Rapid syntactic adaptation in self-paced reading: detectable, but only with many participants

Grusha Prasad[1] and Tal Linzen[2]

[1] Department of Cognitive Science, Johns Hopkins University

[2]Department of Linguistics, New York University

[2]Center for Data Science, New York University

Author Note

Grusha Prasad, Department of Cognitive Science, Johns Hopkins University

The stimuli for all experiments and the scripts used to create the experiments, generate the plots and run all the analyses are available on OSF: `https://osf.io/qd8ye/`

Correspondence concerning this article should be addressed to Grusha Prasad, Department of Cognitive Science, Johns Hopkins University, 3400 North Charles Street, Baltimore MD, 21218.

Contact: grusha.prasad@jhu.edu

Abstract

Temporarily ambiguous sentences that are disambiguated in favor of a less preferred parse are read more slowly than their unambiguous counterparts. This slowdown is referred to as a *garden path effect*. Recent self-paced reading studies have found that this effect decreased over the course of the experiment as participants were exposed to such syntactically ambiguous sentences. This decrease in the magnitude of the effect has been interpreted as evidence that readers calibrate their expectations to the context; this minimizes their surprise when they encounter these initially unexpected syntactic structures. Such recalibration of syntactic expectations, referred to as *syntactic adaptation*, is only one possible explanation for the decrease in garden path effect, however; this decrease could also be driven instead by increased familiarity with the self-paced reading paradigm (*task adaptation*). The goal of this paper is to adjudicate between these two explanations. In a large between-group study (n = 642), we find evidence for syntactic adaptation over and above task adaptation. The magnitude of syntactic adaptation compared to task adaptation is very small, however. Power analyses show that a large number of participants is required to detect, with adequate power, syntactic adaptation in future between-subject self-paced reading studies. This issue is exacerbated in experiments designed to detect modulations of the basic syntactic adaptation effect; such experiments are likely to be underpowered even with more than 1200 participants. We conclude that while, contrary to recent suggestions, syntactic adaptation can be detected using self-paced reading, this paradigm is not very effective for studying this phenomenon.

Rapid syntactic adaptation in self-paced reading: detectable, but only with many participants

Introduction

Humans' ability to extract statistical regularities from their environment plays an important role in language acquisition and processing (Mitchell, Cuetos, Corley, & Brysbaert, 1995; Romberg & Saffran, 2010). In sentence comprehension, in particular, predictable syntactic structures are easier to process than unpredictable ones (MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell, 1996). Under a rational account of sentence comprehension, we would expect these predictability effects to be driven by context-specific statistical regularities (Anderson, 1990): since the distribution of syntactic structures can vary widely across environments and contexts, readers' expectations will only be an accurate reflection of the statistics of the current environment if they can rapidly calibrate their expectations to match those statistics (Fine, Jaeger, Farmer, & Qian, 2013).

In line with this hypothesis, Wells, Christiansen, Race, Acheson, and MacDonald (2009) showed that participants who were exposed to sentences with relative clauses over several experimental sessions read new sentences with relative clauses faster than did participants who were exposed to sentences with other syntactic structures. Building on this finding, Fine et al. (2013) tested whether readers can recalibrate their expectations over the course of a single experimental session, focusing on sentences such as (1):

(1) The experienced soldiers warned about the dangers **conducted the midnight** raid. (Reduced RC; ambiguous)

Sentence (1) is temporarily ambiguous between a main verb reading, where the soldiers warned someone about the danger, and a relative clause reading, where the soldiers were warned by someone about the danger. The sentence is eventually disambiguated in favor of the relative clause reading by *conducted.* This temporary ambiguity is absent from a minimally different sentence with an unreduced relative clause like (2); in this sentence, only the relative clause reading is possible:

(2) The experienced soldiers who were told about the dangers **conducted the midnight** raid. (Unreduced RC; unambiguous)

Across a range of studies, the words of the disambiguating region of (1), marked in boldface, have been shown to be read more slowly than the same words in a matched unambiguous sentence such as (2) (Clifton Jr et al., 2003; Kemper, Crow, & Kemtes, 2004; Liversedge, Paterson, & Clayes, 2002; MacDonald et al., 1994; Trueswell, 1996). We refer to this difference in reading times as the garden path effect.

Fine et al. (2013) interpreted the garden path effect as a consequence of more general word predictability effects (following Hale 2001): when reading the ambiguous region of sentence (1), participants are likely to interpret the verb *warned* as the main verb of the sentence, since verbs like *warned* occur more frequently as matrix clause verbs than as verbs introducing a passive reduced relative clause as in (1). Given this bias towards a main verb reading, words which disambiguate the temporarily ambiguous sentence in favor of the relative clause reading are less expected than the same words when they occur in a sentence like (2), where only a relative clause reading is possible. Since, all else being equal, less predictable words are read more slowly than predictable ones (Ehrlich & Rayner, 1981; Smith & Levy, 2013), the greater frequency of main verb parses can explain the garden path effect.

Fine and colleagues hypothesized that if participants update their expectations to match the statistics of the environment, then, in an experimental context where participants were exposed to several sentences such as (1), with reduced RCs, words that disambiguate the sentence in favor of the relative clause reading would become more predictable over time; this, in turn, would result in a decrease in the garden path effect. We will refer to this hypothesis as the *syntactic adaptation* hypothesis. In line with this hypothesis, Fine et al. (2013) observed a decrease in the garden path effect over the course of a self-paced reading experiment, in which readers press a key to reveal the next word in the sentence. A similar decrease has since been observed in other self-paced reading studies (Fine & Jaeger, 2016; Stack, James, & Watson, 2018).

While the decrease in garden path effect is consistent with the syntactic adaptation account, syntactic adaptation is not the only possible explanation for this finding. In all of the studies mentioned above, as the experiment progressed, reading times (RTs) decreased not only for temporarily ambiguous sentences, but also for sentences in all other conditions, regardless of the syntactic structure of the sentence (Fine & Jaeger, 2016; Fine et al., 2013; Stack et al., 2018). We will refer to the decrease in RTs that is independent of any recalibration of syntactic expectations as *task adaptation*. In the following paragraphs, we explain how task adaptation could result in a decrease in garden path effect, even in the absence of syntactic adaptation.

We assume that task adaptation does not directly depend on the syntactic structure of the sentence, but could depend on the speed with which the sentence is read when encountered early in the experiment. If the rate of task adaptation—the speedup in milliseconds from one trial to the next—is greater for sentences that are read more slowly at the beginning of the experiment (to which we will refer as "difficult sentences" for convenience) than for sentences that are read more rapidly ("easy sentences"), then, over time, the difference in RTs between easy and difficult sentences will decrease, resulting in a decrease in the garden path effect (see Figure 1). Such variability in difficulty across sentences could arise from any number of of factors, including word frequency, plausibility, predictability, and syntactic disambiguation difficulty. We will refer to the class of task adaptation functions that have this property as *start-point dependent task adaptation*. If task adaptation is indeed start-point dependent, then even though the same task adaptation function applies to both reduced and unreduced RCs, the rate of decrease in RTs would be greater for reduced RCs than for unreduced RCs. If that is the case, it is possible that the decrease in garden path effect observed in previous studies was driven by task adaptation alone, or by a combination of task and syntactic adaptation.

There are at least two other possible types of task adaptation functions. First, the

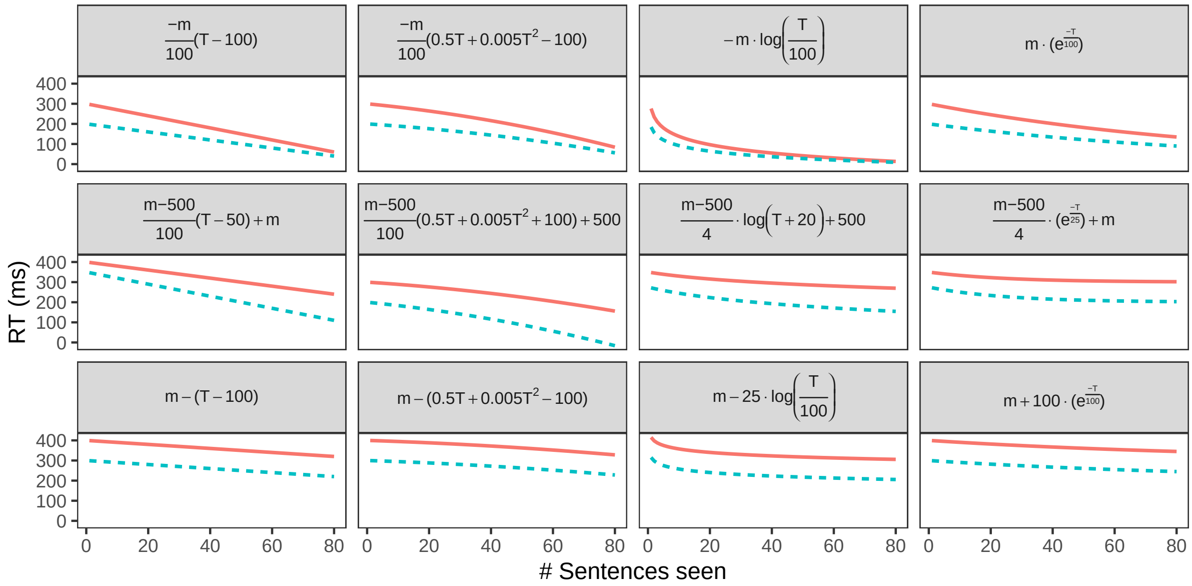


Figure 1. An illustration of some of the possible functions that could describe the decrease in reading time caused by task adaptation for two sentences (red solid and blue dashed) over the course of the experiment. At the beginning of the experiment (at trial 1), the sentence depicted by the red solid line is read more slowly than the sentence depicted by the blue dashed line. The top two rows depict functions that are sensitive to the initial reading times of the sentences (start-point dependent and diverging start-point dependent functions) and the bottom row depicts functions that are not sensitive to these initial reading times (start-point independent functions). The value of the parameter m is 300 for the red line and 200 for the blue one. The difference in RTs between the red solid and blue dashed line decreases only in the start-point dependent functions. These simple functions were chosen to illustrate the three classes of task-adaptation functions rather than for their psychological plausibility. While many of these functions are not psychologically plausible because they predict negative RTs after some trials, they can be modified to be more psychologically plausible (e.g., by enforcing a floor).

rate of task adaptation could be *lower* for difficult sentences than for easy ones (*diverging start-point dependent*). In this intuitively less likely case, the garden path effect would increase over time. Second, the rate of task adaptation could be identical for easy and difficult sentences (*start-point independent*). In this case, task adaptation would not cause the garden path effect to change over time. If task adaptation follows either of these patterns, the decrease in garden path effect observed by previous studies cannot be explained by task adaptation.

Since the form of the task adaptation function that characterizes self-paced reading studies is currently unknown, all of the three alternatives discussed above are possible. Therefore, we cannot know whether the decrease in garden path effect observed in previous studies was driven by start-point dependent task adaptation alone, by syntactic adaptation alone, or by a combination of the two. The goal of this paper is to adjudicate between these three possibilities. Before describing our approach, we briefly discuss previous attempts to do so.

The Fine et al. (2013) experiment mentioned above consisted of two blocks. In the first block, participants ($n = 80$) read either 16 filler sentences (Filler-exposed group), or 16 sentences with RCs, half of which had reduced RCs like (1), and the other half unreduced RCs like (2) (RC-exposed group). Then, in the second block of the experiment, the garden path effect was measured in both groups by comparing the RTs for sentences with reduced RCs and with unreduced RCs (five each).[1] Fine et al. (2013) found that the garden path effect in the RC-exposed group decreased between the first block and the second. In the second block, the garden path effect was smaller in the RC-exposed group than the Filler-exposed group, although this interaction was only marginally significant ($\beta = -5$, $t = -1.7$, $p = 0.08$). Fine and colleagues argued that the decrease in garden path effect they observed was a result of syntactic adaptation: if it had been caused by task adaptation alone, the garden path effect would not differ

[1] Fine et al. (2013) also included a third block with sentences that were disambiguated in favor of the main verb reading, e.g., *The experienced soldiers warned about the dangers before the midnight raid.* We briefly discuss this manipulation in the General Discussion.

across the two groups, both of which were exposed to the same number of sentences.

In a later experiment that used the same design as Fine et al. (2013) but considerably more participants and items (423 participants, 32 sentences in Block 1 and 20 sentences in Block 2), Stack et al. (2018) replicated the decrease in the garden path effect observed by Fine et al. (2013) for the RC-exposed group of participants, but failed to replicate the crucial interaction: the garden path effects in Block 2 did not differ significantly between the RC-exposed and Filler-exposed participants ($\beta = 1.25$, $t = 1.05$, $p > 0.05$).[2] Based on these results, Stack and colleagues argued that the observed decrease in garden path effect was likely driven by task adaptation and not by syntactic adaptation. In a response to Stack and colleagues, Jaeger, Bushong, and Burchill (2019) challenged these conclusions. Based on a reanalysis of the data from Stack et al. (2018) and computational simulations, Jaeger and colleagues argued that Stack et al.'s experiment, far from being a failure to replicate their earlier work, in fact provides evidence for syntactic adaptation.

The present paper aims to clarify the empirical picture regarding syntactic adaptation in self-paced reading. We report on two experiments designed to investigate which of the factors described earlier can drive the decrease in garden path effect observed in self-paced reading experiments: will we observe syntactic adaptation only, task adaptation only, or a combination of the two? Instead of Fine et al. (2013), our design is based on the second experiment of Fine and Jaeger (2016) (henceforth referred to as FJ16); this experiment includes more items and has a simpler design than the earlier study by the same authors.[3] Across three similar experiments, FJ16 presented their participants with 20 sentences with reduced relative clauses (like (3a)) and 20 with unreduced relative clauses (like (3b)); as in Fine et al. (2013), they found a decrease in the garden path effect over the course of the experiment.

[2] The difference in signs is an artifact of how the predictors were coded in the two studies. In both the studies the garden path effect for the RC-exposed group was smaller than that for the Filler-exposed group.

[3] Specifically, FJ16 did not include the manipulation with sentences that were disambiguated in favor of the main verb reading, e.g., *The experienced soldiers warned about the dangers before the midnight raid.*

(3) a. The evil genie served the golden figs went into a trance.
b. The evil genie who was served the golden figs went into a trance.

Experiment 1 of the present paper is a replication of FJ16. This replication had two goals: first, to ensure that the decrease in garden path effect can be replicated with FJ16's simpler design (to our knowledge, ours is the first attempt to replicate FJ16); and second, to investigate whether task adaptation is start-point dependent and, as such, can on its own lead to a decrease in garden path effect. This experiment successfully replicated the results of FJ16 in both direction and magnitude: as in FJ16, the garden path effect in our Experiment 1 decreased by approximately 1% with every additional reduced relative clause sentence encountered by the participant. We also found evidence that task adaptation is start-point dependent—the rate of task adaptation was greater for sentences that were initially read more slowly than for sentences that were initially read more rapidly. These results suggest that the observed decrease in garden path effect does not necessarily reflect syntactic adaptation: in principle, the decrease could have been driven entirely by start-point dependent task adaptation.

Next, Experiment 2 investigates whether syntactic adaptation results in a decrease in garden path effect over and above the decrease caused by start-point dependent task adaptation. Following a similar logic as in Fine et al. (2013) and Stack et al. (2018), we used a between-group blocked design to compare the garden path effect between participants exposed to RRC sentences (RRC-exposed group) and those exposed to filler sentences (Filler-exposed group). As discussed earlier, if syntactic adaptation results in a decrease in garden path effect over and above task adaptation, we expect the garden path effect following exposure to be smaller in the RRC-exposed group than in the Filler-exposed group.

To test this prediction, we first ran a preliminary experiment, Experiment 2a, in which we measured the magnitude of the garden path effect in a Filler-exposed group. We then used this estimate to predict the magnitude of garden path effect that we are likely to observe for the RRC-exposed group. Based on this prediction, we ran a power analysis to estimate the number of participants required to detect between-group difference in the garden path effect. This power analysis indicated that it would be possible to detect such an effect with adequate power with 800 participants. Next, in Experiment 2b, we collected data for both groups, with a sample size based on our power analysis, and found evidence for syntactic adaptation over and above task adaptation.

Finally, based on our data from Experiment 2b, we ran power analyses to estimate the number of participants required for future experiments investigating the effects of syntactic adaptation using similar between-group designs. These simulations suggested that self-paced reading experiments with a blocked between-group design identical to ours will require around 800 participants to detect the basic syntactic adaptation effect with adequate power; experiments aimed at detecting modulations of this basic effect—e.g., determining whether the magnitude of syntactic adaptation varies across RC types—could be underpowered even with 1200 participants. We conclude that while syntactic adaptation can be detected using self-paced reading (contra Stack et al. 2018), this paradigm might not be very effective for studying this phenomenon; this explains the mixed results found in previous studies.

Experiment 1: Does the garden path effect decrease over time? Can task adaptation account for the decrease?

Method

Participants. We recruited 80 participants via Prolific, a crowdsourcing platform. All participants specified on their profile that English was their first language. They were compensated at a rate of $6.51 per hour.

Materials. We used the same 40 critical items and 80 filler sentences as FJ16. Each of the critical items had a reduced form as in (3a) and an unreduced form as in (3b). To avoid the the temporary syntactic ambiguity illustrated in (3a), the main verbs in all filler sentences were verbs like *woke*, which can only be interpreted as a past tense verb (the past participle in this case would be *woken*), rather than verbs like *served*, which are ambiguous between the two forms.

We generated four pseudorandom orders and, for each of the four orders, two lists counterbalanced for sentence type (i.e. if list 1 had the unreduced version of sentence A and the reduced version of sentence B, list 2 would include the reduced version of sentence A and the unreduced version of sentence B). We then generated a reversed version of each of these eight lists, for a total of 16 lists. Each participant was assigned to one of these 16 lists. To ensure that stimuli from the three conditions—RRC sentences, URC sentences and filler sentences—were evenly distributed throughout the experiment, we generated the pseudorandom orders in five blocks, where each block contained four RRCs, four URCs, and 16 filler sentences. Every two critical items were separated by at least one filler, and at most two critical items of the same condition were allowed to follow each other (across filler items).

Procedure. The experiment was hosted on the IbexFarm website (Drummond, 2016). The procedure was standard for self-paced reading experiments. At the beginning of every trial, each of the words of the sentence was replaced by a dash whose length was roughly equivalent to the length of the word. When the participant pressed the space bar, the dash was replaced by the next word in the sentence and the previous word disappeared. At the end of the sentence, the participant was presented with a comprehension question, and used the keys 'z' and 'm' to respond 'yes' and 'no' respectively. We used the same comprehension questions as FJ16. The correct answer was 'yes' half of the time. Before the experiment started, participants were asked to fill out a brief demographic survey, and were given three practice trials.[4]

Results

Data filtering and exclusion. Although we indicated that only workers whose first language is English should participate in the experiment, four participants reported that English was not their first language. We excluded these participants from our

[4] All the experiments described in this paper were approved by The Johns Hopkins University Homewood Institutional Review Board.

analyses. We further excluded three participants whose comprehension question accuracy on filler sentences was lower than 80%; we excluded from this calculation two fillers whose mean accuracy was two standard deviations lower than the mean accuracy across fillers. Since a majority of the comprehension questions did not directly test whether participants correctly parsed RRC sentences, we did not exclude trials in which participants responded incorrectly to the comprehension questions; our results were qualitatively similar when trials with incorrect answers were excluded.[5] Following the data exclusion criteria used by FJ16, all observations (words) with RTs lower than 100 ms or greater than 2000 ms were excluded. This lead to the exclusion of 0.47% of the observations from the participants who were not excluded.

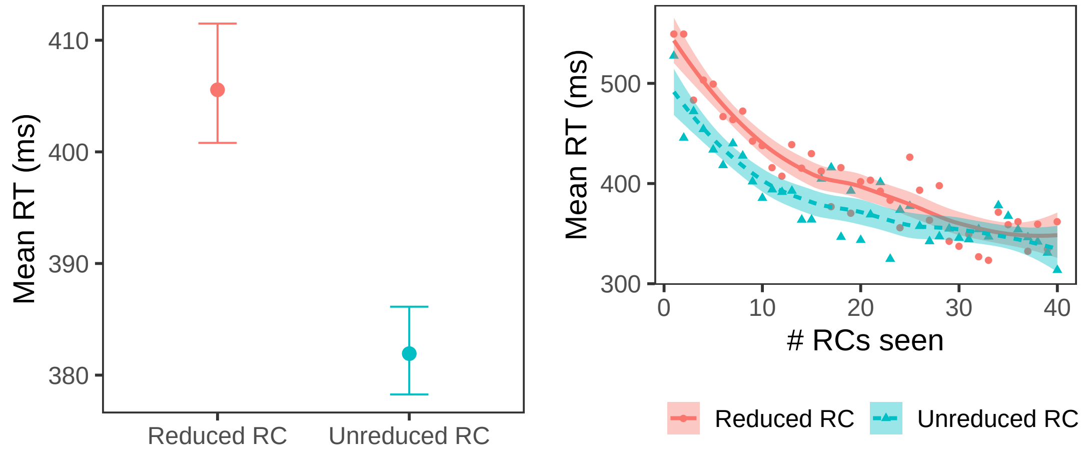
Analysis 1.1: A replication of FJ16's analysis. FJ16 divided each sentence into five regions: subject (*the experienced waitress*), relativizer (*who was*: only URC sentences had this region), ambiguous region (*cooked the grilled chicken*), disambiguating region (*sent her food*) and final word (*back.*). They log-transformed the RTs; further, to control for word length, they fit a linear mixed-effects model predicting log-transformed RTs from word length, and performed all subsequent statistical analyses on the residuals of this model.

Since the garden path effect, which is the focus of interest in the current work, manifests in the disambiguating region, we restricted our analysis of residualized log RTs to this region. We fit a linear mixed-effects model that was nearly identical to the one specified by FJ16 (we modified the random effect structure slightly in order to allow the model to converge).[6] The model included the following predictors:

- Sentence type (referred to as Ambiguity in FJ16): A categorical variable coded as 1 for RRC sentences and -1 for URC sentences.
- Critical item number (Item order in FJ16): The number of critical items (reduced

[5] We provide all details of analyses with the incorrect trials excluded in the following Open Science Framework (OSF) project: `https://osf.io/57ckx/`

[6] Fitting a model with the same random effect structure as in FJ16 yielded nearly identical $\hat{\beta}$ coefficients, but that model, unlike the model we report in this section, failed to converge. Further details can be found in the OSF project.



(a) (b)

Figure 2. Results of Experiment 1. (a) RTs in the disambiguating region for RRC sentences and URC sentences averaged over all participants and items. Error bars represent bootstrapped 95% confidence intervals. (b) RTs as a function of the number of critical items (both reduced and unreduced) seen by the participant, averaged across all participants and items. We fit the data points with a LOESS curve.

and unreduced) that the participant has seen so far.

- *log*(Stimulus number) (Stimulus order in FJ16): The natural log of the total number of sentences (critical items and filler sentences) that the participant has seen so far.
- Interaction between sentence type and critical item number.

Both critical item number and log stimulus order were centered around their mean. The model also included by-item and by-participant random intercepts, along with by-participant slopes for sentence type, critical item number and the interaction between sentence type and critical item number, as well as a by-item slope for sentence type. We estimated p values for the coefficients of this model using Satterthwaite's method, as implemented in the lmerTest package in R (Kuznetsova, Brockhoff, & Christensen, 2017).

The results of this analysis closely replicated FJ16. There was a significant garden path effect ($\hat{\beta} = 0.020$, $SE = 0.005$, $p \ll 0.01$; see Figure 2a). Length-corrected log RTs decreased significantly as a function of both log stimulus number ($\hat{\beta} = -0.083$, $SE = 0.008$, $p \ll 0.01$) and critical item number ($\hat{\beta} = -0.003$, $SE = 0.001$, $p = 0.02$). Crucially, the speedup over the course of the experiment was more pronounced for RRC sentences than for URC sentences ($\hat{\beta} = -0.001$, $SE = 0.0003$, $p < 0.01$; see Figure 2b). The coefficient of this interaction term was identical to that reported by FJ16 ($\hat{\beta} = -0.001$).

Analysis 1.2: Methods. This section reports an alternative analysis that addresses potential limitations of FJ16's analysis replicated in our Analysis 1.1. The first concern is that if word length is collinear with other predictors, then the residualization process used to correct for word length can bias the model's estimates and standard errors for the non-residualized predictors (York, 2012). Length correction is arguably unnecessary with the current design, which is within-item: since the critical region is identical across the URC and RRC versions of the same item, any effect of word length would be canceled out when we estimate the garden path effect. To address this potential issue, in Analysis 1.2 we used log-transformed RTs as the dependent variable instead of residualized length-corrected log transformed RTs used in Analysis 1.1.

A second concern regards the log transformation. The garden path effect is typically calculated by summing or averaging RTs over the disambiguating region. But in Analysis 1.1 we averaged log-transformed RTs, which, when translated to the raw RT scale, is equivalent to *multiplying*, rather than summing, the RTs before dividing the log of the outcome by the number of words in the region. To avoid this counterintuitive arithmetic operation, in Analysis 1.2 we averaged the RTs in the disambiguating region *before* applying the log transformation.

Finally, Analysis 1.1 predicted log-transformed RTs as a linear function of log-transformed stimulus number; this is equivalent to assuming a linear relationship between RTs and stimulus number. Previous work outside of the sentence processing literature, however, suggests that RTs decrease exponentially, not linearly, as a function of practice (Heathcote, Brown, & Mewhort, 2000). In Analysis 1.2, we avoided log-transforming our stimulus number predictor; as a result, this analysis assumes a linear relationship between log-transformed RTs and stimulus number, or, equivalently, an exponential relationship between raw RTs and stimulus number, in line with prior work on the effect of practice.

In summary, the model we fit in Analysis 1.2 included the following predictors: stimulus number, ambiguity, critical item number, and the interaction between ambiguity and critical item number. We centered both stimulus number and critical item number by their mean and scaled them by their standard deviation. The random effect structure for this model included by-item and by-participant random intercepts, along with by-participant and by-item slopes for ambiguity, critical item number and the interaction between the two. We were unable to include by-item and by-participant random slopes for stimulus-number due to model convergence issues.

Analysis 1.2: Results. In this analysis, unlike in Analysis 1.1, the overall decrease in RTs across all conditions was only marginally significant ($\hat{\beta} = -0.158$, $SE = 0.091$, $p = 0.08$). Crucially, however, the magnitude of the garden path effect was greater than in Analysis 1.1, as was the magnitude of the decrease in the garden path effect (garden path effect: $\hat{\beta} = 0.024$, $SE = 0.005$, $p \ll 0.01$; decrease in garden path effect: $\hat{\beta} = -0.014$, $SE = 0.004$, $p \ll 0.01$). If anything, then, addressing our concerns with FJ16's analytical choices caused the effects of primary interest to be more pronounced than they were in Analysis 1.1.

Is task adaptation start-point dependent?

The decrease in RTs across all conditions as a function of stimulus number that was observed in analyses 1.1 and 1.2 suggests, in line with previous studies (Fine & Jaeger, 2016; Fine et al., 2013; Stack et al., 2018), that participants adapt to the self-paced reading paradigm and read sentences more rapidly as the experiment progresses. However, these results do not directly speak to the question of whether task adaptation is start-point dependent or start-point independent[7]—i.e. whether or not the rate of task adaptation is greater for sentences that are read relatively slowly when presented early in the experiment ("difficult sentences") than for those that are read relatively rapidly when presented early in the experiment ("easy sentences"). As discussed earlier, if task-adaptation were indeed start-point dependent, we expect the difference in RTs between easy and difficult sentences to decrease over time, raising the possibility that the decrease in garden path effect observed in Experiment 1 was driven entirely by start-point dependent task adaptation. In this section we investigate whether task adaptation is in fact start-point dependent.

We define the difficulty of a sentence x, which we denote $RT_{start}(x)$, as the time taken to read a word in sentence x, averaged across all the words in x and across all participants, when x was one of the first 24 sentences presented in the experiment (i.e. in the first block of the experiment).[8] Similarly, we define $RT_{end}(x)$ as the average RT on x when x was one of the last 24 sentences presented in the experiment (i.e. in the last block of the experiment). We then define $\Delta RT(x)$, the rate of task adaptation measured on x, as follows:

$$\Delta RT(x) = RT_{start}(x) - RT_{end}(x)$$

If task adaptation is start-point dependent, then for two sentences x and y where $RT_{start}(x) > RT_{start}(y)$ (i.e., x is more difficult than y), we would expect $\Delta RT(x) > \Delta RT(y)$.

To estimate ΔRT for all sentences, we first randomly split our participants into two halves. We used the first half of the participants (the *Difficulty Estimation Group*) to bin sentences according to their difficulty. Then, using the second half of the participants (*Task Adaptation Estimation Group*), we measured the rate of task

[7] Given the data, it is unlikely that task adaptation is characterized by diverging start-point dependent task-adaptation because these functions predict an *increase* in garden path effect over time, whereas we observed a decrease.

[8] Our definition of difficulty is empirical and is agnostic to *why* a particular sentence is difficult a priori. In future work, alternative definitions could categorize sentences based on factors such as word length or frequency, syntactic complexity, and so on.

adaptation by comparing the RTs at the start and end of the experiment for the sentences included in each bin. We used two sets of participants in this manner to avoid a circular analysis where the process of grouping sentences by their difficulty biases our estimates of task adaptation.

The analysis proceeded as follows. Using the RTs for the participants in the Difficulty Estimation Group, we computed RT_{start} for each filler sentence. Then, we binned these sentences into quartiles based on their RT_{start} values only (without taking into account their RT_{end}): for example, the first quartile consisted of the 25% of the sentences that were read most rapidly in Block 1 by the participants in the Difficulty Estimation Group, and the fourth quartile consisted of the 25% of sentences that were read most slowly in Block 1. We repeated this process separately for RRC, URC and filler sentences. Then, using the RTs from the other half of participants—the Task Adaptation Estimation Group—we computed the mean RT_{start} and RT_{end} for each quartile and for each of the three types of sentences by averaging the RTs for all words in all of the sentences included in that quartile. We repeated this process for 1000 random splits of participants, and averaged our RT_{start} and RT_{end} estimates across these random splits.

The results of this analysis indicate that in our data task adaptation was indeed start-point dependent (Figure 3): ΔRT was greater for sentences that were read more slowly when presented early in the experiment than for sentences that were read more rapidly. Difficulty was generally consistent across the Difficulty Estimation Group and the Task Adaptation Estimation Group. This pattern held for filler sentences as well as for RRC and URC sentences. Crucially, on average, ΔRT was greater for RRC sentences then URC sentences; this leads to a decrease in the difference in RTs between RRC sentences and URC sentences over the course of the experiment. In other words, at least some of the decrease in garden path effect over time observed in Experiment 1 can be accounted for by start-point dependent task adaptation.[9]

[9] We observed qualitatively similar results when we repeated the analysis with log transformed RTs. This analysis can be found on OSF.

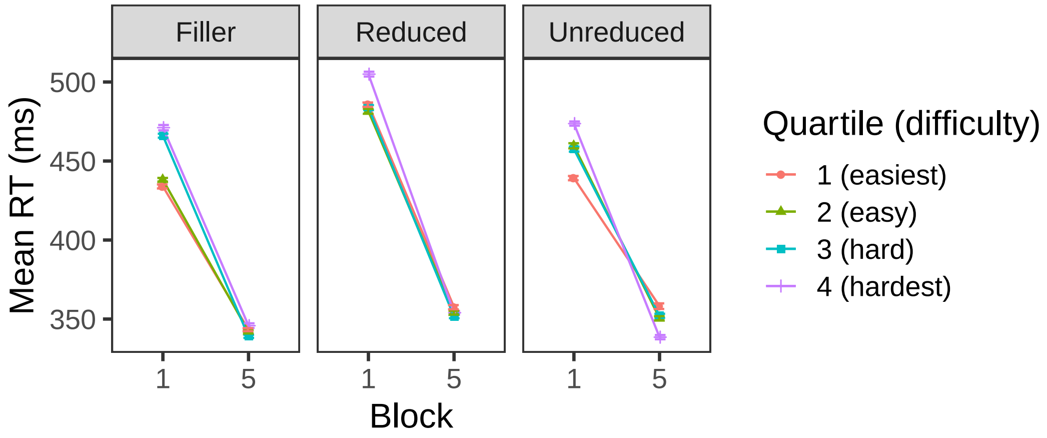


Figure 3. Task adaptation in Experiment 1. We plot RTs for participants in the Task Adaptation Estimation Group averaged across all words in the sentence for all sentences in Block 1 and Block 5. Sentences are binned into quartiles based on the RTs in Block 1 for participants in the Difficulty Estimation Group (binning was performed separately for each of the three classes of sentences). The estimates are averaged across 1000 random splits of participants. Error bars reflect two standard errors above and below the mean.

Discussion

Experiment 1 had two goals. The first was to replicate the decrease in garden path effect observed in previous studies. The second goal was to determine whether task adaptation is start-point dependent: if it is, then it could account at least in part for any decrease in garden path effect. We replicated in both direction and magnitude the decrease over time in garden path effect that was reported by FJ16; the coefficient of the interaction between sentence type and critical item number was -0.001 in both cases. This increases our confidence in the robustness of FJ16's empirical finding. At the same time, we also found that the decrease in RT measured for a particular sentence—whether it was an RRC, URC or filler sentence—depended on its "difficulty", or the time participants took on average to read that sentence when they encountered it early in the experiment. This suggests that at least a part of the observed decrease in garden path effect was driven by start-point dependent task-adaptation. In any study

Group	**Exposure phase**	**Test phase**
RRC-exposed	16 RRC, 16 Fillers	12 RRC, 12 URC, 24 Fillers
Filler-exposed	32 Fillers	12 RRC, 12 URC, 24 Fillers

Table 1

Design of Experiment 2. Experiment 2a only included a Filler-exposed group, whereas Experiment 2b included both groups.

whose goal is to measure syntactic adaptation, then, it is essential to demonstrate that exposure to a certain syntactic structure results in a decrease in garden path effect *over and above* the decrease caused by task adaptation alone. The following section describes experiments motivated by this goal.

Overview of Experiments 2a and 2b

As discussed earlier, the syntactic adaptation account predicts that participants exposed to reduced relative clauses early in the experiment will be less surprised when reading these structures later on in the experiment, and will consequently display a reduced garden path effect compared to participants who are not exposed to sentences without such relative clauses early in the experiment. We test this prediction using a between-subject design with two phases, an exposure phase and a test phase (the division between the two phrases was not indicated to participants). In the exposure phase, participants in the RRC-exposed group read both RRC and filler sentences, whereas participants in the Filler-exposed group read only filler sentences. In the test phase, both groups of participants read RRC sentences, URC sentences, and filler sentences. This design is summarized in Table 1.

We ran two experiments using this design. In Experiment 2a, we collected data from 81 participants, all of which were assigned to the Filler-exposed group. We used this smaller preliminary experiment to obtain an estimate of the garden path effect that arises in a setting where only task adaptation is possible. We then used the results of

Experiment 2a as a basis for simulations whose goal was to predict the garden path effect for the RRC-exposed group, where both task adaptation and syntactic adaptation are at least in principle possible. Based on these estimates, we conducted power simulations whose goal was to estimate the number of participants required to reliably detect a between-group difference in the garden path effect; we then ran that number of participants in Experiment 2b.

Experiment 2a: What is the garden path effect for Filler-exposed participants?

Methods

Participants. We recruited 81 participants from Amazon's Mechanical Turk (one participant recruited unintentionally). This number was nearly identical to the number of participants recruited in FJ16 and in Experiment 1 (80). To limit the number of non-native speakers, participants were only recruited if the home address associated with their Amazon account was located in the United States. We based the compensation for our participants on a \$8/hour rate (which was 75 cents above the US minimum wage at the time the experiment was run). Since the average duration of the experiment was approximately 15 minutes, participants received \$2 for their time.

Materials. Our materials were based on those of FJ16, with two modifications. First, we added the word *the* to the beginning of four of FJ16's original stimuli, to ensure consistency across all items. Second, we replaced 27 of FJ16's original sentences with new ones. We did so because some of FJ16's sentences had verbs with a transitivity bias—that is, verbs that typically occur with a noun phrase (NP) complement—which caused them to be effectively disambiguated before the start of the disambiguating region (cf. Malone & Mauner, 2018). The following sentence from F16's materials, for example, is in practice disambiguated in favor of the relative clause reading at the prepositional phrase (*in the alley*), rather than at the second verb (*ran*), as intended:

(4) The calico cat licked in the alley ran into the street.

After the preposition phrase *in the alley* is encountered, a main verb reading can only be maintained under a heavy NP shift parse (e.g., *the cat licked in the alley the toy*). Since heavy NP shifts are relatively infrequent, the relative clause reading becomes highly probable even before the disambiguating region. This is likely to diminish the garden path effect in the disambiguating region in such sentences, and, consequently, diminish the extent to which they will cause syntactic adaptation—and thereby our power to detect a syntactic adaptation effect. We replaced these items with sentences that included optionally reflexive verbs (5a), ditransitive verbs (5b), or optionally transitive verbs without a strong transitivity bias (5c), where transitivity bias was determined based on estimates from Roland and Jurafsky (2002):

(5) a. The bearded man shaved two weeks ago liked his stylish new look.
b. The helpful librarian lent the frayed book took good care of it.
c. The ferocious lions attacked during the day were unable to escape the hunters.

After both of these modifications, all the sentences had seven words before the disambiguating region: three words in the subject NP, one verb and three words in the NP or prepositional phrase following the verb. We also created 64 filler sentences with similar properties to those we used in Experiment 1: they did not contain any relative clauses, and the main verbs' past participle differed from their past tense form.

Design. Experiment 2a consisted of an exposure phase and a test phase. In the exposure phase, participants read 32 filler sentences. In the test phase, they were presented with 12 RRC sentences, 12 URC sentences and 24 filler sentences (see Table 1). We generated four pseudo-random orders and two lists from each order, counter-balanced for ambiguity in the test phase, as in Experiment 1.

Procedure. The same procedure was used as in Experiment 1.

Results

Data filtering and exclusion. We used the same filtering and exclusion criteria as in Experiment 1. We excluded one participant who reported that English was not their first language. We additionally excluded eight participants whose mean accuracy on filler sentences was lower than 80%. Finally, we excluded all observations (words) with reading times lower than 100 ms or greater than 2000 ms, leading to the exclusion of 0.36% of all observations of the participants who were not excluded.

Estimating the garden path effect in the test phase. For every participant and trial, we averaged the RTs on the words in the disambiguating region. We then used a linear mixed-effects model to predict the log of these averaged RTs from sentence type (coded as 1 for RRC sentences and -1 for URC sentences). As discussed in Analysis 1.2, we did not include word length as a predictor because the critical region contained the same words across the RRC and URC version of a given item. We used the maximal random effects structure: by-participant and by-item random intercepts and a by-participant random slope for sentence type.

This model revealed a significant garden path effect: the disambiguating region was read significantly more slowly in RRC sentences than in URC sentences ($\hat{\beta} = 0.015$, $SE = 0.006, p = 0.02$).

Power analysis for Experiment 2b

Before conducting Experiment 2b, which follows the between-group design described above, we conducted simulations to estimate the number of participants required to obtain at least 80% power in this paradigm. We expect to observe a greater garden path effect when only task adaptation is possible (in the Filler-exposed group) than when both task adaptation and syntactic adaptation are possible (for the RRC-exposed group). To estimate power, we need an hypothesis as to the relative magnitude of the garden path effect size for each group, or the value of Ω in $GPE_{RRC} = \Omega \cdot GPE_{Filler}$, where GPE_{RRC} denotes the garden path effect for the RRC-exposed group, GPE_{Filler} the garden path effect for the Filler-exposed group, and

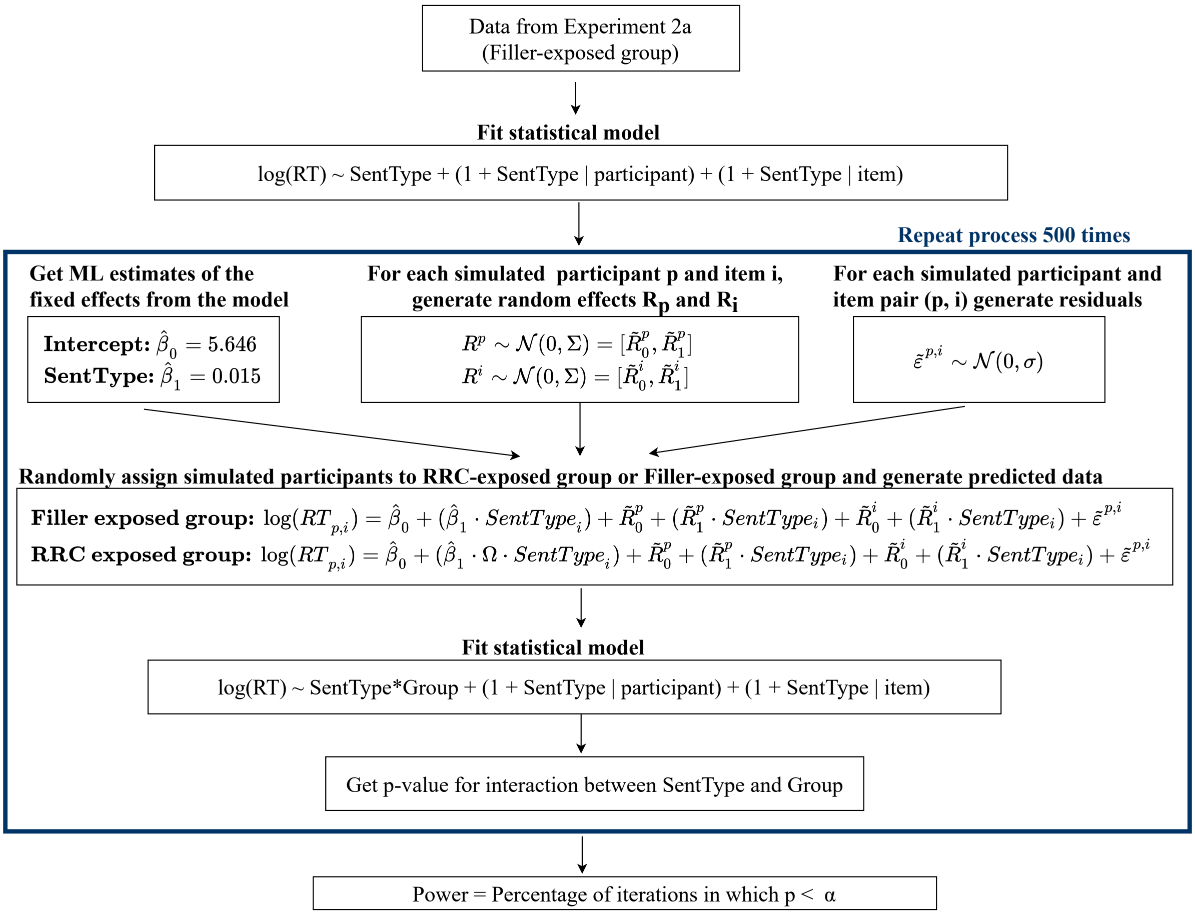

Data from Experiment 2a (Filler-exposed group)
Fit statistical model
log(RT) ~ SentType + (1 + SentType | participant) + (1 + SentType | item)
Repeat process 500 times
Get ML estimates of the fixed effects from the model
Intercept: $\hat{\beta}_0 = 5.646$
SentType: $\hat{\beta}_1 = 0.015$
For each simulated participant p and item i, generate random effects R_p and R_i
$R^p \sim \mathcal{N}(0, \Sigma) = [\tilde{R}^p_0, \tilde{R}^p_1]$
$R^i \sim \mathcal{N}(0, \Sigma) = [\tilde{R}^i_0, \tilde{R}^i_1]$
For each simulated participant and item pair (p, i) generate residuals
$\tilde{\varepsilon}^{p,i} \sim \mathcal{N}(0, \sigma)$
Randomly assign simulated participants to RRC-exposed group or Filler-exposed group and generate predicted data
Filler exposed group: $\log(RT_{p,i}) = \hat{\beta}_0 + (\hat{\beta}_1 \cdot SentType_i) + \tilde{R}^p_0 + (\tilde{R}^p_1 \cdot SentType_i) + \tilde{R}^i_0 + (\tilde{R}^i_1 \cdot SentType_i) + \tilde{\varepsilon}^{p,i}$
RRC exposed group: $\log(RT_{p,i}) = \hat{\beta}_0 + (\hat{\beta}_1 \cdot \Omega \cdot SentType_i) + \tilde{R}^p_0 + (\tilde{R}^p_1 \cdot SentType_i) + \tilde{R}^i_0 + (\tilde{R}^i_1 \cdot SentType_i) + \tilde{\varepsilon}^{p,i}$
Fit statistical model
log(RT) ~ SentType*Group + (1 + SentType | participant) + (1 + SentType | item)
Get p-value for interaction between SentType and Group
Power = Percentage of iterations in which p < α

Figure 4. A schematic of how we calculated the power to detect a significant difference in the garden path effect between the RRC-exposed group and the Filler-exposed group. We use the LMER notation in R for Model$_1$ and Model$_2$. The fixed effects for Model$_2$, ($\hat{\beta}_0$ and $\hat{\beta}_1$), were estimated from Experiment 2a, and correspond to the coefficients of the intercept and sentence type respectively. The by-participant and by-item random intercepts ($\tilde{R}^p_0$, $\tilde{R}^i_0$) and random slopes ($\tilde{R}^p_1$, $\tilde{R}^i_1$), were sampled from the multivariate normal distribution $\mathcal{N}(0, \Sigma)$ where Σ corresponds to the covariance matrix of Model$_1$. The residual error for each observation ($\tilde{\varepsilon}^{p,i}$) was sampled from the normal distribution $\mathcal{N}(0, \sigma)$, where σ corresponds to the residual standard deviation of Model$_1$.

Ω value	# Participants	$p < 0.05$	$p < 0.01$	$p < 0.001$
0.10	200	0.45	0.21	0.05
	400	0.76	0.55	0.22
	800	0.97	0.89	0.68
0.18	200	0.38	0.16	0.04
	400	0.68	0.42	0.14
	800	0.94	0.82	0.54
0.25	200	0.31	0.13	0.04
	400	0.60	0.34	0.09
	800	0.89	0.73	0.44
0.50	200	0.17	0.05	0.01
	400	0.29	0.10	0.01
	800	0.59	0.34	0.09

Table 2

Power to detect a significant difference in the garden path effect between a Filler-exposed group and an RRC-exposed group if the garden path effect of the RRC-exposed group was 0.18 *times that of the Filler-exposed group.*

$\Omega < 1$ is a constant proportion.

The simulations we report below are based on $\Omega = 0.18$; this value was derived from a simple Bayesian belief update model (Fine, Qian, Jaeger, & Jacobs, 2010). After running Experiment 2b, we discovered an error in the calculation; however, post-hoc power calculations with other values of Ω revealed that the estimates for the number of required participants did not change substantially for values up to $\Omega = 0.25$ (see Table 2).

To estimate the power of our paradigm to detect a between-group difference in the garden path effect with n participants and the number of items included in the experiment, we sampled participants and items from the empirical random effect distribution estimated in Experiment 2a. We then randomly assigned half of the simulated participants to the Filler-exposed group and the other half to the RRC-exposed group. For the Filler-exposed group, we generated predicted RTs for each trial by combining the fixed and random effects estimates from Experiment 2a with a sample from the same model's residual distribution. For the RRC-exposed group, we used a similar process but with one difference: we multiplied the coefficient of Sentence type (i.e., the garden path effect) by Ω.

With this simulated dataset in place, we then fit a linear mixed-effects model whose fixed effects included *Sentence type* (coded 1 for RRC sentences and -1 for URC sentences), *Group* (coded 1 for the RRC-exposed group and -1 for the Filler-exposed group), and the interaction between these two predictors. The random effects included intercepts for participants and items, along with a by-item and by-participant slope for Sentence Type. The random effect structure was not maximal because it was not possible to include a by-item slope for group: since Experiment 2a did not include RRC-exposed participants, we could not estimate the by-item variability in the difference between the two groups. Finally, we calculated the p value for the crucial interaction between Sentence Type and Group. For a diagram summarizing this procedure, see Figure 4.

We repeated the above process 500 times each for 200, 400 and 800 participants and for four different values of Ω: 0.10, 0.18, 0.25 and 0.50.[10] Table 2 summarizes the percentage of iterations in which the interaction between Sentence Type and Group was significant for each of the datasets at different p value thresholds for rejecting the null hypothesis (α levels). Our power simulations indicate that for values of Ω up to 0.25

[10] A reviewer points out that 500 iterations for each combination of Ω and n are insufficient to obtain precise estimates—assuming a binomial distribution for the power estimates, with 500 iterations, it is not unlikely that our power estimates differed from the true value by up to 10 percentage points (i.e., if our estimate was 0.8, then it the true power likely lies between 0.9 and 0.7). The lack of precision does not change our conclusions, since even if the true power with 800 participants were 10 percentage points lower, the power would still be greater than 0.8; however, we recommend that in future work a larger number of iterations is used.

(i.e, if the garden path effect of the RRC-exposed group is predicted, under the syntactic adaptation hypothesis, to be a quarter of that of the Filler-exposed group) the power to detect a significant interaction was greater than 0.9 with 800 participants. One striking finding is that at $\alpha = 0.05$, the power to detect a significant interaction was much lower than 0.8 even with 200 participants—far more than typically participate in self-paced reading experiments.

Experiment 2b: Is the garden path effect for the Filler-exposed group greater than that for the RRC-exposed group?

Methods

Participants. We recruited participants on Amazon's Mechanical Turk using Microbatcher (Leonard, 2019). We planned to include in the experiment 800 participants, but ended up recruiting a slightly larger number (828). Only participants whose home address was located in the United States were recruited. Participants received $2 for their time.

Materials and Design. We used the same materials as in Experiment 2a. Filler-exposed participants were randomly assigned to one of the eight lists generated from the four pseudo-random orders used in Experiment 2a. We created eight additional lists for the RRC-exposed group by replacing 16 of the fillers from the exposure phase with RRC sentences. RRC-exposed participants were randomly assigned to one of the latter eight lists.

Procedure. The procedure was identical to Experiments 1 and 2a.

Results

Data filtering and exclusion. We used the same data filtering and exclusion criteria as in Experiment 2a. This led to the exclusion of 11 participants who reported that English was not their first language and 175 participants whose accuracy on filler sentences was lower than 80%. The high proportion of participants with low filler accuracy in comparison to Experiment 2a cannot be attributed to question difficulty: in

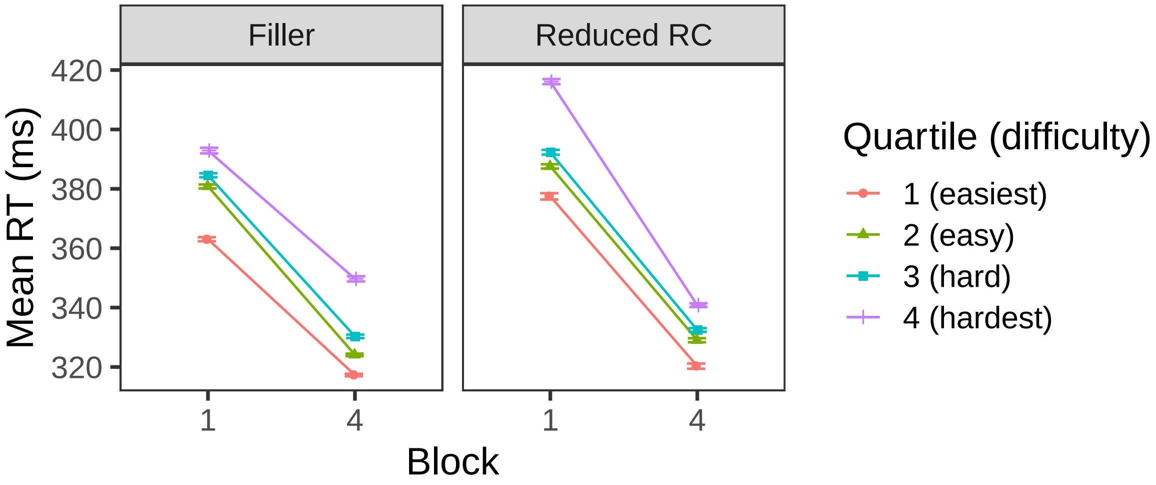


Figure 5. RTs for participants in the Task Adaptation Estimation Group averaged across all words in the sentence for all sentences in Block 1 and Block 5. Sentences (both critical items and filler sentences) are grouped into quartiles based on the RTs in Block 1 for participants in the Difficulty Estimation Group. Estimates are averaged across 1000 random splits of participants, and error bars reflect two standard errors above and below the mean.

both experiments, Filler-exposed participants were presented with the same fillers, yet the proportion of participants with low filler accuracy differed drastically between the two experiments (10% in Experiment 2a and 21% in Experiment 2b). Additionally, even though the RRC-exposed group was presented with just a subset of the fillers presented to Filler-exposed group, the number of participants whose accuracy was low did not differ between the groups (87 in the Filler-exposed group and 88 in the RRC-exposed group), further suggesting that the difference in accuracy was not driven by the presence or absence of specific items. It is possible that the larger sample size of Experiment 2b led to the recruitment of less attentive participants or even bots.

As in the previous experiments, we also excluded observations (words) with RTs less than 100 ms or greater than 2000 ms. This led to the exclusion of 0.48% of all observations for the remaining 642 participants.

Is the rate of task adaptation higher for more difficult items? We used the same method to diagnose start-point dependent task adaptation as in

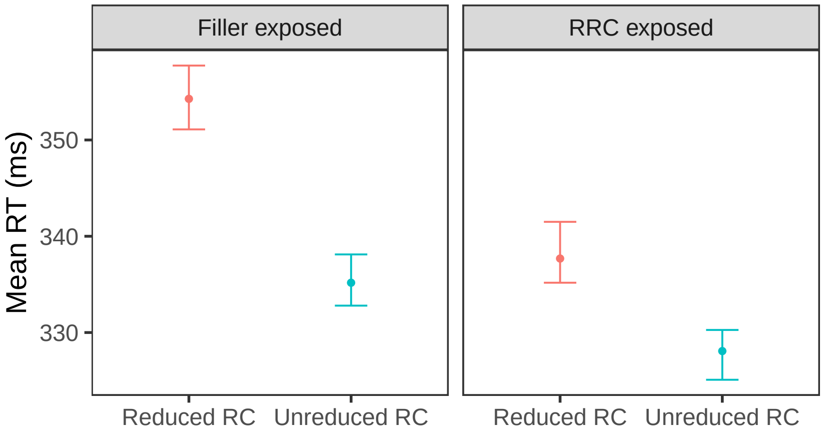


Figure 6. Garden path effect in the test phase for the Filler-exposed group and RRC-exposed group. Error bars reflect bootstrapped 95% confidence intervals.

Experiment 1. We sampled half of the participants, and we divided both RRC and filler sentences into quartiles based on their RTs in this group of participants prior to task adaptation (that is, early in the experiment). Then, using the remaining participants, we estimated the rate of task adaptation for each quartile by comparing the mean RTs, averaged across all sentences in the quartile, before and after task adaptation. We repeated this process for 1000 random splits of participants. As in Experiment 1, in almost all quartiles and types of sentences, sentences that were read more slowly when presented early in the experiment showed a greater task adaptation effect (ΔRT) than sentences that were read more rapidly early in the experiment. This supports the hypothesis that task adaptation is start-point dependent (see Figure 5).[11] As discussed earlier, we expect the rate of start-point dependent task adaptation to be similar across RRC-exposed and Filler-exposed participants. As such, a difference between groups in garden path effect in the test phase can only be attributed to syntactic adaptation.

[11] The only exception were the filler sentences that were read the most slowly (i.e., in the fourth quartile). For these sentences, ΔRT was smaller than for other filler sentences that were read more rapidly. We find qualitatively similar results when we repeat this analysis with log transformed RTs, with the exception of the RRC sentences that were read most rapidly, where ΔRT was larger than for other RRC sentences that were read more slowly. This analysis can be found on OSF.

Is there evidence for syntactic adaptation over and above task adaptation? As in Experiment 2a, we averaged the RTs in the disambiguating region and log-transformed these averaged RTs. We then fit a linear mixed-effects model with the predictors we used in our power simulations. The fixed effects included Sentence Type, Group and the interaction between the two; and the random effects included random intercepts for participants and items, along with a by-participant slope for sentence type and by-item slope for sentence type, group and the interaction between the two.

The model revealed a significant garden path main effect: the words in the disambiguating region were read more slowly in RRC sentences than in URC sentences ($\hat{\beta} = 0.016$, $SE = 0.002, p \ll 0.001$). There was also a main effect of group: Filler-exposed participants read sentences significantly more slowly on average than RRC-exposed participants ($\hat{\beta} = 0.038$, $SE = 0.010$, $p < 0.001$). We briefly discuss this effect, which is not predicted by the syntactic adaptation hypothesis, in the discussion section. Finally, the crucial interaction was significant: the garden path effect was greater for the Filler-exposed group than for the RRC-exposed group ($\hat{\beta} = 0.006$, $SE = 0.002, p = 0.001$), providing evidence for syntactic adaptation over and above task adaptation (see Figure 6).

As was pointed out by a reviewer, by fitting a linear mixed-effects model to log transformed RTs, we made the (standard) assumption that RTs are lognormally distributed, and therefore assumed that the lowest possible RT was 0 ms. This assumption is physiologically implausible: RTs are constrained by factors such as the speed of muscle movements and cannot in practice be as low as 0 or 1 ms. To address this issue, we reanalyzed the data using Bayesian mixed-effects models based on the assumption that RTs follow a shifted log normal distribution (Rouder, 2005), a generalized form of the lognormal distribution with a shift parameter which determines the lowest possible RT value that the model can predict (i.e. the floor). The fixed effect and random effect structure of the shifted model was identical to the unshifted model described above. We allowed the shift parameter of the lognormal distribution to vary across participants. We used weakly informative priors, as recommended by Schad, Betancourt, and Vasishth (2019). These priors expressed the assumptions that RTs are very likely to lie between 100 to 2000 ms, and that the difference in RTs between RRC and URC sentences was likely to lie between -100 and 100 ms, as was the difference in garden path effect between the RRC-exposed and Filler-exposed groups.[12]

The shifted model revealed qualitatively similar effects to the unshifted model, although all of the fixed effects were larger and there was more uncertainty about the estimates: a garden path main effect ($\hat{\beta} = 0.033$, $SE = 0.006$), a main effect of group ($\hat{\beta} = 0.062$, $SE = 0.018$), and an interaction between group and garden path effect ($\hat{\beta} = 0.009$, $SE = 0.004$).

Discussion

As in Experiment 1, we found that the effect of task adaptation was start-point dependent—the rate of decrease in RTs was higher in sentences that were read slowly when presented early in the experiment than sentences that were read rapidly. This supports the hypothesis that task adaptation causes a decrease in the garden path effect over time. At the same time, we also found evidence for a decrease in garden path effect over and above the decrease caused by task adaptation—the garden path effect was greater in participants who were only exposed to filler sentences in the exposure phase than in those who were exposed to 16 RRC sentences. This lends support to the syntactic adaptation hypothesis. However, as we discuss below, this effect is relatively small; this fact, in conjunction with design decisions that could have led to reduced power, may explain the recent failure of Stack et al. (2018) to observe a syntactic adaptation effect.

We also found that Filler-exposed participants read sentences significantly more slowly on average than participants in the RRC-exposed groups (see Figure 6). A similar main effect of group, which is not predicted by the syntactic adaptation account, was observed by both Fine et al. (2013) and Stack et al. (2018). One possible

[12] Further details about the priors can be found on OSF.

explanation for this finding is that extensive exposure to syntactically simple filler sentences, followed by a sudden transition to syntactically challenging RRC sentences in the test phase, causes Filler-exposed participants to slow down and read all test sentences more carefully. Future work can test this hypothesis by determining whether this pattern persists when the Filler-exposed group is exposed to sentences that include temporary syntactic ambiguities other than that used to measure the garden path effect, for example the direct object / sentential complement (NP/S) ambiguity if the target ambiguity is main verb / reduced relative as in the present study.

Exploratory analyses. We now turn to exploratory analyses that further investigate the viability of self-paced reading as a paradigm for studying syntactic adaptation. We estimate the number of participants required for future experiments using this paradigm and compare the magnitude of task adaptation and syntactic adaptation.

How many participants should be recruited for future experiments with the same design? This section reports the results of simulations whose goal was to estimate the power to detect a between-group difference in the garden path effect in future experiments with the same design as Experiment 2b. This approach was similar to the power analysis we conducted using the data from Experiment 2a, with two crucial differences. First, in Experiment 2a, we fit a linear mixed-effects model and calculated the power based on the maximum likelihood estimates of all the parameters. In this analysis, by contrast, we fit a Bayesian version of the linear mixed-effects model and calculated the power based not only on the posterior mean estimates of all parameters, but also several other values of the parameters that have a range of posterior probabilities given the results of Experiment 2b. Second, in Experiment 2a we collected data from only the Filler-exposed group, we used Ω—the hypothesized ratio between the garden path effects shown by the two groups—to generate predictions for the RRC-exposed group. This hypothesized ratio was not required in the present simulations, since Experiment 2b included empirical data collected from the RRC-exposed group.

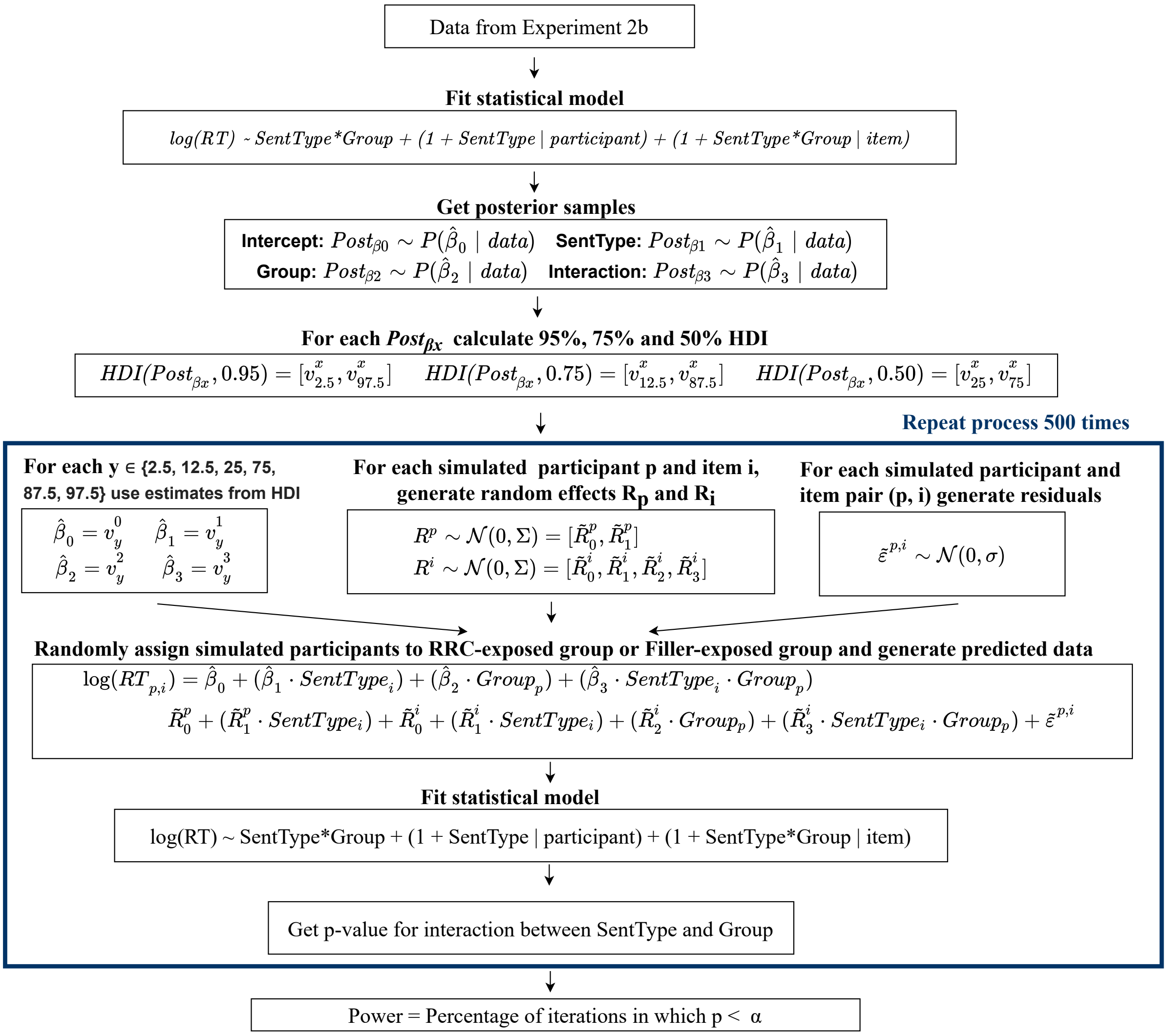

Data from Experiment 2b
Fit statistical model
*log(RT) ~ SentType*Group + (1 + SentType | participant) + (1 + SentType*Group | item)*
Get posterior samples
Intercept: $Post_{\beta 0} \sim P(\hat{\beta}_0 \mid data)$ SentType: $Post_{\beta 1} \sim P(\hat{\beta}_1 \mid data)$
Group: $Post_{\beta 2} \sim P(\hat{\beta}_2 \mid data)$ Interaction: $Post_{\beta 3} \sim P(\hat{\beta}_3 \mid data)$
For each $Post_{\beta x}$ calculate 95%, 75% and 50% HDI
$HDI(Post_{\beta x}, 0.95) = [v^x_{2.5}, v^x_{97.5}]$ $HDI(Post_{\beta x}, 0.75) = [v^x_{12.5}, v^x_{87.5}]$ $HDI(Post_{\beta x}, 0.50) = [v^x_{25}, v^x_{75}]$
Repeat process 500 times
For each y ∈ {2.5, 12.5, 25, 75, 87.5, 97.5} use estimates from HDI
$\hat{\beta}_0 = v^0_y$ $\hat{\beta}_1 = v^1_y$ $\hat{\beta}_2 = v^2_y$ $\hat{\beta}_3 = v^3_y$
For each simulated participant p and item i, generate random effects R_p and R_i
$R^p \sim \mathcal{N}(0, \Sigma) = [\tilde{R}^p_0, \tilde{R}^p_1]$
$R^i \sim \mathcal{N}(0, \Sigma) = [\tilde{R}^i_0, \tilde{R}^i_1, \tilde{R}^i_2, \tilde{R}^i_3]$
For each simulated participant and item pair (p, i) generate residuals
$\tilde{\varepsilon}^{p,i} \sim \mathcal{N}(0, \sigma)$
Randomly assign simulated participants to RRC-exposed group or Filler-exposed group and generate predicted data
$\log(RT_{p,i}) = \hat{\beta}_0 + (\hat{\beta}_1 \cdot SentType_i) + (\hat{\beta}_2 \cdot Group_p) + (\hat{\beta}_3 \cdot SentType_i \cdot Group_p)$
$\tilde{R}^p_0 + (\tilde{R}^p_1 \cdot SentType_i) + \tilde{R}^i_0 + (\tilde{R}^i_1 \cdot SentType_i) + (\tilde{R}^i_2 \cdot Group_p) + (\tilde{R}^i_3 \cdot SentType_i \cdot Group_p) + \tilde{\varepsilon}^{p,i}$
Fit statistical model
log(RT) ~ SentType*Group + (1 + SentType | participant) + (1 + SentType*Group | item)
Get p-value for interaction between SentType and Group
Power = Percentage of iterations in which p < α

Figure 7. A schematic of how we calculated the power to detect a significant difference in the garden path effect between the RRC-exposed group and the Filler-exposed group for future experiments with the same design. We use the LMER notation in R for the statistical models.

We simulated participants and items using the random effects estimated from the model fit to the results of Experiment 2b. This simulation process was identical to the prior power analysis. Then, for any given set of values of the fixed effects—the intercept (β_0), the main effect of sentence type (β_1), the main effect of group (β_2), and the interaction between these two predictors (β_3)—we generated 500 simulated RT datasets by combining the values of these fixed effects with samples from the random effects and residuals. Finally, we fit to each of these 500 datasets a new model similar to one we used to analyze the results of Experiment 2b, and calculated the proportion of simulated datasets in which β_3, the crucial interaction term, reached significance. We repeated this process separately for 200, 400 and 800 participants.

We calculated different sets of values for the fixed effects as follows. First, we fit a Bayesian version of the statistical model used in Experiment 2b. Then, we computed the highest density interval (HDI) for β_0, β_1, β_2 and β_3. An $x\%$ HDI specifies a range of values (a, b) such that $x\%$ of the posterior probability mass falls within this range. For example, if the 95% HDI for β_1 is $(0.001, 0.01)$, then $P_{posterior}(0.001 < \beta_1 < 0.01) = 0.95$. We computed the 95%, 75% and 50% HDIs for each of the predictors and used the lower and upper bounds of these intervals as six sets of values of the fixed effects for the power analysis. For each of these six sets of values, we generated 500 datasets and calculated power as described in the previous paragraph. We also calculated power for the set of values with the posterior mean.

The Bayesian regression model we used for the power analysis differed in two ways from the shifted lognormal Bayesian regression model described above: first, we used the standard unshifted lognormal distribution and second, we used the default priors specified by the brms package (Bürkner et al., 2017): for the fixed effects, a uniform distribution over all real numbers; for the intercept, a Student's t distribution with mean 0, standard deviation 10, and 6 degrees of freedom; for the by-participant and by-item random slopes and intercepts, as well as the parameter for the residual standard deviation, a Student's t distribution with mean 0, standard deviation 10, and 3 degrees of freedom; and for the covariance matrices, LKJ Cholesky priors with $\eta = 1$.

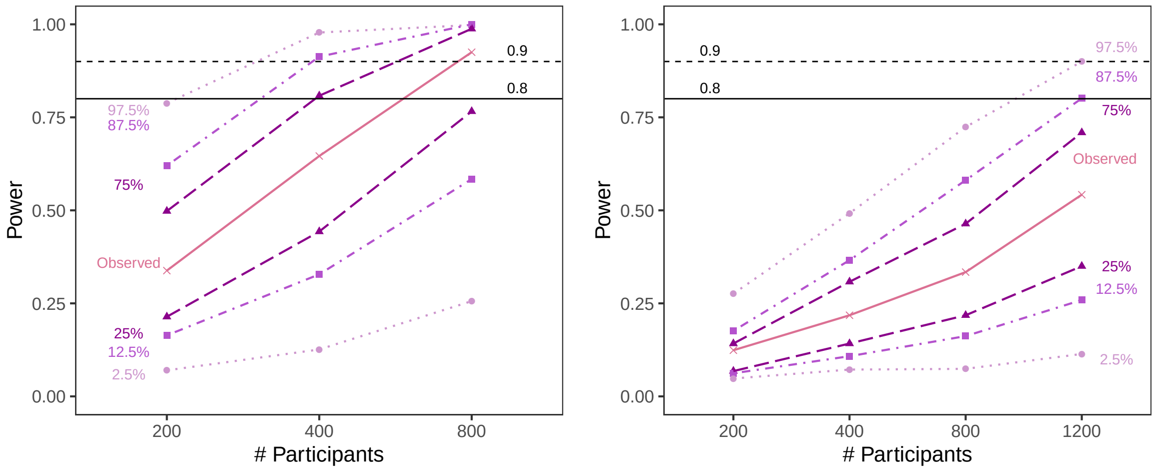
In light of the similarity between the results we obtained from the shifted distribution with informative priors and the current unshifted distribution with uninformative priors, we did not repeat our power analyses with the estimates from the shifted model.

Results. Our power analyses indicated that if the true effect size of syntactic adaptation is the same as that observed in Experiment 2b (the posterior mean estimate), then future experiments with the design of Experiment 2b will require between 400 and 800 participants to detect a significant interaction at the $p < 0.05$ threshold with 80% power (see Figure 8a). If the true effect size is the highest value included in 95% HDI—1.7 times the observed effect size—then 400 participants might be sufficient to detect a significant interaction. On the other hand, if the true effect size is on the lower end of the 95% HDI—0.3 times the observed effect size—then even 800 participants might not be enough.[13]

How many participants would we need to detect modulations of syntactic adaptation? The goal of Experiment 2b was to detect the *presence* of syntactic adaptation. As such, we optimized the design of that experiment to obtain the maximal possible difference in garden path effect between the two groups: in the exposure phase, Filler-exposed participants read sentences that had minimal to no structural overlap with the RRC sentences included in the test phase, whereas RRC-exposed participants were exposed to sentences that had maximal structural overlap with the test sentences.

By contrast, any between-group self-paced reading experiment designed to detect *modulations* of this basic syntactic adaptation effect would likely yield smaller between-group differences than we found in Experiment 2b. Consider, for example, an experiment designed to test whether the garden path effect associated with RRCs can be diminished by repeated exposure to another type of relative clause, such as an unreduced relative clause (URC), and if so, whether the degree of adaptation differs across the two scenarios (RRC in both exposure and test, compared to URC in exposure and RRC in test). Such a hypothetical experiment would include

[13] The posterior mean estimate of the interaction coefficient was 0.006 and the HDI was 0.002–0.010.



(a) (b)

Figure 8. (a) Power to detect a significant interaction between group and sentence type for future studies with the same expected effect size as in Experiment 2b. (b) Power to detect a significant interaction between group and sentence type for future studies with an expected effect size of half of what was observed in Experiment 2b. Lines of the same colour and line type correspond to upper and lower bound of HDI with the same credible interval. For example, the dotted line in lightest purple reflects the upper and lower bound for the 95% HDI.

RRC-exposed, URC-exposed and Filler-exposed groups. Any difference between RRC-exposed and URC-exposed participants is very likely be smaller than the difference between RRC-exposed and Filler-exposed groups; consequently, detecting such a modulation of syntactic adaptation would require even more participants than needed to detect its presence, as in Experiment 2b.

To estimate the power of experiments measuring such modulations of syntactic adaptation, we re-ran all the power analyses after dividing by two the upper bound and lower bound values of β_0, β_1, β_2 and β_3 described above; this expresses the assumption that modulations of the basic syntactic adaptation effect will yield smaller effect sizes than in our Experiment 2b.[14] Under these assumptions, the power analysis based on the

[14] Since we sampled the random effects from the original multivariate normal distributions, dividing the beta coefficients of the lower and upper bounds does not result in a decrease in the uncertainty of

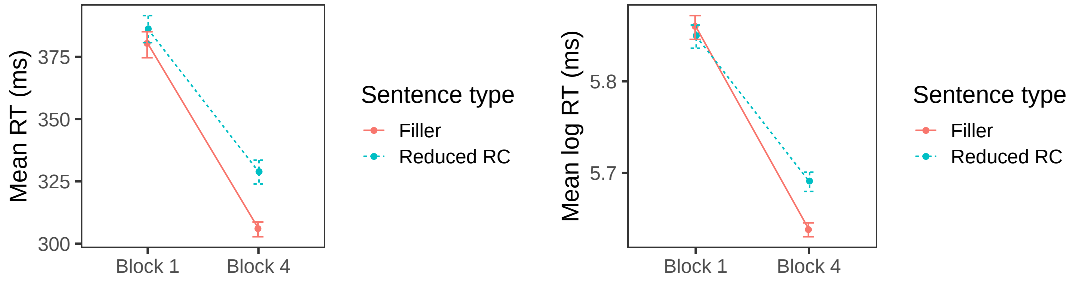
posterior mean estimates indicated that even with 1200 participants the experiment would have only 60% power to detect a significant interaction effect at the $p < 0.05$ threshold (see Figure 8b). In the best case scenario, where the modulation effect size is based on the largest possible effect size contained in the 95% HDI from Experiment 2b, we would have 72% power to detect an interaction at the $p < 0.05$ threshold with 800 participants, and 90% power with 1200 participants. In the worst case scenario, when the effect size is based on the smallest possible effect size within the same 95% HDI, we would have 7% power to detect a significant interaction with 800 participants and 11% power with 1200 participants. In other words, experiments designed to detect modulations of the syntactic adaptation effect using a between-group design could be underpowered even with as many 1200 participants.

Comparing the magnitude of task adaptation and syntactic adaptation. The reduction in the size of garden path effect is caused by task adaptation alone in the Filler-exposed group, and by both task adaptation and syntactic adaptation in the RRC-exposed group. As such, the difference in garden path effect between the two groups can be interpreted as an estimate of the effect of syntactic adaptation over and above task adaptation. In Experiment 2b, the garden path effect was 14.07 ms for the Filler-exposed group and 5.67 ms for the RRC-exposed group, as calculated from the mixed effect model estimates. This suggests that syntactic adaptation resulted in 8.4 ms decrease in the garden path effect over and above task adaptation.

This estimate has a critical limitation: it compares across two sets of participants that differ in their average reading times (see discussion of main effect of group above). To obtain an estimate of the relative magnitude of syntactic and task adaptation within participants, we focused on the RRC-exposed group, and compared the change in RTs over time between RRC sentences and filler sentences: The decrease in RTs for filler sentences is caused by task adaptation, whereas the decrease in RTs for RRC sentences is caused by a combination of task and syntactic adaptation. Therefore, if we assume that the effects of syntactic adaptation and task adaptation are additive, then we can

our estimates.

calculate the within-participant magnitude of syntactic adaptation by subtracting the decrease in RTs observed in RRC sentences from that observed in filler sentences.



(a) (b)

Figure 9. RTs (panel a) and log RTs (panel b) averaged across sentence positions 8–10 for the RRC-exposed group in Block 1 and Block 4 for filler sentences and RRC sentences matched for RTs in Block 1. The mean RTs for all of the items in Block 1 were not greater or less than the mean RTs for all filler sentences across participants in both groups by more than 30 ms. Error bars reflect bootstrapped 95% confidence intervals.

This within-participant comparison is again complicated by a main effect, this time the main effect of condition: because filler sentences were on average read more rapidly than RRC sentences, and because task adaptation is start-point dependent, we could not directly compare the rate of task adaptation for the RRC and filler sentences. To mitigate this, we created a subset of RRC and filler sentences that were roughly matched in difficulty: we only included a sentence if its mean RT, when averaged across all participants who read the sentence as one of the first 20 sentences, was in the range defined by the mean RT for all filler sentences in the first block ±30 ms (350.9–410.9 ms). We focused on the words in positions 8–10 of both filler and RRC sentences; in RRC sentences, these are the words that make up the disambiguating region. We then averaged the RTs on these words across all the items in the subset and across all participants in the RRC-exposed group, separately when the items occurred early in the experiment (first 20 sentences) and when they occurred later in the experiment (last 40 sentences).

If the effects of syntactic adaptation and task adaptation are additive, such that syntactic adaptation results in a decrease in RTs over and above task adaptation, then we would expect a *greater* reduction in RTs for RRC sentences than for filler sentences. Contrary to this prediction, we found that RTs decreased *less* for RRC sentences (57 ms) than for filler sentences RTs (74 ms; see Figure 9a). We repeated this analysis with log transformed RTs and observed qualitatively similar results (see Figure 9b). These surprising results suggest that on both the raw and logarithmic scale, the rate of task adaptation is lower for syntactically complex sentences than syntactically easier sentences, even when the RTs for the complex and simple sentences are matched. This poses a problem for the simplistic notion of task adaptation that we (and others) have adopted, which assumes that the effects of task adaptation and syntactic adaptation are additive and independent of each other.

General Discussion

The garden path effect observed in temporarily ambiguous sentences that are disambiguated in favor of a low-probability parse decreases over the course of a reading experiment (Fine & Jaeger, 2016; Fine et al., 2013). This finding has been interpreted as evidence that participants update their syntactic expectations to match the statistics of the environment (*syntactic adaptation*). But syntactic adaptation is not the only possible explanation for this finding: a decrease over time in the garden path effect can also be driven by the hypothesis we termed “start-point dependent task adaptation”, according to which task adaptation—the decrease in RTs due to increased familiarity with the task—is greater for sentences that are read slowly when encountered early in the experiment (“difficult sentences”) than for sentences that are initially read more rapidly (”easy sentences”). Such start-point dependent task adaptation would result in a decrease over time in the difference in reading times between easier unambiguous sentences and difficult ambiguous sentences—in other words, the garden path effect. The goal of this paper was to investigate whether syntactic adaptation results in a decrease in garden path effect over and above the decrease caused by any such start-point dependent task-adaptation.

In Experiment 1, we replicated the results of one of the experiments from Fine and Jaeger (2016) that have been taken as evidence for syntactic adaptation: as in their experiment, both overall reading times and the garden path effect decreased over the course of Experiment 1. We also found evidence for start-point dependent task-adaptation, suggesting that the observed decrease in garden path effect could, in theory, be entirely driven by a greater rate of task adaptation for ambiguous sentences with reduced RCs (RRC sentences) than unambiguous ones with unreduced RCs (URC sentences).

The main experiment of the paper was Experiment 2b, whose goal was to detect syntactic adaptation over and above task adaptation. This experiment compared the garden path effects in two groups of participants: one exposed to filler sentences only (Filler-exposed group), and the other exposed to both filler and RRC sentences (RRC-exposed group). Following the exposure phase, both groups read RRC and URC sentences. In the Filler-exposed group, only task adaptation was possible, whereas in the RRC-exposed group both task and syntactic adaptation were possible.

Before running Experiment 2b, we ran a preliminary experiment, Experiment 2a, in which we collected data from Filler-exposed participants only, and used it to estimate the number of participants to run in Experiment 2b. We estimated that the number of participants required to reliably detect a significant difference in garden path effect between the two groups can be as high as 800. Consequently, in Experiment 2b, we collected data from 828 participants, 642 of whom were included in the analyses.

Experiment 2b showed that after the exposure phase, the garden path effect for the RRC-exposed group was diminished compared to that of the Filler-exposed group. Since both groups were exposed to the same number of sentences during the exposure phase, the difference in garden path effect between the groups cannot be completely explained by task adaptation, and has to be driven by the difference in the types of sentences that the participants were exposed to (i.e. RRC sentences vs. filler sentences). As such, these results support the hypothesis that syntactic adaptation causes a decrease in the garden path effect over and above the decrease caused by task-adaptation.

We next conducted a Bayesian analysis to estimate the range of effect sizes that are plausible given our data, and used those to estimate the power required to detect an effect in future studies with the same experimental design as Experiment 2b. This power analysis indicated that if the true effect size is equal to the effect observed in our experiment, then future experiments would require between 400 and 800 participants to have 80% power to detect the difference in garden path effect between groups. If the true effect size is smaller than that observed in our experiment, but still within the 95% credible interval given our results, then future experiments with the same design are likely to be underpowered with even 800 participants. Finally, we estimated the power to detect an effect in future between-group studies with similar experimental setup as Experiment 2b aimed at investigating how syntactic adaptation interacts with other factors. Under the assumption that such subtler effects result in an effect size half as large as in Experiment 2b, we found that these experiments could be underpowered even with as many as 1200 participants.

Why are so many participants required to reliably detect effects of syntactic adaptation in self-paced reading?

We discuss two possible answers to this question: first, that a decrease in garden path effect in a self-paced reading experiment is not an ideal dependent measure if the goal is to detect syntactic adaptation; and second, that syntactic adaptation results in very small and hard-to-measure changes to readers' expectations, more generally.

Explanation 1: Decrease in garden path effect in self-paced reading is a dependent measure that is ill-suited for studying syntactic adaptation. It is possible that syntactic adaptation can, in principle, be reliably detected with fewer participants in a between-group design than our power analysis suggests, but that self-paced reading is not an ideal paradigm to do so. As discussed earlier, task adaptation in this paradigm is start-point dependent; this leads to a compression over time of the difference in RTs between "easy" and "difficult" sentences, independently of any syntactic properties of those sentences. This compression causes a reduction in garden path effect. The high rates of task adaptation in self-paced reading therefore lead to smaller garden-path effects overall in the later parts of the experiment. This in turn results in a smaller absolute between-group differences in garden path effect. Since smaller effect sizes are often accompanied by lower power, more participants are likely to be required to detect effects of syntactic adaptation.[15]

This explanation points to two alternative methods of measuring syntactic adaptation that might result in larger effects: first, using a dependent measure that is not confounded with task adaptation; second, using a paradigm where task adaptation is not start-point dependent. It is unclear whether the latter method is currently feasible, since we are unaware of paradigms where task adaptation has been demonstrated to be start-point independent. However, a reviewer pointed out that there is indeed a dependent measure of syntactic adaptation that is not confounded with task adaptation — an *increase* in the garden path effect for sentences disambiguated in favor of the main verb reading, as in (6):

(6) The evil genie served the golden figs before going into a trance.

Since task adaptation results in a *decrease* in garden path effect, it would be possible to circumvent the loss in power due to task-adaptation even in self-paced reading studies, if we used the increase in garden path effect as a dependent measure. A potential concern with using the *increase* in garden path effect as a dependent measure is that, under the expectation adaptation account, after n observations, there is a greater change in surprisal for unexpected structures (reduced RC reading) than for sentences

[15] In principle, it is possible for power to stay the same as the effect size decreases, if the variability in the data also decreases along with the effect size. To test this, we refit the statistical model from Analysis 1.1 separately on the first two and the last two blocks of Experiment 1. If the variability in the data decreased along with the effect size, we would expect both the estimate of garden path effect and the standard error in the last two blocks to be lower than in the first two. In contrast to this prediction, we found that while the estimate of garden path effect decreased (from 0.044 in the first block to 0.007 in the last block), the standard error of the estimates remained the same (0.007 in both blocks).

with expected structure (MV reading) (Jaeger et al., 2019). Therefore, detecting an increase in the garden path effect for sentences with a MV reading can be much more challenging than detecting a decrease in the garden path effect for sentences with reduced RC reading. Further simulations and experiments are required to investigate whether the advantage of using the increase in garden path effects as a dependent measure (it is not correlated with task-adaptation) outweighs the disadvantage (it is predicted to have a smaller effect size).

Explanation 2: Syntactic adaptation results in extremely small changes to our expectations. An alternative explanation, which is also consistent with our results, is that exposure to sentences with unexpected structures in the context of an experiment results in extremely small changes to our expectations. If that is the case, syntactic adaptation may be difficult to observe irrespective of the paradigm or dependent measure we use. If the true effect size of syntactic adaptation is indeed very small, then this raises a broader question: what constitutes a psychologically meaningful effect size? The answer to this question can vary depending on the goals of the research program. If the goal is to apply the findings from the syntactic adaptation literature in a practical context (e.g., in education), then extremely small effect sizes might not be meaningful. On the other hand, if the goal is to build a theory on the basis of syntactic adaptation, then extremely small effect sizes might be meaningful, but not practical to study. Finally, if the goal is to only use syntactic adaptation to verify one of the predictions of a larger theoretical framework, then extremely small effect sizes can be both meaningful and practical.

What properties of RRC sentences are participants adapting to?

Experiment 2b indicated that participants in the RRC-exposed group adapted to some property of the RRC sentences they were exposed to, but did not isolate the property (or properties) of the RRC sentences to which participants were adapting. Following previous papers on syntactic adaptation, we assumed that participants updated their expectations about an abstract grammar rule such as “the subject of the sentence is modified by a reduced relative clause". However, it is also possible that participants were adapting to an accidental property of RRC sentences included in the experiment, such as the fact that the seventh word of the sentence was always a verb; or that they were adapting their parsing strategies to the large number of temporarily ambiguous sentences included in the experiment, for example by maintaining a larger number of potential parses for each sentence (Jurafsky, 1996).

In future work, these possibilities can be distinguished by measuring the magnitude of syntactic adaptation for sentences with varying properties. For example, if syntactic adaptation is weaker when the verbs in the exposure sentence occur in varying positions than when they occur in the same position, we can conclude that participants were adapting to the position of the verb in the sentence. Similarly, if syntactic adaptation is stronger when the exposure phase contains other types of garden path sentences (e.g., *When Anna bathed the baby spit up*) than when it contains filler sentences only, we can conclude that participants were adapting to the prevalence of temporarily ambiguous sentences in the experiment. As discussed earlier, the power of such experiments, which are designed to measure modulations of the syntactic adaptation effect, is likely to be relatively low in self-paced reading studies with designs similar to Experiment 2b.

Conclusion

This study provided evidence for rapid syntactic adaptation in self-paced reading studies using a between-group experimental setup. At the same time, hundreds of participants were required to detect a syntactic adaptation over and above the substantially stronger effect of adaptation to the self-paced reading task. Power analyses indicated that experiments with a similar between-group design whose goal is to study factors that modulate this effect, such as the particular syntactic properties that participants are able to adapt to, will likely require even more participants. We conclude that theoretical questions about syntactic adaptation are likely to be more fruitfully addressed using experimental paradigms that are not confounded with task adaptation, or paradigms in which task adaptation is not start-point dependent (if such paradigms exist).

Acknowledgments

We would like to thank Nicholas Douglass and Brian Leonard for assistance with stimulus creation and data collection respectively. We would also like to thank Duane Watson and the two anonymous reviewers for their extensive constructive feedback which helped us significantly improve the paper. Finally, we would like to thank Brian Dillon, the members of the Computation and Psycholinguistics lab at Johns Hopkins University, and audiences at the 2019 and 2020 CUNY Conferences on Human Sentence Processing for their helpful discussion of earlier versions of this work.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought.* Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc.

Bürkner, P.-C., et al. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Clifton Jr, C., Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K., & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory and Language*, *49*(3), 317–334.

Drummond, A. (2016). *Ibex farm.* `https://github.com/addrummond/ibex`. GitHub.

Ehrlich, S., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655.

Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1362–1376. Retrieved from `http://dx.doi.org/10.1037/xlm0000236`

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

Fine, A. B., Qian, T., Jaeger, T. F., & Jacobs, R. A. (2010). Is there syntactic adaptation in language comprehension? In *Proceedings of the 2010 Workshop on Cognitive Modeling and Computational Linguistics* (pp. 18–26).

Hale, J. (2001). A Probabilistic Earley Parser As a Psycholinguistic Model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1–8). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `https://doi.org/10.3115/1073336.1073357` doi: 10.3115/1073336.1073357

Heathcote, A., Brown, S., & Mewhort, D. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207.

Jaeger, T., Bushong, W., & Burchill, Z. (2019). *Strong evidence for expectation adaptation during language understanding, not a replication failure. a reply to harrington stack, james, and watson (2018).*

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.

Kemper, S., Crow, A., & Kemtes, K. (2004). Eye-fixation patterns of high-and low-span young and older adults: down the garden path and back again. *Psychology and Aging*, *19*(1), 157.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Leonard, B. (2019). *mturk-microbatcher.* `https://github.com/jhupsycholing/mturk-microbatcher`. GitHub.

Liversedge, S. P., Paterson, K. B., & Clayes, E. L. (2002). The influence of only on syntactic processing of “long” relative clause sentences. *The Quarterly Journal of Experimental Psychology Section A*, *55*(1), 225–240.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703. Retrieved from `http://dx.doi.org/10.1037/0033-295X.101.4.676`

Malone, A., & Mauner, G. (2018). What do readers adapt to in syntactic adaptation? In *Poster session presented at the 31st Annunal CUNY Sentence Processing Conference.*

Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1995). Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. *Journal of Psycholinguistic Research*, *24*(6), 469–488.

Roland, D., & Jurafsky, D. (2002). Verb sense and verb subcategorization probabilities. *The lexical basis of sentence processing: Formal, computational, and experimental issues*, 325–346.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition.

Wiley Interdisciplinary Reviews: Cognitive Science, *1*(6), 906–914.

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*(2), 377.

Schad, D. J., Betancourt, M., & Vasishth, S. (2019). Toward a principled Bayesian workflow in cognitive science. *arXiv preprint arXiv:1904.12765*.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

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

Trueswell, J. C. (1996). The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, *35*(4), 566–585. Retrieved from `https://doi.org/10.1006/jmla.1996.0030`

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*, 250–271.

York, R. (2012). Residualization is not the answer: Rethinking how to address multicollinearity. *Social Science Research*, *41*(6), 1379–1386.