
Article:
Avery, Nicholas and Marsden, Emma Josephine orcid.org/0000-0003-4086-5765 (2019) A meta-analysis of sensitivity to grammatical information during self-paced reading: Towards a framework of reference for reading time effect sizes. Studies in Second Language Acquisition. ISSN 1470-1545

https://doi.org/10.1017/S0272263119000196.

Final version accepted for publication in Studies in Second Language Acquisition

(11 February 2019)

Nick Avery & Emma Marsden

University of York

A meta-analysis of sensitivity to grammatical information during self-paced reading in a second language

Abstract

Despite extensive theoretical and empirical research, we do not have estimations of the magnitude of sensitivity to grammatical information during L2 online processing. This is largely due to reliance on null hypothesis significance testing (Plonsky, 2015). The current meta-analysis draws on data from one elicitation technique, self-paced reading, across 57 studies (N = 3,052), to estimate sensitivity to L2 morphosyntax and how far L1 background moderates this. Overall, we found a reliable sensitivity to L2 morphosyntax at advanced proficiencies (d = 0.20, 95% CIs 0.15, 0.25), with some evidence this was reliably lower than for native speakers. These patterns were not generally moderated by linguistic feature or sentence region. However, effects for anomaly detection were larger among native speakers than L2 learners and the effects among L2 learners appeared to show a trend towards L1 influence. Finding smaller effects than in other subdomains, we provide an initial framework of reference for L2 reading time effect sizes.

Keywords: cross-linguistic influence; grammatical sensitivity; meta-analysis; morphosyntax; online processing; reading time; self-paced reading; sentence processing

BACKGROUND

Effective and efficient reading in an L2 partly depends on learners' sensitivity to grammar[1], which encodes crucial information such as plurality or the assignment of subject and object roles. L2 learners[2] may have different knowledge of this grammar, different access to that knowledge, and/or different processing strategies, when compared to speakers with different language backgrounds, such as native speakers (NSs) or learners with higher proficiencies or different first languages. Such L2 phenomena can result in misinterpretations or processing problems. According to some theories, processing problems can even impede acquisition itself (O'Grady, 2005; Sharwood-Smith & Truscott, 2005; VanPatten, 2007). Thus, it is important to understand L2 learners' sensitivity to L2 morphosyntax and the variables that influence their processing performance.

Some suggest that the degree of similarity between L1 and L2 features may explain variation in L2 learners' sensitivity to morphosyntax. If L1 properties do influence L2 processing performance, structures shared between the L1 and L2 would be expected to generate relatively less processing effort than those which are not (Tolentino & Tokowicz, 2011). This view is common to the morphological congruency hypothesis (Jiang, Novokshanova, Masuda, & Wang, 2011), according to which L2 acquisition of morphemes is dependent on L1 properties, and the unified competition model (MacWhinney, 2005), which claims that L1 influence[3] is "pervasive in the arena of sentence interpretation" (p. 77). MacWhinney suggests that when linguistic features are represented in the same way in the L1 and L2, positive transfer occurs. The weight of cues used to signal grammatical functions in the L1 may moderate these effects, as may L2 proficiency level.

Not all theories attribute such importance to L1 influence. For example, the shallow structure hypothesis (Clahsen & Felser, 2006a) foregrounds the universally 'shallow' nature

of processing of L2 syntax. Clahsen and Felser argue that L2 learners are unable to reliably analyse morphosyntactic information in online L2 input in the same way as NSs, and consequently rely more on lexico-semantic and pragmatic cues. This implies that L1 background is (at least relatively) insignificant compared to the general effects common to L2 syntactic processing (see also Bley-Vroman, 1990, 2009). Note, however, that although the shallow structure hypothesis does not readily predict a strong influence of the L1 in the domain of syntax (as noted elsewhere, e.g., Tolentino & Tokowicz, 2011), it does not completely rule out the possibility of L1 influence. Indeed, most recently, Clahsen and Felser (2018) describe their position on the question of L1 influence as "largely agnostic" (p. 697).

A number of methods have been used to investigate L1 influence in L2 processing, including the use of neurolinguistic and eye movement data. For example, in a narrative review of nine studies providing evidence from functional magnetic resonance imaging and event-related potentials, Tolentino and Tokowicz (2011) concluded that L1 features instantiated differently in the L2 and features unique to the L2 generated greater processing difficulty than those shared between the L1 and L2. The authors interpreted their findings as evidence in support of the unified competition model and against the shallow structure hypothesis.

Another, arguably more frequent and accessible, method used to investigate L1 influence on L2 morphosyntactic sensitivity has been the self-paced reading (SPR) paradigm (Just, Carpenter, & Wooley, 1982). Indeed, a recent systematic review of L2 SPR research found 26 studies that were motivated by questions relating to L1 influence during L2 processing (Marsden, Thompson, & Plonsky, 2018). While L2 processing is of course investigated using a variety of methods, the findings from this methodological paradigm are the focus of the current study and, therefore, also of this Background section.

SPR Research Investigating Online L2 Morphosyntactic Sensitivity and L1 Influence

In SPR tasks, sentences are presented word by word (or in groups of words) on a computer screen. The participant taps a key to see the next word (or multi-word segment), and software measures the time taken between each key press. These reading times offer insights into the relative processing effort required to comprehend, with higher reading times usually indicating greater processing effort (for methodological reviews of SPR, see Jegerski, 2014; Keating & Jegerski, 2015; Marinis, 2010; Marsden, Thompson, & Plonsky, 2018; Roberts, 2012, 2016).

Studies using SPR have yielded conflicting findings regarding whether L2 learners are sensitive to L2 morphosyntax, with some suggesting similar sensitivity compared to NSs (e.g., Jackson, 2008; Jackson & Dussias, 2009; Kaan, Ballantyne, & Wijnen, 2015; Song, 2015), and others suggesting less sensitivity (e.g., Jiang, 2004, 2007; Marinis, Roberts, Felser, & Clahsen, 2005). Conflicting findings have also been observed regarding whether L1-L2 similarity facilitates L2 processing: some studies have suggested that differential sensitivity to morphosyntax between learners of typologically different L1s is due to L1-L2 similarity/difference (e.g., Jackson, 2010; Jiang et al., 2011; Roberts & Liszka, 2013); other studies have provided evidence against L1 influence (e.g., Felser, Roberts, Marinis, & Gross, 2003; Gerth, Otto, Felser, & Nam, 2017; Papadopolou & Clahsen, 2003); and others have suggested that the L1 offers only a partial explanation of results (e.g., Juffs, 2005; Perpiñán, 2014).

This mixture of findings may be for a range of reasons. One might be that the kind of processing phenomena that the study taps into may affect the observation of morphosyntactic sensitivity and L1 influence. For example, a group of SPR studies investigating one type of ambiguity resolution - relative clause attachment - has fairly consistently suggested no L1

influence, at least among advanced L2 learners. This group of studies draws on cross-linguistic variation in preferences about relative clause attachment among monolinguals, and then explores whether L2 participants carry those L1 preferences into their L2. The phenomenon has been investigated by making the antecedent of a relative clause temporarily ambiguous until morphology disambiguates it, thus forcing a particular attachment to one of the nouns in the matrix clause. For example, in the sentence 'Someone shot the servant of the actress who was on the balcony', the relative clause ('who was on the balcony') may refer to either the first determiner phrase ('the servant') or the second ('the actress'). Studies have generally found a preference for the latter among English NSs (Papadopolou & Clahsen, 2003). In their study, Papadopolou and Clahsen investigated the preferences of Spanish, German and Russian learners of Greek. Since monolinguals of all these languages have a preference for attaching the relative clause to the first noun in the matrix clause, one would hypothesise that L2 learners process faster in morphosyntactic contexts that force this attachment than those that force the alternative attachment, if L1 preferences are transferred to L2 processing. The reading time data, however, suggested no processing preference for either attachment for any group (see Dussias, 2003; Felser et al., 2003; Pan, Schimke, & Felser, 2015, for similar patterns of findings).

In addition to this type of ambiguity, SPR research has investigated other types of temporary and also global ambiguities (such as role assignment where ambiguity can remain after the sentence is complete). All such ambiguity resolution studies share the common characteristic that sentences are grammatical, and morphosyntactic sensitivity is monitored by whether a processing cost is incurred by morphosyntactic re-analysis at or after the disambiguation point. On the other hand, other SPR studies insert ungrammatical morphosyntax into sentences, and monitor whether an interruption to the parse was caused by detection of the anomaly. The distinction between processing cost caused by grammatical

(though potentially unexpected or incongruous) morphosyntax versus ungrammatical morphosyntax (violations) may lead to a systematic difference in the magnitude of effects, though this issue has not been systematically explored to date.

Another, closely related characteristic that varies between SPR studies is the wide range of linguistic features investigated. Studies investigating L1 influence have spanned filler-gap dependencies (e.g., Marinis et al., 2005), subject-object role assignment (e.g., Gerth et al., 2017; Hopp, 2006, 2009, 2010), tense-aspect agreement (e.g., Roberts & Liszka, 2013), verb-aspect licensing (e.g., Tokowicz & Warren, 2010), reduced relative versus main verb ambiguity (e.g., Rah & Adone, 2010), relative clause attachment (e.g., Papadopolou & Clahsen, 2003), number agreement (e.g., Jiang et al., 2011), gender agreement (e.g., Renaud, 2014), and causative versus inchoative alternators (e.g., Juffs, 1998b). The type of linguistic feature may systematically affect the magnitude of sensitivity to morphosyntax or L1 influence, and, ideally, syntheses should take this variety into account. However, cross-study comparisons at the level of specific linguistic feature are difficult as the number of studies investigating one linguistic feature can be small. This reflects the paucity of replication in L2 research more generally (Marsden, Morgan-Short, Thompson, & Abugaber, 2018; though see Juffs, 2006).

Nevertheless, one point of contrast in terms of linguistic feature that allows for cross-study comparisons with a larger k (k = number of studies) relates to whether the feature represents a processing preference or a grammatical principle that can be perceived as being invariable (often dichotomously, as ‘acceptable’ or ‘unacceptable’). It is conceivable that studies of core, invariable morphosyntax render stronger effects than attachment preference studies, which may in turn render L1 effects more observable in the former studies. Indeed, a good number of such investigations have suggested L1 effects in, for example, tense-aspect agreement (Roberts & Liszka, 2013), number marking (Chan, 2012, Jiang et al., 2011), wh-

dependencies (Jackson, 2010), reduced relative clauses (Rah & Adone, 2010), and gender agreement (Tokowicz & Warren, 2010). For parsing preferences, however, it has been suggested that effects attributable to L1 influence may be confounded (and, arguably, potentially masked) by a complex range of issues, such as prosodic and lexico-semantic information (Jackson & Dussias, 2009) or amount and age of exposure (Jegerski, 2018).

SPR research also varies in terms of the study designs used to investigate L1 influence. Some have used within-subject designs, in which the linguistic conditions manipulated cross-linguistic similarity. Tokowicz and Warren (2010), for example, investigated the sensitivity of beginner L1 English-L2 Spanish learners to violations of four features. Two were cross-linguistically similar (verb progressive aspect licensing), one was cross-linguistically different though present in both the L1 and L2 (possessive and definite determiner-noun number agreement), and one was unique to the L2 (determiner-noun gender agreement). Reading times provided evidence that the L2 learners were sensitive to grammatical violations of both similar and different features, but not those unique to the L2.

Other authors have used a between-group design, in which the performance of L2 groups with typologically different L1s is compared on the same linguistic feature across groups. For example, Juffs (2005) investigated wh-movement with L1 Chinese, Japanese and Spanish learners of English, and NSs. Spanish, like English, instantiates wh-movement, but Japanese and Chinese are wh-in-situ. Reading times might thus reveal L1-attributable effects. Variation might also be caused by L1 word order, given that Spanish and Chinese are predominantly SVO, while Japanese is SOV. Speed differences were found (with Japanese being slowest and Spanish being fastest), but since all L2 learner groups experienced greater difficulty processing subject extractions than object extractions, regardless of L1 background, the author concluded that their parsing profiles could not solely be explained by L1-L2 con-/divergence.

Still further SPR studies do not explicitly set out to test L1 influence within their study design but consider L1-L2 similarity as relevant to their predictions and/or post-hoc interpretation of results (e.g., Rah & Adone, 2010; Song, 2015; VanPatten, Keating, & Leeser, 2012). For example, Jiang (2004) investigated how advanced L1 Chinese learners of English processed the bound inflectional morpheme, plural –s. Reading times for sentences like those in (1) and (2) below were compared:

(1) The key to the cabinet was rusty from years of disuse

(2) The key to the cabinets was rusty from years of disuse

Jiang found that the L1 Chinese learners did not show a statistically significant sensitivity to the plural morpheme, unlike native English speakers who slowed down due to the 'broken agreement effect' in (2) compared to (1). Jiang suggested this might reflect, among other things, the fact that overt plural morphemes are rarely grammaticalized in Chinese (see also Chan, 2012; Jiang, 2007).

It is conceivable that this variation in study design moderates the detection of L1 influence. First, within-subject manipulations of L1-L2 similarity may yield more valid results than between-group comparisons because a single L1 group acts as its own control, lessening the risk of sampling error and extraneous variables (Morris & DeShon, 2002). Second, they may yield larger effects, as illustrated by Plonsky and Oswald's (2014) L2 field-wide meta-analysis that found stronger effects for within-group (median d = 1.06) than between-group contrasts (median d = 0.70). As yet, this issue has not been explored specifically for reading time research, probably largely due to the fact that the use and interpretation of effect sizes in such research has been relatively neglected.

Using and Interpreting Effect Sizes in L2 Reading Time Research

In addition to investigating the substantive issues above, an important, more general purpose of the current study was to investigate the feasibility and usefulness of calculating and interpreting effect sizes in L2 reading time research. To date, null hypothesis significance testing (NHST; e.g., analyses of variance, t-tests and, more recently, regression models with mixed effects) represents the default option for analysing data in L2 SPR (Marsden, Thompson, & Plonsky, 2018) and L2 research generally (Plonsky, 2013, 2015). The main concerns about NHST include: (1) because p is influenced by both the relationship and sample size being measured, it may lead to false conclusions. For example, in L2 SPR research samples are usually small, and so non-significant findings that are used to infer no L1 influence or no difference between L2 learners and NSs may in fact reflect a lack of statistical power rather than an absence of effects in the data; (2) p values channel research towards dichotomous conclusions, such as 'L1 influence' vs. 'no L1 influence', which are less nuanced than conclusions about the relative strength of L1 influence for specific conditions, learners, or features (VanPatten & Jegerski, 2010); (3) the cut-off values for rejecting the null hypothesis, such as 0.5, are arbitrary and yet contribute to the oft-cited "file drawer problem" (Rosenthal, 1979) whereby results not meeting this threshold are not submitted to and/or published by journals, leading to published findings that underrepresent the totality of evidence (Marsden, Morgan-Short et al., 2018; Norris & Ortega, 2006; Plonsky & Oswald, 2014); (4) an over-reliance on NHST can also exacerbate a temptation to p-hack, whereby participants or items are added or removed in order to obtain a statistically significant result, a questionable practice which does not necessarily reveal a more informative observation.

Effect size reporting could be a step towards more nuanced evidence regarding sensitivity to L2 morphosyntax and L1 influence, providing a) information about the

magnitude of difference between two means; b) evidence that is less confounded by sample size; and c) standardised units of measurement (standard deviations) that can be aggregated systematically across studies. Standardised units are necessary because the raw data reported in two SPR tasks are difficult to compare due to differences between participants, stimuli, processing issues, and general processing speeds. Means in raw milliseconds (also a 'standardised unit', of course) could arguably be used to compare across studies. However, this would present several problems. First, it ignores variation in standard deviations, yielding a distorted interpretation of the size of differences found in different studies. (For example, two comparisons between means may render an identical raw difference, yet if one mean has a greater standard deviation, confidence in that mean is weaker. Since the calculation for d uses the standard deviation as its denominator, it expresses difference in terms of standard deviation units.) Second, and relatedly, different overall reading speeds between participant groups (due to, for example, different proficiencies or literacy skills) distort the interpretation of raw millisecond reading time data. That is, the difference between two conditions within one participant group may be relatively small simply due to their overall fast reading speeds, whereas in another group it is much larger simply due to their slower reading speed; yet, the 'small difference' may be meaningful whilst the 'larger difference' may not. Third, some SPR tasks use residual reading times, which control for word length differences between stimuli (see Keating & Jegerski, 2015), and could not be directly compared with the raw reading time data from another study. Effect sizes are unaffected by this since the type of data (raw or residual) is constant within each dataset that yields an effect size.

Despite these potential benefits, the d family of effect sizes has rarely been reported in L2 reading time studies to date, so we do not have a clear picture of what constitutes a 'small' or 'large' effect. (For exceptions, see Brysbaert & Stevens, 2018; Jegerski, 2018; Marsden, Thompson, & Plonsky, 2018.) Illustrating two consequences of this problem, we note that in

Adesope, Lavin, Thompson, & Ungerleider's (2010) meta-analysis, data had to be extracted to calculate effect sizes for the two reading time studies they included and, potentially more concerning, these effect sizes were aggregated with effects from a different type of data, test scores (accuracy, judgements etc.). An illustration of how effect sizes can vary across data types is found in Hedge et al.'s (2018) meta-analysis of task performance, which observed reliably different effect sizes (with no correlation) when performance cost was measured by reaction times versus errors. As their comparability with other data types is not well understood, effect sizes from measurements with reading time data may constitute a moderator variable ('data type') that should be considered when aggregating effects across data types.

Towards a framework of reference.

To begin to address this gap, we sought to capture a 'framework of reference' within which to interpret effect sizes for our main substantive areas of interest as discussed above (i.e., sensitivity to L2 morphosyntax and L1 influence during online reading). Of course, for interpreting d, we could turn to Cohen's (1969) general guidelines, suggesting that an effect size of 0.2 is small, 0.5 is medium, and 0.8 is large, yet these were only intended as a rule of thumb for the social sciences (Howell, 2013). In L2 research, Plonsky and Oswald (2014) calculated generic medians, based on 25th, 50th, 75th percentile cut-offs, that might reflect patterns of small (0.4), medium (0.7) and large (1.0) effects for between-subject differences, and 0.6 (small), 1.0 (medium), and 1.4 (large) for within-subject differences. However, given that effect sizes for reading time data are rarely reported, in the current study we needed to identify and contextualise the usefulness, magnitude, and meaningfulness of the effects found for our domain of interest.

One step towards gauging the usefulness of effect sizes is to explore the extent to which effects might vary as a function of corresponding results from NHSTs that were reported as 'statistically significant' versus 'not significant'. First, for example, d may reveal a reliable effect (albeit small) where NHST suggested a 'non-significant difference'. Or alternatively, d may reveal an unreliable or negligible difference where NHST results suggested a 'significant difference.' Second, the 95% confidence intervals (CIs) of effects for 'significant' versus 'non-significant' findings may overlap considerably or one effect may sit within the CI of the other effect, suggesting that although the effects may be different descriptively (as well as according to NHST), the difference between them may not be reliable (see Larson-Hall, 2016; Plonsky, 2015).

Another way of interpreting the magnitude and meaningfulness of effect sizes in SPR research is to ascertain the magnitude of differences in reading speeds between participant groups. Regardless of L1-L2 similarity of specific morphosyntax, different individuals use different amounts of time to interpret lexical and morphosyntactic information (Roberts, 2013). Despite variable reading speeds being a widely acknowledged phenomenon (see, e.g., Rayner, Schotter, Masson, Potter, & Treiman, 2016), we do not have any general estimate of the magnitude of difference that might be expected between the kinds of participant groups commonly used in L2 SPR studies. First, it is unclear to what extent L2 learners who are proficiency-matched but have different L1s (and different L1 scripts) vary in their reading speed. While some studies report significant group differences in reading speed (e.g., Pan et al., 2015), others do not, even when L1 scripts differ (e.g., Fender, 2003). Second, it is unclear to what extent reading speed is associated with proficiency (compare, for example, Roberts & Felser, 2011 with Kaan et al., 2015). In terms of speed differences between advanced L2 learners and NSs, it has often been observed that L2 learners read more slowly (e.g., Marinis et al., 2005; Roberts & Liszka, 2013), although some studies suggest otherwise

(e.g., Bel, Sagarra, Comínguez, & García-Alcaraz, 2016; Jegerski, 2016; Kaan et al., 2015). Thus, although not the main focus of our study, estimating the magnitude of speed differences in reading was not only useful methodologically (to identify a frame of reference for our main effect size estimates), but also of substantive interest.

THE PRESENT STUDY

In sum, the current study is a meta-analysis of morphosyntactic sensitivity and L1 influence during L2 SPR. There is sufficient primary evidence for such a meta-analysis, and existing results have not been consistent, possibly reflecting cross-study variation such as study design, processing issue, and linguistic feature, variables that can be submitted to systematic moderator analyses. More generally, given the potential benefits of effect size reporting, the present study is a preliminary attempt to calculate and aggregate effect sizes from reading time data. With a view to identifying a preliminary framework of reference, we investigate the magnitude of differences between: a) effects reported as 'statistically significant' and 'non-significant'; b) reading speeds of proficiency-matched L2 learner groups with different L1s; and c) reading speeds of L2 learner and NS groups. Critically, by using data from a single type of instrument (SPR) we improve cross-study comparability by reducing the effects of heterogeneity due to instrumentation, a frequent concern acknowledged by meta-analysts as they interpret their findings (see Borenstein, Hedges, Higgins, & Rothstein, 2009, pp. 379-80; Norris & Ortega, 2006, p. 16).

RESEARCH QUESTIONS

Our study addressed two substantive research questions:

RQ1) How sensitive are L2 learners and NSs to morphosyntactic violations and disambiguating information during self-paced reading?

RQ2) To what extent is sensitivity to L2 morphosyntactic violations and disambiguating information influenced by L1 background?

It was anticipated that effects for RQ1 and 2 might vary as a function of four moderator variables: study design, processing phenomenon, sentence region, and linguistic feature.

Our third objective (henceforth RQ3) was to capture a frame of reference within which to consider effect sizes during SPR; to address this, we asked the following three sub-questions:

RQ3a) To what extent do effect sizes differ for NHST results reported as 'statistically significant' versus 'not statistically significant'?

RQ3b) What is the average magnitude of difference in online reading speeds for proficiency-matched learners of the same L2 with different L1s?

RQ3c) What is the average magnitude of difference in online reading speeds between L2 learners (of any L1 background) and NSs?

METHOD

Study Eligibility Criteria and Data Collection

For inclusion in our meta-analysis, we applied the following eligibility criteria:

- The study used SPR.
- Raw or residual group mean reading times, standard deviations and n were provided, or t statistics that could be converted to d.[4]
- Participants were adult L2 learners. Studies with young bilinguals or L2 learners (e.g., Marinis, 2007) and heritage speakers raised in a bilingual environment (e.g., Keating, Jegerski, & VanPatten, 2016) were excluded.

- The study focused on morphosyntax. Investigations into other areas such as the lexicon (e.g., Bultena, Dijkstra, & van Hell, 2014) were excluded.
- The potential effect of L1-L2 similarity/difference of the target morphosyntax could be ascertained from the author's predictions or post-hoc explanations (i.e., the study's own design was not required to explicitly test L1 influence; see Supplementary Material 1).
- For inclusion in RQ3b, the study had to have at least two L2 groups with different L1s, with no statistically significant proficiency differences reported between the groups, as proficiency may be a confound (as noted by Hopp, 2006; Juffs, 2005). Non-matching groups were excluded.
- For inclusion in RQ3c, the study had to have at least one L2 group (of any L1 background and any proficiency) and a NS comparison/control group.

The search for studies involved two phases. First, we extracted studies that met the criteria above from Marsden, Thompson and Plonsky's (2018) methodological synthesis of L2 SPR research, for which the search finished in June 2016. Our second phase involved both updating that body of studies to the following 12 months[5] and extending it, given that it had only included journal articles. For our meta-analysis we wanted to increase inclusivity to generate better representativeness and reduce vulnerability to any bias towards the publication of statistically significant results (as recommended by Plonsky & Oswald, 2014). Thus, we included both published and unpublished material, with a view to determining the possibility of publication bias in our sample.

Our search drew on a range of search techniques and online databases, as recommended by Plonsky and Brown (2015). The online databases were ERIC, LLBA, PsycINFO, ProQuest, Cambridge Core, Google and Google Scholar. Relevant authors' websites were also checked and a backwards search of one extensive reference list was also carried out (Juffs & Rodriguez, 2015). The LingRef (Cascadilla Proceedings Project)

database was also consulted, as were the resources of the authors' university library. The search terms "self-paced reading", "moving window" and "subject-paced reading" were cross-referenced with "language", "processing", "learning", "second", "acquisition" and "cross-linguistic influence".

We attempted to address the "missing data problem" (Plonsky, 2013, p. 613) by contacting individual researchers, five of whom provided their data. However, 18 studies had to be excluded because either descriptive statistics or t statistics were not available or not reported for the sentence regions of interest. For three studies that reported mean reading times without standard deviations, we decided, since inclusivity was a priority and our study was exploratory (and following advice from a meta-analyst), to estimate their standard deviations based on those found for the same processing phenomenon and linguistic feature (see Supplementary Material 1).

In the end, for RQ1 and RQ2, which minimally required a single L2 learner group with any L1, 57 studies met the inclusion criteria (i.e., 57 study reports; see Supplementary Material 2). 54 of these studies also included a NS group. 27 manipulated cross-linguistic similarity as either a within- ($k = 11$) or a between-subject ($k = 16$) variable. The remaining 30 studies involved at least one L2 learner group (each with one L1), using L1-L2 similarity/difference of the target feature(s) to either predict or interpret findings.

For the analyses for RQ3a, we used the same sample ($K = 57$) as for RQ1 and 2. For RQ3b, 27 studies met the criteria, all of which used two or more groups of L2 learners with different L1s. Due to reported proficiency differences between groups, four L2 groups were excluded from four separate studies within this set (see Table S2, Supplementary Material 1). All but one of the studies included in RQ3b was also used for RQ3c ($k = 26$), which required the inclusion of a NS group.

Coding the Studies and Extracting Secondary Data

A coding scheme (see Supplementary Material 3) was developed to record data about: bibliographic information (author; year; publication type); participant characteristics (proficiency as reported; L1; L2); potential moderator variables (linguistic feature; processing phenomenon; sentence region, study design for manipulating L1-L2 similarity, i.e., within-subject, between-subject, none). To calculate effect sizes, we extracted sample sizes, mean group reading times for each condition, and standard deviations. For our analysis of the moderator variable 'sentence region', we coded whether each reading time was on the critical, post-critical, or wrap-up region.[6, 7] Finally, to investigate how effect sizes related to the findings from the studies' NHST, we extracted: whether the result was reported as 'significant' or 'not significant'; the t statistic (or equivalent); and corresponding p value. When available, these were recorded separately for the subject and item analyses.

Calculating Effect Sizes

Sensitivity to morphosyntax and L1 influence (RQ1 & RQ2). For all effect sizes, Cohen's d was calculated to measure the mean difference in every paired comparison. For the denominator, we used the pooled standard deviation (the average of two groups' standard deviations weighted according to sample size [Lipsey & Wilson, 2001]), rather than the standard deviation of a control group, as the pooled standard deviation more accurately represents variance when two sample sizes are fairly similar (Coe, 2002; Norris & Ortega, 2006), as was most often the case in our study sample.

To determine sensitivity to grammatical information during SPR, comparisons were drawn between the reading times for an experimental condition and its corresponding 'baseline' condition. To identify each comparison, we adhered strictly to the primary study's within-subject experimental manipulations between conditions. This was done separately for

each morphosyntactic feature investigated. Effect sizes were calculated for L2 learners (for RQ1 & RQ2) as well as for the NSs (for RQ1).

We illustrate how this was done with two examples. First, in a study of anomaly detection with L1 French and L1 German learners of English, the reading times for past simple sentences with tense-aspect agreement (as in 3) were compared (within-subject) with ungrammatical equivalents (as in 4). If L2 learners are insensitive to the tense-aspect violation in (4), their reading times for the two conditions would be similar and the corresponding effect size would be relatively small/negligible. If, on the other hand, L2 learners detect the tense-aspect violation, a slowdown would be expected on the critical and, possibly, post-critical regions in (4) compared to (3). The increased processing cost in (4) would yield a meaningful and reliably larger effect size between the two conditions relative to that of participants who were insensitive to the morphology. Furthermore, as the primary study's authors predict, if L2 learners' sensitivity to tense-aspect violations is influenced by their L1, greater effect sizes would be observed for the French participants, whose L1 grammar marks tense-aspect agreement in the past simple, than for the Germans, whose L1 grammar does not. The size of the difference between reading times for the underlined regions in sentence (3) compared to (4) is given in [ ] for the L1 French and L1 German groups respectively:

(3) Since / last / week / James / <u>has</u> / <u>gone</u> / <u>swimming</u> / every / day

(4) *Last / year / James / <u>has</u> [critical region d = 0.06; 0.21] / <u>gone</u> [post-critical d = 0.00; 0.18] / <u>swimming</u> [post-critical d = 0.58; 0.12] / every / day

(Roberts & Liszka, 2013, p. 421 [d not in original])

A similar type of comparison of effects is illustrated via a different processing phenomenon: ambiguity resolution. L2 learners' sensitivity to morphosyntax that

disambiguates meaning during online processing can be tested by comparing their reading time in a temporarily ambiguous experimental condition with that in a comparison condition. For example, Juffs (2005) investigated L2 learners' assignment of subject (S) and object (O) roles in sentences with wh-interrogatives. L1 Chinese, L1 Japanese and L1 Spanish learners of English read sentences such as (5) and (6):

(5) Who$_i$ / did / the / woman / suggest / the / manager / liked e_i / at / the / office?

(6) Who$_i$ / did / the / woman / suggest e_i / liked / the / manager / at / the / office?

In (5), the parser extracts an object in the embedded clause after 'liked', whereas in (6) it extracts a subject in the embedded clause after 'suggest'. Since the parser seeks to build sentence structure as early as possible, it is possible that in (6), 'who' is initially analysed as the object of the main clause ('Who did the woman suggest'). This initial interpretation needs to be revised on reaching the verb of the embedded clause ('liked') as this verb now requires a complementiser phrase. This would result in a slowdown in (6) compared to (5) at word six. If L2 learners' sensitivity to wh-dependencies is influenced by L1-L2 correspondence, greater effect sizes would be observed for the Chinese and Japanese participants, whose L1s do not instantiate wh- movement, than for Spanish participants, whose L1 does. As per the primary study's author's predictions, syntactic reanalysis should be less costly for participants of Spanish, an L1 that builds wh- chains in a similar way to English. See sentences (7) and (8) to illustrate, with the d that we calculated provided in [ ] for L1 Japanese, L1 Chinese and L1 Spanish, respectively.

(7) Who$_i$ / did / the / woman / suggest / the / manager / liked e_i / at / the / office?

(8) Who$_i$ / did / the / woman / suggest e_i / liked [critical region d = 0.66; 0.63; 0.36] / the / manager / at / the / office?

(Juffs, 2005, pp. 136-7 [d not in original])

Thus, to investigate whether L1 background moderates the magnitude of effect sizes, each comparison between conditions was coded for 'potential L1 effect' or 'no potential L1 effect', as per the primary study's author's manipulations and argumentation. These terms were used to code the predicted magnitude of effects based on L1-L2 correspondence of the target grammar feature: 'potential L1 effect' refers to conditions in which L1-L2 convergence[8] (e.g., L1 French, Roberts & Liszka, 2013) or non-convergence (e.g., L1 Japanese, Juffs, 2005) was predicted to render a larger effect size, whereas comparisons coded for 'no potential L1 effect' concerned conditions in which L1-L2 convergence (L1 Spanish, Juffs) or non-convergence (L1 German, Roberts & Liszka) was predicted to render a smaller effect size. Note, L1-L2 convergence can facilitate sensitivity that either reduces or increases processing cost, depending on the study in question. Since larger effect sizes are not necessarily more desirable or indicative of 'positive' L1 influence, we decided against using the terms 'positive' and 'negative' to describe potential and no potential L1 effect.

If a study did not provide the necessary information to allow coding for '(no) potential L1 effect', data for that feature and participant group were excluded. As a result, data from eight studies were excluded for these comparisons (see Table S1, Supplementary Material 1). Further exclusions were also made for two studies in which effect sizes could not be calculated due to the absence of a comparison condition in the within-subject experimental manipulation.

An experienced second coder (the second author) checked (a) the coding for '(no) potential L1 effect' and (b) the suitability of conditions being compared in 14 studies (25% of the total sample). Once the coding scheme had been piloted (with approximately nine studies) and finalised, the extraction of data was then an objective process (e.g., extraction of bibliographic, unequivocal study design features, reading times, and reported statistics).

In sum, if L1-L2 convergence or non-convergence affect online sensitivity to an L2 feature, reliably larger effect sizes are potentially expected for comparisons coded as 'potential L1 effect' than for those coded as 'no potential L1 effect'. If, on the other hand, the L1 does not influence online sensitivity to an L2 feature, effect sizes coded for 'potential L1 effect' and 'no potential L1 effect' would not differ reliably.

The size of effects *corresponding to* 'statistically *significant' and 'non-significant'* NHST results (RQ3a). For this analysis, each effect size extracted for RQ1 and RQ2 above was coded for whether the authors reported a difference that was statistically significant (or not) by both participants and items, by participants, and by items.[9] (Only two studies reported NHSTs that were only significant by items and these are therefore not reported here; see Supplementary Material 5.)

Between-group comparisons for speed differences (RQ3b & RQ3c). To estimate speed differences between different learner groups for RQ3b, we compared the reading times of two groups of L2 learners with different L1s on the same SPR task for every condition and every critical, post-critical or wrap-up region, as reported. Since we were interested in the size rather than the direction of difference, we recorded the absolute value of d (i.e., removed minus signs); whether one L1 group was faster or slower than another was not of interest here. Instead, we were interested in ascertaining preliminary estimates of basic speed differences between groups to contextualise the results from RQs 1 and 2.

To estimate speed differences between L2 learners and NSs for RQ3c, comparisons were drawn between the reading times of each learner group and the NS group on the same SPR task for every condition and every critical, post-critical or wrap-up region. Since we wanted to ascertain whether any difference between L2 learners and NSs reflected faster or slower reading, the direction of difference was also recorded (i.e., keeping minus signs).

To assess the reliability of both the individual effect sizes and of the grand weighted mean (see below), 95% CIs were recorded. If the CIs of an effect size pass through zero, the effect size is unlikely to be reliably different to zero (Cumming, 2012; Larson-Hall, 2016). When comparing two effect sizes, the difference between them is sometimes considered reliable if their CIs do not overlap (see, e.g., Lee & Huang, 2008; Lee, Jang, & Plonsky, 2015). Other guidelines suggest that "if the two CIs just touch, p is about 0.01, and the intervals can overlap by as much as about half the length of one CI arm before p becomes as large as 0.05" (Cumming, 2009, p. 205; see also Cumming, 2012; Greenland et al., 2016, p. 344). According to a third approach, the difference between two means is reliable if "the mean of one group falls … outside the CI for the other group's mean" and vice versa (Plonsky, 2015, p. 40; see also Greenland et al., 2016). In describing and interpreting differences between our effect sizes, we considered these three criteria together. Furthermore, in Supplementary Material 6, we provide detailed interpretations of the results of the present study according to these three criteria.

Averaging and Weighting

For within-subject comparisons (RQ1 and RQ2), an average effect size estimate was first calculated based on comparisons for each unique sample. We repeated this for each unique sample and then averaged the results to calculate an overall estimate of learners' sensitivity to L2 morphosyntax and an overall estimate of NS sensitivity to morphosyntax. For RQ1, these averages were based on 130 effect size estimates for the L2 learners and 61 effect size estimates for NSs.[10, 11] Effects were coded for linguistic feature, processing issue, and sentence region, as potential moderators of sensitivity to L2 morphosyntax.

Having calculated an overall effect size estimate for RQ1, the effect sizes were then separated, for each unique sample, according to 'potential L1 effect' (91 effect sizes) and 'no

potential L1 effect' (62 effect sizes) for RQ2. Effects were coded for study design, linguistic feature, processing issue, and sentence region, as potential moderator variables of the average 'potential L1 effects' and 'no potential L1 effects'.

For RQ3a, which investigated the size of within-subject effects corresponding to 'statistically significant' and 'non-significant' NHST results, the effect sizes, calculated by unique samples, were averaged to provide a grand estimated effect for L2 participants (and also separately for 'potential L1 effect', 'no potential L1 effect') and for NSs. The grand estimated effects for differences reported significant by 'participants and items', 'significant by participants', and 'not significant' were based, respectively, on 16, 21, and 42 effect sizes for L2 learners, and on 21, 14, and 15 effect sizes for NSs.

For RQ3b, we averaged effect sizes of speed differences between two groups of L2 learners of different L1s on the same SPR task. We note that for those studies with three or more L1 groups (k = 8), this method used the results from the same L1 group more than once (e.g., L1a vs. L1b, L1a vs. L1c). Although this may underestimate the overall level of sampling error in summary figures (Borenstein et al., 2009), it was deemed necessary in order to retain the relevant sample-level variability. The alternative, which would have involved aggregating all comparisons to calculate a single effect size for each study (Lipsey & Wilson, 2001), would have reduced the detail in our data by collapsing the size and L1 characteristics of each L1 subsample. Our process yielded 55 effect size estimates for RQ3b.

For RQ3c, an average was calculated from each set of effect sizes for an L2 vs NS comparison on the same SPR task. This generated 73 effect size estimates, a higher figure than for RQ3b because (a) comparisons in studies with two L2 groups (k = 13) contributed two effect sizes in RQ3c (L1a vs. NSs, L1b vs. NSs), but only one in RQ3b (L1a vs. L1b) and

(b) as noted above, effect sizes were not calculated between L2 groups for which the authors had reported proficiency differences.

In meta-analytic research, statistical models (e.g., with fixed and random effects) may be chosen to combine effects across studies (see Borenstein, Hedges, & Rothstein, 2007). However, this was not deemed to be important in the present study, as recommendations for L2 meta-analysis downplay the importance of homogeneity tests, citing their low degree of statistical power and consequent risk of inaccurate conclusions (Oswald & Plonsky, 2010). We opted instead for an approach based on averaging, weighting, and theoretically motivated moderator analyses.

In line with good practice in meta-analysis, effect sizes from larger samples should represent a greater proportion of the grand effect size (Borenstein et al., 2007). To this end, we used inverse variance weighting, whereby the weight of a given effect size is proportionate to the inverse of the sampling error variance of the two groups.[12]

Moderator Analyses

Following the grand effect size calculations for RQ1 and RQ2, moderator analyses were undertaken, as follows:

(a) **Processing issue (RQ1 & 2):** To investigate whether morphosyntactic sensitivity and L1 influence varied as a function of different processing phenomena, our moderator analysis isolated anomaly detection (involving violations) and ambiguity resolution (involving grammatical sentences only).

(b) **Linguistic feature (RQ1 & 2):** Online sensitivity to morphosyntax may vary as a function of linguistic feature. In our study sample, 18 types of linguistic feature were investigated and we calculated average effect sizes for each. This allowed us to isolate studies that investigated morphosyntax that aligns (or not) with parsing preferences (in studies of

relative clause attachment), from other studies where the morphosyntax was operationalised as being invariably grammatical or not (i.e., less reliant on preferences). As noted above, results from parsing preference studies often seem to have suggested no L1 influence, whilst results from the latter may be more likely to observe L1 effects.

For RQ2, we were able to explore the extent to which eight linguistic feature types moderated L1 influence, i.e., those that had been investigated both in contexts where there was potential and no potential L1 influence, though k was small for some of these features.

(c) **Sentence region (RQ1 & 2):** To date, we have little general sense of the extent to which heavier processing costs are observed on one region compared to another, although L2 processing effects have often been thought to manifest themselves in 'spillover' sentence regions after the critical region. Thus, we separated critical, post-critical and wrap-up regions as a potential moderating variable.

(d) **Study design (RQ2 only)**: Within-subject manipulations of cross-linguistic similarity, where the same participants yield effects for both 'potential L1 effect' and 'no potential L1 effect', may yield larger indicators of L1 influence than between-subject manipulations (in between-group and single group designs), where the comparisons for each unique sample were coded as either 'potential L1 effect' or 'no potential L1 effect'. Thus, our moderator analysis investigated the extent to which aggregated results for potential versus no potential L1 effects were influenced by these different study designs.

Although proficiency level was also of interest it could not be investigated as a moderator because diverse proficiency measures were used (see Supplementary Material 4), an issue also noted in a recent meta-analysis of lexical activation in bilingual sentence processing (Lauro & Schwartz, 2017; see also Thomas, 2006). Furthermore, the majority of our studies in our sample (k = 44) involved L2 participants at advanced or near-native level,

reflecting researchers' interest in the later stages/end-state of L2 acquisition and their concern that L2 participants comprehend the sentences (Marsden, Thompson, & Plonsky, 2018).

Visual Data Inspection for Influence of Publication Bias and Sample Size on Effects

The distribution of effects was visually inspected via a funnel plot of the effect sizes for sensitivity to L2 morphosyntax (RQ1), where asymmetries may indicate the presence of (unplanned) moderator variables in the data. Figure 1 shows that a greater number of effect sizes appear to the left of the overall unweighted mean (d = 0.21). This plot does not very clearly reflect the traditional funnel shape with effects falling evenly on either side of the mean. Instead, it suggests a tendency for studies to have small effects with effect sizes clustering below the mean effect size (as also found by Plonsky, 2011), and a mean that is affected by a smaller number of studies with large effect sizes. The plot illustrates the expected influence of sample size, in that effects found for small samples are more widely distributed (with a wider range of effects) than those of larger samples. However, it also suggests a dispersion of effects that is not similar either side of the mean, with a smaller range in the dispersion found to the left of the mean compared with the range of effects found above the mean. This might indicate some kind of suppression of effects that might be observed in the 'true' population (for further explanation see Oswald & Plonsky, 2010). However, it seems unlikely that this is due to a publication bias that generally suppresses small or negative effects. Figure 1 shows that effects from published sources were unlikely to be larger than those found for unpublished sources and, in fact, that fewer unpublished studies found effects below the mean (k [unique samples] = 7) than above the mean (k [unique samples] = 14).

<<INSERT FIGURE 1 ABOUT HERE>>

We also checked for publication bias by carrying out a moderator analysis of RQ2 results, relating to the influence of L1, as this is a controversial theoretical issue that might conceivably engender a publication bias (one way or the other). The overall mean effect found from published sources (journal articles, edited book chapters) was d = 0.18 (0.13, 0.24), k (unique samples) = 109 and this was similar to the mean from unpublished sources (i.e., conference proceedings, dissertations, working papers; d = 0.29 [0.15, 0.43], k [unique samples] = 21), with the mean from published sources sitting within the 95% CIs of the mean from unpublished sources and substantial overlap in the CIs of both means. In line with what Figure 1 indicated, this did not suggest a reliable effect of publication type in the expected direction of larger effects in published compared to unpublished work. However, the slightly larger effect in unpublished sources (with a mean sitting above the 95% CIs of the mean from published sources) may tentatively indicate a borderline trend to publish or submit for publication findings that suggest little or no influence of L1 compared to findings that suggest a larger influence of L1 on morphosyntactic sensitivity during reading. On the other hand, it may also suggest other tendencies, such as a greater use of anomaly detection in unpublished studies (see Results for RQ2).

Another indication of other types of bias relate to the reporting of non-significant effects and the relationship between the results of NHST in the primary studies and the effect sizes that we extracted – issues that are examined for RQ3a.

RESULTS

Sensitivity to L2 Morphosyntax (RQ1)

For RQ1, summary results are reported first, followed by moderator analyses. Our general estimate of sensitivity to morphosyntactic information during L2 SPR was d = 0.20 (0.15, 0.25), based on 130 unique samples and 3,052 L2 participants. The 95% CIs did not pass

through zero, suggesting a reliable effect that would normally fall around one fifth of a standard deviation unit. Expressed in another way, this suggests that advanced L2 learners changed their processing speeds by about one fifth of a standard deviation unit when elements in a sentence coerced sensitivity to morphosyntactic information (overall, across anomaly detection and ambiguity resolution). Among the NS groups, we found the magnitude of overall sensitivity to be d = 0.28 (0.21, 0.35), drawn from 61 unique samples and 1,494 participants. Note that the means of the L2 learners and NSs fall just outside each other's CIs and the CIs overlapped by approximately half an average CI arm, suggesting that the two grand effect sizes are likely to be reliably different.

These are general estimates, irrespective of potential L1 effect, processing phenomenon, linguistic feature, or sentence region. L1 influence is examined in RQ2 (below). In Table 1 we present analyses for the other three potential moderators. As shown in Table 1, for L2 participants we did not find processing phenomenon was a reliable moderator as CIs for the effects overlapped considerably and the means sat within each other's CIs. For NS groups, on the other hand, processing phenomena did seem to moderate effects, with evidence of a greater sensitivity during anomaly detection than ambiguity resolution – note that the means sat outside each other's CIs and the CIs only overlapped very slightly. Furthermore, the difference between NSs and L2 speakers for anomaly detection seemed to be reliable, with means that sat outside each other's CIs and CIs that only touched end-to-end. For ambiguity resolution, however, there was no evidence of a reliable difference in effects between L2 learners and NSs.

In terms of linguistic feature, studies were highly heterogeneous, with 18 different linguistic features investigated across our sample. Results are presented in full in Supplementary Materials 5, showing a low number of unique samples for some features. In Table 1 we present the five most frequently investigated. Of these five features, only relative

Table 1. Results of moderator analyses for morphosyntactic sensitivity

Moderator variable	Sub-category	L2 learners Cohen's d [95% CIs] (k unique samples)	Native speakers Cohen's d [95% CIs] (k unique samples)
Processing phenomenon	anomaly detection	0.19 [0.09, 0.29] (38)	0.41 [0.29, 0.54] (21)
	ambiguity resolution	0.19 [0.12, 0.25] (84)	0.23 [0.13, 0.32] (34)
Linguistic feature	gender agreement	0.23 [0.10, 0.37] (13)	0.44 [0.24, 0.64] (7)
	subject-object role assignment	0.28 [0.14, 0.42] (25)	0.37 [0.14, 0.59] (7)
	number agreement	0.21 [0.09, 0.32] (20)	0.49 [0.32, 0.67] (9)
	wh-dependencies	0.25 [0.14, 0.35] (28)	0.18 [0.01, 0.35] (11)
	relative clause attachment	0.02 [-0.15, 0.18] (13)	0.10 [-0.10, 0.31] (9)
Sentence region	critical	0.19 [0.14, 0.24] (130)	0.26 [0.19, 0.33] (61)
	post-critical	0.20 [0.13, 0.27] (63)	0.31 [0.21, 0.41] (35)
	wrap-up	0.10 [-0.01, 0.21] (24)	0.21 [0.02, 0.41] (8)

clause attachment studies were not reliable, with CIs that passed through zero for both L2 learners and NSs, suggesting that core, invariable morphosyntax might have yielded stronger effects than attachment preference studies. Notably, two of the highest effect sizes were 0.49 and 0.44, for number and gender agreement respectively, among NSs. However, these effects were largely derived from anomaly detection studies (k [unique samples] = 15 out of 16), suggesting these higher effects may (also) be attributable to the nature of the processing

phenomenon investigated (i.e., sensitivity to violations) rather than to (only) the specific type of morphosyntax.

In terms of sentence region, there was tentative evidence that this moderated effects for L2 learners, given that effect sizes for critical and post-critical regions were reliable (and similar to each other) as their CIs that did not pass through zero, in contrast to sentence-final wrap-up segments which had CIs passing through zero. However, the effect sizes for critical and post-critical regions did not fall outside the CIs of wrap-up regions and their CIs overlapped those of wrap-up regions by almost one average arm length, suggesting a lack of reliable difference. For NSs, effect sizes for all regions were reliable and not reliably different to each other.

Morphosyntactic Sensitivity Moderated by L1 Background (RQ2)

Regarding the extent to which L1 background moderated the online sensitivity of L2 learners, effect sizes were averaged separately, and by unique samples, for conditions coded as 'potential L1 effect' (d = 0.23 [0.17, 0.29], k [unique samples] = 90), and 'no potential L1 effect' (d = 0.19 [0.12, 0.26], k [unique samples] = 64). The two means fell within each other's CIs and overlapped considerably, suggesting no reliable difference, and therefore, no reliable, general effect of cross-linguistic influence.

The results of three of our four moderator analyses for RQ2 are presented in Table 2. Effects from studies with a within-subject design were descriptively, but not reliably higher than those from studies with a between-subject design, both for potential L1 effect and no potential L1 effect, with at least one mean that fell inside the CIs of the other and CIs that overlapped considerably. For studies with a within-subject design, effect sizes did not differ reliably as a function of potential L1 effect versus no potential L1 effect, suggesting no reliable L1 influence. However, for studies with a between-subject design, effect sizes with

potential L1 effect appeared to be reliably greater than those with no potential L1 effect, with means that fell outside the other's CIs and CIs that overlapped by around half an average arm length. With regard to processing issue, effects in anomaly detection studies for conditions with potential L1 effect were reliably larger than those with no potential L1 effect, with means that fell outside the other's CIs and CIs that only overlapped slightly. In contrast, effect sizes for ambiguity resolution studies did not appear to differ reliably. Comparing these two processing phenomena, effect sizes with potential L1 effect for anomaly detection appeared to be reliably greater than those with potential L1 effect for ambiguity resolution, with means that sat outside the CIs of the other and CIs that overlapped by about half an average arm length.

The other analyses investigated whether effect sizes for specific sentence regions and linguistic features were moderated by potential L1 effect versus no potential L1 effect. In terms of sentence region, effects did not appear to be moderated by L1 background as the means for potential L1 effect and no potential L1 effect overlapped considerably for every sentence region, though there was tentative evidence of a trend toward larger effects on post-critical regions with potential L1 effect, given that each mean fell outside each other's CIs. In the analyses by linguistic feature, it was possible to analyse eight of the 18 linguistic features for potential L1 effect versus no potential L1 effect (the remaining ten could be coded only for either potential L1 effect or no potential L1 effect; see Table S1, Supplementary Materials 5). For every feature, the effect sizes for potential L1 effect versus no potential L1 effect clearly overlapped in their CIs, suggesting little meaningful difference. Two findings aligned with a pattern that could tentatively support a prediction that the L1 influences effects: tense-aspect agreement yielded a reliable effect for potential L1 effect (d = 0.41 [0.10, 0.73], k [unique samples] = 4), but not for no potential L1 effect (d = 0.11 [-0.23, 0.46], k [unique samples] = 3); similarly, number agreement yielded a reliable effect size for

Table 2. Results of moderator analyses for potential L1 effect and no potential L1 effect

Moderator variable	Sub-variable	d [95% CIs] for potential L1 effect (k unique samples)	d [95% CIs] for no potential L1 effect (k unique samples)
Study design	within-subject manipulations of L1-L2 similarity	0.34 [0.22, 0.45] (26)	0.27 [0.16, 0.39] (26)
	between-subject manipulations of L1-L2 similarity[a]	0.31 [0.18, 0.44] (22)	0.13 [-0.01, 0.27] (18)
Processing phenomenon	anomaly detection	0.34 [0.21, 0.48] (15)	0.17 [0.08, 0.26] (34)
	ambiguity resolution	0.19 [0.12, 0.27] (67)	0.22 [0.10, 0.33] (30)
Sentence region	critical	0.22 [0.16, 0.28] (90)	0.18 [0.11, 0.25] (65)
	post-critical	0.28 [0.19, 0.37] (40)	0.18 [0.09, 0.27] (34)
	wrap-up	0.14 [0.01, 0.27] (17)	0.06 [-0.09, 0.22] (13)

Note: [a]These results were very similar when split into single versus multi-group designs.

potential L1 effect (d = 0.30 [0.15, 0.44], k [unique samples] = 11) but not for no potential L1 effect (d = 0.10 [-0.07, 0.26], k [unique samples] = 10). However, the considerably overlapping CIs do not provide convincing evidence of a reliable influence of L1 for either feature, and mean effects for tense-aspect agreement were based on small number of samples. We also note that the effect size for relative clause attachment with potential L1 effect (d = 0.02 [-0.15, 0.18], k [unique samples] = 13) was not reliable, with CIs that passed through zero. This differed reliably from subject-object role assignment with potential L1 effect (d = 0.31 [0.17, 0.46], k [unique samples] = 21), with means that fell outside the CIs of the other and CIs that only overlapped marginally, suggesting that potential L1 influence might have

been moderated by the type of morphosyntax (core, invariable morphosyntax versus morphosyntax used for interpretation preferences).

Summing up the results for RQ2, no reliable difference was found between our grand effect sizes for 'potential L1 effect' and 'no potential L1 effect'. However, the results of moderator analyses only aligned to some extent with this. In particular, the effects for anomaly detection studies seemed to be moderated by potential L1 effect as those effects were reliably larger than those for no potential L1 effect. The average effect size for anomaly detection also seemed reliably larger than for ambiguity resolution, at least for conditions with potential L1 influence. In the analyses by study design, potential L1 effect appeared to yield reliably larger effects than no potential L1 effect for between-subject study designs, but not for within-subject study designs. Finally, the magnitude of effects for different sentence regions and linguistic features was largely unaffected by L1 background, though two linguistic features with potential L1 effect, relative clause attachment and subject-object role assignment, did differ reliably, in line with the suggestion that the type of morphosyntax might moderate L1 influence.

An Initial Contextual Framework for Interpreting Effect Sizes of Reading Times in L2 SPR (RQ3).

RQ3 sought to explore the usefulness and meaningfulness of the effect sizes found above, and in SPR research more generally, through three analyses.

First, for RQ3a, we estimated the average effect sizes for within-group contrasts of overall sensitivity to morphosyntax that had been reported as 'statistically significant' or 'not significant'. For contrasts reported as significant by both participants and items analyses, the mean effect size was d = 0.64 (0.49, 0.79), k (unique samples) = 16 and by participants d = 0.55 (0.44, 0.67), k (unique samples) = 23; both these effects were reliable. For findings

reported as not statistically significant, mean d = 0.17 (0.08, 0.25), k (unique samples) = 42. That is, the effect was considerably smaller and, reassuringly, its CIs did not overlap with those of the mean effect for 'statistically significant' findings. This suggests a reliable difference between the magnitudes of differences that had been originally reported as statistically significant and non-significant. However, note that the CIs for the effects for 'non-significant' differences did not pass through zero, suggesting some small but reliable, and therefore potentially meaningful, effect, of approximately one sixth of a standard deviation unit. This pattern was the same regardless of whether the context was 'potential L1 effect' or 'no potential L1 effect' (see Table S2, Supplementary Material 5). For NSs the results patterned similarly: for comparisons reported as statistically significant by both participants and items analyses, d = 0.64 (0.52, 0.76), k (unique samples) = 21 and by participant analyses, d = 0.46 (0.29, 0.62), k (unique samples) = 14. Again, these effects were reliably larger than for comparisons reported as not significant, d = 0.18 (0.02, 0.33), k (unique samples) = 15. Again, however, the 95% CIs did not pass through zero, suggesting a small (just under a fifth of a standard deviation unit) but reliable difference for within-subject comparisons where 'null' NHST results could be (mis-)represented as evidence of no differences.

For RQ3b we estimated the size of difference between the processing speeds of two groups of participants with different L1 backgrounds learning the same L2 undertaking the same SPR task. A grand weighted average effect size of d = 0.35 (0.26, 0.43) was found, from 55 effect size estimates from 75 unique samples with a total of 1,540 participants. The CIs suggest that this comparison is statistically reliable. This is a relatively crude between-group comparison of basic speed differences, regardless of the experimental manipulation or L1-L2 convergence. The result also appears to be reliably greater than the estimated mean magnitude of general L2 morphosyntactic sensitivity, with non-overlapping CIs.

RQ3c addressed the general magnitude of difference between the processing speeds of an L2 learner group (of any L1) and an NS group undertaking the same SPR task. We found a weighted average effect size of d = 0.53 (0.46, 0.60), comprising 73 effect size estimates from 99 unique samples with 2,145 participants. Strikingly, this grand effect size is reliably larger (in raw terms, by about a fifth of a standard deviation) than that found for differences between learner groups with different L1s.

DISCUSSION

In answer to RQ1, we found that learners showed a reliable sensitivity to L2 morphosyntax of about one fifth of a standard deviation unit. When set against benchmarking used both traditionally (Cohen's) and in the specific field of L2 research (Plonsky & Oswald's), this is a small to very small effect. However, given that two of our largest average effect sizes were for NSs processing anomalies in gender agreement (d = 0.44) and number agreement (d = 0.49) we cannot conclude that 0.20 is necessarily 'small' in the context of reading times during SPR. The magnitude of sensitivity was different between L2 learners (generally advanced/near-native) and NSs, with means that fell outside of each other's CIs and CIs that overlapped by around half an average arm length. In answer to RQ2, our overall results did not suggest any kind of general L1 influence during L2 SPR, and in this sense they are consistent with observations that L1 effects in online L2 morphosyntactic processing are "elusive" (VanPatten & Jegerski, 2010, p. 9) and "more limited than one might expect" (Clahsen & Felser, 2006b, p. 565).

The main findings, both for overall morphosyntactic sensitivity and L1 effects, were partially altered by our moderator analyses. On the one hand, with regard to study design, the results for within-subject manipulations of L1-L2 similarity, anticipated to be the most likely to yield observable and reliable results for RQ2, did not vary reliably as a function of

'potential L1 effect', and since 25 unique samples were used in this moderator analysis, the finding cannot be entirely attributed to a lack of statistical power. Nevertheless, we did find a reliable effect of the L1 in our studies with between-subject designs. This might possibly reflect systematic 'cluster' effects due to various known and unknown shared group characteristics, as our grand effect size for speed differences between L2 learners of different L1s (d = 0.35) also suggests. The results of moderator analyses by processing issue also diverged to some extent from the overall findings; on the one hand, L2 learners' general sensitivity seemed to be unaffected by processing issue, yet their sensitivity was reliably greater during anomaly detection than ambiguity resolution in comparisons with potential L1 effect, suggesting that the type of processing phenomenon under investigation may have moderated L1 influence.

One challenge that arose in discussing our results was how to interpret whether two effect sizes were similar or different. Different guidelines rendered different conclusions for a number of our analyses, and this is clearly indicated in Supplementary Materials 6. For example, for L2 learners versus NSs during anomaly detection, the CIs touched end-to-end. If entirely non-overlapping CIs are used to determine a difference between means (e.g., Lee & Huang, 2008), this difference was not reliable. However, if overlap of about half an average CI arm (Cumming, 2009, 2012) or one mean falling outside the 95% CIs of the other (Plonsky, 2015) is used, then the difference was reliable (see Supplementary Material 6). Similarly, for L2 learners with potential L1 effect versus no potential L1 effect during anomaly detection, the difference was unreliable according to non-overlapping CIs, but reliable according to the other two criteria. In other words, according to two of the three criteria, the results suggested a reliable native/non-native difference and reliable L1 influence for L2 learners during anomaly detection.

Interestingly, the three criteria above also yield different interpretations for results in the moderator analyses by linguistic feature. Although these analyses did not generally yield many meaningful differences, the CIs of our grand effect size for L2 learners in studies of relative clause attachment overlapped only very slightly with that those of wh-dependencies and subject-object role assignment (see Supplementary Material 6). These differences were not reliable according to the criteria of non-overlapping CIs, but were reliable according to the criteria of Cumming (2009) and Plonsky (2015). Consequently, we do not rule out the possibility that sensitivity differed for core morphosyntax versus processing preferences, at least for these features, though we remain cautious given that these analyses were based on relatively few unique samples. More generally, our findings suggest that the use of non-overlapping CIs to determine the reliability of a difference between two means is relatively conservative compared to other criteria (Schenker & Gentleman, 2001). We hope that future meta-analyses, given a wider body of sample studies and greater consensus in the field on interpreting reliable differences between effect sizes, can reach firmer conclusions about the processing phenomena investigated in this paper.

In terms of informing theories about the nature of online processing among advanced L2 learners, we offer several insights, albeit tentatively. Overall, we found different morphosyntactic sensitivity among advanced L2 learners and NSs and our analysis of the subset of anomaly detection studies also found a reliable difference between advanced L2ers and NSs, lending some evidence to the notion that shallow morphosyntactic processing may happen more often among L2 learners than NSs (Clahsen & Felser, 2006a, 2018; Clahsen, Felser, Neubauer, Sato, & Silva, 2010). Also aligning to some extent with the idea that "the hypothesized tendency [for L2 learners] to underuse grammatical information … is independent of a learner's L1" (Clahsen & Felser, 2018, p. 697), we found no reliable evidence for general L1 influence on morphosyntactic sensitivity. However, shedding some

light on the current agnostic position of the Shallow Structure Hypothesis, we found that in this set of anomaly detection studies, L1-L2 similarity tended to slightly, but reliably, increase sensitivity to morphosyntax compared to when the L2 feature did not share characteristics with the L1, at least among the advanced L2 learners in our study sample. Future studies are required to further probe these tendencies, including evidence from different data elicitation techniques.

Another avenue for further investigation might be the extent to which making an explicit judgement about grammaticality after having read the sentence affects online reading times. Most (k = 40) studies included only a comprehension question at the end of (some) sentences, but 12 sought grammaticality judgements after reading the sentence (Dekydtspotter, Edmonds, Fultz, & Renaud, 2010; Jackson & Dussias, 2009; Jackson & van Hell, 2011; Juffs, 1998a, 1998b; 2005; 2006; Juffs & Harrington, 1995; Jung, 2010; Lakshamanan et al., 2009; Perpiñán, 2014; Renaud, 2014). A further two studies used the 'stop-making sense' paradigm that elicits an explicit judgement about meaning during reading. These 14 studies cover a wide range of processing and grammatical phenomena and present too small a sample for a useful moderator analysis in the current study. However, we urge future researchers to explore the possibility that explicit judgements alter differences between NSs and L2 learners or any L1 effects.

We acknowledge that our substantive findings for RQ1 and RQ2 remain somewhat tentative given that they draw on results from one data elicitation method among a number used in L2 processing and from learners of advanced L2 proficiency, who were also very often highly literate, educated, and proficient readers in their L1 and L2 (see Hulstijn, 2018; Marsden, Thompson, & Plonsky, 2018). Indeed, the lack of strong and consistent L2 learner-NS differences or L1 influence across all our moderator analyses (only really emerging for anomaly detection) may in part be due to the high levels of proficiency and literacy of L2

participants in our study sample. Such high-proficiency L2 learners may reach a stage where L1 effects are only very selectively observable, at least on SPR tasks. Indeed, a number of studies using SPR have found that proficiency moderated the sensitivity of learners to features that do not exist or are different in their L1. For example, Sagarra and Herschensohn (2010, 2011) found that intermediates and NSs experienced slowdowns for gender agreement violations, while beginners did not. Similarly, Jackson (2008) found that advanced learners of German, unlike intermediates, were able to use case marking to determine the subject of a sentence, even though their L1 English largely relies on word order for this. These findings are broadly consistent with those of Bel et al. (2016), Hopp (2010), Jackson and Dussias (2009), Jackson and van Hell (2011) and Lee, Lu, and Garnsey (2013). We acknowledge, nevertheless, that as most studies in our synthesis involved advanced and near-native proficiencies, further research with other proficiencies is needed to ascertain the extent to which L1 influence during L2 SPR is attenuated by proficiency. Such an agenda will certainly require increased parity in the proficiency measures used across studies (Thomas, 2006).

A further issue to be considered when accounting for the effects found for RQ1 and RQ2 is the extent to which general reading speeds play a role. Evidence from Kaan et al. (2015) suggested that faster readers showed smaller and more delayed grammaticality effects than slower readers, both among NSs and L2 learners. After controlling for differences in reading speed between the L2 learners, who read faster overall than NSs, the authors found that L2 learners and NSs patterned similarly in their sensitivity to grammaticality. Our findings of slower reading among L2 learners and overall lower sensitivity among L2 learners versus NSs do not seem to align with Kaan et al.'s more fine-grained study. More research is needed to understand the interaction of reading speed and sensitivity to morphosyntax. The individual difference of reading speed may in fact exert more of an

influence over online reading behaviours than 'native versus L2 learner' status or specific L1-L2 morphosyntactic relations.

Our RQ3 explored a broader framework for interpreting effect sizes in SPR research. First, effect sizes for results reported as 'statistically significant' were reliably greater than those reported as 'not significant', both for L2 and NSs, with the upper-bound 95% CI reaching a maximum of d = 0.76 for NSs. Interestingly, however, results reported as 'not significant' rendered a small but reliable effect, with a maximum upper-bound CI of about one third of a standard deviation unit, suggesting the presence of small differences where NHST results were, for all intents and purposes, broadly interpreted as indicating 'no difference'. This pattern of findings was similar regardless of L1 background. Since non-significant effects were not widely reported for our comparisons (L2 learners: k [unique samples] = 42 out of 130; native speakers: k [unique samples] = 15 out of 61), we wonder how much this figure would have changed if authors had reported all non-significant effects. This highlights the need for effect size reporting to move SPR research beyond the arbitrary dichotomy and sample-size dependency inherent in NHST towards more nuanced interpretations. It may also indicate the need for an increased use of Bayes factors to enable L2 researchers to make more accurate inferences about accepting null hypotheses (discussed by Dienes, 2014, and see Morgan-Short et al., 2018, for an example). Our effect sizes can also be used to establish 'priors' for future Bayesian analyses (that is, estimates of intervals of effect sizes within which future results may occur [Norouzian, de Miranda, & Plonsky, 2018]).

The other analyses for RQ3 provided general estimates of processing speed differences between groups that are typically included in SPR research. The grand effect size for proficiency-matched L2 learner groups of different L1s (d = 0.35) may be small according to existing L2 benchmarks for between-group comparisons, and yet it was larger than our

estimated grand effect for overall morphosyntactic sensitivity among L2 learners. As previously noted, such between-group speed differences may possibly reflect 'cluster' effects, highlighting the importance and difficulty of establishing parity between learner participant groups.

The grand effect size for comparisons between L2 learners (of any L1) and NSs (d = 0.53) was reliably greater than differences between L2 groups. This supports the notion of a "general L2 processing effect" (Roberts, 2007, p. 124) where L2 processing tends to be slower than L1 processing, regardless of variables such as linguistic feature, processing issue, and cross-linguistic similarity.[13] These findings are noteworthy; even though our dataset drew largely on participants with advanced or near-native L2 proficiency, they still exhibited overall slower processing on the sentence regions examined. This likely reflects the demands of real-time access to and integration of morphosyntactic, lexico-semantic and pragmatic information, whilst activating/inhibiting representations of several languages. Our non-overlapping CIs for speed differences among L2 learners compared to L2 learners versus NSs suggest that this general effect may exist regardless of individual differences between L2 learners, but, as noted above, this remains to be investigated further.

Overall, the most basic but perhaps the most significant of our findings is that SPR reading time data seem to generate considerably smaller effect sizes than those for L2 research more generally, as observed in the many meta-analyses to date. One explanation for low effect sizes may be the stage of maturity of the subdomain (discussed by Plonsky & Oswald, 2014). This yields two opposing accounts, depending on where the subdomain is in its temporal trajectory: differences between SPR effect sizes might become more apparent as hypotheses, design, and instrumentation are refined; or the effect sizes may be nearing the end of a process of regressing to the mean. Given a) the relatively small number of studies to date (compared to other subdomains, such as motivation or grammar instruction), b) the

clustering of our sample of studies within the last decade, and c) Marsden, Thompson & Plonsky's (2018) observations that refinement of design and instrumentation is required, it seems that SPR effect sizes may be likely to grow. Given that SPR use in L2 research is relatively new, future research might also consider the magnitude of effects in L1 research, although we remain unaware of any systematic synthesis to date in mainstream psycholinguistic research. Our study is one example, among others in SLA research, where L2 researchers are at the cutting edge of exploring innovative approaches to improving our toolkit for understanding language learning phenomena.

However, another, or possibly additional, explanation for smaller effects might be the observation that variation in reaction time data is often large relative to the mean; tasks with a mean of 600ms often have a standard deviation in the region of 150-200ms (Brysbaert & Stevens, 2018). Thus, larger standard deviations, serving as the denominator, decrease effect sizes. Furthermore, a positive correlation tends to exist between the size of a mean and standard deviation (Brysbaert & Stevens), and so studies that report mean reading times by multi-word segments, which are inevitably longer than by single word times, may increase variance / decrease precision of observed effects. Since nearly half of our SPR studies elicited reading times on multi-word segments (k = 26 out of 57), this may have attenuated effects.

In sum, our findings suggest that caution should be exercised when comparing effect sizes between measurement types if the size of the standard deviation relative to the mean varies systematically as a function of measurement type. On the basis of the low effect sizes found in our meta-analysis, this may be the case with reading times from SPR compared to measurements used more widely in the field, such as judgment or production tests.

Effect sizes are of course not new to L2 research; several journals including Language Learning, Language Learning and Technology, Studies in Second Language Acquisition,

TESOL Quarterly, and The Modern Language Journal require their reporting, and The American Psychological Association Publication Manual (2010, p. 34) states that they are "almost always necessary" to communicate the practical significance of results. We note, however, that some researchers have expressed concern that the d family of effect sizes are rendered redundant or inappropriate by the increasing, and highly appropriate, use of linear mixed effects models (MEMs) for analysis of SPR data. Nonetheless, we suggest that the d family is still meaningful and appropriate, for reasons that include: 1) Marsden, Thompson and Plonsky (2018) found only seven such SPR studies to date, illustrating that there is some way to go before these models are incorporated into the field's statistical toolkit; 2) linear mixed effect models, although bringing many benefits to quantitative analyses, represent a type of NHST and so could still raise some of the concerns outlined above and elsewhere; 3) reporting d effect sizes allows meta-analyses to be done on paired variables of theoretical and practical interest; 4) there is no mathematical reason why data cannot be set out differently for separate analyses, one using models and one calculating means for the d family of effect sizes. To expand on this last point, for MEMs, data are laid out with each data point from each item (trial) from each participant in a separate row. For the d family of effect sizes, data from individual items (trials) are aggregated across individual learners (for 'by participant' analyses) and across individual items (for 'by item' analyses). Each approach (MEM and d family of effect sizes) has advantages and disadvantages. Aggregating across items or participants loses the capacity to calculate random effects of items or participants (though in a carefully designed experiment, the stimuli are created in such a way so as to reduce the effect of 'item' and participants are selected to a degree that is at least theoretically and ecologically satisfactory, e.g., by proficiency, age, L1). The advantage is that aggregating across items or individuals allows us to infer the magnitude of effects between pairs of theoretically interesting variables.

Thus, the two analyses are complementary and such a belts and braces approach can only strengthen argumentation.

The use of effect sizes such as d appears particularly compelling given the current lack of meta-analysis in L2 processing research, leaving effects in individual studies somewhat "quarantined from related research data" (Ellis, 2006, p. 303). As such, our study makes a first step towards a more contextualised understanding of effect sizes based on reaction time data, and suggests that effects fall somewhat lower than those found by other syntheses of L2 and social science research. Providing more fine-grained benchmarks, such as effects at the 25th, 50th and 75th percentiles, would require a larger set of studies than our sample, as low numbers of studies within each percentile could render misleading benchmarks at this stage. We leave that task to future research, but we suspect that the ranges will be considerably tighter than other benchmarks that have been proposed.

Finally, we note that a limitation of the present study, and of meta-analyses in general, is that the quality of meta-analytic findings is affected by cross-study heterogeneity in areas such as study design, sample demographics, instrument construction, data cleaning procedures (e.g., outlier removal, treatment of non-normal distributions), analysis, and transparency. As Marsden, Thompson & Plonsky (2018) found, there is a pressing need to standardise the use and reporting of SPR research. While improving our understanding of L2 processing, it will also facilitate future research syntheses. However, in defence of our approach, note that our effect sizes were calculated for each experimental manipulation of conditions within each study, which means that decisions about data cleaning (such as the use of trimmed, transformed, residual, or aggregated data) were all held constant for each effect size calculation. This reduces some potential threats caused by inter-study methodological heterogeneity.

CONCLUSION

As effect sizes are best interpreted in the context of a specific domain, the results reported in the present study could offer a preliminary reference framework for future SPR research. These include: (1) an overall reliable sensitivity to morphosyntax for L2 learners of d = 0.20 (0.15, 0.25), which was likely meaningfully less than for NSs (d = 0.28 [0.21, 0.35]); (2) a reliable sensitivity to anomaly detection for L2 learners of d = 0.19 (0.09, 0.29), which seemed reliably less than for NSs (d = 0.41 [0.29, 0.54]); (3) larger effect sizes in anomaly detection studies in conditions with L1-L2 similarity (d = 0.34 [0.21, 0.48]); (4) a lack of L1 influence in ambiguity resolution studies; (5) speed differences between L2 learner groups with different L1s (d = 0.35 [0.26, 0.43]); (6) speed differences between L2 learner groups and NSs (d = 0.53 [0.46, 0.60]).

These findings should be understood as estimates of processing performance at advanced stages of L2 learning. They can help orient researchers towards situating future SPR research within a preliminary framework of reference. For example, the average effect sizes we found may inform subsequent power-analyses to determine sample sizes required to obtain a particular effect size. Principally, we hope our findings help contextualise future within- and between-subject comparisons for a range of different research questions that are unlikely to lend themselves to binary distinctions. Further synthesis of the growing body of L2 processing research will certainly be helped by the field's gradually increasing disposition to report and interpret effect sizes and share raw data.

SUPPLEMENTARY MATERIAL

To view supplementary material for this article, please visit

NOTES

[1]We use 'morphosyntax' and 'grammar' synonymously and in a broad sense to cover any element of syntax or morphology.

[2]'L2 learners' are taken to be individuals who learnt or are learning an L2 during adulthood, as opposed to those raised in a bilingual environment (Papadopolou, 2005).

[3]We recognise that existing terminology regarding cross-linguistic influence is debated (for discussions, see Odlin, 2005, 2012, and Jarvis & Pavlenko, 2008).

[4]The t statistic from the participants' analysis was used.

[5]The search ended in June 2017.

[6]These exact terms were not always used, or used systematically, by authors (as also found by Marsden, Thompson, & Plonsky, 2018). Where authors were not explicit, the regions of anticipated processing cost could still be inferred from authors' choice of sentence regions for analysis.

[7]We did not include reading time data from post-stimulus comprehension questions, as very few studies have done this (see, e.g., Jegerski, 2016).

[8]This included obligatory and optional instantiation. Coding decisions were uncontroversial except in the case of one feature – plural number marking – in four studies, which we specify here: We considered plural number marking to be cross-linguistically different for Chinese learners in Chan (2012) and Jiang (2004, 2007), and for Japanese learners in Jiang et al. (2011). Although the authors of these studies described the feature as optional in Chinese and Japanese, they argued clearly that plural number marking was rarely used and might explain learners' insensitivity to it.

[9]For all studies, the authors reported a result as 'statistically significant' if p values were below 0.05 (though several did not explicitly state this alpha level). Mueller and Jiang (2013), included in this analysis for RQ3a, reported NHST results for the same paired comparisons that we made for effect sizes, but their NHST were based on the median group differences in each case (not the mean).

[10]Four studies reported the results of more than one SPR task, but each time with unique participant samples (Foote, 2011, two tasks; Havik, Roberts, van Hout, Schreuder, & Haverkort, 2009, two tasks; Jiang, 2004, three SPR tasks; Renaud, 2014, two tasks) and so the effect sizes for each unique sample were kept separate. In contrast, two studies used the same participants for more than one SPR task (Aldawayan, Fiorentino, & Gabriele, 2010; Chan, 2012). For the former study, results were drawn together across the two SPR tasks, which investigated the same linguistic feature, drawing multiple effect sizes from the same sample (see Norris & Ortega, 2006). For the latter, the results were not drawn together because the two tasks investigated different linguistic features; thus, we avoided collapsing data that was valuable for the moderator analysis by linguistic feature.

[11]In Jegerski (2012), reading times were reported based on a median split between faster and slower readers. In the absence of reported n for these subgroups, we calculated the number of participants for each subgroup by dividing the number of participants in each group (L2 learner group [n = 23] and native speaker group [n = 35]) by two and rounding to the nearest single figure.

[12]This is a more precise method than weighting by sample size alone (Lipsey & Wilson, 2001). Each mean effect size between two different groups (i) was given the following weight (w): $w_i = 1 / v_i$ (Borenstein et al., 2007, p. 7), whereby the weight of each effect size (w_i) is calculated by dividing 1 by the within-study variance (v_i) of the effect size.

All of the resulting effect sizes are multiplied by their respective weights, then divided by the sum of the weights, to reach a weighted grand average effect size.

[13]L2 learners were slower than NSs in almost all cases, with a very small number being marginally faster (k [unique samples] = 7 out of 73).

REFERENCES

Adesope, O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. Review of Educational Research, 80, 207-245.

Aldawayan, S., Fiorentino, R., & Gabriele, A. (2010). Evidence of syntactic constraints in the processing of wh-movement: A study of Nadji Arabic learners of English. In B. VanPatten & J. Jegerski (Eds.), Research in second language processing and parsing (pp. 65-86). Amsterdam: John Benjamins.

American Psychological Association. (2010). Publication manual of the American Psychological Association (6th ed.). Washington, DC: Author.

Bel, A., Sagarra, N., Comínguez, J., & García-Alcaraz, E. (2016). Transfer and proficiency effects in L2 processing of subject anaphora. Lingua, 184, 134-159.

Bley-Vroman, R. (1990). The logical problem of foreign language learning. Linguistic Analysis, 20, 3-49.

Bley-Vroman, R. (2009). The evolving context of the fundamental difference hypothesis. Studies in Second Language Acquisition, 31, 175-198.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). Introduction to meta-analysis. Chichester: Wiley.

Borenstein, M., Hedges, L., & Rothstein, H. (2007). Meta-analysis: Fixed effect vs. random effects. Retrieved from www.meta-analysis.com.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. Journal of Cognition, 1, 1-20.

Bultena, S., Dijkstra, T., & van Hell, J. (2014). Cognate effects in sentence context depend on word class, L2 proficiency, and task. The Quarterly Journal of Experimental Psychology, 67, 1214–1241.

Chan, H. (2012). Tense-aspect processing in second language learners (unpublished doctoral dissertation). University of Pittsburgh.

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

Clahsen, H., & Felser, C. (2006b). How native-like is non-native language processing? Trends in Cognitive Sciences, 10, 564-570.

Clahsen, H., & Felser, C. (2018). Some notes on the shallow structure hypothesis. Studies in Second Language Acquisition, 40, 693-706.

Clahsen, H., Felser, C., Neubauer, C., Sato, M., & Silva, R. (2010). Morphological structure in native and nonnative language processing. Language Learning, 60, 21–43

Coe, R. (2002). It's the effect size, stupid: What the effect size is and why it is important. Paper presented at the British Educational Research Association annual conference, Exeter, 12-14 September, 2002.

Cohen, J. (1969). Statistical power for the behavioural sciences (2nd ed.). Hove: Laurence Erlbaum.

Cumming, G. (2009). Inference by eye: Reading the overlap of independent confidence intervals. Statistics in Medicine, 28, 205-220.

Cumming, G. (2012). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York: Routledge

Dekydtspotter, L., Edmonds, A., Fultz, A., & Renaud, C. (2010). Modularity of L2 sentence processing: Prosody, context and morphology in relative clause ambiguity in English-French interlanguage. In M. Iverson., I. Ivanov., T. Judy., J. Rothman., R. Slabakova., & M. Tryzna (Eds.), Proceedings of the 2009 mind/context divide workshop (pp. 13-27). Somerville, MA: Cascadilla Proceedings Project.

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. Frontiers in Psychology, 5, 781.

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

Ellis, N. (2006). Meta-analysis, human cognition and language learning. In J. Norris & L. Ortega (Eds.), Synthesizing research on language learning and teaching (pp. 301-322). Amsterdam: John Benjamins.

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

Fender, M. (2003). English word recognition and word integration skills of native Arabic and Japanese-speaking learners of English as a second language. Applied Psycholinguistics, 24, 289-315.

Gerth, S., Otto, C., Felser, C., & Nam, Y. (2017). Strength of garden path effects in native and non-native speakers' processing of subject-object ambiguities. International Journal of Bilingualism, 21, 125-144.

Greenland, S., Senn, S., Rothman, K., Carlin, J., Poole, C., Goodman, S., & Altmann, D. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. European Journal of Epidemiology, 31, 337-350.

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.

Hedge, C., Powell, G., Bompas, A., Vivian-Griffiths, S., & Sumner, P. (2018). Low and variable correlation between reaction time costs and accuracy costs explained by two accumulation models: Meta-analysis and simulations. Psychological Bulletin, 144, 1200-1227.

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

Hopp, H. (2009). The syntax-discourse interface in near-native L2 acquisition: Off-line and on-line performance. Bilingualism: Language and Cognition, 12, 463-483.

Hopp, H. (2010). Ultimate attainment in L2 inflection: Performance similarities between non-native and native speakers. Lingua, 120, 901-931.

Howell, D. (2013). Statistical methods for psychology (8th ed.). Wadsworth: Cengage Learning.

Hulstijn, J. (2018). An individual differences framework for comparing non-native with native speakers: Perspectives from basic language cognition theory. Language Learning, 1-27.

Jackson, C. (2008). Proficiency level and the interaction of lexical and morphosyntactic information during L2 sentence processing. Language Learning, 58, 875-909.

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

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

Jackson, C., & van Hell, J. (2011). The effects of L2 proficiency level on the processing of wh-questions among Dutch second language speakers of English. International Review of Applied Linguistics, 49, 195-219.

Jarvis, S., & Pavlenko, A. (2008). Cross-linguistic influence in language and cognition. London: Routledge.

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

Jegerski, J. (2014). Self-paced reading. In J. Jegerski & B. VanPatten (Eds.), Research methods in second language psycholinguistics (pp. 20-49). New York: Routledge.

Jegerski, J. (2016). Number attraction effects in near-native Spanish comprehension. Studies in Second Language Acquisition, 38, 5-33.

Jegerski, J. (2018). Sentence processing in Spanish as a heritage language: A self-paced reading study of relative clause attachment. Language Learning, 68, 598-634.

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

Jiang, N. (2007). Selective integration of linguistic knowledge in adult second language learning. Language Learning, 57, 1-33.

Jiang, N., Novokshanova, E., Masuda, K., & Wang, X. (2011). Morphological congruency and the acquisition of L2 morphemes. Language Learning, 61, 940-967.

Juffs, A. (1998a). Main verb versus reduced relative clause ambiguity resolution in L2 sentence processing. Language Learning, 48, 107-147.

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

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

Juffs, A. (2006). Processing reduced relative vs. main verb ambiguity in English as a Second Language: A replication study with working memory. In R. Slabakova., S. Montrul., & P. Prevost (Eds.), Inquiries in linguistic development in honor of Lydia White (pp. 213-232). Amsterdam: John Benjamins.

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., & Rodriguez, G. (2015). Second language sentence processing. New York: Routledge.

Jung, S. (2010). Second language processing of wh-movement in English: The effects of first language and learning environment (unpublished MA dissertation). Michigan State University.

Just, M., Carpenter, P., & Woolley, J. (1982). Paradigms and processes in reading comprehension. Journal of Experimental Psychology: General, 111, 228-238.

Kaan, E., Ballantyne, J., & Wijnen, F. (2015). Effects of reading speed on second-language sentence processing. Applied Psycholinguistics, 36, 799-830.

Keating, G., & Jegerski, J. (2015). Experimental designs in sentence processing research: A methodological review and user's guide. Studies in Second Language Acquisition, 37, 1-32.

Keating, G., Jegerski, J., & VanPatten, B. (2016). Online processing of subject pronouns in monolingual and heritage bilingual speakers of Mexican Spanish. Bilingualism: Language and Cognition, 19, 36-49.

Lakshmanan, U., Kim, K., McCreary, R., Park, K., Suen, U., & Lee, S. (2009). L2 learners' sensitivity to strong and weak subjacency-violations in online processing. In M. Bowles., T. Ionin., S. Montrul., & A. Tremblay (Eds.), Proceedings of the 10th

generative approaches to second language acquisition conference (GASLA 2009) (pp. 136-143). Somerville, MA: Cascadilla Proceedings Project.

Larson-Hall, J. (2016). A guide to doing statistics in second language research using SPSS and R. New York: Routledge.

Lauro, J., & Schwartz, A. (2017). Bilingual non-selective lexical access in sentence contexts: A meta-analytic review. Journal of Memory and Language, 92, 217-233.

Lee, E., Lu, D., & Garnsey, S. (2013). L1 word order and sensitivity to verb bias in L2 processing. Bilingualism: Language and Cognition, 16, 761-775.

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. Applied Linguistics, 36, 345-366.

Lee, S., & Huang, H. (2008). Visual input enhancement and grammar learning: A meta-analytic review. Studies in Second Language Acquisition, 30, 307-331.

Lipsey, M., & Wilson, D. (2001). Practical meta-analysis. Thousand Oaks, CA: Sage.

MacWhinney, B. (2005). Extending the competition model. International Journal of Bilingualism, 9, 69-84.

Marinis, T. (2007). On-line processing of passives in L1 and L2 children. In A. Belikova., L. Meroni., & M. Umeda (Eds.), Proceedings of the 2nd conference on generative approaches to language acquisition North America (GALANA) (pp. 265-276). Somerville, MA: Cascadilla Proceedings Project.

Marinis, T. (2010). Using on-line processing methods in language acquisition research. In S. Blunsworth., & E. Blom (Eds.), Experimental methods in language acquisition research (pp. 139-162). Amsterdam: John Benjamins.

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

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. Language Learning, 68, 321-291.

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. Applied Psycholinguistics, 39, 861-904.

Morgan-Short, K., Marsden, E., Heil, J., Issa II, B., Leow, R., Mikhaylova, A., Mikołajczak, S., Moreno, N., Slabakova, R., & Szudarski, P. (2018). Multi-site replication in second language acquisition research: Attention to form during listening and reading comprehension. Language Learning, 68, 392-437.

Morris, S., & DeShon, R. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent–groups designs. Psychological Methods, 7, 105-125.

Mueller, J., & Jiang, N. (2013). The acquisition of the Korean honorific affix (u)si by advanced L2 learners. The Modern Language Journal, 97, 318-339.

Norris, J., & Ortega, L. (Eds.) (2006). Synthesizing research on language learning and teaching. Amsterdam: John Benjamins.

Norouzian, R., de Miranda, M., & Plonsky, L. (2018). The Bayesian revolution in second language research. Language Learning, 68, 1032-1075.

Odlin, T. (2005). Cross-linguistic influence. In C. Doughty., & M. Long (Eds.), The handbook of second language acquisition (pp. 436-486). Oxford: Blackwell.

Odlin, T. (2012). Cross-linguistic influence in second language acquisition. In C. Chappelle (Ed.), The encyclopaedia of applied linguistics (pp. 1562-1568). Malden, MA: Blackwell.

O'Grady, W. (2005). Syntactic carpentry: An emergentist approach to syntax. Mahwah: Lawrence Erlbaum

Oswald, F., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. Annual Review of Applied Linguistics, 30, 85-110.

Pan, H., Schimke, S., & Felser, C. (2015). Referential context effects in non-native relative clause ambiguity resolution. International Journal of Bilingualism, 19, 298-313.

Papadopolou, D. (2005). Reading time studies of second language ambiguity resolution. Second Language Research, 21, 98-120.

Papadopoulou, D., & Clahsen, H. (2003). Parsing strategies in L1 and L2 sentence processing: A study of relative clause attachment in Greek. Studies in Second Language Acquisition, 25, 501-528.

Perpiñán, S. (2014). L2 grammar and L2 processing in the acquisition of Spanish prepositional relative clauses. Bilingualism: Language and Cognition, 18, 577-596.

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. Language Learning, 61, 993-1038.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses and reporting practices in quantitative L2 research. Studies in Second Language Acquisition, 35, 655-687.

Plonsky, L. (Ed.) (2015). Advancing quantitative methods in second language research. New York: Routledge.

Plonsky, L., & Brown, D. (2015). Domain definition and search techniques in meta-analyses of L2 research (Or why 18 meta-analyses of feedback have different results). Second Language Research, 31, 267-278.

Plonsky, L., & Oswald, F. (2014). How big is 'big'? Interpreting effect sizes in L2 research. Language Learning, 64, 878-911.

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.

Rayner, K., Schotter, E., Masson, M., Potter, M., & Treiman, R. (2016). So much to read, so little time: How do we read, and can speed reading help? Psychological Science in the Public Interest, 17, 4-34.

Renaud, C. (2014). A processing investigation of the accessibility of the uninterpretable gender feature in L2 French and L2 Spanish adjective agreement. Linguistic Approaches to Bilingualism, 4, 222-255.

Roberts, L. (2007). Investigating real-time sentence processing in the second language. Stem-, Spraak- en Taalpathologie, 15, 115-127.

Roberts, L. (2012). Psycholinguistic techniques and resources in second language acquisition research. Second Language Research, 28, 113-127.

Roberts, L. (2013). Sentence processing in bilinguals. In R. van Gompel (Ed.), Sentence processing (pp. 221-246). Hove: Psychology Press.

Roberts, L. (2016). Self-paced reading and L2 grammatical processing. In A. Mackey., & E. Marsden (Eds.), Advancing methodology and practice: the IRIS repository for instruments for research into second languages (pp. 58-72). New York: Routledge.

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

Roberts, L., & Liszka, S. (2013). Processing tense/aspect-agreement violations on-line in the second language: A self-paced reading study with French and German L2 learners of English. Second Language Research, 29, 413-439.

Rosenthal, R. (1979). The "file-drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-461.

Sagarra, N., & Herschensohn, J. (2010). The role of proficiency and working memory in gender and number agreement processing in L1 and L2 Spanish. Lingua, 120, 2022-2039.

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

Schenker, N. & Gentleman, J. (2001). On judging the significance of differences by examining the overlap between confidence intervals. The American Statistician, 55, 182-186.

Sharwood-Smith, M., & Truscott, J. (2005). Stages or continua in second language acquisition: A MOGUL solution. Applied Linguistics, 26, 219-240.

Song, Y. (2015). L2 processing of plural inflection in English. Language Learning, 65, 233-267.

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. Norris & L. Ortega (Eds.), Synthesising research on language learning and teaching (pp. 279-298). Amsterdam: John Benjamins.

Tokowicz, N., & Warren, T. (2010). Beginning adult L2 learners' sensitivity to morphosyntactic violations: A self-paced reading study. European Journal of Cognitive Psychology, 22, 1092-1106.

Tolentino, L., & Tokowicz, N. (2011). Across languages, space and time: A review of the role of cross-language similarity in L2 (morpho)syntactic processing as revealed by fMRI and ERP Methods. Studies in Second Language Acquisition, 33, 91-125.

VanPatten, B. (2007). Input processing in adult second language acquisition. In B. VanPatten., & J. Williams (Eds.), Theories in second language acquisition: An introduction (pp. 115-135). Mahwah: Lawrence Erlbaum.

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

VanPatten, B., Keating, G., & Leeser, M. (2012). Missing verbal inflections as a representational problem. Linguistic Approaches to Bilingualism, 2, 109-140.