Quantifying individual differences in native and non-native sentence processing

Quantifying individual differences in native and non-native sentence processing

Ian Cunnings

&

Hiroki Fujita

University of Reading

Address for correspondence:

Dr Ian Cunnings

School of Psychology and Clinical Language Sciences

The University of Reading

Reading

RG6 7BE

Abstract

Research in sentence processing has increasingly examined the role of individual differences in language comprehension. In work on native and non-native sentence processing, examining individual differences can contribute crucial insight into theoretical debates about the extent to which nativelike processing is possible in a non-native language. Despite this increased interest in individual differences, whether commonly used psycholinguistic tasks can reliably measure individual differences between participants has not been systematically examined. As a preliminary examination of this issue in non-native processing, we report a self-paced reading experiment on garden-path sentences in native and non-native comprehension. At the group level we replicated previously observed findings in native and non-native speakers. However, while we found that our self-paced reading experiment was a reliable way of assessing individual differences in overall reading speed and comprehension accuracy, it did not consistently measure individual differences in the size of garden-path effects in our sample (N = 64 native and 64 non-native participants, and 24 experimental items). These results suggest that before individual differences in sentence processing can be meaningfully assessed, the question of whether commonly used tasks can consistently measure individual differences requires systematic examination.

Keywords: individual differences; measurement reliability; bilingual sentence processing; sentence comprehension; garden-path sentences

Introduction

There is increasing interest in the role that individual differences may play in contributing to theoretical debate in language acquisition and processing (e.g. Kidd et al., 2018). For example, examining individual differences can elucidate the extent to which language processing is governed by domain-specific or domain-general mechanisms (e.g. Caplan & Waters, 1999; MacDonald & Christiansen, 2002), or in teasing apart different accounts of the role of working memory in sentence processing (e.g. Daneman & Carpenter, 1980; Van Dyke et al., 2014). Within the context of multilingual language processing, examining individual differences can inform debate about the similarities and differences between native (L1) and non-native (L2) processing (Clahsen & Felser, 2006, 2018; Cunnings, 2017; Hopp, 2018). Indeed, theories that predict L1 and L2 processing are qualitatively similar (e.g. Hopp, 2018) would predict that L2 processing may be nativelike if appropriate individual differences are accounted for.

Despite this interest in individual differences, the extent to which psycholinguistic tasks can consistently index individual differences in sentence processing has not been systematically examined, and to date we are unaware of any study that has examined this issue from the perspective of L2 processing. This is in stark contrast to other areas of bilingualism research, for example research on L2 proficiency (e.g. Leclercq et al., 2014), where assessing the reliability of measures of individual variation is routine. However, assessing how well psycholinguistic tasks can measure individual differences is crucial, as one can only meaningfully correlate two variables when it can be shown that there is systematic variance in the individual measures to begin with and that this variance can be reliably measured (Spearman, 1904). In L2 processing research, one can only measure shared individual variation between a measure of individual differences and a measure of sentence processing if there is systematic variation in *both* individual measures to begin with.

This paper has two aims. Firstly, we highlight the importance of measurement reliability when studying individual differences in native and non-native sentence processing, as low measurement reliability has adverse effects on the interpretability of any statistical inferences (Parsons et al., 2019). Secondly, we report a self-paced reading experiment examining garden-path effects in L1 and L2 readers and assess the extent to which our experiment consistently measured individual differences. Our results indicated that we did not observe consistent individual differences in the size of garden-path effects between participants in our sample, which limited any conclusions we could draw about individual differences in self-paced reading. This case study in garden-paths highlights how L2 researchers need to consider measurement reliability when examining individual differences.

We begin by outlining issues in using tasks from psycholinguistics and experimental psychology more broadly to examine individual differences, before discussing the scant existing research on this issue in sentence processing. Although we focus on the L1/L2 comparison in sentence processing, the issues we discuss also arise in comparisons between other multilingual populations and in other experimental settings.

The reliability paradox

It has long been noted that research in psychology is conducted within two broad traditions of correlational and experimental research (Hedge et al., 2018). Hedge et al. noted how ‘reliability’ is defined differently in these two traditions. In correlational research, a ‘reliable’ task is one that can consistently rank individuals along a continuum. Reliability here thus refers to *measurement reliability*. For individual differences tasks to be reliable, there must be larger variance between individuals than the error variance in measuring a construct within individuals (for further discussion, see e.g. Hedge et al., 2018, Parsons et al., 2019). Such tasks rely on high levels of systematic between-subject variance to consistently rank individuals. If there is little between-subject variance, in relation to the error variance in estimating a construct within individuals, it becomes difficult to rank individuals consistently. On the other hand, in experimental research, a 'reliable' task is one where an effect replicates between participants and across studies (i.e. consistently returns p values below .05). Replicability here will be more likely when between-subject variability in a within-subject effect is low.

As such, a 'reliable' task from the perspective of the experimental approach, with low between-subject variability, may have properties that are the opposite of what is desired of a 'reliable' task in the correlational approach, which requires high levels of systematic between-subject variability (Hedge et al., 2018). This in turn entails that tasks from the experimental tradition may not, or indeed are unlikely to be, good measures of individual differences, an issue Hedge et al. dubbed 'the reliability paradox'. They tested a number of commonly used cognitive tasks, including the Flanker task, Stroop, stop-signal and go/no-go, amongst others, and found low test-retest correlations for most tasks, indicating they did not consistently measure individual differences. These results suggest widely used cognitive tasks are not ideal measures of individual differences. The more general point that Hedge et al. make about the different definitions of 'reliability' in correlational and experimental research, also potentially has consequences for the use of any 'reliable' task from the experimental psychology tradition as a measure of individual differences (see also Elliott et al., 2020).

Quantifying individual differences in sentence processing

These issues have important consequences for research in L1 and L2 processing. Research in psycholinguistics, and by extension L2 processing, has almost exclusively used tasks from the experimental psychology tradition. In psycholinguistic experiments, the researcher typically aims to minimise between-subject variability as much as possible. The fact that research in L2 sentence processing relies on tasks from the experimental tradition, entails that the L2 researcher cannot take for granted that such tasks should be able to consistently measure individual differences.

One challenge to studying individual differences in sentence processing is that experiments typically adopt a Latin-square design in which participants do not see all versions of the same stimuli. For example, in a study on garden-path sentences, participants read temporarily ambiguous sentences like (1/2a) and unambiguous controls like (1/2b). The temporary ambiguity arises in (1/2a) because “the boy/the reporter” may be initially interpreted as the direct object of the preceding verb, when it is in fact the subject of the main clause. (1/2b) are unambiguous controls disambiguated by a comma. The temporary ambiguity leads to longer reading times at the disambiguating verb (“ate/denounced”), where the direct object interpretation becomes impossible, in (1/2a) compared to (1/2b) (e.g. Frazier & Rayner, 1982; Sturt et al., 1999).

(1a) While Mary washed the boy in the room ate an apple.

(1b) While Mary washed, the boy in the room ate an apple.

(2a) While Simon helped the reporter in the library denounced the government.

(2b) While Simon helped, the reporter in the library denounced the government.

Although such effects may be observed at the group level, the nature of a Latin-square design means that individual participants may not clearly show the effect due to the different lexical material used across the sets of sentences. For example, a subject who sees ambiguous (1a) but unambiguous (2b) may not individually clearly show the expected garden-path effect, merely because “denounced” is longer and less frequent than “ate”. Across participants, the Latin-square design should eliminate such issues for group-level analyses, but analyses that do not take into account this by-item variation will complicate any assessment of subject-level individual differences. Mixed-effects models with crossed random effects for subjects and items (see Baayen et al., 2008; Barr et al., 2013; Cunnings, 2012; Linck & Cunnings, 2015) can help provide a solution for this issue, as the by-item adjustments made by the model will help provide better estimates of the by-subject variance. Note that usually in individual differences research all participants see the same items, though this approach is rarely adopted in psycholinguistic studies of individual differences (but see Swets et al., 2007).

To date, little is known about whether typical psycholinguistic experiments can in fact consistently measure individual differences in L1 and L2 processing. Indeed, a first step in examining individual differences in L2 sentence processing is to assess whether tasks used to study sentence processing can indeed reliably assess individual differences. In individual differences research, two ways of assessing the reliability of individual differences measures are test-retest reliability and split-half correlations (see e.g. Parsons et al., 2019). Test-retest reliability involves participants completing a task twice, and then correlating performance between each attempt. If participants have only completed a task once, split-half correlations can be calculated by randomly dividing the data in half, and correlating participants' performance in each half of the data. Both assess measurement reliability by measuring a task's internal consistency. Typically a correlation above .7 is taken to indicate a consistent measure (Nunnally, 1978), and although this should not be taken as a strict cut-off, high measurement error (i.e. low test-retest or split-half reliability) is detrimental to any inferences that can be drawn (Parsons et al., 2019). Given that participants in psycholinguistics experiments typically do not complete each task twice, calculating test-retest reliability is not often practical, but split-half correlations can usually be computed.

James et al. (2018) recently investigated this issue in L1 sentence processing. In a self-paced reading experiment, they investigated individual differences in the use of verb-bias information in the resolution of syntactic ambiguity (e.g., Lee et al., 2013) and the increased processing difficulty observed in processing object compared to subject relative clauses (e.g. Gibson et al., 2005). They also investigated individual differences in offline measures investigating comprehension accuracy of the verb-bias and relative clause sentences, and also additionally tested interpretive preferences in ambiguous relative clause attachment.

For self-paced reading, although James et al. replicated previously observed effects of verb bias and relative clauses at the group level, in a series of split-half analyses they found little evidence that their experiment consistently measured individual differences. For the verb-bias sentences and those testing the subject/object asymmetry, James et al. reported high split-half correlations (up to and above > .9) for measures indexing overall participant reading speed and overall comprehension accuracy across conditions. This indicates that their self-paced reading experiment could consistently distinguish slower from faster readers. However, split-half correlations for effects of ambiguity, verb bias (and their interaction), and the subject/object relative clause asymmetry were all very low (all < .427), both offline and online, indicating little internal consistency in assessing individual differences in these effects. For ambiguous relative clauses, their offline data reasonably consistently measured individual differences in attachment preferences, with a split-half correlation of .678. Swets et al. (2007) also previously reported high levels of consistency in an offline study on individual differences in relative clause attachment, suggesting offline tasks may be a suitable measure of individual differences for this phenomenon.

In sum, James et al. found little evidence that their self-paced reading experiment consistently measured by-subject individual differences, especially in terms of online reading, beyond effects related to overall reading speed. They concluded that either there is simply no meaningful individual variation in the studied effects to systematically investigate individual differences in the first place, or the task did not reliably measure them. Either way, these results raise serious challenges to the study of individual differences during sentence processing (for similar concerns raised in other psycholinguistic tasks, see Arnon, 2020; Brown-Schmidt & Fraundorf, 2015; Ryskin et al., 2015).

A case study on garden-path effects in L1 and L2 processing

Against this background, we aimed to provide an initial examination of the quantifiability of individual differences in L2 sentence processing using a widely used sentence processing task and a well-studied phenomenon. Specifically, we conducted a self-paced reading experiment examining garden-path effects. A number of studies have examined L2 processing of garden-path sentences. These have shown that both L1 and L2 speakers show garden-path effects, with longer reading times in sentences like (1/2a) than (1/2b) at the disambiguating verb (Gerth et al., 2017; Hopp, 2006; Jacob & Felser, 2016). One recent study, Jacob and Felser (2016), however reported smaller garden-path effects in L2ers than L1ers, which they took to interpret that L2ers may not initiate/complete reanalysis as consistently as L1ers.

Garden-path sentences are also sometimes misinterpreted by both L1 and L2 speakers (Christianson et al., 2001; Jacob & Felser, 2016; Pozzan & Trueswell, 2016). Christianson et al. (2001) showed that in sentences like (1a), the initial direct object misinterpretation ('Mary washed the boy') lingers after disambiguation, with participants answering comprehension questions probing interpretation of the temporarily ambiguous noun phrase ('Did Mary wash the boy?) incorrectly more often in ambiguous compared to unambiguous sentences. This is known as 'lingering misinterpretation', as the initial direct object interpretation is not fully erased. L2 learners have been shown to have more difficulty recovering from garden-paths than L1 speakers. For example, in an act out task Pozzan and Trueswell (2016) found that L2 learners misinterpreted temporarily ambiguous sentences approximately 50% of the time, while L1 speakers made errors on only 10% of trials. Although there are different potential accounts of this L1/L2 difference (see Cunnings, 2017) for present purposes suffice to say that if L2 learners have more difficulty than L1 speakers in reanalysing garden-path sentences, we would expect L2 learners to have lower comprehension accuracy of garden-path sentences than L1 speakers, while accuracy for unambiguous sentences between the two groups should be similar.

How individual differences may influence garden-paths and lingering misinterpretation in L1 and L2 speakers has also been examined. In the L1 processing literature, individual differences in the ability to recover from garden-path sentences has been linked to individual differences in cognitive control (Kan et al., 2013; Novick et al., 2014; Vuong & Martin, 2014). In the L2 processing literature, theoretical accounts that predict that L1/L2 processing are qualitatively similar (Hopp, 2018), would predict that any L1/L2 differences in garden-path effects or lingering misinterpretation should be influenced by individual differences. For example, such theories would predict that any L1/L2 differences should be attenuated as L2 proficiency increases. Existing studies have however not provided consistent results in this regard. For example, Gerth et al. (2017) reported some effects of individual differences in proficiency in that L2ers with higher proficiency had higher comprehension accuracy and generally quicker reading times, but proficiency did not interact with their ambiguity manipulations.

In this study, we examine garden-path effects in co-ordinated sentences like (3). In (3a) “the cat” is temporarily ambiguous and may be interpreted as being co-ordinated with “the dog” as the direct object of “washed”, or as the subject of the second clause. The latter interpretation turns out to be correct at “played”, but readers initially compute the co-ordinated interpretation during incremental processing (Engelhardt & Ferreira, 2010; Hoeks et al., 2002). (3b) is an unambiguous control in which the temporal conjunction “while” removes the temporary ambiguity. Thus, longer reading times are expected at “played” in (3a) than (3b). Additionally, if initial misinterpretations linger, readers may answer the comprehension question less accurately in (3a) than (3b).

(3a) Ken washed the dog and the cat in the garden played with a ball.

(3b) Ken washed the dog while the cat in the garden played with a ball.

Did Ken wash the cat?

Witzel et al. (2012) examined L1 and L2 processing of temporarily ambiguous sentences like (3a) compared to an unambiguous control containing a comma after "the dog". They found that both L1 and L2 readers had longer reading times for the temporarily ambiguous than unambiguous sentences, indicating garden-path effects online. Kaan et al. (2019) compared sentences like (3a) to an unambiguous control containing the co-ordination "but" in a self-paced reading experiment. Both L1ers and L2ers showed longer reading times at the disambiguating verb, as evidence of garden-path effects. Kaan et al. also examined offline comprehension and found some evidence of lingering misinterpretation. However, comprehension questions that tapped lingering misinterpretation specifically were only asked after a small minority of trials, and as such it is difficult to assess the extent to which misinterpretations linger in L1 and L2 processing of co-ordinated ambiguities based on this study. Kaan et al. also tested for individual differences in L2 processing by proficiency, but did not find any significant effects.

The present study

We utilised the co-ordination ambiguity to investigate the quantifiability of individual differences in sentence processing in L1 an L2 readers. Participants read ambiguous sentences like (3a) and unambiguous controls like (3b) in a self-paced reading experiment, and answered comprehension questions that probed for lingering misinterpretation. We expected to find garden-path effects during online reading and lingering misinterpretation in offline comprehension. We also examined whether individual differences in L2 proficiency influenced the size of these effects. Our primary aim however was to assess the reliability of our self-paced reading experiment as a measure of individual differences, as a case study in highlighting issues of measurement reliability in L1 and L2 processing research. As such, we also conducted a series of split-half correlations to assess the internal consistency of individual differences in our sample.

Participants

Participants included 64 L1 English speakers (24 males, mean age = 31, range = 18–61) and 64 L2 English speakers (21 males, mean age 29, range = 18–51), from different language backgrounds. Participants were recruited online, through either email or via Prolific (https://prolific.co/). L1 English speakers identified English as their only first language, while L2 speakers identified language(s) other than English as their first language(s). Some L1 English participants had knowledge of other languages, but we only included participants who did not consider themselves bilingual (defined as having nativelike command of multiple languages). L1 English speakers had UK or US nationality, while L2 participants were from other nationalities. Although we refer to our L2 participants as non-natives, as we did not gather information regarding the onset of English exposure, we do not know whether they are 'early' or 'late' bilinguals. We recruited participants until we had 64 in each group, on the criterion that participants scored over 75% correct on comprehension questions to filler trials. Participants took part voluntarily or were paid a small fee.

L2 participants completed the Quick Placement Test. Their mean score was 50/60 (SD = 8, range 21-60). The majority of L2ers were classed as intermediate to advanced learners.

Materials

Experimental items consisted of 24 sentences as in (3). Each sentence manipulated ambiguity to be either temporarily ambiguous or unambiguous as in (3a) and (3b) respectively. All sentences were followed by a yes/no comprehension question that probed interpretation of the initially assigned misinterpretation. The correct answer was always ‘no’.

In addition to the 24 experimental sentences, 56 filler sentences were also created. These included a variety of syntactic structures and were always followed by a yes/no question. Half of the fillers required a ‘yes’ response and the other half ‘no’. The full list of items can be found at the Open Science Framework (OSF) website (https://osf.io/hmuvz/).

Procedure

The experiment was administered online using Ibexfarm (Drummond, 2003; for evidence of the validity of internet-based data collection in psycholinguistics, see Enochson & Culbertson, 2015; Keller et al., 2009). The experiment used non-cumulative, word-by-word self-paced reading. Participants read each sentence one word at a time, pressing the space bar to move on to the next word. After finishing each sentence, a yes/no comprehension question appeared on a separate screen, which participants responded to by pressing the ‘1’ or ‘2’ keys. Experimental and filler items were pseudo-randomised such that no two experimental items appeared next to each other and distributed across two lists in a Latin-square design that were completed by the same number of participants.

The experiment began with an information page and consent form. Participants completed three practice trials before the main experiment. After the main experiment, L2 participants completed the Quick Placement Test.

Data Analysis

Before data analysis, comprehension accuracy to filler items was checked as a measure of participant attention. Participants were only included if they scored at least 75% correct on the filler questions. This led to the removal of two L2 speakers, who were replaced with two L2ers who reached the threshold value. For the remaining 128 participants, mean comprehension accuracy to fillers was 96% and 95% for L1ers and L2ers respectively (L1 range = 86–100; L2 range = 84–100). Additionally, all reaction times less than 100ms or greater than 10,000ms were removed prior to analysis, as these likely indicate lapses in attention. This led to the removal of less than 0.1% of the data.

To check for ambiguity effects, we first conducted a standard between-groups analysis of reading times at the critical disambiguating verb and a spillover region, defined as the next word in the sentence, along with comprehension accuracy. Reading times were log-transformed to remove skew and to normalise model residuals (Vasishth & Nicenboim, 2016). The reading time data were analysed using linear mixed-effects models with crossed random effects for subjects and items (Baayen et al., 2008), using the maximal random effects structure that converged (Barr et al., 2013). Fixed effects included sum-coded main effects of group (L1/L2), ambiguity (unambiguous/ambiguous) and their interaction. For each fixed effect, p values were calculated using the Satterthwaite approximation implemented by the lmerTest package (Kuznetsova et al., 2017). Analysis of comprehension question accuracy followed a similar procedure, but instead used generalised mixed-effects models with a binomial distribution. The data and analysis code can be found at the OSF website (https://osf.io/hmuvz/).

Results

Group analyses

Mean reading times at the critical and spillover regions, along with comprehension accuracy, can be found in Table 1. Table 2 provides a summary of the inferential statistics.

TABLES 1 AND 2 HERE

For reading times, L2ers were generally slower readers than L1ers, although the main effect of group was significant at the critical region but not the spillover region. There were significant main effects of ambiguity at both regions, with longer reading times in ambiguous than unambiguous sentences. The group by ambiguity interaction was not significant at either region.

For comprehension accuracy, there was a significant main effect of ambiguity, with lower comprehension accuracy in the ambiguous compared to unambiguous condition, and a significant interaction between group and ambiguity. Separate analyses indicated significant ambiguity effects for both groups, with a larger effect for the L2ers (L1 estimate = -0.739, 95% CI [-0.977, -0.501], z = -6.20, p < .001); L2 estimate = -1.373, 95% CI [-1.683, -1.062], z = -8.85, p < .001). However, the difference here seems to be driven by the fact that L1ers are significantly less accurate than L2ers in the unambiguous condition (estimate = 0.321, 95% CI [0.036, 0.606], z = 2.26, p = .024). Although L1ers have numerically higher accuracy in the ambiguous condition than L2ers, this comparison was not significant (estimate = -0.207, 95% CI [-0.552, 0.139], z = -1.20, p = .231).

This analysis indicates a standard garden-path effect in reading times and lingering misinterpretation in comprehension accuracy. Although we did observe a significant group by ambiguity interaction in comprehension accuracy, the nature of the interaction did not suggest L2ers had more difficulty than L1ers in ambiguous sentences only, contrary to some previous findings (Pozzan & Trueswell, 2016). We return to this issue in the General Discussion, but for present purposes suffice to say that our data at the group level indicate both garden-path effects during online reading and lingering misinterpretation in offline comprehension.

Individual differences in L2 proficiency

We examined whether individual differences in L2 proficiency influenced L2ers' performance in our experiment. For this, we treated proficiency as a (centred) continuous predictor in an analysis of the L2 data, along with the effect of ambiguity and the proficiency by ambiguity interaction. Figure 1 illustrates proficiency effects during online reading and offline comprehension.

FIGURE 1 HERE

The main effect of ambiguity was significant at the critical and spillover regions (critical region estimate = 0.058, 95% CI [0.031, 0.085], t = 4.23, p < .001; spillover region estimate = 0.063, 95% CI [0.041, 0.085], t = 5.66, p < .001), with longer reading times in ambiguous sentences. There was a significant main effect of proficiency at the critical region (estimate = -0.012, 95% CI [-0.022, -0.002], t = -2.43, p = .018), with shorter reading times as proficiency increased, but not the spillover region (estimate = -0.006, 95% CI [-0.013, 0.001], t = -1.62, p = .110). The interaction was significant at the spillover region (estimate = 0.003, 95% CI [0.0004, 0.006], t = 2.33, p = .026), but not the critical region (estimate = 0.001, 95% CI [-0.002, 0.004], t = 0.45, p = .655). At the spillover region, the effect of ambiguity appears to get larger as proficiency increases (see Figure 1). This is however driven by reading times becoming quicker in the unambiguous condition as proficiency increases. The wide shaded area for the ambiguous condition at the lower end of the proficiency scale indicates wide confidence intervals, suggesting difficulty in calculating precise estimates of the ambiguity effect in lower proficiency L2ers (note that only 7 out of 64 participants scored in the 18–39/60 proficiency range).

For comprehension accuracy, the main effects of proficiency and ambiguity were significant (proficiency estimate = 0.099, 95% CI [0.051, 0.147], t = 4.19, p < .001; ambiguity estimate = -1.334, 95% CI [-1.639, -1.028], t = -8.73, p < .001), indicating higher accuracy as proficiency increases and lower scores for ambiguous sentences. The interaction was not significant (estimate = 0.022, 95% CI [-0.006, 0.051], t = 1.52, p = .127).

Split-half correlations

To illustrate by-subject variance in our experiment, Figures 2/3 show the by-subject random effects extracted from a series of mixed-effects models containing a fixed effect of ambiguity only, fit separately to the L1 and L2 data respectively across the different reported measures. Each panel in Figures 2/3 shows the model estimates for each subject's intercepts (overall reading speed or comprehension accuracy) and their ambiguity effect (i.e. the by subject slopes for ambiguity). The estimates are centred around the average effect, and thus show how each subject deviates from this average. For each measure, the intercepts suggest there is overall more variation between subjects than there is error in estimating the intercept for each subject. However, the slopes suggest there is generally as much error in estimating each subject's ambiguity effect as there is variation between subjects in the size of the ambiguity effect. This observation in itself suggests that it may be difficult to consistently assess individual differences in the size of the ambiguity effect in our sample, while it may be possible to consistently measure individual differences in overall reading speed (or accuracy).

FIGURES 2/3 HERE

To examine the consistency of our self-paced reading experiment in measuring individual differences, we conducted a split-half analysis similar to James et al. (2018). We conducted separate analyses for each group, for reading times and comprehension accuracy. In each analysis, we randomly split the data in half, equally divided between the 24 experimental items. We fit a separate mixed-effects model to each half of the data, containing a fixed effect of ambiguity, and crossed random effects for subjects and items with by-subject and by-item random intercepts and slopes for ambiguity. We used mixed-models rather than simply calculating by-subject average reading times and ambiguity effects for each half of the data, as these would be confounded by the different items that a participant sees as a result of our Latin-square design. The by-item adjustments in the mixed-effect model will help deal with this issue.

We then extracted the values for the by-subject random intercepts, as an estimate of overall participant reading speed (or comprehension accuracy), and the by-subject random slopes for ambiguity, as an estimate of each participant's ambiguity effect from the two models fit to the two halves of the data. We then correlated the intercepts and slopes from each half of the data, and repeated this process 100 times, as permutation-based approaches provide a better estimate of split-half reliability than conducting a single split-half (Parsons et al., 2019).[1] The average correlations across these 100 iterations, along with 95% confidence intervals, are reported in Table 3.[2]

[1] We also computed split-half correlations by calculating by-subject means for each effect, which led to similar results. We do not report this analysis here, but provide the code for it at the OSF website (https://osf.io/hmuvz/).

[2] All models were able to estimate by-subject intercepts, and the models for accuracy also estimated by-subject slopes. For reading times, some models estimated zero variance in the by-subject random slope for ambiguity, which made it impossible to correlate this effect between

Note that we do not report p values for the split-half correlations (see Parsons et al., 2019: 8–9), as we are interested in the size of the correlation in our sample, not its statistical significance. Similarly, one might consider comparing two mixed-effects models, one containing by-subject random slopes for an effect, and the other not, and conclude that there are ‘reliable’ individual differences if the by-subject slope significantly improves fit. The significance of this comparison however provides no information about the size of any by-subject variation and does not test whether there is more variance between subjects than within subjects, which is crucial for assessing measurement reliability. For this reason, we recommend against using the ‘significance’ of a model comparison alone in assessing individual differences.

TABLE 3 HERE

The split-half correlations corroborate our observations from Figures 2/3. In reading times at both regions of text, there are high correlations (> .9) between intercepts for both groups. This indicates that our self-paced reading experiment was able to consistently measure individual differences in overall reading speed. However, the split-half correlations for ambiguity are far lower, and are all below < .2, suggesting our experiment was not able to consistently assess individual differences in the size of ambiguity effects. For comprehension accuracy, the correlation for intercepts, an index of overall accuracy, is approximately .76 for both groups, which can be considered consistent (Nunnally, 1978). For ambiguity, the comprehension questions have a higher correlation (~ .4) than the reading times (< .2), but this is still far lower than would typically be desired. Also note that the values for all correlations

the two split-halves. In this case the models were discarded, and the relevant values in Table 3 are based on the remaining models.

are very similar for L1ers and L2ers, suggesting our experiment is roughly as good (or bad) at estimating individual differences in both groups. We also calculated split-half reliability of the L2 proficiency test using the permutation method described above, which yielded a high split-half correlation (.806, 95% CI [.800, .813]).

The split-half analyses suggest our self-paced reading experiment can consistently measure individual differences in overall reading speed and comprehension accuracy and, as the split-half correlation for the proficiency test was also high, the significant main effects of L2 proficiency that we observed are interpretable. The very low levels of consistency in the ambiguity effect however, especially during reading, indicates it is difficult to meaningfully interpret any individual differences in the size of ambiguity effects in our sample. In particular, it raises questions about the interpretability of the ambiguity by proficiency interaction at the spillover region. We discuss this issue in more detail in the General discussion.

General discussion

In this study, we provided a case study of how measurement reliability affects the quantifiability of individual differences in sentence processing by examining garden-path effects offline and online in self-paced reading. At the group level, we replicated previous results (Jacob & Felser, 2016; Kaan et al., 2019; Witzel et al., 2012), indicating garden-path effects in online reading and lingering misinterpretation in offline comprehension. L2 proficiency also influenced the results. In terms of measurement reliability, although our experiment provided a consistent measure of individual differences in overall reading speed and comprehension accuracy, it did not consistently measure individual differences in the size of garden-path effects. We discuss our results in terms of garden-path effects, and in terms of measurement reliability in L1 and L2 sentence processing, in turn below.

Garden-path effects in L1 and L2 processing

Our online results replicated previous studies indicating garden-path effects resulting from the co-ordination ambiguity in L1 and L2 readers (Kaan et al., 2019; Witzel et al., 2012). We also found clear evidence of lingering misinterpretation in both groups, with lower comprehension accuracy in ambiguous than unambiguous sentences, extending previous L2 findings from Jacob and Felser (2016) on subject-object temporary ambiguities to co-ordination. Although we did not observe any significant differences in the size of the garden-path effect between the two groups during online reading, we did find a significant group by ambiguity interaction in offline comprehension. Here, the ambiguity effect, defined as the difference between unambiguous and ambiguous sentences, was larger for L2 speakers than L1 speakers. This might, we believe wrongly in this case, be taken as evidence that L2ers have more difficulty revising garden-path sentences than L1ers (e.g. Pozzan & Trueswell, 2016). However, here the nature of the interaction needs to be considered, as the L1/L2 difference was largely driven by L2ers having significantly higher accuracy than L1ers in the unambiguous condition, while differences in the ambiguous condition were not significant. We thus do not interpret this interaction as indicating more difficulty in reanalysis in L2 learners. We also do not draw any strong conclusions about why the L1ers were less accurate than the L2ers in the unambiguous condition, as other studies (e.g. Pozzan & Trueswell, 2016) have not reported this finding. Rather, we note how this limits our interpretation of the group by ambiguity interaction. Indeed, these results highlight a broader issue in interpreting interactions. Sometimes such effects are interpreted in terms of difference scores, but these are difficult to interpret in the case of baseline differences between groups, as in our study (see Hedge et al, 2018 for discussion).

We also found some evidence of individual differences in L2 processing. Reading times at the critical region were generally faster as proficiency increased, and overall offline comprehension accuracy was positively correlated with proficiency. That we observed high split-half correlations for proficiency, overall reading speed and accuracy suggests these results are interpretable. The same cannot be said for individual differences in garden-path effects however. The results of the inferential statistics at the spillover region in online reading might be taken to suggest larger garden-path effects as proficiency increased. However, given the wide confidence intervals of the ambiguity effect in participants with lower proficiency (see Figure 1), we are cautious in interpreting this effect. The low split-half correlations for the garden-path effect (discussed in more detail below) also raise concerns about the interpretability of this finding given our sample size. Consideration of measurement reliability is thus crucial here in assessing what inferences can be drawn from our data.

Note also that in their study of co-ordination, Kaan et al., (2019) did not find any significant proficiency effects. Some other previous studies on ambiguity resolution have also either reported overall individual differences in reading speed as a result of proficiency (Gerth et al., 2017; Roberts & Felser, 2011) or no significant effects of individual differences in proficiency (Jacob & Felser, 2016). That studies do not consistently find an interaction between ambiguity and proficiency is not surprising if standardly used reading tasks do not consistently measure by-subject individual differences in garden-path effects.

Quantifying individual differences in bilingual sentence processing

The results of our split-half correlations raise concerns about what our sample can tell us about individual differences. Although we observed high split-half correlations for measures of overall reading speed and overall comprehension accuracy, the split-half correlations for the ambiguity effect were well below accepted standards to be considered consistent measures of individual differences. These results are comparable to James et al.’s (2018) findings, who similarly reported high split-half correlations for overall reading speed, but low split-half correlations for experimental manipulations in their self-paced reading study on L1 readers.

We emphasise that measurement reliability is a property of a sample (Parsons et al., 2019) and as such, while our own sample yielded low reliability for garden-path effects, we cannot rule out that reliability might be higher in a different sample. Large sample sizes are of course required for individual differences research, and we do not claim that our own sample was large in this regard. Parsons et al. (2019) discuss the relationship between reliability and statistical power and demonstrate the adverse effect that low reliability has on power. They calculated the sample sizes required to test a range of correlations between two variables (r = .3, .5 and .7) with 80% power, assuming varying degrees of reliability. Based on their estimates (Parsons et al., 2019: 5), assuming reliability of one individual differences measure of .8 and another .2, similar to our split-half correlations for L2ers' OPT scores and their garden-path effects during reading respectively, we would require over 400 participants to test a correlation between the two variables if the true correlation between them was r = .3, while over 200 participants would be required if we assume a larger correlation of .5. However, if reliability for both measures were high (.8), our tested sample size would be adequate if the true correlation was r = .5, while over 100 participants would still be required for r = .3. Thus, it is imperative to consider measurement reliability when assessing the sample size required to examine individual differences.

Rouder et al. (2019) distinguish "tasks" from "measures". In experimental tasks, researchers typically calculate difference scores between conditions, for example the garden-path effect in our study, while in measures, such as a proficiency test, there are no manipulations and scores are based on overall performance. Measures are thus typically more reliable than tasks because measures do not involve subtraction. Rouder et al. claim that for some experiment tasks the size of the between-subject difference may be so small, relative to trial-level noise, that it will always be difficult to assess individual differences. As such, they recommend researchers focus on measures, rather than tasks, when examining individual differences. Our results, with high reliability for measures of overall reading speed and comprehension accuracy but not garden-path effects, are consistent with this claim.

Although our results suggest little between-subject variability in the size of garden-path effects in our sample, note that we cannot distinguish between whether this indicates little systematic variation in garden-path effects between individuals, or little systematic variation specifically in self-paced reading. Further assessment of measurement reliability with other tasks such as eye-tracking is required to examine this issue. Adaptations to the self-paced reading task will also influence reliability. For example, increasing the number of observations per participant helps better estimate the effect at the participant level by minimising trial-level noise (Rouder et al., 2019). Thus, having more observations per participant will help here. Note that in some psycholinguistic tasks, including self-paced reading, participants speed-up over the course of an experiment, which attenuates ambiguity effects (Fine et al., 2013; Harrington Stack et al., 2018; Prasad & Linzen, 2019). This speed-up over time may make it difficult to estimate the subject-level effect with any certainty. Thus, while having more observations per participant will increase measurement reliability, the benefit of more observations needs to be weighed against potential changes in behaviour over the course of a longer experiment, at least in reading time studies. Irrespective of this issue, our results nevertheless indicate that high measurement reliability cannot be taken for granted.

One cause for concern might be that our experiment yielded low split-half correlations for the ambiguity effect because it was conducted online. However, if the data collection method was the source of the low correlations, we should not have observed high split-half correlations in our analyses for the by-subject intercepts. Additionally, the split-half correlations we found are comparable to those reported by James et al. (2018), who conducted their experiment in a lab setting. Thus, we do not believe that the online setting of our study has impacted on our results.

Although we would contend that based on our results, James et al. (2018) and the logic of the reliability paradox as described by Hedge et al. (2018), it is entirely possible that the problem of low measurement reliability is an issue beyond the garden-path effects we studied, future research will need to assess the consistency of any individual differences in other linguistic phenomena. In addition to the garden-path effects we examined, L2 researchers often use other diagnostics, such as plausibility effects (e.g. Williams et al., 2001) or grammaticality effects (e.g. Jiang, 2004), amongst others, to compare L1 and L2 processing. Future research will need to assess the consistency with which these other phenomena vary between L1 and L2 speakers.

Further research is also required to assess how well other commonly used tasks are able to consistently assess individual differences. Consideration should be given to offline and online tasks, as our offline task gave the best split-half correlation for the ambiguity effect, although this was still below accepted standards. Some studies have reported high levels of consistency for offline tasks that have measured attachment preferences in ambiguous relative clauses in particular (James et al., 2018; Swets et al., 2007), suggesting offline tasks may be better suited for assessing individual differences in specific phenomena. Given the variability that has been observed in relative clause attachment across decades of psycholinguistic research (see research following Carreiras & Clifton, 1993), it is perhaps not surprising that this specific phenomenon is subject to measurable individual variation. More generally, these findings suggest it might be frugal to first assess individual differences in offline measures before attempting to examine them during online processing.

In sum, our results highlight how assessing measurement reliability is an important issue in L2 research. James et al. (2018) recommended that researchers report reliability measures when examining individual differences in sentence processing, while Parsons et al. (2019) made similar recommendations for research in psychology more broadly. We echo these proposals, and recommend that researchers report reliability measures when studying individual differences, describe how they were calculated, and consider what they mean for the interpretability of any statistical inferences that may be drawn from a study. Given the importance of measurement reliability for statistical inference (Parsons et al., 2019), L2 researchers must seriously consider what inferences can be made about individual differences when measurement reliability is low.

Conclusion

Although individual differences are of increasing theoretical interest in L1 and L2 sentence processing, there has thus far been little research on measurement reliability in the L2 processing literature, despite its importance in assessing individual differences. We examined garden-path effects in L1 and L2 speakers in a self-paced reading experiment and tested the extent to which our sample consistently measured individual differences. Although at the group-level we found evidence of garden-path effects during processing and lingering misinterpretation in offline comprehension, we did not observe consistent individual differences in the size of the garden-path effect, either online or offline. This raises questions about whether the task we used is able to consistently measure individual differences. These results are in-line with other recent research which more broadly questions whether cognitive tasks can be used to examine individual differences (Elliott et al., 2020; Hedge et al., 2018), and highlights the need to report measurement reliability (James et al., 2018; Parsons et al., 2019). Before searching for individual differences in L1 and L2 sentence processing, we must first assess the extent to which psycholinguistic tasks are able to reliably gauge such differences.

References

Arnon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*, *52*(1), 68–81. https://doi.org/10.3758/s13428-019-01205-5

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Brown-Schmidt, S., & Fraundorf, S. H. (2015). Interpretation of informational questions modulated by joint knowledge and intonational contours. *Journal of Memory and Language*, *84*, 49–74. https://doi.org/10.1016/j.jml.2015.05.002

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*(01). https://doi.org/10.1017/S0140525X99001788

Carreiras, M., & Clifton, C. (1993). Relative Clause Interpretation Preferences in Spanish and English. *Language and Speech*, *36*(4), 353–372. https://doi.org/10.1177/002383099303600401

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic Roles Assigned along the Garden Path Linger. *Cognitive Psychology*, *42*(4), 368–407. https://doi.org/10.1006/cogp.2001.0752

Clahsen, H., & Felser, C. (2006). Continuity and shallow structures in language processing. *Applied Psycholinguistics*, *27*(1), 107–126. https://doi.org/10.1017/S0142716406060206

Clahsen, H., & Felser, C. (2018). Some notes on the shallow structure hypothesis. *Studies in Second Language Acquisition*, *40*(3), 693–706. https://doi.org/10.1017/S0272263117000250

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*(3), 369–382. https://doi.org/10.1177/0267658312443651

Cunnings, I. (2017). Parsing and Working Memory in Bilingual Sentence Processing. *Bilingualism: Language and Cognition*, *20*(4), 659–678. https://doi.org/10.1017/S1366728916000675

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, 095679762091678. https://doi.org/10.1177/0956797620916786

Engelhardt, P. E., & Ferreira, F. (2010). Processing Coordination Ambiguity. *Language and Speech*, *53*(4), 494–509. https://doi.org/10.1177/0023830910372499

Enochson, K., & Culbertson, J. (2015). Collecting Psycholinguistic Response Time Data Using Amazon Mechanical Turk. *PLOS ONE*, *10*(3), e0116946. https://doi.org/10.1371/journal.pone.0116946

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PLoS ONE*, *8*(10), e77661. https://doi.org/10.1371/journal.pone.0077661

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, *14*(2), 178–210. https://doi.org/10.1016/0010-0285(82)90008-1

Gerth, S., Otto, C., Felser, C., & Nam, Y. (2017). Strength of garden-path effects in native and non-native speakers' processing of object–subject ambiguities. *International Journal of Bilingualism*, *21*(2), 125–144. https://doi.org/10.1177/1367006915604401

Gibson, E., Desmet, T., Grodner, D., Watson, D., & Ko, K. (2005). Reading relative clauses in English. *Cognitive Linguistics*, *16*(2). https://doi.org/10.1515/cogl.2005.16.2.313

Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*(6), 864–877. https://doi.org/10.3758/s13421-018-0808-6

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hoeks, J. C. J., Vonk, W., & Schriefers, H. (2002). Processing Coordinated Structures in Context: The Effect of Topic-Structure on Ambiguity Resolution. *Journal of Memory and Language*, *46*(1), 99–119. https://doi.org/10.1006/jmla.2001.2800

Hopp, H. (2006). Syntactic features and reanalysis in near-native processing. *Second Language Research*, *22*(3), 369–397. https://doi.org/10.1191/0267658306sr272oa

Hopp, H. (2018). The Bilingual Mental Lexicon in L2 Sentence Processing. *Second Language*, *17*, 5–27. https://doi.org/10.11431/secondlanguage.17.0_5

Jacob, G., & Felser, C. (2016). Reanalysis and semantic persistence in native and non-native garden-path recovery. *Quarterly Journal of Experimental Psychology*, *69*(5), 907–925. https://doi.org/10.1080/17470218.2014.984231

James, A. N., Fraundorf, S. H., Lee, E.-K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language*, *102*, 155–181. https://doi.org/10.1016/j.jml.2018.05.006

Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics*, *25*(4), 603–634. https://doi.org/10.1017/S0142716404001298

Kaan, E., Futch, C., Fernández Fuertes, R., Mujcinovic, S., & Álvarez De La Fuente, E. (2019). Adaptation to syntactic structures in native and nonnative sentence comprehension. *Applied Psycholinguistics*, *40*(1), 3–27. https://doi.org/10.1017/S0142716418000437

Kan, I. P., Teubner-Rhodes, S., Drummey, A. B., Nutile, L., Krupa, L., & Novick, J. M. (2013). To adapt or not to adapt: The question of domain-general cognitive control. *Cognition*, *129*(3), 637–651. https://doi.org/10.1016/j.cognition.2013.09.001

Keller, F., Gunasekharan, S., Mayo, N., & Corley, M. (2009). Timing accuracy of Web experiments: A case study using the WebExp software package. *Behavior Research Methods*, *41*(1), 1–12. https://doi.org/10.3758/BRM.41.1.12

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Sciences*, *22*(2), 154–169. https://doi.org/10.1016/j.tics.2017.11.006

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). **lmerTest** Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Leclercq, P., Edmonds, A., & Hilton, H. (Eds.). (2014). *Measuring L2 proficiency: Perspectives from SLA*. Multilingual Matters.

Lee, E.-K., Lu, D. H.-Y., & Garnsey, S. M. (2013). L1 word order and sensitivity to verb bias in L2 processing. *Bilingualism: Language and Cognition*, *16*(4), 761–775. https://doi.org/10.1017/S1366728912000776

Linck, J. A., & Cunnings, I. (2015). The Utility and Application of Mixed-Effects Models in Second Language Research: Mixed-Effects Models. *Language Learning*, *65*(S1), 185–207. https://doi.org/10.1111/lang.12117

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*(1), 35–54. https://doi.org/10.1037/0033-295X.109.1.35

Novick, J. M., Hussey, E., Teubner-Rhodes, S., Harbison, J. I., & Bunting, M. F. (2014). Clearing the garden-path: Improving sentence processing through cognitive control training. *Language, Cognition and Neuroscience*, *29*(2), 186–217. https://doi.org/10.1080/01690965.2012.758297

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

Pozzan, L., & Trueswell, J. C. (2016). Second language processing and revision of garden-path sentences: A visual word study. *Bilingualism: Language and Cognition*, *19*(3), 636–643. https://doi.org/10.1017/S1366728915000838

Prasad, G., & Linzen, T. (2019). *Do self-paced reading studies provide evidence for rapid syntactic adaptation?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/9ptg4

Roberts, L., & Felser, C. (2011). Plausibility and recovery from garden paths in second language sentence processing. *Applied Psycholinguistics*, *32*(2), 299–331. https://doi.org/10.1017/S0142716410000421

Rouder, J., Kumar, A., & Haaf, J. M. (2019). *Why Most Studies of Individual Differences With Inhibition Tasks Are Bound To Fail* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/3cjr5

Ryskin, R. A., Benjamin, A. S., Tullis, J., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General*, *144*(5), 898–915. https://doi.org/10.1037/xge0000093

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72. https://doi.org/10.2307/1412159

Sturt, P., Pickering, M. J., & Crocker, M. W. (1999). Structural Change and Reanalysis Difficulty in Language Comprehension. *Journal of Memory and Language*, *40*(1), 136–150. https://doi.org/10.1006/jmla.1998.2606

Swets, B., Desmet, T., Hambrick, D. Z., & Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: A psychometric approach. *Journal of Experimental Psychology: General*, *136*(1), 64–81. https://doi.org/10.1037/0096-3445.136.1.64

Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, *131*(3), 373–403. https://doi.org/10.1016/j.cognition.2014.01.007

Vasishth, S., & Nicenboim, B. (2016). Statistical Methods for Linguistic Research: Foundational Ideas - Part I: Statistical Methods for Linguistics - Part I. *Language and Linguistics Compass*, *10*(8), 349–369. https://doi.org/10.1111/lnc3.12201

Vuong, L. C., & Martin, R. C. (2014). Domain-specific executive control and the revision of misinterpretations in sentence comprehension. *Language, Cognition and Neuroscience*, *29*(3), 312–325. https://doi.org/10.1080/01690965.2013.836231

Williams, J. N., Möbius, P., & Kim, C. (2001). Native and non-native processing of English *wh* - questions: Parsing strategies and plausibility constraints. *Applied Psycholinguistics*, *22*(4), 509–540. https://doi.org/10.1017/S0142716401004027

Witzel, J., Witzel, N., & Nicol, J. (2012). Deeper than shallow: Evidence for structure-based parsing biases in second-language sentence processing. *Applied Psycholinguistics*, *33*(2), 419–456. https://doi.org/10.1017/S0142716411000427

Table 1. *Average reading times and comprehension accuracy (SDs in parenthesis).*

	L1 Speakers		*L2 Speakers*	
Critical Region				
Unambiguous	451	(284)	513	(340)
Ambiguous	533	(569)	621	(524)
Spillover Region				
Unambiguous	418	(253)	445	(198)
Ambiguous	492	(317)	535	(417)
Comprehension Accuracy				
Unambiguous	0.80	(0.40)	0.87	(0.50)
Ambiguous	0.57	(0.50)	0.50	(0.34)

Table 2. *Summary of between-groups statistical analysis.*

	Estimate (95% CI)	*t (z)*	*p*
Critical Region			
Group	0.080 [0.016, 0.144]	2.51	.013
Ambiguity	0.044 [0.025, 0.062]	4.79	< .001
Group*Ambiguity	0.014 [-0.004, 0.033]	1.53	.134
Spillover Region			
Group	0.047 [-0.002, 0.097]	1.91	.058
Ambiguity	0.064 [0.049, 0.079]	8.43	< .001
Group*Ambiguity	-0.001 [-0.018, 0.016]	-0.14	.893
Comprehension Accuracy			
Group	0.078 [-0.201, 0.358]	0.56	.575
Ambiguity	-1.040 [-1.265, -0.815]	-9.26	< .001
Group*Ambiguity	-0.275 [-0.431, -0.120]	-3.54	< .001

Table 3. *Split half correlations for L1 speakers and L2 speakers.*

	L1 Speakers		*L2 Speakers*	
	r	*95% CI*	*r*	*95% CI*
Critical Region				
Intercept	.919	(.916, .922)	.922	(.920, .925)
Ambiguity	.174	(.153, .195)	.190	(.172, .209)
Spillover Region				
Intercept	.917	(.914, .919)	.918	(.915, .920)
Ambiguity	.188	(.171, .204)	.170	(.153, .187)
Comprehension Accuracy				
Intercept	.766	(.758, .774)	.764	(.755, .773)
Ambiguity	.410	(.395, .426)	.416	(.398, .433)