CrossMark

Studies in Second Language Acquisition, 2015, **37**, 1–32.
doi:10.1017/S0272263114000187

EXPERIMENTAL DESIGNS IN SENTENCE PROCESSING RESEARCH

A Methodological Review and User's Guide

Gregory D. Keating
San Diego State University

Jill Jegerski
University of Illinois at Urbana-Champaign

Since the publication of Clahsen and Felser's (2006) keynote article on grammatical processing in language learners, the online study of sentence comprehension in adult second language (L2) learners has quickly grown into a vibrant and prolific subfield of SLA. As online methods begin to establish a foothold in SLA research, it is important that researchers in our field design sentence-comprehension experiments that adhere to the fundamental principles of research design typical of sentence processing studies published in related subfields of the language sciences. In this article, we discuss and review widely accepted principles of research design for sentence processing studies that are not always followed in L2 sentence processing research. Particular emphasis is placed on the design of experimental items and distractors, the choice and design of the poststimulus distractor task, procedures for presenting stimuli to participants, and methods for trimming and analyzing online data, among others.

Correspondence concerning this article should be addressed to Gregory D. Keating, Department of Linguistics and Asian/Middle Eastern Languages, San Diego State University, San Diego, CA 92182-7722. E-mail: gkeating@mail.sdsu.edu

© Cambridge University Press 2014

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

Interest in how adult learners of second languages (L2s) interpret sentences in their nonnative language dates back to the mid-1980s (Gass, 1989; LoCoco, 1987; Musumeci, 1989; VanPatten, 1984). Until recently, most sentence processing studies conducted in the field of SLA have used offline methods such as pencil-and-paper questionnaires, in which information about sentence interpretation is gathered after a participant hears or reads a sentence in its entirety. In contrast to offline methods, online methods gather information about sentence interpretation as each word or phrase is read or heard in real time. In the field of mainstream psycholinguistics, online methods have been the tools of choice to study monolingual sentence processing for more than 50 years. By comparison, real-time sentence processing and the methods used to measure it were unfamiliar to most SLA researchers until the publication of Juffs and Harrington's (1995) self-paced reading study on the processing of *wh*-movement violations in adult learners of English as a L2. Since then, the field has witnessed a notable increase in the number of published studies of L2 sentence processing that use online methods such as self-paced reading, eye-tracking, and event-related brain potentials (ERPs).

Online techniques for studying sentence processing are preferred to offline methods for two reasons. First, online methods provide fine-grained information about moment-by-moment sentence comprehension. Because data are gathered as each word or phrase in a sentence is perceived, researchers can examine what happens at precise points in a sentence and at the exact moment those sites are perceived. Second, because online methods measure interpretation in real time, they are believed to tap participants' implicit knowledge of language. That is, online methods minimize the use of explicitly learned knowledge and allow little time for conscious linguistic problem solving. However, the advantages of online methods only hold when experiments adhere to accepted principles of research design. The simplest of design flaws, such as improper randomization of test sentences or poorly designed comprehension questions, can unintentionally reveal the purpose of a study to participants or invite the use of explicit knowledge or strategic processing (i.e., unusually slow or careful reading), all of which can render the otherwise valuable moment-by-moment processing data uninformative.

In the last 10–15 years, SLA researchers have gained access to and expertise in using many of the online methods used by psycholinguists in mainstream psychology. However, although the methods have been successfully imported into our field, some aspects of the research designs required of sentence processing studies have not. Some of the first published studies of online L2 sentence processing did not follow fundamental design principles used in mainstream psycholinguistics research. Although the studies of the past cannot be held to the standards

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

of the present, many papers submitted to SLA journals for peer review continue to ignore such principles or fail to report necessary information about research design. Most of these papers are not published, despite the sound ideas tested in the research. This is unfortunate given the amount of time and effort it takes to design and conduct an online study. Furthermore, journal referees who themselves are unfamiliar with sentence processing research sometimes suggest practices and analyses that are not appropriate for online processing research.

The purpose of this article is to review practices and guidelines common to the design of most sentence processing experiments,[1] regardless of the particular online method chosen. Where differences in design principles exist among methods, readers are referred to relevant sources for consultation. This article describes a set of best practices, while acknowledging that the aims of a particular study may justify practices that differ from those reported herein. The article is aimed at new and veteran researchers in the field of SLA who want to include online sentence processing experiments in their research agenda. Additionally, this article will be useful to reviewers of SLA research who lack expertise in sentence processing but who are, nonetheless, asked to review a paper in their area of expertise that reports the results of a sentence processing experiment.

ONLINE METHODS AND READING PARADIGMS

The methods used to examine real-time sentence processing number close to a dozen. To date, most L2 sentence processing studies focus on the comprehension of written sentences using one of three methods: self-paced reading, eye-tracking, or ERPs. What follows are intentionally brief descriptions of each technique that provide a backdrop against which to discuss principles of research design. For a thorough account of each method and its use in L2 processing research, we refer readers to the relevant chapters in Jegerski and VanPatten (2014).

In a self-paced reading study, participants read sentences on a computer one word or phrase (i.e., one *segment*) at a time. The first press of a designated button displays the first segment. With each subsequent press of the button, the next segment in the sentence appears, and the one before it disappears from view. The time it takes a participant to read each segment is recorded for analyses. In an eye-tracking study with text, participants read a text (i.e., a sentence or larger discourse) on a computer in its entirety. While the participant is reading, an infrared camera tracks pupil movement as the eyes move across the text. The time a reader spends fixating a word or series of words and the time taken to move the eyes forward or backward in the text are recorded for analyses. In an ERP study, participants read sentences on a computer

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

one word or phrase at a time while wearing a skull cap equipped with electrodes that measure brain wave activity. Wave patterns generated after reading each segment are recorded for analyses.

The methods described previously are commonly used in three experimental paradigms: anomaly detection, ambiguity resolution, and syntactic dependency formation, each of which relies on a different type of sentential stimulus to induce processing effects that are measured relative to a baseline condition. In anomaly detection, participants read sentences with some type of grammatical error or semantic or pragmatic inconsistency, and the baseline condition is comprised of the same set of sentences without the anomaly. Ambiguity paradigms identify processing effects by comparing results for ambiguous versus unambiguous sentences (in the case of temporary ambiguities or "garden path" sentences) or for a forced dispreferred reading versus a forced preferred reading (in the case of global ambiguities). Finally, dependency paradigms look for evidence of linking grammatical features over distance, such as the filler-gap dependency that occurs with *wh*-phrases. For all three paradigms, the processing effects that are elicited tend to be similar. With self-paced reading and eye-tracking, processing effects are typically manifest as longer reading times for an experimental condition versus a baseline condition (e.g., a set of stimuli with an anomaly versus the same set of sentences converted to well-formed versions). With ERPs, a comparison of the average waveform for the experimental condition versus the baseline condition typically reveals a difference in amplitude.

FUNDAMENTAL CONSIDERATIONS IN RESEARCH DESIGN

To make meaningful inferences about the nature of L2 sentence processing, experiments must be designed and conducted with great care. Failure to follow basic principles of research design increases the probability of committing Type I and Type II errors. A Type I error consists of rejecting a null hypothesis that is true. For example, a researcher might claim that L2 processing of feature *x* is not nativelike—a rejection of the null hypothesis that states that there is no difference between first language (L1) and L2 processing—when, in fact, it is. A Type II error consists of not rejecting a null hypothesis that is false. In this case, a researcher claims that L2 processing of feature *x* is nativelike when, in fact, it is not. Although the probability of committing either type of error can never be zero, both types of error are problematic for the advancement of theories of L2 sentence processing and should be avoided. The sources of Type I and Type II errors are many, one being flaws in research design. In what follows, we discuss fundamental considerations in research design that are critical to limiting the probability

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

of committing Type I and Type II errors in sentence processing research. Space does not allow us to exhaust all of the possible considerations; rather, we underscore those that are most fundamental to conducting sound sentence processing research and that have not always been followed in previously published L2 processing research and in studies that continue to be received in the review process.

The Experimental Item

The single most important consideration in sentence processing experiments is the design of the sentences that will be used to test the phenomenon under study. Most experimental paradigms test one or two independent variables, each of which has two (and occasionally more) levels. For example, a basic anomaly-detection experiment consists of one independent variable (grammaticality) composed of two levels (grammatical and ungrammatical). Taking English subject-verb agreement as an example, to test whether participants are sensitive to person-number violations during online sentence comprehension, a researcher needs to include grammatical and ungrammatical sentences. One possible—but methodologically unsound—way of doing this is to create one list of grammatical sentences (e.g., *John and Tom live in the dorms*; *Every day the students walk to the gym on campus*) and another list of ungrammatical sentences (e.g., **The man park his car on the street*; **It's great that Sara bake cookies on the weekend*) and compare the processing of the verbs in each type of sentence. The results of such a study would be difficult to interpret because person-number inflections on verbs are not the only differences between the two sets of sentences. The verbs in the ungrammatical set are less frequent than those of the grammatical set and the position of the verb differs within and across the two sets of sentences, not to mention the fact that all lexical items differ, and so forth.

To rule out the possibility that factors other than those studied by the researcher are responsible for the obtained results, psycholinguists create one sentence, called an *item*, that appears in a separate *version* or *condition* for each level of the independent variable. A sample item for the hypothetical subject-verb agreement experiment appears in (1).

(1) a. The student studies in the library on campus.
b. *The student study in the library on campus.

Item (1) appears in two versions that reflect the two levels of the independent variable tested: a grammatical version (1a) and an ungrammatical version (1b). The two versions are *lexically matched* (i.e., they contain

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

the same words) such that the only difference between the two sentences is the violation of person-number agreement between the subject and verb. Paired sentences such as (1) are called experimental *doublets*.

The number of versions of an item is determined by the number of independent variables tested and the number of levels of each. For example, a balanced study of English subject-verb agreement requires testing not only singular subjects, such as those in (1), but also plural subjects. This means that, in addition to grammaticality, a second variable—subject number—is required. Assuming a design in which subject number has two levels (singular and plural), an experimental item that crosses grammaticality (grammatical and ungrammatical) and subject number (singular and plural) requires four versions, as illustrated in (2):

(2) a. The student studies in the library on campus. (grammatical, singular subject)
b. *The student study in the library on campus. (ungrammatical, singular subject)
c. The students study in the library on campus. (grammatical, plural subject)
d. *The students studies in the library on campus. (ungrammatical, plural subject)

Item (2) depicts an experimental *quadruplet*. Versions (2a) and (2b) test sensitivity to subject-verb agreement violations when the subject is singular and versions (2c) and (2d) test violations with plural nouns. This design is commonly referred to as a 2 × 2 design. Although more complicated designs exist, most sentence processing studies test a maximum of two independent variables with two levels each.

Item Consistency. As evinced in examples (1) and (2), lexical matching creates a high level of internal consistency within an item. The target or critical region in the sentence (i.e., the verb *study/studies* in [1]) is as similar as possible in each version. Additionally, the words following the verb (. . . *in the*) are also the same. This is necessary given that the processing of a critical region in a sentence oftentimes continues or "spills over" onto the words immediately following the critical region. This is known as the *spillover effect* (e.g., Rayner & Duffy, 1986), and the regions that follow the critical region are called *spillover regions*. In some sentence processing studies, including those conducted on monolinguals, the effect of a tested variable first appears in a spillover region, not in the target region. Thus, the design of the spillover region is as important as the design of the critical region. Likewise, regions prior to the critical region should be as similar as possible across

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

conditions given that any difficulties inherent in processing precritical regions could cause unintended spillover effects in critical regions. For example, in a study of English subject-verb agreement conducted with English monolinguals, Wagers, Lau, and Phillips (2009, Experiment 1) found that adverbs that immediately followed subject nouns took longer to read when the subject was plural (e.g., *The old keys unsurprisingly were . . .*) versus when it was singular (**The old key unsurprisingly were . . .*). Without the intervening adverb, the spillover processing would have occurred on the critical verb, which would have made the results difficult to interpret. These findings suggest that all versions of item (2) need an intervening word or phrase between the subject and verb.

Just as it is important to create consistency within an item, it is equally important to maintain consistency across all experimental items in a study, especially with respect to the critical and spillover regions. For example, the critical verbs to be included in the hypothetical subject-verb agreement study should appear in the same position in every item and, to the extent possible, should be similar in length (i.e., number of characters or syllables) and frequency. The simple reason for this is that longer words and less frequent words take longer to read than shorter words and more frequent words.[2] Likewise, the nouns that serve as subjects of the target verbs should also be similar in length and frequency. Additionally, the words that make up the spillover regions should be as similar as possible, if not identical, across items. That is, in the subject-verb agreement study, it would be ideal for the critical region to be *in the* in every item. Where this is not possible, alternate words should be of the same grammatical category and identical in length (e.g., *at the*, *on the*, *by the*). An additional point to consider when testing L2 learners is knowledge of the target words. In our view, far too few L2 processing studies provide independent assessments of learners' knowledge of the critical words used in experimental items. Using only high-frequency target words—as determined by a corpus or frequency dictionary—reduces the chance of including words that are unknown to participants but is not a substitute for independently verifying word knowledge, especially when testing lower proficiency learners.

Sentence Norming Studies. For most anomaly-detection studies, following the guidelines discussed so far will yield a methodologically sound set of experimental items that will elicit the anticipated effects in a control group without the need for pilot testing. This is because most (morpho)syntactic anomalies are categorically unacceptable and easy to create. In contrast, when the experimental manipulation involves a lexical, semantic, pragmatic, or plausibility bias, it may be useful to conduct a sentence *norming study* to ensure that the experimental items work as intended. In a norming study, participants recruited from the same population as the control group—but who do not participate in

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

the main experiment—are asked to rate or judge sentences on a psychometric scale, usually a three-, five-, or seven-level Likert-type scale.[3] Depending on the purpose of the norming study, the sentences may be identical to those intended for use in the main study or may contain keywords or sentence frames that will be used to create the experimental items for the main study. The numerical scores associated with participants' ratings are averaged, and the means are used to determine whether the sentences (or parts thereof) are suitable for inclusion in the main study. If not, revisions are made, and a second norming study with a new group of participants from the same population is conducted to test the revised items.

Roberts and Felser (2011) used a norming procedure to confirm the reliability of the plausibility manipulation used in their experiment. To summarize, their study investigated the influence of plausibility information on the resolution of temporary subject-object ambiguities (i.e., garden paths) in English. Their stimuli consisted of paired items such as those in (3). In the (a) versions, the critical noun phrase was a plausible direct object for the preceding verb, and in the (b) versions, it was not (compare *song* in [3a] to *beer* in [3b]).

(3) a. While the band played the song pleased all the customers.
b. While the band played the beer pleased all the customers.

To ensure that the (b) versions of items were considered less plausible than the (a) versions, Roberts and Felser administered an offline plausibility rating questionnaire in which a subgroup of native English controls rated the plausibility of sentences such as *The band played the song* and *The band played the beer* on a scale from *1* "very plausible" to *7* "very implausible." The analyses confirmed a reliable difference between the two conditions. Additional examples of norming procedures used in L2 processing research can be found in Havik, Roberts, van Hout, Schreuder, and Haverkort (2009) and Siyanova-Chanturia, Conklin, and Schmitt (2011).

Number of Items. The quantity of items to include in a study is an important point to consider. Regardless of online method, some data will be lost due to participant error or distraction. Furthermore, as is discussed in a later section of this article, some data must be removed before conducting statistical analyses. Researchers using behavioral methods such as self-paced reading and eye-tracking typically lose 8–15% of collected data, and significantly more data are lost in ERP studies (see Morgan-Short & Tanner, 2014). Additionally, participants in sentence processing studies never read more than one version of the same item. That is, a participant in the subject-verb agreement study described previously would read either version (1a) or version (1b) of

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

item (1), but not both. Reading both versions of the same item would cause *repetition effects*, which refer to when participants respond to a stimulus in an unnatural way (e.g., reading it superficially or deliberately slowly) because they have seen it before. To minimize such effects, researchers divide the different versions of the experimental items into separate presentation lists. A participant only reads sentences from one of the possible lists. The number of lists equals the number of versions in the items; that is, a study that uses experimental doublets like (1) requires two presentation lists, and one that uses experimental quadruplets like (2) requires four presentation lists. Each list must contain an equal number of sentences from each condition. The practice of rotating different versions of items across different presentation lists and then rotating the lists across participants is called *counterbalancing*. The following list presents the division of items for a study that tests the experimental doublets that appear in (1). The (a) versions are grammatical and the (b) versions are ungrammatical.

List 1: 1a, 2b, 3a, 4b, 5a, 6b, etc.
List 2: 1b, 2a, 3b, 4a, 5b, 6a, etc.

The division of items for a study that tests the experimental quadruplets in (2) is as follows (with versions [a] and [c] being grammatical and versions [b] and [d] being ungrammatical):

List 1: 1a, 2b, 3c, 4d, 5a, 6b, 7c, 8d, etc.
List 2: 1b, 2c, 3d, 4a, 5b, 6c, 7d, 8a, etc.
List 3: 1c, 2d, 3a, 4b, 5c, 6d, 7a, 8b, etc.
List 4: 1d, 2a, 3b, 4c, 5d, 6a, 7b, 8c, etc.

Psycholinguists usually create enough items for participants to read 8–12 items per condition, the goal being to have data for at least 6–10 items per condition after accounting for data loss. Therefore, studies that test items with two conditions require 16–24 items. Studies testing items that appear in four conditions require 32–48 items. In contrast to self-paced reading and eye-tracking, ERP studies require 30–40 items per condition due to the larger amounts of data loss associated with this method (Morgan-Short & Tanner, 2014). Therefore, an ERP study that tests experimental doublets requires 60–80 items, and one that tests experimental quadruplets requires 120–160 items. Complete sets of experimental items used in sentence processing research do not always appear in published articles, due to space limitations; however, it is helpful to reviewers to include them in articles submitted for peer review. For examples of complete sets of experimental doublets that follow the design considerations discussed so far, see the appendices in Frenck-Mestre and Pynte (1997), Williams (2006), and Roberts and

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

Felser (2011), among others. Complete sets of experimental quadruplets can be found in the appendices of Cunnings and Felser (2013), Dussias and Piñar (2010), and Jackson and Dussias (2009), among others.[4]

Describing Item Creation. Even when items are lexically matched in appropriate ways, researchers in our field sometimes describe them in ways that suggest otherwise. For example, when discussing the creation of experimental doublets, authors often provide a description and example such as the following: "Our study tested 24 grammatical sentences such as (4) and 24 ungrammatical sentences such as (5) for a total of 48 test sentences.

(4) The student studies in the library on campus.
(5) *The student study in the library on campus."

The consecutive numbering of the examples suggests that the two sentences belong to separate items when, in fact, they are different versions of the same item. This particular study tests 24 items (not 48), each of which appears in two versions: (a) and (b). In sum, to avoid confusion, stimuli should be described as consisting of some number of items (not sentences), each of which appears in some number of versions or conditions. The example should show one item in all of its conditions (a, b, c, d), as shown previously in (1) and (2).

The Poststimulus Distractor Task

The second most important consideration in the design of a sentence processing study is the design of the poststimulus distractor task. Although the primary measure in sentence processing studies is obtained during the reading of each stimulus (i.e., in real time or online), the poststimulus distractor task provides a secondary, offline measurement that is related to each item and is usually a binary-choice question or decision that appears after some or all of the stimuli. A critical item and the question or decision probe that immediately follows it compose a *trial*. In the most basic sense, the purpose of this poststimulus task is to give participants a clear purpose for reading the stimuli so that they pay attention to them for the duration of the experimental session. It also seems that, in practice, research participants often assume that this secondary task is a primary measure of interest. As an added bonus to the distractor function, participants' responses to the poststimulus questions and the time taken to select responses have the potential to be informative with regard to later or delayed stages of sentence processing.

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

Distractor Task Types. Two of the most common types of poststimulus distractor questions are acceptability judgments and meaning-based comprehension questions, though other types of binary-decision tasks have been employed as well. Judgment tasks have enjoyed long-standing popularity in psycholinguistics, in part because they yield additional offline data that is interesting to most researchers, but this practice has recently been questioned by some SLA scholars. Their concern arises in part from the assumption that metalinguistic tasks such as grammaticality or acceptability judgments can increase the application of explicit knowledge during the primary experimental task of reading. It is difficult to argue against this point of view, particularly when an acceptability judgment is used as the distractor task for an experiment that tests L2 learners' sensitivity to violations of grammatical rules that are taught and tested in language classes, as is the case for most structures that are the focus of anomaly-detection experiments. Indeed, it has been shown empirically with self-paced reading that the online processing behavior of L2 learners can vary according to whether the distractor task is an acceptability judgment or a meaning-based comprehension question. In one instance, Leeser, Brandl, and Weissglass (2011) found that intermediate-level L2 learners of Spanish exhibited online sensitivity to violations of noun-adjective gender agreement when the secondary task was a grammaticality judgment but not when it was a meaning-based comprehension question. The findings of the Leeser et al. study also speak to the importance of the instructions that participants receive in advance of completing an online task. If one's aim is to test for implicit sensitivity to a particular type of violation, participants should not be told or led to believe that they will read erroneous sentences.

Another concern with acceptability judgments is regarding the validity of the task itself as an activity that falls outside the realm of daily language comprehension in the real world. In other words, real-time measurement of the behavior and brain activity of L2 learners while they are determining the acceptability of a sentence is probably not as widely generalizable as is the real-time measurement of the behavior of L2 learners who are engaged in meaning-driven language comprehension. Still, there is evidence that the secondary task does not always affect experimental outcomes in self-paced reading so dramatically. For instance, Jackson and Dussias (2009) and, later, Jackson and Bobb (2009) studied the processing of complex *wh*-questions with case marking in L2 German. The first study used a grammaticality judgment as a distractor task and the second used meaningful comprehension questions, but both studies found a nativelike initial preference for object extraction among L2 participants. Despite the overall similarity between the outcomes of the two studies, however, only the L2 participants in the investigation that

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

used the grammaticality judgment probe exhibited nativelike reading time effects later on in the sentence while reading a second, subordinate clause. It is also worth noting that all of the critical stimuli in both of these studies were grammatical, unlike the agreement stimuli for the Leeser et al. (2011) study described previously, so the type of distractor task may be less critical with sentence processing paradigms that do not involve errors or other easily recognizable phenomena. It also appears that the waveforms observed in ERP studies, typically obtained when sentences are processed in conjunction with an acceptability judgment task, can also be observed when the only task is to passively read for comprehension and there is no acceptability judgment (Osterhout, Allen, McLaughlin, & Inoue, 2002). Thus, there is the potential for the distractor task to affect experimental outcomes, and these effects may vary further according to research method and the linguistic structure tested. With self-paced reading in particular, the evidence suggests that acceptability judgments are inappropriate as a distractor task, and eye-tracking would probably be similar in this regard, but with ERPs, the available evidence regarding the potential role of the distractor task is much more limited. The choice of poststimulus task should therefore be intentional, so that the task is consistent with the objectives of the investigation and with prior work using the same research method.

Frequency of Distractors. Another consideration regarding the distractor task is whether the distractor probes appear after all of the stimuli for an experiment or after just a certain percentage of stimuli chosen at random. Both formats are common, and the choice to include poststimulus questions or decisions in some or all experimental trials depends on the purpose of the distractor task in a given experiment. If the only reason for including the secondary task is to keep participants engaged in reading the stimuli, then it is probably not necessary for questions or decisions to appear with every single trial, and limiting the number of poststimulus queries to one for every three or four stimuli, for instance, can save quite a bit of time (if that is a concern). In contrast, the inclusion of distractor questions or decisions consistently across all trials in an experiment can have some important advantages, including improved distraction from the primary experimental measure and research objectives, better face validity (because most linguistically naïve participants would accept answering questions as a language-related task but probably do not see the reading of sentences as a language measure in and of itself), and more consistent attention paid to stimuli throughout the experiment. Last, and most important, accuracy and reaction time data from the poststimulus task items can be additionally informative with regard to sentence processing behavior, but only if sufficient data are available. Thus, if this is an objective, then

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

there should be distractor task questions or decisions after every stimulus.

Of course, the advantages of having a poststimulus distractor task can be inadvertently undermined if the task items are not appropriate for the experiment at hand, so the creation of individual questions or decisions for the task is very important. Every effort should be made to create and pilot poststimulus queries that do not introduce confounding effects or extraneous variability in the data. Specifically, distractor task items should be in the same language as the stimuli to avoid the increased crosslinguistic influence that may arise in bilingual or mixed-language contexts (Grosjean, 2008). Additionally, it is most typical to have just one type of distractor task in a given experiment because the inclusion of more than one type of task (e.g., some stimuli followed by a plausibility decision and others followed by an argument-interpretation question) could introduce task-switching costs as a confounding variable. Furthermore, the content of the poststimulus items should also be neutral with regard to the linguistic variables that are manipulated in the stimuli. For instance, in two self-paced reading experiments, Havik et al. (2009) found that the processing preferences of L2 learners of Dutch reading temporary subject-object ambiguities like those in (6) were influenced by the type or frequency of comprehension questions that followed the stimuli.

(6) a. Subject Relative: Short
Daar is de machinist die de conducteurs heeft bevrijd uit het brandende treinstel.
"That is the engine-driver who the guards has saved from the burning train carriage."
b. Object Relative: Short
Daar is de machinist die de conducteurs hebben bevrijd uit het brandende treinstel.
"That is the engine-driver who the guards have saved from the burning train carriage."

(7) Poststimulus Comprehension Question (True-False Semantic Verification)
De machinist bevrijdde de conducteurs.
"The engine-driver saved the conductors."

Specifically, these researchers observed a nativelike subject relative preference among the L2 participants—who also would be expected to show the same preference in their native German—only when each of the experimental stimuli was followed by a true/false semantic verification like that in (7), and even then only among those L2 participants with a higher working memory span and only with short relative clauses. When a second experiment was run with the same materials but with only 25% of stimuli followed by a semantic verification and with only 25% of those items specifically targeted at the interpretation of subjects

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

and objects, the L2 participants did not show any online processing preference for the subject or object relative clause. Another notable difference between the two experiments was that the first contained 64 experimental stimuli and only 16 distractors, whereas the second included 64 critical items and 64 distractors. It is interesting to note that the native Dutch participants in the study did not appear to be affected by the type and frequency of poststimulus distractor questions, as their reading behavior was consistent across both experiments, which suggests that these aspects of experimental methodology may be more crucial in L2 research than in mainstream psycholinguistics.

Another way in which extraneous effects might inadvertently affect experimental data is if distractor task items repeat the target form from the stimuli that precede them or if they draw participants' attention to the semantic content of the target form. To illustrate, if an experiment targets verbal agreement with an error recognition paradigm, as exemplified in (8) (repeated from [1] for convenience), and if the poststimulus query repeats the target form without the error, as the examples in (9) do, this may draw additional attention to the error present in the stimulus, it may provide an external reminder of the correct form, or it can otherwise affect participant behavior while reading subsequent stimuli. The question in (10) is similarly inappropriate, in this case because answering it correctly requires participants to focus on the semantic content of the verb, thus encouraging participants to selectively pay more attention to the verbs in sentences. This is especially true when the same type of question appears after stimuli throughout the experiment, as it encourages participants to develop a task-specific reading strategy. Another problem with comprehension questions that repeat part or all of the stimulus verbatim is that they are not a good measure of *sentence* comprehension because it is possible to respond correctly without processing word meaning and because they target the meaning of only one part of the sentence. The comprehension questions in (11), however, use synonyms and paraphrasing to go beyond the surface form of the stimulus to test higher level processing (while still using language that is no more complex than that of the stimuli). This type of question avoids pitfalls associated with repeating or targeting a specific linguistic form in the stimulus and also is a better gauge of meaningful comprehension. (Note that the question format, such as yes/no, true/false, or other binary-choice options, is not relevant here because these and other formats are all acceptable.)

(8) Agreement Stimulus from (1): Grammatical and Ungrammatical Conditions
a. The student studies in the library on campus.
b. *The student study in the library on campus.

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

(9) Comprehension Questions: Inappropriate, Target Form Repetition
   a. Does the student study in the library?
      Yes No
   b. This person studies in the library at school.
      True False

(10) Comprehension Questions: Inappropriate, Focus on Verb Semantics
   a. Does the student sleep in the library?
      Yes No

(11) Comprehension Questions: Appropriate
   a. This library is at a school, probably at a university.
      True False
   b. Does the event described take place in a restaurant?
      Yes No

A couple of final methodological points regarding the poststimulus distractor task are the timing of responses to these items and the provision of feedback on the accuracy of responses. Although many published papers do not indicate whether distractor questions were timed, it is common to limit the amount of time that a participant can take to answer a question or make a decision to 10,000 or 15,000 ms or even less, depending on how much additional reading is required. This can be done in one of two ways: Either the experiment can be programmed such that the question screen times out and the experiment continues after a set time limit, such as 10,000 ms, or the participants can take unlimited time to respond and the data can be trimmed later on. Building a time limit into the experiment is probably preferable, as it discourages excessive contemplation or reflection during the experiment, and it improves the efficiency of research sessions. Some L1 researchers also opt to provide feedback when responses to poststimulus queries are inaccurate (e.g., Gibson & Wu, 2013), presumably to maximize participants' level of attention while reading stimuli, but it is not clear yet what effect feedback may have with L2 learners.

Distractors and Fillers

Along with a number of critical stimuli, participants in L2 sentence processing research also read a number of mostly unrelated sentences as part of an experiment. The purpose of these *noncritical items* is to obscure the critical items and thus the specific research objectives from participants. Distractors and fillers are also included to minimize task effects; if critical test items were simply administered one after the other, this might lead to repetition effects or to participants adopting an unnatural processing strategy (e.g., the structure of sentences becomes predictable and therefore they are

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

processed only very superficially). The term *filler* is sometimes used interchangeably with the word *distractor*,[5] but the two terms can also be used by psycholinguists to distinguish different types of noncritical items. Distractors, like critical stimuli, are intentionally designed to contain a specific linguistic form or structure, either as critical items for another experiment or to counterbalance some characteristic of the critical stimuli that might otherwise make them stand out to the participant. Fillers, in contrast, are unrelated sentences that are not intended to elicit any specific type of processing effects. For example, with a subject-verb agreement experiment as exemplified in (2), the distractors could be for another experiment, such as one that examines the processing of temporary subject-object ambiguities of the type in (12), adapted from Juffs (2004). (Note that distractors such as these with two conditions can be combined with target stimuli with any even number of conditions.) If the distractors are for another experiment, it should not be so closely related to the primary experiment as to involve similar stimuli (and the two experiments may be published separately if deemed appropriate in light of the APA guidelines on piecemeal publication). The distractors would also have to appear in different conditions, counterbalanced across the different presentation lists. The fillers, in contrast, would each appear in only one condition and would be identical across all presentation lists, because there is no experimental manipulation with fillers. Some examples are given in (13)–(15). Given that their purpose is to divert participants' attention from the experimental stimuli, both fillers and distractors should be superficially similar to the target items, particularly with regard to sentence length. In addition, other salient sentence characteristics such as interrogative versus declarative status should be balanced across critical and noncritical items. In other words, if all of the critical stimuli for an experiment are questions and all of the fillers are declarative sentences, this might draw participants' attention to the critical items while they are reading.

(12) a. Before the student guessed the answer appeared on the next page. (Distractor Item)
b. Before the student spoke the answer appeared on the next page. (Distractor Item)
(13) Yesterday, there was a book on the table in the hallway. (Filler Item)
(14) The bank usually closes early on Wednesday afternoons. (Filler Item)
(15) The clerk changes the sign outside the store every day. (Filler Item)

As already mentioned, the number of critical stimuli is determined by the experimental design (i.e., the number of variables and levels of each variable) and by the specific measure used (i.e., self-paced reading,

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

eye-tracking, or ERPs). The number of fillers and distractors is, in turn, determined by the number of critical stimuli. Although there is no single accepted ratio for critical to noncritical items, there is some evidence that including less than 50% noncritical items can affect the outcome of a self-paced reading experiment (Havik et al., 2009). This makes sense, as even with 50% noncritical items, the critical stimuli for the experiment would be separated by only one noncritical item, and the pattern of alternation between the two sentence types could be quite predictable. It is, thus, highly desirable to have greater than 50% noncritical items. In fact, it could be said that the greater the proportion of noncritical items, the better, were it not for the practical limitations on the number of total sentences that participants can read before fatigue or boredom begins to affect their performance. In striking a balance between the contradictory objectives of concealing critical stimuli within a large number of fillers and limiting the length of an experimental session to avoid fatigue, self-paced reading, eye-tracking, and ERP experiments of simple design (i.e., one linguistic variable with two levels) typically include 75% noncritical sentences. So, a self-paced reading or eye-tracking study that tests 24 experimental doublets, as in (1), requires approximately 72 noncritical items (for a total of 96 items all inclusive). A similar study composed of 32 experimental quadruplets, as in (2), requires approximately 96 noncritical items (for a total of 128 items all inclusive). In ERP experiments that have a single linguistic variable with more than two levels or in those that cross more than one linguistic variable, the need for 30–40 stimuli per condition or cell usually means that it is not feasible to have greater than 50% fillers.

Sequencing Trials

After creating the critical stimuli and matched conditions (including any norming), the counterbalanced presentation lists, the filler sentences, and the poststimulus distractor task, the last step in materials development for a sentence processing experiment is to determine the combination and ordering of sentences within each presentation list. The sequencing of critical stimuli and fillers is important for several reasons and should be regarded from two different perspectives. First, the placement of each item within the broader presentation list (e.g., the 3rd or the 49th trial of 120 total trials) should be balanced such that no one stimulus item or condition appears consistently in approximately the same position in the sequence for the entire experiment. This is because data from any cognitive task are potentially affected by both (lack of) task familiarity and fatigue effects, which would most likely occur toward the beginning and the end of an experimental

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

session, respectively. One way to avoid the influence of familiarity and fatigue effects on the data is to randomize the stimuli, with a unique randomization for each participant, an easy option that is built into most stimulus presentation software packages. However, simple randomization can result in the appearance in succession of two or more sentences of the same type or even condition, which is problematic because repetition priming is known to affect the processing of a variety of linguistic stimuli. To illustrate, the appearance of several sentences in a row with errors of subject-verb agreement, as in (1b)/(8b), may affect how participants process subsequent stimuli, possibly making them more likely than normal to notice errors. Reading several sentences of the same type together may also draw participants' attention to the target form or structure, especially in the case of unusual constructions such as disambiguated relative clauses or complex *wh*-sentences. Thus, the second important consideration in sequencing the presentation lists is the positioning of sentence items relative to other items of the same type, with special attention given to the sentences that are read immediately before and after a given item. This type of balancing is usually achieved by using modified randomization or *pseudorandomization* rather than simple randomization. Although many software programs are capable of creating a unique pseudorandomized list for each participant, researchers usually create a limited set of pseudorandomizations (i.e., oftentimes just one per list). The concern for avoiding the placement of each item consistently in the same position within the broader presentation sequence can be addressed by grouping sentences in equivalent sets or *blocks*, pseudorandomizing the sentences within each block, and then varying the order of the blocks within the broader presentation list.

Beyond the presentation order for the critical stimuli and fillers used for a sentence processing experiment, there is also a series of 5–10 practice items—unrelated to the critical stimuli—that appear at the beginning of the session and serve to further minimize task familiarity effects on both the primary experimental measure and the distractor task responses. Additionally, fatigue effects in general can be minimized by limiting the number of trials administered in a single session, which usually does not exceed a total of 120–160 sentences read (one exception being ERP studies), including both experimental stimuli and fillers but not practice items, and by requiring participants to take one or more breaks from reading during the session. Finally, repetition priming can also affect responses to the distractor task (e.g., a *yes* response is generally faster when it follows another *yes* response than when it follows a *no* response), so the ordering of items according to the expected response is also a potential concern, at least when the task involves some type of binary-choice decision.

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

Analyses

Data collected in sentence processing experiments are treated and analyzed differently than data obtained from other tasks traditionally used in mainstream SLA research, such as speech-elicitation tasks and offline sentence-interpretation and judgment tasks. In what follows, we review analysis procedures that are currently in widespread use with the three online techniques most commonly employed in L2 processing research: self-paced reading, eye-tracking, and ERPs. Although there are some similarities in the way data are analyzed across the three methods, we opt to discuss each separately because of the differences in the types of data each elicits from research participants. Additionally, because current trends suggest that L2 psycholinguists will, in the future, prefer mixed-effects models as an alternative to traditional ANOVAs, we include an overview of mixed-effects analysis as well.

Self-Paced Reading. In selecting which self-paced reading data to analyze and report, one important decision for the researcher is whether or not to include reading time data from trials with incorrect distractor task responses. It has, thus far, been fairly standard practice in sentence processing research to eliminate such data, under the assumption that inaccurate responses to basic comprehension questions (i.e., ones that all normal adults would be expected to respond to accurately) and similarly transparent queries reflect either a lack of attention during the experiment or comprehension behavior that is otherwise nonstandard. There are necessary exceptions to this assumption, however, when there is no expectation for consistently accurate responses, perhaps because of task difficulty level or because the type of distractor task is such that there is no objectively correct response (e.g., when the task is an ambiguity interpretation), so this type of data cleaning is not always appropriate. Additionally, when the data corresponding to incorrect responses are substantial in number, as is more common with L2 learners, these can be reserved for independent analyses (see, e.g., Juffs & Harrington, 1996), or, when the data from incorrect responses are not as numerous, statistical analyses can be run once on data from both correct and incorrect trials and once without data from the incorrect trials, to see if there are any differences. Although such exploratory analyses are not as common in sentence processing research with normal adult populations of native speakers (but see, among others, Gibson, Desmet, Grodner, Watson, & Ko, 2005; Hsiao & Gibson, 2003) and thus far have not been very common with research on L2 sentence processing either, they are of particular interest with any population that may regularly exhibit comprehension difficulty, as is the case with L2 learners, and thus should not be discounted a priori.

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

The next step in preparing the reading time data for statistical analyses with traditional ANOVAs is trimming to minimize the effects of outliers or extreme data points. (If residual reading times are to be calculated, as described later in this section, that calculation can come before or after trimming, depending on the trimming method used. With mixed-effects models, described at the end of this section, data trimming is not as important and is therefore minimal to nonexistent.) There is no single acceptable technique for data trimming, but it usually involves the deletion of reading times of less than 100–200 ms, which are quite rare, as well as the replacement of very high values with more moderate ones. Outlying high values can be designated in one of three ways: (a) via an absolute cutoff in the range of 2,000–6,000 ms (depending on the length of the stimulus region), (b) with a variable cutoff that is 2–3 standard deviations above the mean reading time for each group or individual participant in each stimulus region and in each condition, or (c) with a combination of both methods. Where intersubject variability is high, as is often the case with nonnative processing, the identification of outliers by individual participant and item—rather than by group—will lead to greater experimental power (i.e., will reduce the chances of avoiding a Type II error; Ratcliff, 1993) but takes more time. Once the outliers have been identified, the values are then either replaced by a more moderate number that is typically the same as the cutoff or removed entirely, leaving missing values. One potential concern with data trimming is that it can result in too many missing values for a given participant or item in any of the stimulus conditions (i.e., less than six remaining reading times on which to base an aggregate mean, described in the next paragraph), particularly when data points from trials with inaccurate distractor task responses have already been deleted, so it is usually preferable to replace outliers rather than delete them.

Once the reading time data have been selected and trimmed, they can be averaged and submitted to statistical analyses. In both cases, each region of interest from the stimuli is treated as a distinct measure, with separate descriptive and inferential statistics. One notable point of difference between psycholinguistic reading time studies of L2 acquisition and most other types of SLA research is the addition of items analysis to the traditional subjects analysis, which entails double the amount of averaging and analyzing. To begin the analysis by subject, aggregate means are calculated for the critical stimulus region, once for each participant for each stimulus condition. Each aggregate mean is computed by averaging the reading times for all the stimuli read in a given condition by the participant, so a simple two-condition experiment like that illustrated in (1) would have two aggregate means per participant (one for all of the grammatical items read and another for all of the ungrammatical items read), whereas a four-condition experiment like that in (2)

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

would have four means per participant. Descriptive statistics normally include grand means for each group in each stimulus condition, which are averages of all the individual means from the participants in each group. To complete the subjects analysis, the processes of calculating aggregate means and then group grand means are repeated for each stimulus region that is of potential interest, which often includes a precritical region in which no effects are expected; the critical region, in which the initial processing effect is predicted to occur; the spillover region; the sentence-final region; and any intervening regions. The spillover region is of potential interest because reading time effects can carry over from one stimulus region to the next, whereas the last region of a sentence is the site of additional processing known as *sentence wrap-up* (Just & Carpenter, 1980). In addition to causing generally longer reading times, sentence wrap-up can also be the site of delayed or echoic reading time effects caused by the experimental manipulation in the stimuli. For the items analysis, the whole process is repeated for each stimulus region of interest, but, instead of calculating aggregate means for each participant, means are calculated for each stimulus item for each participant group in each stimulus condition. Thus, an experiment with two stimulus conditions and three participant groups would have six aggregate means per stimulus item.

Statistical analyses of reading time data are traditionally ANOVAs and *t* tests, which are conducted separately for each stimulus region of interest, once by subject and once by item. Analyses by subjects are distinguished from analyses by items by the use of subscripted numerals. F_1 and t_1 refer to analyses conducted by subjects, and F_2 and t_2 refer to analyses conducted by items. The purpose of such statistical tests is to determine whether there are real differences in reading times among stimulus conditions and whether these effects vary among participant groups. Thus, a typical set of ANOVAs by subject will each have repeated measures and will include the one or two linguistic variables manipulated in the stimuli as within-subjects variables (e.g., from example [2], grammaticality: grammatical and ungrammatical; subject number: singular and plural) plus the participant grouping variable as a between-subjects variable (e.g., group: L1, L2 advanced, and L2 near-native). For the corresponding ANOVAs by item, the variables and levels of variables would remain the same, but they would all be repeated-measures or within-subjects ("within-items") variables because the same item contributes to all levels of all variables. Generally speaking, a main effect is deemed significant when *F* and *t* tests are significant by subjects and by items (a variation of the technique proposed by Clark, 1973), but it is common in sentence processing studies to obtain a significant effect in one of the two analyses and a marginally significant or insignificant effect in the other. With mixed-effects models, subjects and items are both incorporated into a single set of tests, so the interpretation of statistical

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

significance is more straightforward because there is only one *p* value to consider.

In some studies, it may not be possible for the critical or spillover region to be equal in length across the different versions of the experimental items. This is true of the subject-verb agreement experiments depicted in (1) and (2), in which the critical verb varies in length from five to seven characters (*study* vs. *studies*). Differences in length introduce confounds that could make the results difficult to interpret. For example, reading times on the ungrammatical verb *study* in (2b) may not be longer than reading times on the grammatical verb *studies* in (2a), as expected if readers display online sensitivity to such errors, because the ungrammatical version is shorter than its grammatical counterpart. Similarly, participants may spend more time reading the ungrammatical verb *studies* in (2d) relative to the grammatical verb *study* in (2c) because it is longer, not because they have detected a violation of person-number agreement. Additionally, participants (including monolinguals) read at different rates, which introduces interparticipant variability to the computation of group means. One way to control for differences in length across otherwise comparable conditions and individual differences in reading rates is to calculate *residual reading times*, also called *deviations from regressions* (Ferreira & Clifton, 1986; Trueswell, Tanenhaus, & Garnsey, 1994). Residual reading times are calculated separately for each participant and sentence region by submitting the raw reading times from each participant for all segments of all items (i.e., critical and noncritical) to a regression equation with word length as the predictor variable. Reading times predicted by the participant's regression equation are then subtracted from the raw reading times obtained in the experiment proper. Positive numbers refer to reading times that were slower than expected, and negative values refer to reading times that were faster than expected. The resulting residual reading times are then trimmed and statistically analyzed using the same procedures as for the raw times.[6] So, when it is necessary to adjust for length or individual reading rates, researchers should conduct both sets of analyses. In contrast, individual reading speed is known to correlate with some other individual differences, such as L2 proficiency, working memory, lexical access, and semantic integration (Hopp, 2013), so residual reading times may be inappropriate for the investigation of individual differences. In studies in which both raw and residual times are analyzed and the patterns of significance are identical or highly similar, or in which manuscript space is a concern, researchers often report one set of statistical analyses (i.e., either the raw or the residual reading times), noting any differences in the other set as appropriate (for an example of this practice, see Chen, Gibson, & Wolf, 2005).

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

Finally, the data from the offline distractor task are also of interest and therefore analyzed and reported. Aggregate means for global accuracy, with data from both critical stimuli and fillers included in the scores, are reported and compared statistically by subject and by item using *t* tests or ANOVAs, usually to show the overall range of comprehension among the L2 participants and how their performance compared to that of native speakers. Where there are group differences, these should be taken into consideration later on, when making a final interpretation of the overall outcome of the experiment. Additionally, similar sets of statistical tests are performed on the reaction times and accuracy scores from only those distractor questions that followed the experimental stimuli. Experimental power can be improved with these tests by trimming the response time data from the distractor questions with the same method used to trim the reading time data from the stimuli, as already discussed. Analysis of poststimulus distractor question data is potentially informative because the question is essentially a spillover region for the experimental stimulus, so there may be delayed processing effects that surface in the response times or accuracy. For instance, meaning-based comprehension questions can sometimes be read more slowly when they follow ungrammatical sentences than when they follow grammatical sentences, and accuracy rates for meaning-based comprehension questions can also be slightly but significantly lower with ungrammatical sentences than with grammatical sentences (e.g., Jegerski, 2012). As with the primary reading time data from the experimental stimuli, these effects can also vary among participant groups, which would be evident in statistical interactions.

Although L2 psycholinguistics has, thus far, relied primarily on the type of ANOVAs just described for statistical hypothesis testing, recent developments in the field of statistics have catalyzed a fairly rapid shift in mainstream psycholinguistics toward mixed-effects modeling as the standard method of analysis, and there is good reason to think that a similar shift is underway in L2 processing research. There are several important advantages to mixed-effects modeling that make it especially appealing for experimental research on human language. Perhaps most important, it provides a considerable improvement on previous methods for treating experimental items as well as participants as random effects (the importance of which was articulated by Clark, 1973), especially with independent variables that are continuous as opposed to categorical. Additionally, as compared to traditional parametric testing, mixed-effects modeling is less affected by missing data points and outliers and can be used to analyze both normal and binomial data distributions (with linear and logit models, respectively). Finally, from the practical standpoint that is the focus of this article, mixed-effects modeling is also more efficient because it requires fewer steps and less time to prepare raw data for analysis. There is no need

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

for data trimming or for the calculation of aggregate means by participant and by item. Data output files are simply compiled into a master sheet and imported into an appropriate statistical software program, such as R (R Core Team, 2013), in which any necessary sorting to remove trials with incorrect comprehension question responses or transformations can be carried out prior to mixed-effects analysis (described in greater detail by Cunnings, 2012).

Eye-Tracking. Analysis of eye-tracking data involves the same steps required of self-paced reading, albeit with a considerably larger number of dependent variables. In contrast to self-paced reading, which provides one online reading time measure for each region of interest, eye-tracking provides two broad categories of data: (1) reading/fixation times (reported in ms), which can be fractionated into various subcomponents that include early and later measures of processing, and (2) regressive (i.e., backward) eye movements out of or into a region of interest, which are typically reported as the proportion of the sentences in an experimental condition in which such movements occurred out of or into the critical region (for details about the various eye-movement measures, see Keating, 2014). Eye-tracking studies usually report one or two of the so-called early measures and one or two later measures. Thus, for each region of interest, researchers analyze several dependent variables, the choice of which is determined by the goals of the study and the point in the eye-movement record at which predicted effects are likely to show up.

Other than the differences in the total number and types of dependent variables available for analysis, self-paced reading and eye-tracking data are handled in similar ways. First, when analyzing fixation times and regressions out of or into a region, it is standard practice to remove from analyses all trials for which the poststimulus distractor question was answered incorrectly (barring the same exceptions mentioned previously for self-paced reading). It is also common practice to screen fixation time data for each measure (separately by subject and by item) to identify outlying values. Outliers can be identified by establishing cutoff values or using the standard deviation method. Fixations shorter than 50–100 ms are usually removed or merged with a neighboring fixation under the assumption that such short fixations are not indicative of any meaningful cognitive behavior. Establishing an absolute cutoff for an outlying high value depends on the type of measure (early vs. later) and other factors such as length of the region and so forth (see Keating, 2014). In addition to using absolute cutoffs to remove extreme values, researchers may also use the standard deviation method (calculated separately by subject and by item) to remove values in a given condition that exceed some number of standard deviations above or below the participant or item mean (usually 2 or 2.5 *SD*s) for each condition in

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

the study. Once identified, outlying fixation times are either replaced with an alternate value (e.g., the cutoff value, when using an absolute cutoff procedure, or the participant or item mean for the condition, when using the standard deviation method) or removed entirely and left as missing values. Once outliers are treated, the data for each dependent variable are analyzed using ANOVAs and *t* tests, which are conducted separately for each region of interest, once by subjects and once by items, as explained previously for self-paced reading. The practice of deriving and analyzing length-corrected residual reading times is not common in eye-tracking studies. When this technique is employed, it is usually only conducted on gaze durations or first-pass reading times. As with self-paced reading, mixed-effects modeling is fast becoming the standard method of analyzing eye-tracking data in studies of monolingual sentence processing (e.g., Cunnings, Patterson, & Felser, 2014). Finally, data from the offline distractor task are analyzed both for global comprehension accuracy and as a locus for potential spillover processing from the experimental sentences, as described previously for self-paced reading.

ERPs. The first step toward summary and analysis of ERP data from a sentence processing experiment is to select the subset of data that is of interest. The researcher must decide both how many electrodes to record data from during the experiment, typically 19–64 (Morgan-Short & Tanner, 2014), and which of those electrodes to include when reporting the research. It is common to report data from only those electrodes predicted to be most informative, which usually number around nine, although data from as few as three electrodes can be sufficient in shorter articles (e.g., Tanner, Osterhout, & Herschensohn, 2009). The investigator must also decide whether to include data from all trials, regardless of response accuracy, or to exclude data from those trials for which the participant provided an incorrect response to the distractor question. The inclusive method is more common in L2 research.

The next step is to filter and clean the electroencephalography (EEG) data to generate the ERPs for each electrode. Raw EEG recordings are, by nature, very noisy because each electrode measures total activity from an area of the cerebral cortex that contains numerous neurons, each a potential source of electricity. Even something as minute and unconscious as blinking or making saccadic eye movements during a linguistic experiment entails brain activity that generates extraneous voltage fluctuations in the data. Thus, prior to presenting summarized data or conducting statistical analyses, the ERP—the electrical activity associated with a particular cognitive event of interest—is extracted through a series of steps, the first of which is cleaning and filtering the EEG recordings. For the amplified data from each electrode, frequency filters are applied and the specific time interval of interest surrounding

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

each experimental stimulus for each participant is separated out into what is known as an *epoch*. Epochs containing *artifacts*, or extraneous waveforms, especially those associated with any type of muscle movement (e.g., blinking), are excluded from further analysis. Next, the EEGs are averaged across all trials of the same type (i.e., stimulus condition) for each participant, for each time point at each electrode. These subject means are then averaged within each participant group to generate the grand mean waveforms for each stimulus condition at each electrode and are also plotted across time relative to the onset of the stimulus.

Statistical analyses of ERP data from sentence processing research seek to compare sets of individual mean waveforms from several different electrodes, in two or more different stimulus conditions, usually for two or more participant groups. This is most often accomplished via the mean amplitude method (although others are available), which involves the calculation of the mean EEG values across select time periods of interest, such as 300–500 ms poststimulus for an N400 and 500–800 ms poststimulus for a P600 (Tanner et al., 2009). After the mean amplitude is calculated for each participant across each time window, at each electrode, and in each stimulus condition, the means are compared statistically with a separate ANOVA for each time window. Thus, a typical ANOVA may include stimulus condition and electrode as within-subjects variables and group as a between-subjects variable.

Finally, as with self-paced reading and eye-tracking, ERP sentence processing experiments typically yield additional data from an offline distractor task. In ERP research, this is referred to as the *behavioral* data, which stands in contrast with the neural ERP data, and it most often takes the form of an acceptability judgment of the stimuli. The mean accuracy scores of participants are submitted to ANOVAs to determine whether there were any differences, especially between groups. Any effects or differences in the behavioral data can then be compared to those found in the primary ERP data, particularly when discussing and interpreting the overall outcome of a study (e.g., Tokowicz & MacWhinney, 2005).

Participants

The ideal number of participants needed for a sentence processing study is a function of the number of conditions in the experimental items. Given that participants only see one version of each item, for both versions of an experimental doublet to be read just once, two participants per group are needed. Likewise, four participants per group are needed for all versions of an experimental quadruplet to be read

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

once. The tendency in the monolingual processing literature is to aim for a minimum of 12 participants per condition after attrition. This equates to a minimum of 24 participants per group for studies that test experimental doublets and a minimum of 48 participants per group for studies that test experimental quadruplets. To be fair, our colleagues in psychology often have access to large subject pools that are homogeneous with respect to the characteristics that matter for processing research. This is hardly the case in most types of research on adult L2 learners, particularly when the aim is to recruit advanced or near-native speakers, as is often the case in L2 sentence processing research.[7] For this reason, we advocate holding researchers to reasonable expectations for subject recruitment given the L2 chosen, the environment in which recruitment takes place, and so forth. To reduce the chance of committing a Type II error, researchers who anticipate recruiting small sample sizes may wish to conduct a power analysis (Cohen, 1988).

Total numbers aside, to ensure that each condition has equal weight in statistical analyses, it is desirable to have an equal number of participants exposed to each version of each item (although deviations of one or two participants per condition are not likely to impact results in a large sample). This is not easy to achieve, given that one cannot know a priori how many participants will be removed from a study before analyses are conducted and whether discarded participants will be equally distributed across the different experimental lists in which the items appear. At the very least, researchers must be meticulous in ensuring the proper rotation of experimental lists within groups and should make efforts to recruit additional participants if the number of participants exposed to each condition is grossly uneven.

Another important point concerns the linguistic naïveté of research participants tested in L2 processing research. A fundamental assumption of psycholinguistics research is that participants are naïve to the purpose of the study at the outset. Every principle of research design discussed so far contributes to assuring that participants remain naïve throughout the course of a study. In the L2 processing literature, it is common to find studies that do not exclude participants with backgrounds in the language sciences or participants who teach the language tested in the study (a specific example with a self-paced reading experiment on L2 Spanish is discussed by Foote, 2011). Testing such participants greatly increases the chances of committing a Type II error. In our view, it is better to have a small sample of naïve participants than to have a larger sample that is contaminated in this way. This consideration is especially important with L2 participants, who are typically more likely to have thought about the grammar of the language in question.

Finally, in some L2 processing studies, L2 participants' proficiency in the target language is determined on the basis of unstandardized proficiency tests (e.g., a cloze test or C-test designed by the researcher) that

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

are only administered to the L2 participants. When unstandardized tests are used, they should be administered to participants in the monolingual/native speaker control group to provide a baseline against which to compare the scores of the L2 learners.

COMMON PITFALLS

In this final section, we list some common mistakes that are seen in many of the manuscripts of L2 sentence processing research that are received for review. In most cases, a L2 processing study receives an unfavorable review for one (or both) of the following reasons: (a) the design of the study violates a fundamental principle of research design that renders analyses of the online reading time data uninterpretable, or (b) the manuscript does not provide enough detail about research design for reviewers to understand the design of the study and to properly interpret its results. The first reason is by far the most serious, given that it usually results in a paper that cannot be published without conducting a new experiment. The following flaws in research design are among the most common:

- Designing experimental items that are not lexically matched across conditions
- Not controlling for length, frequency, or other variables where appropriate
- Including too few items per condition to conduct statistical analyses by items
- Choosing a distractor task that is metalinguistic in nature for use in a self-paced reading or eye-tracking study, particularly for one that tests tutored learners on a rule or structure that is typically covered in formal language instruction
- Designing poststimulus distractor probes that repeat or draw participants' attention to the target form in the experimental items
- Choosing inappropriate types of fillers or distractors vis-à-vis the target structure
- Using a trial randomization procedure that allows sentences of the same type or condition to appear consecutively
- Administering the online task after another experimental task that is metalinguistic in nature, such as an acceptability judgment task or a proficiency test
- Using F_2 and t_2 to refer to something other than analyses conducted by items

In other cases, it is not clear whether researchers follow fundamental design considerations, because many manuscripts do not report all of the required information about research design in the method section. Many L2 processing studies submitted for review omit one or more of the following items:

- Appropriate background information on L2 participants
- A description of the instructions given to participants in advance of completing the online task (i.e., what participants were told regarding the purpose of the study and whether they were oriented to read for comprehension or meaning vs. something else)

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

- An example of the poststimulus distractor probes used in the study
- A description of the number and types of distractor or filler items included in the study
- Information about how critical and noncritical items were divided into presentation lists and how items within each list were (pseudo)randomized
- Information about how the online data were trimmed and about the amount of data (expressed as a percentage) removed before conducting statistical analyses

CONCLUSION

Second language acquisition researchers working in a variety of theoretical frameworks (e.g., generative, emergentist, functionalist, and so forth) have begun to recognize the potential that online sentence processing research holds to shed new light on long-standing issues of importance in SLA studies, such as the nature of L2 development, the role of the L1 in L2 acquisition, and L2 ultimate attainment, among others. In this article, we have reviewed fundamental methodological and design principles that underpin sentence processing studies conducted using self-paced reading, eye-tracking, and ERP techniques. We hope that this information encourages and assists SLA researchers new to online sentence processing research to conduct methodologically sound sentence processing experiments that will make solid contributions to our understanding of language comprehension in nonnative languages.

Received 2 February 2014

NOTES

1. It is not the aim of this article to critique published L2 processing studies that do not adhere to design principles discussed herein. Many such studies were the first of their kind in our field and made valuable contributions to our understanding of the nativelikeness of L2 processing.

2. Information about word frequency in different languages is much more easily available than it was just a few years ago. Routledge Press publishes a series of frequency dictionaries of core vocabulary for learners of various languages. In mainstream psycholinguistics, researchers have recently turned to corpora based on movie subtitles for information about word frequency (e.g., Brysbaert & New, 2009).

3. Rating sentences is not the only type of norming procedure available and may not be the most appropriate one for every experimental manipulation. For example, if a researcher needs to know whether the verbs to be included in a study bias for a direct object versus a complement clause, participants in a norming study may be asked to complete partially formed sentences (e.g., Wilson & Garnsey, 2009).

4. For complete sets of items that appear in three conditions (i.e., experimental triplets), see Roberts, Gullberg, and Indefrey (2008) and Rah and Adone (2010).

5. Generally speaking, the term *distractor* is also used to refer to an incorrect answer option with multiple-choice items; so, for a binary-choice item, the participant would choose between one correct response and one distractor.

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

6. It is not always stated in the monolingual processing literature whether residual reading times are calculated on the basis of the raw reading times prior to trimming or afterward. In our view, when researchers intend to use an absolute cutoff only, or an absolute cutoff in combination with the standard deviation method, it makes the most sense to apply the absolute cutoff method to the raw reading times before deriving the residual reading times (and then apply the standard deviation method to the residual reading times when using the combination of both methods). Handling the data in this way prevents extreme raw reading times from adversely altering the regression equation. A second consideration in self-paced reading studies is whether or not to include the raw reading times from the sentence-final region in the calculation of the regression equation, given that reading times in this region may be artificially high relative to its length due to sentence-final wrap-up effects.

7. Published studies of disordered populations such as the language impaired also tend to have smaller samples typical of those seen in L2 processing research for similar reasons.

REFERENCES

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990.

Chen, E., Gibson, E., & Wolf, F. (2005). Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language*, *52*, 144–169.

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*, 369–382.

Cunnings, I., & Felser, C. (2013). The role of working memory in the processing of reflexives. *Language and Cognitive Processes*, *28*, 188–219.

Cunnings, I., Patterson, C., & Felser, C. (2014). Variable binding and coreference in sentence comprehension: Evidence from eye movements. *Journal of Memory and Language*, *71*, 39–56.

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.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*, 348–368.

Foote, R. (2011). Integrated knowledge of agreement in early and late English-Spanish bilinguals. *Applied Psycholinguistics*, *32*, 187–220.

Frenck-Mestre, C., & Pynte, J. (1997). Syntactic ambiguity resolution while reading in second and native languages. *Quarterly Journal of Experimental Psychology*, *50A*, 119–148.

Gass, S. (1989). How do learners resolve linguistic conflicts? In S. Gass & J. Schacter (Eds.), *Linguistic perspectives on second language acquisition* (pp. 183–199). Cambridge, UK: Cambridge University Press.

Gibson, E., Desmet, T., Grodner, D., Watson, D., & Ko, K. (2005). Reading relative clauses in English. *Cognitive Linguistics*, *16*, 313–353.

Gibson, E., & Wu, H. H. I. (2013). Processing Chinese relative clauses in context. *Language and Cognitive Processes*, *28*, 125–155.

Grosjean, F. (2008). *Studying bilinguals*. Oxford, UK: Oxford University Press.

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.

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

Hopp, H. (2013). Individual differences in the second language processing of object-subject ambiguities. *Applied Psycholinguistics*. Advance online publication. doi:10.1017/S0142716413000180

Hsiao, F., & Gibson, E. (2003). Processing relative clauses in Chinese. *Cognition*, *90*, 3–27.

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

Jackson, C. E., & Dussias, P. E. (2009). Cross-linguistic differences and their impact on L2 sentence processing. *Bilingualism: Language and Cognition*, *12*, 65–82.

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

Jegerski, J., & VanPatten, B. (Eds.). (2014). *Research methods in second language psycholinguistics*. New York, NY: Routledge.

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

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

Juffs, A., & Harrington, M. (1996). Garden path sentences and error data in second language sentence processing. *Language Learning*, *46*, 283–323.

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

Keating, G. D. (2014). Eye-tracking with text. In J. Jegerski & B. VanPatten (Eds.), *Research methods in second language psycholinguistics* (pp. 69–92). New York, NY: Routledge.

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

LoCoco, V. (1987). Learner comprehension of oral and written sentences in German and Spanish: The importance of word order. In B. VanPatten, T. R. Dvorak, & J. F. Lee (Eds.), *Foreign language learning: A research perspective* (pp. 119–129). New York, NY: Newbury House.

Morgan-Short, K., & Tanner, D. (2014). Event-related potentials (ERPs). In J. Jegerski & B. VanPatten (Eds.), *Research methods in second language psycholinguistics* (pp. 127–152). New York, NY: Routledge.

Musumeci, D. (1989). *The ability of second language learners to assign tense at the sentence level: A crosslinguistic study* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Osterhout, L., Allen, M. D., McLaughlin, J., & Inoue, K. (2002). Brain potentials elicited by prose-embedded linguistic anomalies. *Memory and Cognition*, *30*, 1304–1312.

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.

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

Rayner, K., & Duffy, S. A. (1986). Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory and Cognition*, *14*, 191–201.

R Core Team. (2013). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

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

Roberts, L., Gullberg, M., & Indefrey, P. (2008). Online pronoun resolution in L2 discourse: L1 influence and general learner effects. *Studies in Second Language Acquisition*, *30*, 333–357.

Siyanova-Chanturia, A., Conklin, K., & Schmitt, N. (2011). Adding more fuel to the fire: An eye-tracking study of idiom processing by native and non-native speakers. *Second Language Research*, *27*, 251–272.

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press

Tanner, D., Osterhout, L., & Herschensohn, J. (2009). Snapshots of grammaticalization: Differential electrophysiological responses to grammatical anomalies with increasing L2 exposure. In J. Chandlee, M. Franchini, S. Lord, & G.-M. Rheiner (Eds.), *Proceedings of the 33rd Boston University Conference on Language Development* (pp. 528–539). Somerville, MA: Cascadilla Press.

Tokowicz, N., & MacWhinney, B. (2005). Implicit and explicit measures of sensitivity to violations in second language acquisition in second language grammar: An event-related potential investigation. *Studies in Second Language Acquisition*, *27*, 173–204.

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic influences on parsing: Use of thematic role information in syntactic disambiguation. *Journal of Memory and Language*, *33*, 285–318.

VanPatten, B. (1984). Learner comprehension of clitic pronouns in Spanish: More evidence for a word order strategy. *Hispanic Linguistics*, *1*, 57–67.

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*, 206–237.

Williams, J. N. (2006). Incremental interpretation in second language sentence processing. *Bilingualism: Language and Cognition*, *9*, 71–88.

Wilson, M. P., & Garnsey, S. M. (2009). Making simple sentences hard: Verb bias effects in simple direct object sentences. *Journal of Memory and Language*, *60*, 368–392.

https://doi.org/10.1017/S0272263114000187 Published online by Cambridge University Press