language Journal*, 100, 538–553. doi:10.1111/modl.12335
- Plonsky, L., & Gass, S. (2011). Quantitative research methods, study quality, and outcomes: The case of interaction research. *Language Learning*, 61, 325–366. doi:10.1111/j.1467-9922.2011.00640.x
- Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning*, 65, 9–36. doi:10.1111/lang.12111
- Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics*, 36, 73–97. doi:10.1017/S0267190516000015
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64, 878–912. doi:10.1111/lang.12079
- Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, 35, 579–592

- * Rah, A., & Adone, D. (2010). Processing of the reduced relative clause versus main verb ambiguity in L2 learners at different proficiency levels. *Studies in Second Language Acquisition*, 32, 79–109. doi:[10.1017/S027226310999026X](https://doi.org/10.1017/S027226310999026X)
- Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, 63, 595–626. doi:[10.1111/lang.12101](https://doi.org/10.1111/lang.12101)
- * Renaud, C. (2014). A processing investigation of the accessibility of the uninterpretable gender feature in L2 French and L2 Spanish adjective agreement. *Linguistic Approaches to Bilingualism*, 4, 222–255. doi:[10.1075/lab.4.2.04ren](https://doi.org/10.1075/lab.4.2.04ren)
- Roberts, L. (2012). Psycholinguistic techniques and resources in second language acquisition research. *Second Language Research*, 28, 113–127. doi:[10.1177/0267658311418416](https://doi.org/10.1177/0267658311418416)
- Roberts, L. (2016). Self-paced reading and L2 grammatical processing. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 58–73). New York: Routledge.
- * Roberts, L., & Felsler, C. (2011). Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics*, 32, 299–331. doi:[10.1017/S0142716410000421](https://doi.org/10.1017/S0142716410000421)
- * Roberts, L., & Liszka, S. (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, 413–439. doi:[10.1177/0267658313503171](https://doi.org/10.1177/0267658313503171)
- Roberts, L., & Siyanova-Chanturia, A. (2013). Using eye-tracking to investigate topics in L2 acquisition and L2 sentence and discourse processing. *Studies in Second Language Acquisition*, 35, 213–235. doi:[10.1017/S0272263112000861](https://doi.org/10.1017/S0272263112000861)
- Roehr, K. (2008). Metalinguistic knowledge and language ability in university-level L2 learners. *Applied Linguistics*, 29, 173–199. doi:[10.1093/applin/amm037](https://doi.org/10.1093/applin/amm037)
- Sagarra, N. (2008). Working memory and L2 processing of redundant grammatical forms. In Z. Han & E. S. Park (Eds.), *Understanding second language process* (pp. 113–147). Clevedon: Multilingual Matters.
- * Sagarra, N., & Herschensohn, J. (2010). The role of proficiency and working memory in gender and number agreement processing in L1 and L2 Spanish. *Lingua*, 120, 2022–2039. doi:[10.1016/j.lingua.2010.02.004](https://doi.org/10.1016/j.lingua.2010.02.004)
- * Sagarra, N., & Herschensohn, J. (2011). Proficiency and animacy effects on L2 gender agreement processes during comprehension. *Language Learning*, 61, 80–116. doi:[10.1111/j.1467-9922.2010.00588.x](https://doi.org/10.1111/j.1467-9922.2010.00588.x)
- * Sagarra, N., & Herschensohn, J. (2012). Processing of gender and number agreement in late Spanish bilinguals. *International Journal of Bilingualism*, 17, 607–627. doi:[10.1177/1367006912453810](https://doi.org/10.1177/1367006912453810)
- Seidenberg, M., & MacDonald, M. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science*, 23, 569–588. doi:[10.1207/s15516709cog2304_8](https://doi.org/10.1207/s15516709cog2304_8)
- * Song, Y. (2015). L2 processing of plural inflection in English. *Language Learning*, 65, 233–267. doi:[10.1111/lang.12100](https://doi.org/10.1111/lang.12100)
- Spada, N. (2017). *Exploring second language learners' grammaticality judgment performance in relation to task design features*. Unpublished manuscript.
- Suda, K. (2015). The influences of proficiency levels and working memory capacities on sentence comprehension by Japanese learners of English. *EUROSLA Yearbook*, 15, 143–163.
- * Tokowicz, N., & Warren, T. (2010). Beginning adult L2 learners' sensitivity to morphosyntactic violations: A self-paced reading study. *European Journal of Cognitive Psychology*, 22, 1092–1106. doi:[10.1080/09541440903325178](https://doi.org/10.1080/09541440903325178)
- Tunmer, W., & Nicholson, T. (2010). The development and teaching of word recognition skill. In M. Kamil, P. Pearson, E. Moje, & P. Afflerbach (Eds.), *The handbook of reading research*. Oxford: Routledge.
- * Vafae, P., Suzuki, Y., & Kachisnke, I. (2016). Validating grammaticality judgement tests. *Studies in Second Language Acquisition*. Advance online publication.

- Van-Gompel, R. P. (2013). *Sentence processing*. London: Psychology Press.
- VanPatten, B. (2015). Foundations of processing instruction. *International Review of Applied Linguistics in Language Teaching*, 53, 91–109. doi:[10.1515/irral-2015-0005](https://doi.org/10.1515/irral-2015-0005).
- * VanPatten, B., Keating, G. D., & Leeser, M. J. (2012). Missing verbal inflections as a representational problem: Evidence from self-paced reading. *Linguistic Approaches to Bilingualism*, 2, 109–140. doi:[10.1075/lab.2.2.01pat](https://doi.org/10.1075/lab.2.2.01pat)
- 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. Neil (Eds.), *The generative study of second language acquisition* (pp. 111–131). New York: Psychology Press.
- * 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, 509–540. doi:[10.1017/S0142716401004027](https://doi.org/10.1017/S0142716401004027)
- Witzel, N., Witzel, J., & Forster, K. (2012). Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, 41, 105–128. doi:[10.1007/s10936-011-9179-x](https://doi.org/10.1007/s10936-011-9179-x)
- Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, 46, 680–704. doi:[10.1111/flan.12063](https://doi.org/10.1111/flan.12063)
- * Xu, Y. (2014). Processing relative clauses in Chinese as a second language. *Second Language Research*, 30, 439–461. doi:[10.1177/0267658313511485](https://doi.org/10.1177/0267658313511485)
- * Yamashita, J., & Ichikawa, S. (2010). Examining reading fluency in a foreign language: Effects of text segmentation on L2 readers. *Reading in a Foreign Language*, 22, 263–283.
- Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, 33, 497–528. doi:[10.1177/0265532215594643](https://doi.org/10.1177/0265532215594643)
- * Yang, P.-L., & Shih, S.-C. (2013). A reading-time study of the main verb versus reduced relative clause ambiguity resolution by English learners in Taiwan. *Applied Psycholinguistics*, 34, 1109–1133. doi:[10.1017/S0142716412000343](https://doi.org/10.1017/S0142716412000343)