Strong evidence for expectation adaptation during language understanding, not a replication failure. A reply to Harrington Stack, James, and Watson (2018)

Jaeger, T. Florian[1,2,3]; Burchill, Zachary[1,2]; Bushong, Wednesday[1,2]

[1]Department of Brain and Cognitive Sciences, University of Rochester, NY 14627, USA

[2]Center for Language Sciences, University of Rochester, NY 14627, USA

[3]Department of Computer Science, University of Rochester, NY 14627, USA

Word count main text: ~ 9,500

Including supplementary materials: 16538

Submitted to Memory & Cognition as a reply to Harrington Stack, James, and Watson (2018)

Corresponding author:

T. Florian Jaeger
fjaeger@ur.rochester.edu
Meliora Hall,
University of Rochester,
NY 146267-0268
USA

Abstract

Recent input affects subsequent language processing. One explanation for this holds that comprehenders adapt their implicit linguistic expectations based on the input, so as to facilitate efficient processing. A number of studies have identified support for this hypothesis. Harrington Stack, James, and Watson (2018) report a failure to replicate one of these studies (Fine, Jaeger, Farmer, and Qian, 2013). We show that, to the contrary, the data from Harrington Stack and colleagues constitute strong support for the hypothesis of expectation adaptation. Several factors contribute to the difference in conclusions. For example, Harrington Stack and colleagues argue based on differences in *p*-values, which is known to be problematic. We instead employ well-formed Bayesian measures of evidentiary support to assess replication success. Most critically though, the new experiments by Harrington Stack and colleagues differ in design from the experiments they aim to replicate. We correct for these differences by means of the same single-parameter belief-updating model previously employed by Fine and colleagues. The model provides trial-level predictions for the surprisal that comprehenders experience, based on previous input within and outside of the experiment. Trial-level analyses find that surprisal based on adapted expectations strongly predicts reading times in both the original and the replication data. In fact, once the differences in design are corrected for, the two data are *highly* similar; replication tests estimate the posterior probability of a replication *success* to be $\gg .9999$. We show how the same belief-updating model also predicts trial-to-trial priming, cumulative priming, and the inverse preference effect in priming.

Keywords: expectation adaptation; sentence processing; Bayesian; belief-updating; replication test; reading times

We reply to a reported failure (Harrington Stack, James, & Watson, 2018, henceforth HS18) to replicate a study out of our lab (Fine, Jaeger, Farmer, & Qian, 2013, henceforth F13). Both studies test a hypothesis about the adaptivity of implicit expectations during incremental sentence understanding—how comprehenders react to changes in the statistics of the input when they enter a new linguistic environment (such as an experiment or an encounter with an unfamiliar talker).

Using Bayesian replication tests, we show that the new study is unlikely to constitute a replication failure. Rather, the design of the new study differs from the original study in ways that are predicted by the tested theory to lead to different effects. Once these differences are taken into account—by means of the same 1-parameter computational model that motivated the original experiments (Fine, Qian, Jaeger, & Jacobs, 2010)—replication tests show that the new study by Harrington Stack and colleagues is *extremely* likely to constitute a close replication of the original study by Fine and colleagues (the posterior probability of replication, compared to a null effect, is estimated to be $\gg .9999$). Our analyses show that both the F13 and HS18 data provide clear and significant support the hypothesis of expectation adaptation. Indeed, the new experiments by HS18 shed further light on the speed of expectation adaptation, and thus the degree to which implicit expectations during sentence understanding are continuously adjusted to facilitate efficient language understanding. We find that the effects of expectation adaptation are substantial, reducing per-word reading times by more than 50ms over the course of the experiment—a change of over 80% in the net effect of syntactic expectations on reading times.

Encouragingly, the model also *predicts* small effect sizes precisely where we observe them empirically. We discuss when small effect sizes are to be expected, and why (for reasons that link the present findings to research on error-based implicit learning accounts of syntactic priming). This, we show, helps to reconcile the conclusions offered by Harrington Stack and colleagues with those that we reach here—specifically, we agree that one of the qualitative conclusions made by Fine and colleagues actually has insufficient support in both F13 and HS18. We thus end with a concrete design proposal for future work based on our model. We begin by providing theoretical context.

Theoretical background: Expectations and expectation adaptation

One important step in the decoding of meaning from linguistic input is parsing—the inference of latent grammatical relations between the words in the input. This is a computationally complex problem: although we typically do not become consciously aware of it, there are many possible ways to combine a sequence of words into a grammatical parse. Comprehenders seem to overcome this problem at least in part by drawing on implicit gradient expectations based on implicit probabilistic knowledge about the relative (co-)occurrence statistics of words and latent syntactic structures. This is evident, for example, in effects of contextual predictability on processing times. Words and structures that are contextually predictable are typically processed more quickly (e.g., Boston, Hale, Kliegl, Patil, & Vasishth, 2008; Demberg & Keller, 2008; MacDonald, Pearlmutter, & Seidenberg, 1994; McDonald & Shillcock, 2003; Trueswell, Tanenhaus, & Garnsey, 1994; for a recent review, see MacDonald 2013). Expectation-based processing can be a rational response to the inevitable uncertainty involved in parsing (John T. Hale, 2001; Levy, 2008; Narayanan & Jurafsky, 2004), allowing more efficient processing (Smith & Levy, 2013; for review, see Kuperberg & Jaeger, 2016).

Expectation-based processing raises questions about how comprehenders deal with situations in which their implicit expectations do not match the statistics of the linguistic input. Talkers/writers can differ in their lexical and structural preferences. Lexical and syntactic preferences can also differ based on the register (formal vs. informal) or genre (scientific writing vs. poetry). For example, passive structures are more frequent in formal, compared to informal, registers (Weiner & Labov, 1983). One question then is whether comprehenders ignore the resulting differences in the input statistics, or whether they somehow adjust their expectations towards the statistics of the current input. On the one hand, such expectation adaptation might require additional cognitive and neural resources. On the other hand, a failure to adjust one's

expectations towards the *actual* statistics can annul the advantage that is afforded by expectation-based processing: the efficiency of expectation-based processing depends on the extent to which the expectations reflect the *actual* statistics of the input.

These considerations motivate the hypothesis of expectation adaptation developed by Fine and colleagues (Fine et al., 2013, 2010). Fine and colleagues hypothesized that comprehenders implicitly adapt their expectations about the relative frequency of words and structures in the current environment based on the previously experienced input within this environment. For example, a participant in an experiment would enter the experiment with more or less typical expectations based on the relative frequency of linguistic structures in previously experienced input. But as the experiment progresses, participants might adapt their expectations for the current environment based on the relative frequency of the linguistic structures *in the input received so far during the experiment*. This adaptation is hypothesized to proceed by incrementally and rationally integrating prior expectations (or, in Bayesian terminology, prior beliefs) with the observations made in the current environment, leading to adapted posterior expectations/beliefs after each new relevant observation. We emphasize this point as it is of direct relevance to the interpretation of the study we respond to here: adapted expectations are predicted to reflect a weighted combination of observations made in the present environment and expectations based on relevant previous language experience. Rational expectation adaptation predicts that a comprehender's beliefs about the probability of a structure in a given linguistic context, $p_{posterior}$(structure | context), are a function of:

(i) the relative probability of that structure in that context in the comprehender's language experience prior to the experiment, p_{prior}(structure | context),[1]
(ii) the relative proportion of occurrences of the structure in the relevant context within the experiment so far, count(structure, context) / count(context),
(iii) the strength of the prior beliefs as transferred to the current experiment, and
(iv) the absolute number of times the relevant context has been observed within the experiment, count(context)

where (iii) and (iv) together determine the relative weighting of (i) prior beliefs about the relative probability of structures against (ii) the input from the present environment (with (iii) being the only degree of freedom, as we elaborate below). Bayesian belief-updating provides a way to model rational expectation adaptation, making it possible to compare human processing behavior against the predictions of a normative model (Bushong, Burchill, & Jaeger, n.d.; Fine et al., 2010; Kleinschmidt, Fine, & Jaeger, 2012; Myslín & Levy, 2016). The hypothesis of expectation adaptation thus constitutes a hypothesis about both *why* and *how* recent experience affects language processing (for related proposals, see also Chang, Dell, & Bock, 2006; Dell & Chang, 2013; Garrod & Pickering, 2009; Jaeger & Snider, 2013; Reitter, Keller, & Moore, 2011).

Here, we focus on the two reading experiments described in F13 and the conflicting evidence in a recent large-scale replication attempt by HS18. We emphasize, however, that support for expectation adaptation now comes from a number of studies on spoken and visual language processing (Farmer, Fine, Yan, Cheimariou, & Jaeger, 2014; Fine & Jaeger, 2016a; Fine et al., 2010; Fraundorf & Jaeger, 2016; Kamide, 2012; Kaschak, 2006; Kaschak & Glenberg, 2004; Ryskin, Qi, Duff, & Brown-Schmidt, 2017; Yan, Farmer, & Jaeger, 2018), including some studies that have begun to identify the potential limits of such adaptation (Liu, Burchill, Tanenhaus, & Jaeger, 2017; ongoing work by Rachel Ryskin and Sarah Brown-Schmidt). Since this body of evidence is largely not mentioned in HS18, we include summaries below.

[1] We use "context" throughout to refer strictly to the *linguistic* context in which, say, a word or syntactic structure occurs. "Environment", by contrast, captures indexical variables such as talker, experiment, genre, register, etc. (for related discussion about the relation between context and environment, see Fine et al., 2013; Kleinschmidt & Jaeger, 2015; Qian et al., 2012)

Three qualitative predictions of expectation adaptation

The reading experiments in F13 and HS18 investigate changes in so-called *garden path* effects—slow-downs in reading times when temporarily ambiguous sentence onsets are disambiguated towards the *less* expected parse—through the course of an experiment. The magnitude of such garden-path effects has been linked to the relative expectedness of the parse: the less expected the parse, the larger the slow-down in reading times at the disambiguation point (e.g., Garnsey, Pearlmutter, Myers, & Lotocky, 1997; McRae, Spivey-Knowlton, & Tanenhaus, 1998; Wilson & Garnsey, 2009). This makes garden-path sentences a suitable test bed for the study of implicit changes in expectations.

Specifically, the experiments in F13 and HS18 focus on the garden path ambiguity in (1), between a matrix verb (MV) and a reduced relative clause (RC) continuation. The temporary ambiguity in (1) starts at the ambiguous verb form *warned* and is ended at the disambiguation point—*before* in (1a) or *conducted* in (1b).

(1) The experienced soldiers *warned about the dangers* …

a. ... before the midnight raid. (disambiguation towards MV)

b. ... conducted the midnight raid. (disambiguation towards RC)

(2)

a. The experienced soldiers spoke about the dangers before the midnight raid. (unambiguous MV)

b. The experienced soldiers who were told about the dangers conducted the midnight raid. (unambiguous RC)

Continuations like (1b) consistently elicit what is known as ambiguity or garden-path effects: reading times spike at the disambiguation point (e.g., *conducted* ... in (1b)) where the ambiguity is resolved towards the relative clause interpretation, compared to unambiguous RCs (2b) (F. Ferreira & Clifton, 1986; MacDonald, Just, & Carpenter, 1992; Tabossi, Spivey-Knowlton, McRae, & Tanenhaus, 1994; Trueswell et al., 1994). No such ambiguity effect is found at the same region for ambiguous compared to unambiguous MVs, e.g., on *before* ... in (1a) compared to (2a). Expectation-based accounts attribute this ambiguity effect to the fact that the ambiguous verb form *warned* is overwhelmingly more likely to occur with MVs (67% of the time) than RCs (.8%) in readers' previous language experience (based on Roland, Dick, & Elman, 2007). According to these accounts, the slow-down in reading times at the disambiguation point is a function of the relative degree of expectation violation (the prediction error in the sense of, e.g., Dell & Chang, 2013; Fine & Jaeger, 2013). In line with this prediction, the magnitude of the ambiguity effect—i.e., the slow-down at the disambiguation region—has been found to be smaller for ambiguous verb forms that are less biased towards the MV continuation in everyday language use (e.g., Hare, Tanenhaus, & McRae, 2007).

This, Fine and colleagues argued, makes the magnitude of a garden-path effect a behavioral signature of the relative (un)expectedness of RCs and MVs at that point in the experiment. This in turn makes it possible to test whether comprehenders' expectations for either structure *change* throughout the experiment. Specifically, Fine and colleagues tested three qualitative predictions of expectation adaptation. The first two predictions directly follow from the idea that comprehenders incrementally adapt their expectations towards the statistics of a novel environment as they are exposed to it:

Prediction 1: With sufficient repeated observations of an *a priori* highly unexpected structure—such as RCs in contexts like (1) above—the surprisal of the *a priori* expected structure—such as MVs—should eventually increase so much that comprehenders now garden path for the *a priori* expected structure.

Prediction 2: At the same time, the surprisal of the *a priori* highly unexpected structure will decrease, thus reducing any garden path effects for that structure.

With sufficient exposure within an experiment, it should thus be possible to adapt comprehenders' environment-specific expectations to be opposite of *a priori* expectations based on everyday language input. The third prediction tested by Fine and colleagues relates to the idea that only *relevant* observations should affect a given expectation for a specific structure:

Prediction 3: (After discounting for changes in reading speed due to unrelated factors) the magnitude of the garden path experienced on an *a priori* infrequent structure should only depend on observations that speak to the relative frequency of RCs and MVs in the relevant contexts. Thus, the number of intervening fillers, which do not provide linguistic contexts that inform the relative probability of RCs and MVs in contexts like (1), should *not* affect the magnitude of the garden path. Next, we summarize how Fine and colleagues set out to test these three predictions.

Summary of the original study: Fine, Jaeger, Farmer, and Qian (2013)

In order to test predictions about implicit expectations, or changes thereof, it is also necessary to make assumptions about how these expectations link to observable measures of processing difficulty. Although not critical to the hypothesis of expectation adaptation, Fine and colleagues assumed a log-reciprocal (surprisal) link between expectations and reading times during self-paced reading (Boston et al., 2008; Demberg & Keller, 2008; Frank & Bod, 2011; John T. Hale, 2001; Levy, 2008; Linzen & Jaeger, 2015; Smith & Levy, 2013). Here we follow this assumption. Indeed, as we show later, the data from HS18 and F13 exhibit almost identical increases in RTs for each 1 bit increase in surprisal.

Fine and colleagues present two self-paced reading experiments designed to test the three predictions outlined in the previous section. They found all three predictions supported by their data, though the strength of the support varied between the three predictions. Experiment 1 employed a 2x2 within-participant design, exposing participants to a uniform mixture of 50% RCs and 50% MVs, half of each with the type of temporary ambiguity shown in (1) above and half without. Structure and ambiguity were randomly distributed across trials and counter-balanced across participants via Latin-square design, with fillers interspersed between critical RC and MV items. Figure 1 shows the design and the corresponding predicted change of RC and MV surprisal across the course of the experiment. The numerical predictions are derived from a Bayesian belief-updating model first introduced for this purpose in Fine et al. (2010), and described later in this article. For now, the purpose of this visualization is merely to provide readers with an intuition of the predicted trial-by-trial effects of expectation adaptation. Following Fine and colleague, we assume that participants' prior expectations at the beginning of the experiment reflect the relative proportion of RCs (~.008) and MVs (~.67) in natural language use. These estimates are obtained from a large sample of automatically parsed written British English (Roland et al., 2007), and are thus best thought of as course-grained approximation of the RC/MV statistics that an average (American English speaking) participant has experienced.

Of note in Figure 1 is that RC surprisal is predicted to change much more than MV surprisal. As we show later, this is a direct consequence of the fact that RCs, compared to MVs, are unexpected based on everyday language experience, $\hat{p}(RC) = .008$, and thus highly surprising at the beginning of the experiment, $\hat{I}(RC) = -\log_2 \hat{p}(RC) = 6.97$ bits. We return to this asymmetry below, as it is of relevance to understanding when expectation adaptation is predicted to lead to detectable behavioral signatures.

*** Figure 1 APPROXIMATELY HERE ***

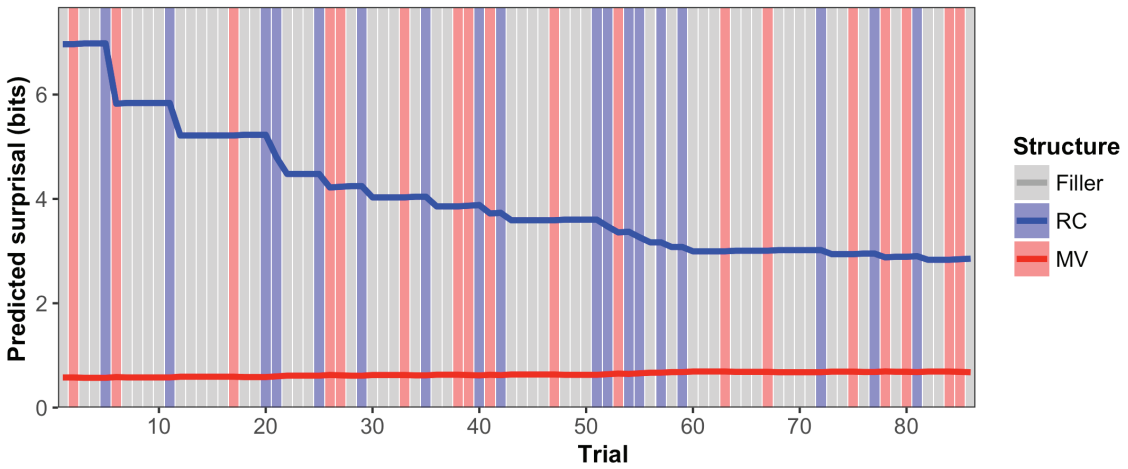

Figure 1 Design (shaded bars) and predicted changes in RC and MV surprisal (lines) over the course of the experiment for Experiment 1 in Fine et al. (2013). Bars show sentence types that participants would read in one of the Latin-square design pseudo-randomized lists. The model from which the predictions are derived is the same as in Fine et al. (2010) and described below. Although hard to discern in this plot, MV surprisal is slowly increasing throughout the experiment.

In line with Figure 1, F13 found a significant decrease in the ambiguity effect for RCs, and an insignificant *in*crease in the ambiguity effect for MVs. The decrease in the garden path effect for RCs was replicated in three self-paced reading experiments with the same design and different materials, reported in Craycraft (2014). The same effect was also replicated in three eye-tracking reading experiments with the same design and different items reported in (Yan et al., 2018). The decrease in the garden path effect for RCs was also observed in three further self-paced reading experiments on only RCs reported in Fine and Jaeger (Fine & Jaeger, 2016a). Additional conceptual replications of this finding—reduction in garden-path effects for an *a priori* unexpected structure when it is observed much more frequently in the current environment—come from reading experiments on other syntactic structures (e.g., Fine & Jaeger, 2011; Fine et al., 2010; Fraundorf & Jaeger, 2016; Kaschak, 2006; Kaschak & Glenberg, 2004; Wells, Christiansen, Race, Acheson, & MacDonald, 2009). Together, the evidence of over 10 self-paced and eye-tracking reading experiments on three different syntactic structures provides strong support for Prediction 2.

F13's first experiment also provides some support for Prediction 1—a numerical, but insignificant, *in*crease in the ambiguity effect for the *a priori* more expected MV structure. Later studies on the RC/MV ambiguity with the same design have replicated this numerical trend (Craycraft, 2014; Yan et al., 2018). On the one hand, it is not surprising that this trend for MV garden paths only shows numerically: the hypothesis of expectation adaptation *predicts* a very small change in MV surprisal, compared to RC surprisal, throughout the experiment (see the red line in Figure 1). Nevertheless, this pattern arguably constitutes weaker support for Prediction 1, compared to the repeated and clear replication of Prediction 2.

To put Prediction 1 to a stronger test, F13 conducted a second self-paced reading experiment on the RC/MV ambiguity. This second experiment employed a between-participant design. Figure 2 (top) illustrates the stimuli seen by the two exposure groups, and the corresponding predicted changes in surprisal. Invisible to participants, the experiment consisted of three blocks. The RC-First group read only RCs in Block 1, a total of 16 RCs. The Filler-First group read only fillers (also 16). In Block 2, both groups read 10 RCs interspersed with 20 fillers. In Block 3, both groups read 10 MVs interspersed with 15 fillers, resulting in a total of 71 sentence trials across the entire experiment. Half of the MV/RCs in each block contained the ambiguity, half did not. This was counter-balanced across participants using a Latin-square design.

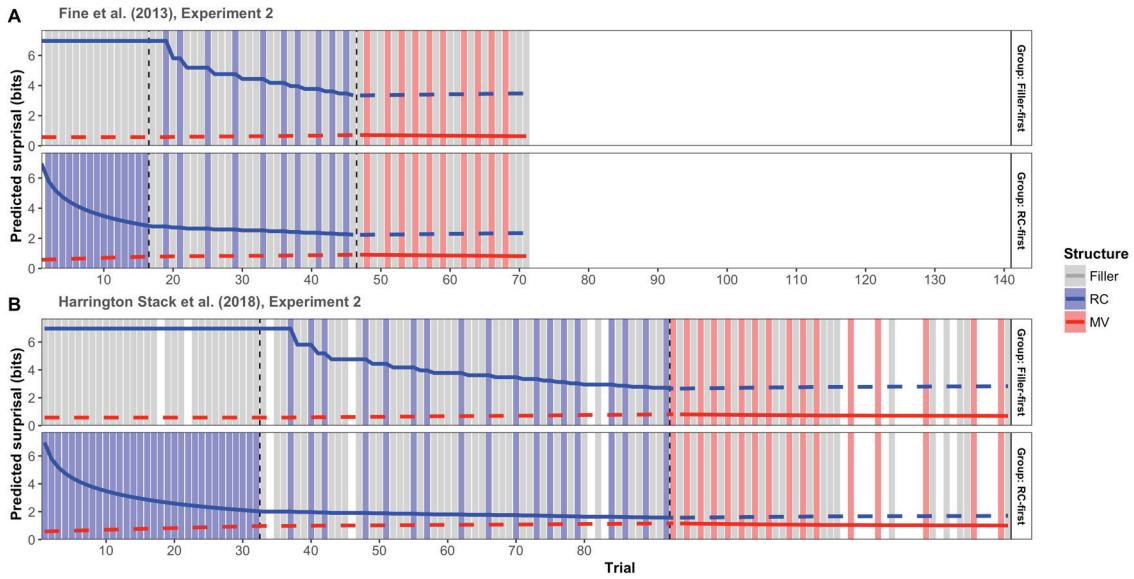

Figure 2 *Design (shaded bars) and predicted changes in RC and MV surprisal (lines) over the course of the experiment for Experiment 2 in Fine et al. (2013), Panel A, and Harrington Stack et al. (2018), Panel B. Predictions for structures that are not present in a block, and thus cannot be tested in that block, are shown dashed. Critical and filler trials that Harrington Stack et al. (2018) removed from analysis are shown in white.*

This design affords qualitative tests of all three predictions made above. In line with Prediction 1, F13 reported a clearly significant ambiguity effect for MVs in Block 3: after repeated exposure to the *a priori* less frequent RC structure, disambiguation towards the MV structure led to the type of processing slow-down typically associated with RCs. This main effect of ambiguity across the two exposure groups could, however, be confounded: F13 did not conduct a control condition in which both Blocks 1 and 2 had only fillers. It is thus unclear whether the MV stimuli in Block 3 just happened to exhibit garden path effects because of their lexical properties, or whether they exhibited the garden path because of the preceding exposure to RCs. Critically, F13 also found a significant interaction between ambiguity and exposure group for Block 3: The garden-path for MVs in Block 3 was larger for participants who had seen more RCs in Blocks 1 and 2 (the RC-first group), compared to participants in the Filler-first group. This provided more decisive support for Prediction 1.

F13's Experiment 2 also provided additional support for Prediction 2, and initial support for Prediction 3. In support of Prediction 2, F13 found that the ambiguity effect on RCs decreased throughout Blocks 1 and Block 2 in the RC-first group. This decrease was significant in a trial-level analysis that let structure, ambiguity, and their two-way interaction interact with item order (the number of RCs observed in the preceding stimuli). The effect did, however, not reach significance in a more coarse-grained block-level analysis that compared the ambiguity effect on RCs in Block 1 against Block 2. Given that the block-level analysis has less power, and given that Prediction 2 has now been confirmed by a large number of studies, we do not discuss this further (incl. HS18, see below).

Finally, in support of Prediction 3, F13 found that the ambiguity effect on RCs in Block 2 was marginally smaller for the RC-first group, compared to the Filler-first group (p=.08). Since both groups of participants read equally many sentence trials preceding Block 2, and equally many RCs (and fillers) during Block 2, this difference cannot be explained by adaptation to the experimental task.

Summary of Stack Harrington, James, and Watson (2018)

HS18's first experiment was intended as an extension based on the design of F13's Experiment 1 but failed to replicate the basic reduction of the RC ambiguity effect observed in F13 (Prediction

2). We do not discuss HS18's Experiment 1 further here: in addition to F13 (Experiment 1 and 2) the reduction in the RC ambiguity effect has been successfully replicated in at least ten other experiments using both self-paced reading (Craycraft, 2014; Farmer, Fine, & Jaeger, 2011; Fine & Jaeger, 2016a) and eye-tracking reading paradigms (Farmer et al., 2014; Yan et al., 2018). Indeed, *all* six attempts to replicate this effect in our lab have found it, and several of these studies arguably had higher power than HS18's Experiment 1. All of these experiments had similar numbers of items and participants as HS18's Experiment 1. Across studies, replications of F13's Experiment 1 cover a variety of different items and additional controls. This includes counterbalancing of item order and item identity across experimental lists (Craycraft, 2014; Fine & Jaeger, 2016a). These replications thus avoid confounding effects of expectation adaptation with well-documented item effects based on, for example, the item's matrix verb (Hare et al., 2007; McRae et al., 1998; Trueswell et al., 1994). This control was absent in Experiment 1 of both F13 and HS18. In short, it is very likely that HS18's Experiment 1 constitutes a Type II error (false negative).

We thus focus on HS18's second experiment, which also forms the bulk of their argumentation. This experiment followed the same between-participant design as F13's Experiment 2 but used twice as many items in each block as well as about five times as many participants (see Figure 2). HS18 analyzed their data using the same block-level analyses that constituted the planned analysis in F13. HS18 found that the ambiguity effect for RCs decreased from Block 1 to Block 2 in the RC-first group. Whereas this effect was only marginal in the block-level analysis of F13, the effect was significant in HS18. This replicates the support for Prediction 2. With regard to Predictions 1 and 3, however, the block-level analyses do not seem to support F13's conclusions. Only a marginal ambiguity effect was observed for the MVs in Block 3, and there was no interaction of ambiguity and exposure group. Similarly, the ambiguity effect on RCs in Block 2 did not differ between the two exposure groups.

HS18 also conducted power analyses of both F13 and HS18, based on which they concluded that the original study was under-powered, and that their replication attempt had drastically higher power (but see our Appendices A.2, B.2, and C.2). HS18 take that their results to show that (1) the results reported in F13 do not replicate, and that (2) there either is no expectation adaptation during sentence understanding, or expectation adaptation proceeds slowly, perhaps taking multiple days of exposure. Next, we show why we think these conclusions are not warranted.

A failure to replicate? Probably not. Rather a new data point predicted by expectation adaptation

HS18 do not report any direct test that assesses the extent to which their data replicates F13's. Instead, HS18 argue based on the failure or success of replicating *significance patterns*. Since this is known to be problematic (Kass & Raftery, 1995; Raftery, 1995; Wagenmakers, 2007), we employ Bayesian replication tests to assess whether the results observed in HS18 constitute a failure or success to replicate the findings in F13 (Verhagen & Wagenmakers, 2014). For completeness' sake, we first present a replication test under the assumption that Harrington Stack and colleagues indeed replicated the design of the original study—i.e., assuming that the hypothesis of expectation adaptation would actually *predict* the same effect for both experiments. However, as we then discuss, the two studies, while similar, differ in their design in ways that change the predicted effects. We thus use an existing model of expectation adaptation (Bushong et al., n.d.; Fine et al., 2010) to derive predictions for both the design in F13 and the design in HS18. We then ask how likely it is that *the data in HS18 constitute a replication with regard to these predictions*—put differently, we ask how (in)compatible the two data sets are in the degree to which they follow the hypothesis of expectation adaptation. Throughout, we talk about the *likelihood of replication* to acknowledge that assessments of replication are a statistical inference, something that is easily forgotten when comparing dichotomous differences in significance patterns. We find that the new data in HS18 are exceedingly likely to constitute a replication of F13. Both data sets are highly similar, and both provide clear support for the predictions of the hypothesis of expectation adaptation.

Replication test under (wrong) assumption of identical designs: Neither clear failure, nor clear success

We repeated the linear mixed-effects analyses reported by Harrington Stack and colleagues over both the F13 and HS18 data. For each of Predictions 1-3, we ask whether the results from the HS18 data replicate the results from the F13 data. This is achieved by calculating the replication Bayes Factor, BF_{r0} for each of the three pairs of analyses (for details, see Verhagen & Wagenmakers, 2014).[2] This BF_{r0} quantifies the extent to which the effect size observed by HS18 replicates the effect observed by F13, as contrasted with the hypothesis that the true effect size is zero (the skeptic's null hypothesis).

Figure 3 summarizes the results of the three replication tests. Neither the replication, nor the null hypothesis, has a posterior probability larger than .95 (BF > 20). Following conventions proposed in Rafferty (1995), we find "positive" support for the hypothesis that Prediction 2 replicates (BF_{r0} = 5.6). Conversely, for Predictions 1 and 3 the replication test returns weak to positive support for the null hypothesis (over the replication hypothesis; for Prediction 1: BF_{0r} = 1 / BF_{r0} = 16.7; Prediction 3: BF_{0r} = 3.8). Bayesian replication tests thus seem to provide decisive support neither for nor against a replication. This might be surprising given the lack of significant effects for Predictions 1 and 3 in HS18's data despite the large number of participants. However, unlike the Bayes factor used here, *p*-values do not provide a well-formed measure of evidentiary support (Kass & Raftery, 1995; Raftery, 1995), and arguments based on significance patterns can be misleading (Vasishth & Gelman, 2017; Wagenmakers, 2007).

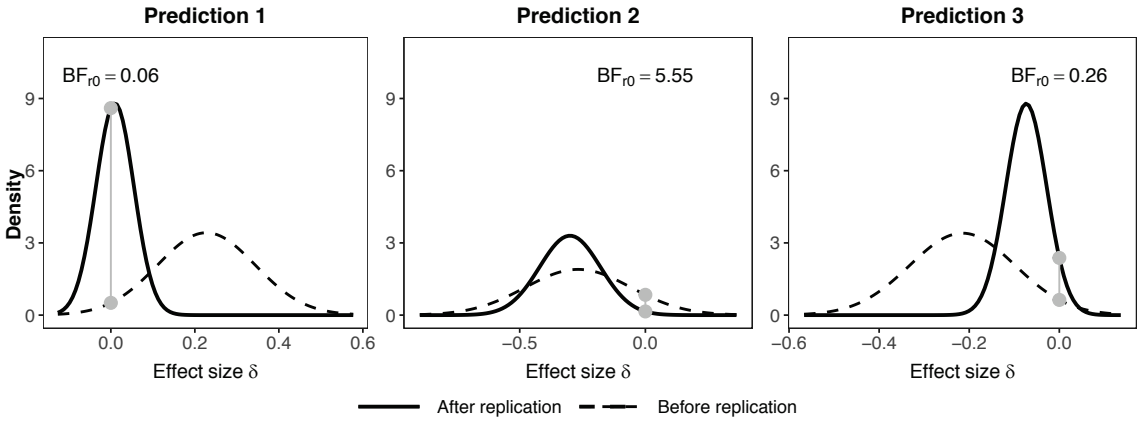

Figure 3 Replication Bayes Factors, BF_{r0}, for each of the three predictions. This test assumes that the predicted *effect sizes for the original F13 study and the new HS18 study are identical (as we discuss below, this is not the case). Dotted lines show the posterior distribution of effect sizes from the original study (Fine et al., 2013, Experiment 2), which is used as prior for the effect size in the replication test. Solid lines show the posterior distribution of effect sizes after the replication study is taken into account (Harrington Stack et al., 2018). The gray dots indicate the probability density for an effect size of zero (the skeptic's null hypothesis). The ratio of these two values—a measure of the decrease in the probability of the null hypothesis after seeing the replication data—gives the result of the replication test, BF_{r0}. BF_{r0} < 1 indicates that the*

[2] Specifically, we calculated replication Bayes Factors with the R code provided by Josine Verhagen (http://josineverhagen.com/?page_id=76). Effect sizes were estimated as the *t*-values of the critical interactions for Predictions 1-3 in the respective linear mixed-effects regressions over the original (F13) and replication experiment (HS18). The number of observations that went these *t*-values were estimated as the degrees of freedoms for the critical interactions, as estimated via Satterthwaite's approximation for HS18 and Kenward-Roger's approximation for F13 to avoid approximation errors on the smaller data set (implemented in *lmerTest*, Kuznetsova, Brockhoff, & Christensen, 2017). Following Verhagen and Wagenmakers (2014), we use BF_{r0} to refer to Bayes Factor comparing the hypothesis of replication against the hypothesis of a null effect, and BF_{0r} for the inverse.

probability of the skeptic's null hypothesis has increased after considering the results of the HS18. $BF_{r0} > 1$ indicates that the probability of the null hypothesis has decreased.

The replication test reported above takes for granted the premise that HS18 constitutes a replication of the original design—that the analyses of the two experiments are *predicted* to yield the same effect. As we discuss next, this assumption is not warranted, leading us to assess the degree to which the results of the two experiments are predicted to differ.

Changing the number of RCs and MVs in the experiment changes the predictions

In Appendix C, we summarize differences in the design and materials of F13's and HS18's experiment that limit the extent to which HS18 is expected to replicate the findings of F13. Here, we focus on the difference that is perhaps most obviously relevant to the hypothesis of expectation adaptation. This difference is evident when one compares the two panels of Figure 2 above: the new experiment in HS18 doubled the number of RCs and MVs in each block. As a result, the mean surprisal for RCs and MVs in each block, and thus also the critical differences in surprisal across groups and blocks relevant to Predictions 1-3—i.e., the predicted effects—differ between the two studies. This follows from the hypothesis of expectation adaptation: the predicted change in RC and MV surprisal is a result of weighing prior expectations (based on previous experience outside of the experiment) against the local statistics of input in the current environment (the experiment). As outlined in (i)-(iv) above, adapted expectations are thus predicted to be a function of not only the *relative* probabilities of the structures in the experimental input so far, but also their absolute frequencies (the latter affect the weight of recent experience against prior language experience; see (iii) and (iv) above). The design changes made in HS18 change the absolute frequency of the structures within the experiment and thus the predictions made by the hypothesis of expectation adaptation.

It is thus possible that the inconclusive outcome of the replication test reported in the previous section is at least in part due to the fact that we would *predict* the two studies to yield different results. To address this possibility, we draw on an existing computational model of expectation adaptation to derive trial-level predictions that correct for differences in the design of the two studies.

Belief-updating as a model of expectation adaptation

The specific model we employ here is developed and tested in Bushong, Burchill, and Jaeger (n.d.).[3] Appendix A.1 provides a summary of the model. The model uses Bayesian belief-updating to predict changes in the probability of RCs, MVs, and other structures in the context of an RC/MV garden-path inducing verb like "warned". This is the model we used to derive the predictions in Figures 1 and 2. The belief-updating model describes comprehenders' implicit probabilistic knowledge (or in Bayesian terminology, "beliefs") as probability distributions over probabilities of possible syntactic structures. These distributions capture a rational comprehender's uncertainty about the true probability of RCs, MVs, and other structures. The model has three free parameters that describe the relative count of RCs, MVs, and other structures that comprehenders are assumed to have experienced prior to the experiment. Bayesian belief-updating tells us how a new observation—such as an RC or MV—changes these prior beliefs to the updated posterior beliefs: by adding a count of 1 to whatever structure was observed. This updating process adds zero degrees of freedom to the model. Specifically, Bayesian belief-updating constitutes a normative model in that it describes belief-updating if the integration of prior experience and the current input proceeds

[3] The model extends the original model by Fine et al. (2010), where it was used to analyze expectation adaptation for another type of garden path structure. Whereas the original model considered only two outcomes (e.g., RC and MV), the model we employ here captures changes in the probability of RCs, MVs, but—more appropriately—also the probability of other possible structures that can occur after syntactically ambiguous verb forms. This does not lead to any qualitative changes in the predictions relevant for the current purpose, compared to the original model from Fine and colleagues, but moves towards a model with fewer simplifying assumptions. For further discussion, we refer to Bushong, Burchill, and Jaeger (n.d.).

rationally, making it a useful 'ground truth' model for expectation adaptation in human comprehenders (on the utility of normative models, see also Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Massaro & Friedman, 1990).

We can reduce the degrees of freedom for the belief-updating model further by fixing the prior beliefs about the relative probability of structures to what is observed in language corpora (as an approximation of participants' prior language experience). That is, we do not allow the model to freely choose how frequent RCs, MVs, or other structures are in previous experience. Instead, we force the model to believe that—prior to seeing any evidence to the contrary—RCs, MVs, and other structures have probabilities that reflect their relative frequencies in language use (in the relevant context of ambiguous verb forms like "warned"). This limits the flexibility of the model, and follows the common assumption that language users acquire implicit statistical knowledge about the relative frequency of linguistic structures in the input (Chang et al., 2006; Dell & Chang, 2013; Elman, 1990; Elman, Hare, & McRae, 1996; MacDonald, 1999, 2013; Tabor & Tanenhaus, 1999). This leaves our belief-updating model with only one degree of freedom: the *strength* of the prior belief, τ. Intuitively, it serves as a crude approximation of participants' estimate of the relevance of previous language experience to the current environment. As we discuss later in more depth, τ also co-determines how *quickly* expectations—and thus surprisal—are predicted to change throughout the course of the experiment.

With sufficient data from suitable designs, τ can be inferred from comprehension data (see Fine et al., 2010; Kleinschmidt et al., 2012). Here we instead committed *a priori* to a weak to moderately strong prior, $\tau = 100$, similar in magnitude to priors employed in our previous work. For example, for a different garden path structure, Fine et al. (2010) found τ=178 to provide the best fit against self-paced reading data, with similarly good fits for $\tau \geq 100$. A τ of 100 means that the model adapts within the experiment as if it has experienced a total of 67 MVs and .8 RCs prior to the experiment (recall that $\mathrm{p}_{\mathrm{prior}}(\mathrm{RC}) \approx .008$ and $\mathrm{p}_{\mathrm{prior}}(\mathrm{MV}) \approx .67$ based on corpus counts in Roland et al., 2007). After 100 relevant new data points,[4] this belief-updating model predicts expectations that reflect a balanced 50/50 mixture of prior beliefs and the statistics observed in the 100 observations made within the experiment.

Figure 4 illustrates how the model allows us to derive predictions for any sequence of sentence types, as we did for Figures 1 and 2. Panel B shows the changing beliefs about the relative probability of RCs and MVs predicted by the belief-updating model as a function of the inputs in Panel A. From these posterior beliefs, we can obtain predicted environment-specific expectation—and thus surprisal—of, for example, RCs on each trial in the experiment by marginalizing over the uncertainty about p(RC) at that trial. Panel C shows the predicted environment-specific expectations for RCs, MVs, and other structures than can occur in the ambiguous context (i.e., after a verb like *warned*) across all trials of the experiment. Finally, Panel D shows the resulting surprisal that would be experienced for a structure if it were observed on that trial.

*** Figure 4 APPROXIMATELY HERE ***

[4] We follow Fine and colleagues (2013) and assume that verb forms ambiguous between the RC and MV interpretation (such as *warned*) constitute the relevant context.

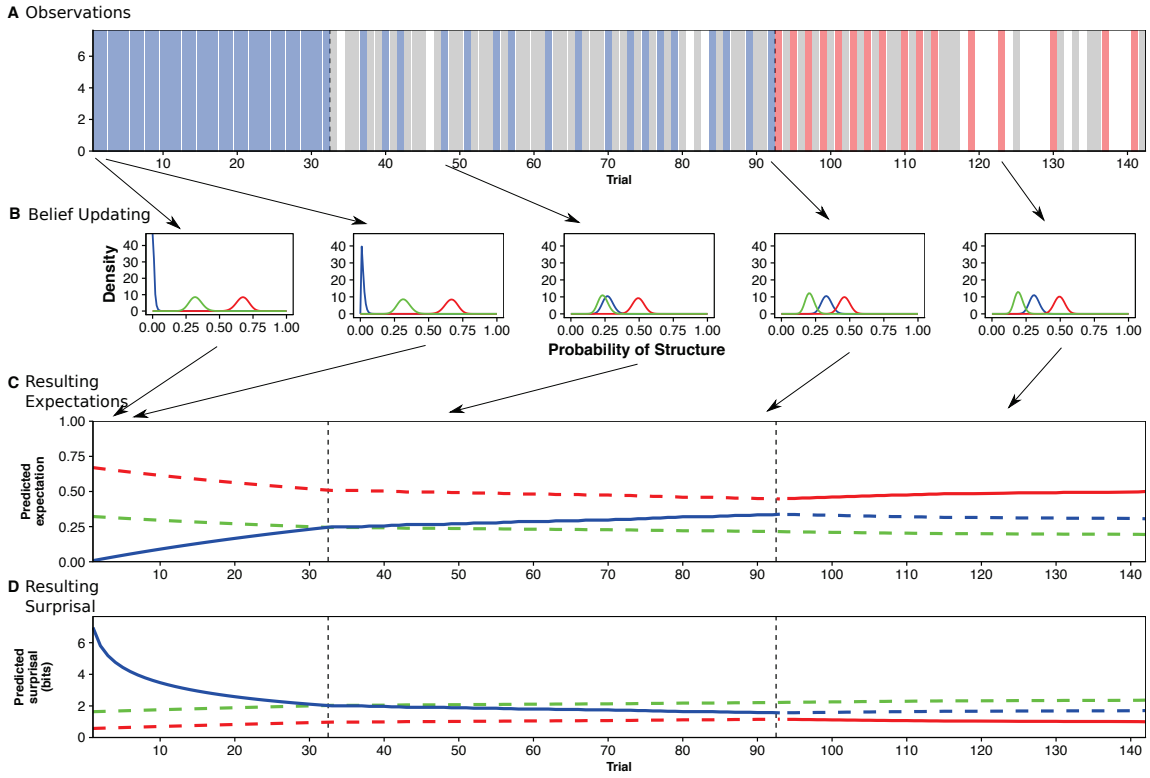

Figure 4 *Illustration of expectation adaptation modeled as Bayesian belief-updating with* $\tau = 100$. *Panel A: As an example, we show the input experienced by participants in the RC-first group in Harrington Stack et al. (2018). Panel B: Model predicted changes in participant's beliefs—and uncertainty— about the relative probability of RCs, MVs, and other structures, based on the inputs shown in Panel A. Fillers are stimuli without the relevant context (here: an ambiguous verb form like* warned*); RCs, MVs, and other structure all contain the relevant context. Panel C: Model predicted environment-specific probability of structures over the course of the experiment (marginalizing over uncertainty in Panel B). Panel D: Resulting surprisal of structures. Predicted expectations and surprisal for structures not present in a block are shown as dashed lines.*

A few basic observations are worth highlighting. As can be seen in Figure 4, the belief-updating model predicts that an observation of a structure reduces the surprisal for the next observation of that same structure. In short, although this is not our focus here, the belief-updating model predicts trial-to-trial syntactic priming in comprehension (as observed in, e.g., Arai, Van Gompel, & Scheepers, 2007; Thothathiri & Snedeker, 2008; Traxler, 2008). The model also predicts that these changes in surprisal persist over intervening trials, unless a competing structure intervenes (Pickering, McLean, & Branigan, 2013; as observed in, e.g., Tooley, Swaab, Boudewyn, Zirnstein, & Traxler, 2013), and that the effects of previous observations accumulate (e.g., Fine & Jaeger, 2016b). As we show more directly later, the model also predicts the inverse preference effect of syntactic priming: on a trial-by-trial basis, less expected structures lead to more change in expectation and surprisal (as observed in, e.g., Arai & Mazuka, 2013; Fine & Jaeger, 2013). Indeed, this is the explanation for the important asymmetry we noted earlier: while the surprisal of RCs (the *a priori* unexpected structure) is predicted to change rapidly, the surprisal of MVs (the *a priori* expected structure) changes much more slowly.

For comparison to our parametrization of $\tau = 100$, we also present predictions for $\tau = 1$, 10, 1,000, 10,000. This allows us to compare our predictions both to the hypothesis that adaptation essentially ignores prior beliefs based on previous language experience ($\tau = 1$), and to the hypothesis that there is no expectation adaptation ($\tau = 10{,}000$). For $\tau = 1$, the model's beliefs will be based to 90% on the statistics of the experiment after just nine relevant observations within the

experiment. For $\tau = 10{,}000$, on the other hand, even the 52 RCs and 20 MVs in HS18's design will lead to little change from the model's prior beliefs. We briefly illustrate this latter point so as to allow readers to develop intuitions about the model. For $\tau = 10{,}000$, exposure to 52 RCs within the experiment increases the prior count of RCs from 80 RCs assumed to have been observed before the experiment (=10,000 * .008) to 132 RCs (=80 + 52) at the end of the experiment. Conversely, the 20 MVs within the experiment increase the prior count from 6,700 to 6,720 MVs. For $\tau = 10{,}000$, the model thus predicts RC surprisal to decrease by .71 bits over the entire experiment (compared to 5.3 bits change in surprisal for $\tau = 100$), and for MV surprisal to increase by .006 bits (compared to .4 bits for $\tau = 100$). That is, for $\tau = 10{,}000$, RC and MV surprisal is close to constant—the predicted changes are at least an order of magnitude smaller than for $\tau = 100$.

Replication test after correction for differences in design: a clear replication success

Now that the belief-updating model is introduced, we return to HS18's decision to double the number of RCs/MVs in each block, compared to F13's design. The belief-updating model allows us to correct for this difference in design by comparing the trial-level effects of predicted surprisal. This also affords an opportunity to break new ground in the study of expectation-based processing—and specifically surprisal effects—during reading. Previous tests of the surprisal link hypothesis have been conducted under the simplifying assumption that expectations, and thus surprisal, are constant (e.g., Boston et al., 2008; Frank & Bod, 2011; John T. Hale, 2001; Linzen & Jaeger, 2015; Smith & Levy, 2013; but see Fine et al., 2010; Kleinschmidt et al., 2012; Myslín & Levy, 2016). Here we tested whether surprisal that results from *adapted* expectations based on recent input can explain reading times in the disambiguation region of a garden path structure.

Method. Specifically, we employed linear mixed-effects regression to analyze RTs in the disambiguation region as a function of the structure's predicted surprisal (in bits, centered) based on the belief-updating model with $\tau = 100$, ambiguity (sum-coded: 1 = ambiguous, -1 = unambiguous), and their interaction. We predict effects of surprisal for the disambiguation region of temporarily ambiguous structures. We also predict small effects of surprisal for the unambiguous structures. This latter prediction follows, for example, under the hypothesis that sentence understanding involves inference under uncertainty over noisy input (Bicknell & Levy, 2011; Levy, 2011). In short, we predict an interaction between surprisal with ambiguity, and possibly a main effect of surprisal.

The linear mixed-effects regression included the maximal converging random effect structure justified by the design.[5] To be maximally conservative, we analyze residual logarithm-transformed RTs that were corrected for both linear effects of word length and non-linear effects of trial position. These corrected log-RTs were obtained by fitting two separate generalized additive mixed models (GAMM, Wood, 2017; as implemented in *mgcv*, Wood, 2016) to the filler RTs of F13 and HS18. The predictions of this GAMM were then removed from the RTs on the disambiguation region of RCs and MVs to obtain the corrected (residual) RTs. Unlike the standard approach to RT correction, this GAMM-based RT correction removes large effects of adaptation to the task of self-paced reading, as participants get used to pressing the space bar to read the next word. Such task adaptation can have substantial effects, as shown in Panel A of Figure 5, with average per-word RTs decreasing by about 150ms (33%) over the course of the experiment. A failure to remove effects of task adaptation would otherwise confound our analysis in favor of our surprisal predictions. The GAMM-based approach also corrects for a second consequence of HS18's decision to double the number of stimuli per block: task adaptation affects the RTs in the different blocks differently in F13 and HS18 (a problem also acknowledged by Harrington Stack and colleagues, p. 875). In short, our approach ensures that we are comparing likes with likes when

[5] The full formula: residual RTs ~ Surprisal * Ambiguity + (1 | Subject) + (0 + Surprisal || Subject) + (0 + Ambiguity || Subject) + (0 + Surprisal:Ambiguity || Subject). Analyses that additionally included a main effect of structure (sum-coded: 1 = MV, -1 = RC) to capture possible differences in the processing of these structures that are independent of surprisal yield the same result.

comparing the replication data from HS18 to the original study. Appendix B provides additional information on the GAMM-based RT correction, and further clarification as to why we log-transformed RTs.

For the sake of comparison, we also present analyses of the standard untransformed word length-corrected RTs that were employed in both F13 and HS18 (derived from a linear mixed model, as implemented in lme4, Bates, Maechler, Bolker, & Steven Walker, 2015). Like the corrected log-RTs, these standard corrected RTs were obtained by fitting two separate models to the filler RTs of F13 and HS18. Details for both correction procedures are provided in Appendix B.3. All analyses were conducted in R (R Core Team, 2017). Figure 5 shows that the GAMM-based approach successfully captures task adaptation (Panel C), whereas the standard residualization approach does not (Panel B). We note that, as is indeed the case, the corrected RTs for RCs and MVs are not expected to pattern with those of fillers. RT correction is meant to remove solely the effects of task adaptation.

*** Figure 5 APPROXIMATELY HERE ***

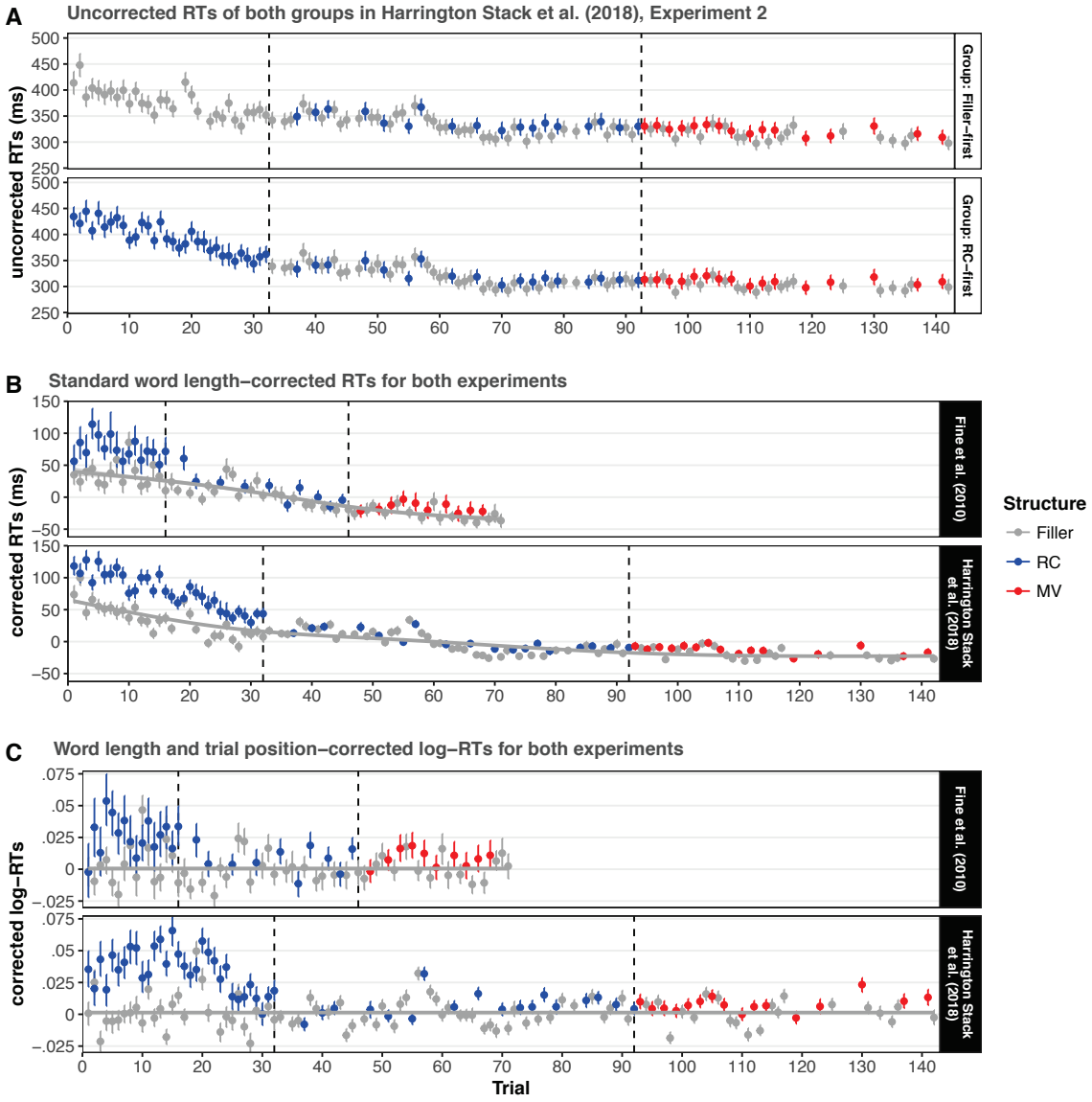

Figure 5 *Panel A: Uncorrected RTs decrease quickly as participants adapt the unfamiliar task of self-paced reading (here shown for Harrington Stack et al., 2018). Points reflect average RTs obtained by first averaging all per-word RTs across all regions within a sentence, and then averaging these mean per-word RTs across participants. Error bars show 95% bootstrapped CIs based on the by-subject mean per-word RTs. Panel B: standard word length-corrected RTs employed in the analyses reported in Fine et al (2013) and Harrington Stack et al. (2018). Panel C: residuals of GAMM-based trial and word length-corrected log-RTs that remove effects of adaptation to the self-paced reading task, thereby de-confounding comparison across the two experiments. Gray lines show smooths through the corrected filler RTs.*

Results. Figure 6 summarizes the results for both F13 and HS18 for $\tau = 100$, for both trial and word length-corrected log-RTs and standard-corrected RTs. For both data, and for both dependent variables, we find a significant main effect of adapted surprisal ($p < .0001$ for both F13 and HS18), and a significant interaction with ambiguity indicating that the effect of surprisal was larger for ambiguous structures ($p < .002$ for F13, $p < .0001$ for HS18). Simple effects analysis found that the effect of adapted surprisal was significant for both ambiguous (ps < .0001) and unambiguous structures (ps < .01), with similar effect sizes across the two data sets. For ambiguous structures, our analyses find an increase in 25.4 and 27.5ms in per-word RTs for every bit of surprisal in F13 and HS18, respectively. For unambiguous structures, the increase per bit of surprisal is 16.4 and 17.7ms in F13 and HS18, respectively. Interestingly, this striking similarity in the effects of surprisal held despite the fact that the two data sets exhibited ambiguity effects of slightly different magnitude—a difference that is to be expected given that the F13 and HS18 differed in the lexical materials they employed in the sentence garden path stimuli.

*** Figure 6 APPROXIMATELY HERE ***

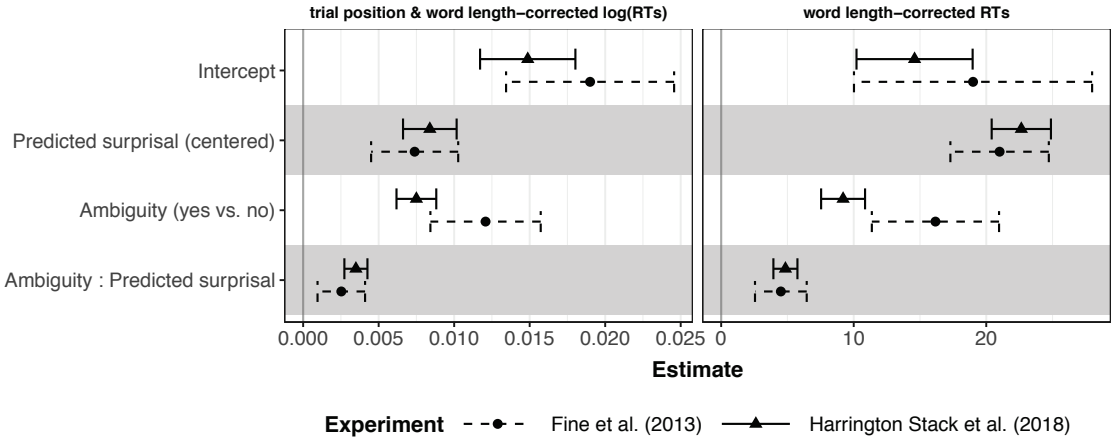

Figure 6 *Coefficient estimates and confidence intervals for our re-analyses of Fine et al. (2013) and Harrington Stack et al. (2018), correcting for differences in the designs of the two experiments. Both the results for trial position and word length-corrected log-RTs (left panel) and the results for standard word length-corrected RTs (right panel) are shown ($\tau = 100$).*

To assess whether the results from HS18's new design replicates the results from the original F13 design when the data is analyzed at this level—i.e., when correcting for the differences in the design, we calculate the replication Bayes factor BF_{r0} for the effect of adapted surprisal and its interaction with ambiguity. As shown in Figure 7, we find "very strong" support for the hypothesis that the effects of both adapted surprisal and its interaction with ambiguity replicate from F13 to HS18 (all BF_{r0}s > 150). Indeed, the BF_{r0}s indicate that the posterior probability that HS18 constitutes a replication of F13 is $\gg .9999$, and the posterior probability that HS18 constitutes a null effect is $\ll .0001$.

*** Figure 7 APPROXIMATELY HERE ***

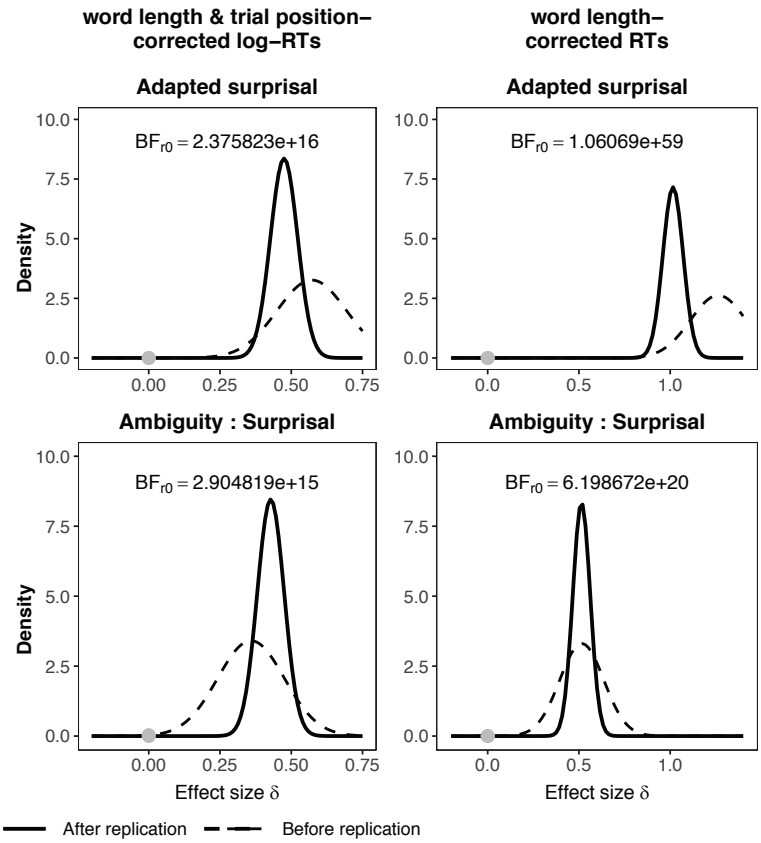

Figure 7 Replication Bayes Factors, BF_{r0}, for the effect of surprisal (top) and its interaction with ambiguity (bottom) for both word length and trial position-corrected log-RTs (left) and standard word length-corrected RTs (right). Unlike the tests in Figure 3, the replication tests shown here take into account that the design employed by Harrington Stack and colleagues (2018) differs from the original design employed by Fine and colleagues (2013).

In summary, once we correct for differences in the design, the results of two experiments are *very* compatible, with both experiments yielding the predicted surprisal effect and the predicted interaction with ambiguity. Indeed, for the word length and trial position-corrected log-RTs the posterior distribution after taking into account the replication data from HS18 suggest *larger* effect sizes for the critical interaction, compared to F13.

Both experiments also present clear evidence that *adapted* surprisal provides the best fit against the data. This is shown in Figure 8. For both F13 and HS18, a prior strength that allows adaptation within the experiment ($\tau = 100$) present a better model of human RTs than large τs for which barely any change in RC and MV surprisal is predicted.[6] Figure 8 further shows that comprehenders' expectations are not *solely* based on the statistics of the input within the experiment (e.g., the fits for $\tau = 1$ are worse than the fits for $\tau = 100$). Rather, comprehenders integrate the statistics of recent input with prior beliefs based on previous language experience—as predicted by rational expectation adaptation.

*** Figure 8 APPROXIMATELY HERE ***

[6] Since all analyses compared here have the same degrees of freedom, better data (log) likelihoods also indicate better Aikeken and Bayesian information criteria (AIC, BIC).

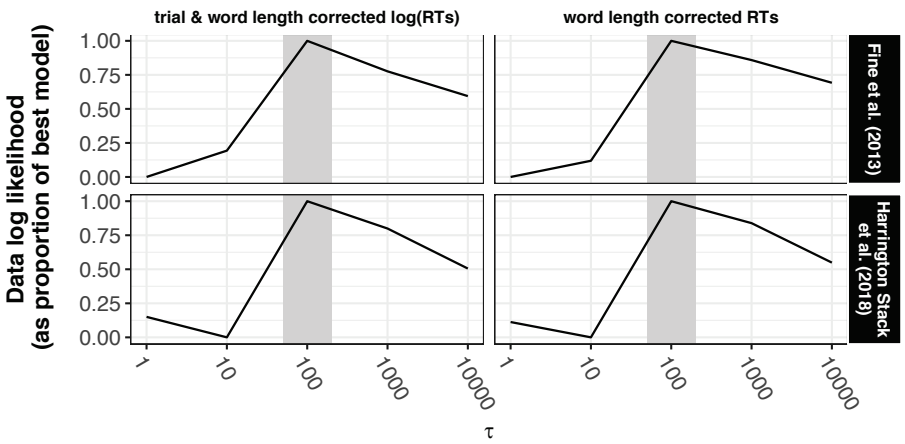

Figure 8 Log data likelihood of the observed RTs depending on τ (higher likelihood indicates better fit against the RTs from the experiments). Both the observed GAMM-based trial position and word length-corrected RTs (left column) and the observed standard word length-corrected RTs (right column) in the two experiments are more likely when adaptation can occur within the experiment (e.g., $\tau = 100$) than when expectations are assumed to be essentially stationary ($\tau = 10{,}000$).

Our trial-level re-analysis thus concludes with high confidence that HS18 constitutes a clear replication. What does all of this mean for the three qualitative predictions made by Fine and colleagues? Arguably, one important result of F13 was that they found evidence that an *a priori* highly expected structure (here MVs) could eventually become sufficiently unexpected to elicit a garden path (Prediction 1). The analyses we have conducted so far do not directly address this question. Indeed, the belief-updating model *predicts* smaller effects of surprisal for MVs, compared to RCs (recall Figure 4 above). Before we assess whether F13 and/or HS18 provide evidence of changes in, specifically, MV surprisal, we thus briefly take a step back and ask *how much* of an effect we expect to see for MV vs. RC surprisal. As we discuss next, this also holds the answer to an important question raised by Harrington Stack and colleagues about the speed of adaptation.

How rapidly do comprehenders adapt their expectations?

Of the five values for τ that we have considered here, the best-fitting τ is the same for both F13 and HS18. This value, $\tau = 100$, is of similar magnitude as the best-fitting prior strength in the only previous work that has explored this question for another syntactic structure (Fine et al., 2010). As mentioned above, the strength of the prior beliefs τ co-determines how quickly expectations can change. With a τ of 100, it takes 100 occurrences of the relevant context within the experiment (here: verb forms that allow both the RC or MV parse) until comprehenders' beliefs are predicted to be a 50/50 mixture of their prior beliefs and the input statistics observed in the experiment. At first blush, this might sound slow and thus in line with the concluding consideration of Harrington Stack et al. (2018): "one possible explanation for these results is that adaptation does, in fact, occur, but that it does not do so rapidly" (p. 876). This interpretation would be misleading, however, for reasons we lay out next.

The reason for this is that τ is only one of two properties that together determine the speed with which the implicit expectations for a structure are predicted to change. The second property is the prior probability of the structure. This is already apparent in the predictions visualized in Figure 4, where the surprisal for the *a priori* less expected structure changes much more quickly. Figure 9 more directly illustrates the relation between the prior probability of a structure (at any point during the experiment) and the change in the structure's probability (Panel A) and surprisal (Panel B) following one observation of that structure. What this figure shows is the typical signature of error-based learning: less expected observations lead to more change in expectation (Chang et al., 2006; Chang, Janciauskas, & Fitz, 2012; Dell & Chang, 2013; Fine & Jaeger, 2013; Jaeger & Snider, 2013). This is precisely the pattern that is observed in studies on trial-to-trial syntactic priming or

persistence, where it is often coined the "inverse preference" or "inverse frequency" effect (e.g., Arai & Mazuka, 2013; Bernolet & Hartsuiker, 2010; V. S. Ferreira, 2003; Fine & Jaeger, 2013; Jaeger & Snider, 2008). Of note is that the inverse preference effect is a *prediction* of the belief-updating model, where it follows from the ideal integration of new evidence with beliefs based on previously experienced inputs. This differs from error-based models, in which the inverse preference effect is explicitly coded into the learning rule (such as Chang et al., 2006; for discussion, see also Jaeger & Snider, 2013).[7]

Panel B shows how the hypothesized surprisal link between expectations and reading times further exaggerates the inverse preference effect: the predicted change in surprisal after an observation of a structure is an exponential function of its prior probability.

*** Figure 9 APPROXIMATELY HERE ***

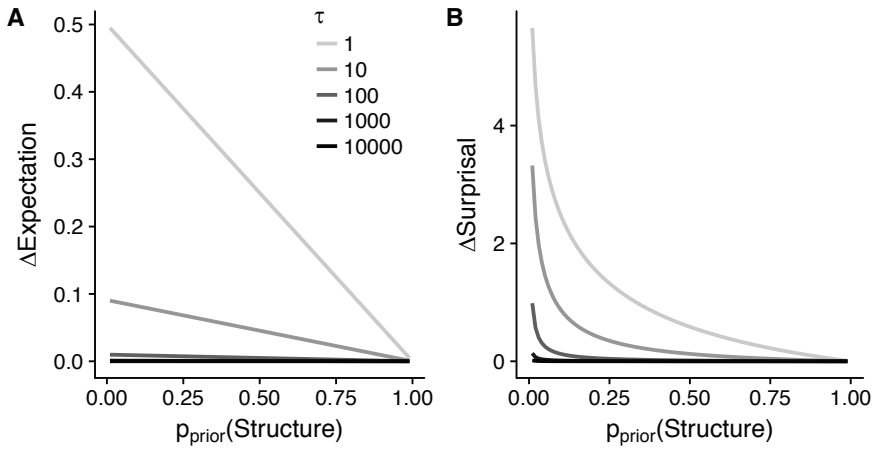

Figure 9 Change in the marginal probability (Panel A) and surprisal (Panel B) of a structure after a single observation of that structure, depending on the prior probability of the structure and the strength of the prior beliefs, τ.

Figure 9 and Figure 4 thus point to an important distinction between the speed of belief-updating and the resulting change in surprisal. The former is determined by τ, and in this sense held constant for RCs and MVs. The latter, however, depends both on τ and the prior probability of a structure. Crucially, for the type of experiment discussed here, it is the latter—the change in surprisal—that is predicted to determine the ability to *detect* an effect of expectation adaptation. Rational expectation adaptation thus predicts that it is much harder to detect the predicted changes in surprisal for an *a priori* highly expected structure (e.g., MVs), compared to the predicted changes for an *a priori* highly unexpected structure (e.g., RCs).

What does all of this mean for the experiments in F13 and HS18? For an *a priori* highly unexpected structure, a τ of 100 actually predicts rapid changes in that structure's surprisal. For example, for $\tau = 100$ the observation of a single RC will decrease the surprisal of the next RC by 1.16 bits from 6.97 bits to 5.81 bits. That is a decrease in surprisal of 17% after a single observation in the experiment. For MVs, on the other hand, it should be much harder to detect changes in

[7] Harrington Stack et al. (2018) seem to assume that prediction errors are only encountered in ambiguous structures: "when comprehenders encounter the same verbs in unambiguous contexts, the lack of ambiguity means that no error signal is generated" (ibid, p. 875). This only follows under the assumption that the mechanisms underlying parsing ignore uncertainty about the lexical input (noise-free input; e.g., in certain discrete non-cascading models), and proceed radically incremental by integrating new input into globally-coherent parses (for relevant discussion, see Bicknell et al., 2010; Kukona, Cho, Magnuson, & Tabor, 2014; Levy, 2011; Tabor & Tanenhaus, 1999). We do not make these assumptions. Expectation adaptation is thus also predicted to occur after unambiguous structures, though not necessarily to the same extent.

surprisal: for $\tau = 100$ our model predicts that the observation of a single RC will increase the surprisal of the next *MV* by only .014 bits from .58 bits to .59 bits, a change of only 2.4% (if an MV, instead of an RC, is observed the predicted decrease in MV surprisal is even smaller, .004 bits or 0.7%).

Figure 10 relates these differences in predicted surprisal effects to predicted effects on RTs. Specifically, Figure 10 shows the predicted effects of expectation adaptation on RTs in the RC-first group of HS18 (estimates for F13 are nearly identical). The estimated effects are based on the word length and trial position-corrected log-RTs, back-transformed into RTs. This means that any effect that is ambiguous between *task* adaptation and the adaptation of linguistic expectations is attributed to task adaptation. The estimates of the surprisal effects shown in Figure 10 are thus likely conservative. Even under such a conservative estimation, we see that surprisal effects on RCs on per-word RTs decrease by about 55ms (83%) over exposure to the first 80 trials (incl. 52 RCs). In contrast, the effects of surprisal on MVs between trials 92-142 barely decrease by 4ms (19%).

*** Figure 10 APPROXIMATELY HERE ***

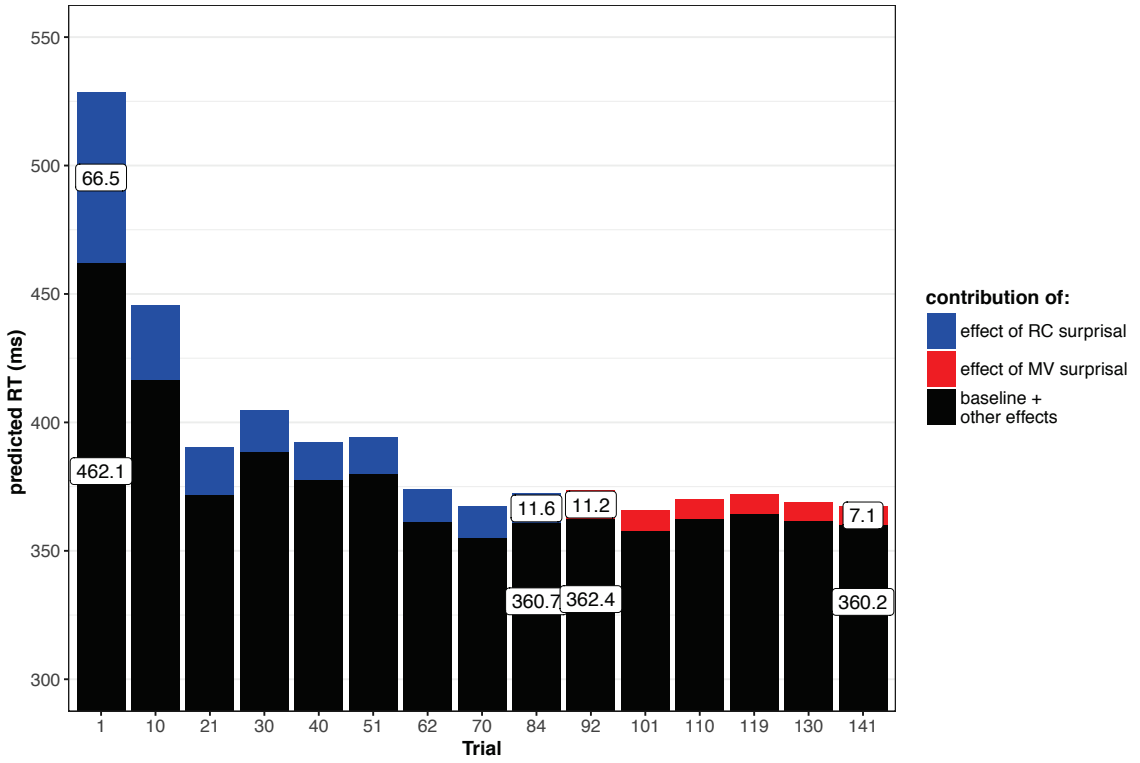

Figure 10 Predicted effects of surprisal on average per-word RTs in the disambiguation region of ambiguous RC. The estimates of the surprisal effects were obtained after removing effects of task adaptation (non-linear effects of trial position) and word length, as well as individual differences between subjects and items.

That is, with the same single value for the its only parameter, the belief-updating model predicts both fast changes in the RTs of RCs and slow changes in the RTs of MVs. An important consequence of this is that it is more difficult—at least if reading times are used as dependent measure—to detect changes in expectations for the *a priori* more expected structure (Prediction 1), compared to changes in expectations for the *a priori* less expected structure (Prediction 2; Appendix A.2 visualizes the consequences of this asymmetry for the type of block-level analysis conducted by Harrington Stack and colleagues; Appendix C.2 gives additional reasons why this holds in particular for the RC/MV ambiguity).

To illustrate this, we repeated the trial-level analysis reported above on only Block 3 which contains only MVs. Specifically, we analyzed the effect of adapted surprisal (centered within Block 3) and its interaction with ambiguity (again sum-coded: 1 = ambiguous, -1 = unambiguous) for only Block 3. We again conducted separate analyses for both standard word length-corrected RTs and log-RTs corrected for trial position and word length, and did so for both F13 and HS18. The interaction between ambiguity and MV surprisal did not reach significance in any of the analyses on Block 3 (there was a marginal effect in the predicted directed for F13, but only over standard word length-corrected RTs, $p < .08$). The main effect of MV surprisal was significant, but only for standard word length-corrected RTs ($ps < .0002$), which can be confounded by task adaptation. As is expected, the confidence intervals of this analysis are very wide (cf. Figure 6): the analysis is based on less than one third of the items as the full trial-level analysis presented above. For the trial position and word length-corrected log-RTs, the confidence intervals of both the surprisal effect and its interaction with ambiguity include the (significant) effect sizes found in the trial analyses that included all blocks (cf. Figure 6).

*** Figure 11 APPROXIMATELY HERE ***

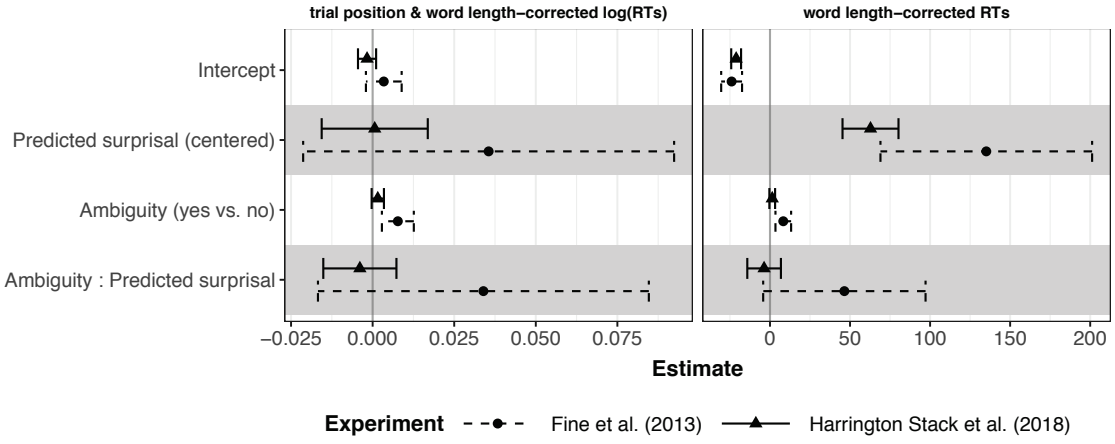

Figure 11 Coefficient estimates and confidence intervals for our re-analyses of Block 3 in Fine et al. (2013) and Harrington Stack et al. (2018), correcting for differences in the designs of the two experiments. Both the results for trial position and word length-corrected log-RTs (left panel) and the results for standard word length-corrected RTs (right panel) are shown ($\tau = 100$).

In short, even if considered in isolation, the MV reading times from Block 3 are compatible with the hypothesis of expectation adaptation. We are not aware of any reason why surprisal—which has been found to be a good predictor of per-word reading times across lexically and structurally heterogenous materials (Boston et al., 2008; Demberg & Keller, 2008; Linzen & Jaeger, 2015; Smith & Levy, 2013)—would affect RTs in MVs and RCs differently. Given the strong support for expectation adaptation from the RC data over Blocks 1 and 2, and given that the small effects on Block 3 are *predicted* by the same model under the same parameterization, it thus seems parsimonious to conclude that the reading of MVs is affects by expectation adaptation in the same way as the reading of RCs. At the same time, the trial-level analysis of surprisal effects in Block 3 confirms an important contribution of Harrington Stack and colleagues: it does indeed seem that the marginal effect in the block-level analysis in F13 not reliable (though we note that we predict the effect to be larger in F13, compared to HS18, see Appendix A.2). We thus close by briefly illustrating how future work could test Prediction 1 more effectively.

A proposal for future work

Consider a simple exposure-test design that maximizes the power to test Prediction 1. One group of participants is exposed only to RCs and the other group only to MVs (optionally with identical intervening fillers for both groups). Following exposure, both groups are tested on a block of MVs. As in the design of F13's and HS18's Experiment 2, we assume that ambiguity is counter-balanced across participants within each group. Figure 12 shows the predicted between-group difference in the average MV surprisal during the text block—a measure of the predicted effect size. Additionally, Figure 12 uses transparency to indicate an approximation of power for a block-level analysis based on the number of items in the test block. While Figure 12 is not intended to substitute a full power simulation, it suggests that at least 64 RCs/MVs would be required during exposure to achieve 1 bit of surprisal difference during test. The code and data for our models are shared via OSF at https://osf.io/4vxyp/, so that other researchers can conduct their own simulations (Appendix C discusses additional design considerations for future studies).

*** Figure 12 APPROXIMATELY HERE ***

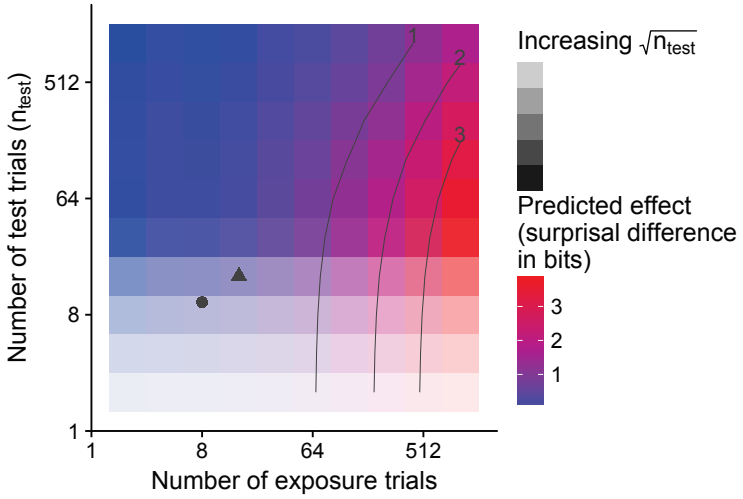

Figure 12 Predicted difference in average MV surprisal during a test block with y MVs after exposure to either only x RCs or only x MVs (interspersed with fillers). As a rough estimate of statistical power, opaqueness scales proportional to the square root of the number of test trials. As a point of reference, we indicate the predicted surprisal difference between participant groups in Block 3 (i.e., Prediction 1) for both Fine et al. (2013, circle) and Harrington Stack et al. (2018, triangle). The vertical position of those points reflects the number of informative (MV) test trials). The horizontal position reflects the predicted effect (see also Appendix A.2) prior to taking into account further properties of Harrington Stack et al.'s experiment that we predict to reduce power (Appendices B.2 and C.2).

Conclusion

Harrington Stack, James, and Watson provide an interesting new data point to the study of expectation adaptation during sentence processing. Contrary to their conclusion that their results constitute a failure to replicate Fine et al. (2013), we find that the results of Harrington Stack and colleagues are exceedingly likely to constitute a replication of Fine et al. (2013). In fact, the effects of expectation-based processing and expectation adaptation observed in the two studies are remarkably similar. Both the original study and the replication by Harrington Stack and colleagues match the predictions of the hypothesis of expectation adaptation. Both studies exhibit similar effect sizes of adapted expectations (once differences in design between the studies are taken into account). The data from both studies argue for a similar strength of prior beliefs—the only free parameter in the model we used to fit both data—and thus the speed of expectation adaptation. Together with other recent studies on expectation adaptation (e.g., Farmer et al., 2014; Fine &

Jaeger, 2016a; Fine et al., 2010; Fraundorf & Jaeger, 2016; Kamide, 2012; Kaschak, 2006; Kaschak & Glenberg, 2004; Ryskin et al., 2017; Yan et al., 2018), the present findings thus suggest that comprehenders can adapt implicit expectations that guide the incremental integration of words into sentence interpretations.

Our reply illustrates the value of computational models in guiding design and analysis. The belief-updating model of expectation adaptation (Bushong et al., n.d.; Fine et al., 2010) provides quantitative predictions for changes in expectations and surprisal given a sequence of observations. This allowed us to correct for differences in the design between the original experiment in Fine et al. (2010) and the replication in Harrington Stack et al. (2018). The same model facilitates predictions for possible future studies, including experiments on other structures. One specific consideration for future studies derived from the model is that structures with lower prior probability are predicted to result in larger and quicker changes in surprisal (Appendix C summarizes additional considerations).

The single-parameter belief-updating model we have employed here is not intended as a full model of expectation adaptation. For discussion of the inferences underlying adaptation, and how the brain might strike a rational balance between flexible and yet stable expectations, we refer to Kleinschmidt and Jaeger (2015; see also Qian, Jaeger, & Aslin, 2012 and Appendix C.3). Future work might also compare mechanistic models that implement approximations to rational or rationally bounded expectation adaptation, to the type of simple normative model we have employed here. For example, certain exemplar-based models can be computationally equivalent to the type of Bayesian inference model we have presented here (Shi, Griffiths, Feldman, & Sanborn, 2010; see also discussion in Kleinschmidt & Jaeger, 2015: 183-184). Similarly, we project that error-based implicit learning models predict at least qualitatively similar effects (Chang et al., 2006; Dell & Chang, 2013; for discussion, see also Jaeger & Snider, 2013).

We would like to close by acknowledging how helpful we have found the willingness of Harrington Stack and her colleagues to share ideas, comment, and to ask/answer questions. The present reply elaborates and clarifies the proposal put forward in Fine et al. (2013). This would not have been possible without the work by Harrington Stack and colleagues. On a personal note, we have found the discourse surrounding this replication—in particular, with Harrington Stack, James, and Watson—to be intellectually rewarding, and hope that the results presented here contribute to further our understanding of adaptivity during language understanding.

Acknowledgments

We are grateful to Caoimhe Harrington Stack, Ariel James, and Duane Watson for responding to our queries about their sentence materials, exclusion criteria, and analyses. We appreciate that Harrington Stack and colleagues reached out to the first author of the original study, Alex B. Fine, to make sure that the analyses they conducted mirror those conducted in the original paper, and that Harrington Stack and colleagues shared results and data with us prior to publication. We thank Alex B. Fine for patiently answering questions about the original study, for sharing code and data with us—all despite having left academia. We also thank Alex B. Fine and the members of the Human Language Processing Lab—in particular, Shaorong Yan—for insightful discussion of earlier write-ups and presentations of the work summarized here.

Contributions

- TFJ prepared the data from both studies for analysis with help from Alex B. Fine, Ariel James, and Caoimhe Harrington Stack.
- WB implemented the beta-binomial belief-updating model and created all prediction graphs with input from TFJ.
- ZB developed and validated the GAMM used to correct RTs for linear effects of word length and non-linear effects of trial position, with input from TFJ and WB

- ZB, with input from TFJ and WB, conducted power analyses (not reported here) that confirmed that lognormal analyses of (corrected or uncorrected) RTs have higher statistical power, than standard normal analyses.
- TFJ implemented and conducted the Bayesian replication tests.
- TFJ wrote the paper with input from WB and ZB. WB wrote Appendix A.1 with input from TFJ.
- TFJ, WB, and ZB created all figures and tables.

References

Arai, M., & Mazuka, R. (2013). The development of Japanese passive syntax as indexed by structural priming in comprehension. *Quarterly Journal of Experimental Psychology*, 3–8. http://doi.org/10.1080/17470218.2013.790454

Arai, M., Van Gompel, R. P. G., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology*, *54*, 218–250. http://doi.org/10.1016/j.cogpsych.2006.07.001

Baayen, R. H., & Milin, P. (2010). Analyzing Reaction Times. *International Journal of Psychological Research*, *3*(2), 12–28.

Bates, D., Maechler, M., Bolker, B., & Steven Walker. (2015). lme4: Linear mixed-effects models using Eigen and S4. *Journal of Statistical Software*, *67*(1).

Bernolet, S., & Hartsuiker, R. J. (2010). Does verb bias modulate syntactic priming ? *Cognition*, *114*, 455–461. http://doi.org/10.1016/j.cognition.2009.11.005

Bicknell, K., & Levy, R. (2011). Why readers regress to previous words: A statistical analysis. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 931–936).

Bicknell, K., Slattery, T., Rayner, K., Johnson, M. D. L., Coggan, K. A., Sperlazza, J. R., … He, C. (2010). Correction for Levy et al., Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the National Academy of Sciences*, *107*(11), 5260–5260. http://doi.org/10.1073/pnas.1000194107

Boston, M. F., Hale, J. T., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, *2*(1), 1–12.

Burchill, Z., Bushong, W., & Jaeger, T. F. (2018). Wasted power in reading time analyses (and a simple fix).

Bushong, W., Burchill, Z., & Jaeger, T. F. (n.d.). Bayesian belief-updating captures syntactic adaptation in garden path environments.

Chang, F., Dell, G. S., & Bock, J. K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–72. http://doi.org/10.1037/0033-295X.113.2.234

Chang, F., Janciauskas, M., & Fitz, H. (2012). Language adaptation and learning: Getting explicit about implicit learning. *Linguistics and Language Compass*, *6*, 259–278. http://doi.org/10.1002/lnc3.337

Chodroff, E., Golden, A., & Wilson, C. (2016). An empirical and computational study of generalized adaptation to natural talker-specific VOT. In *LabPhon*.

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics*, *61*, 30–47. http://doi.org/10.1016/j.wocn.2017.01.001

Chodroff, E., & Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, *4*(s2). http://doi.org/10.1515/lingvan-2017-0047

Craycraft, N. (2014). Distributional learning in rapid syntactci adaptation. *Journal of Undergraduate Research*, 34–40.

D'Agostino, R. (1970). Transformation to Normality of the Null Distribution of g1. *Biometrika*, *57*(3), 679–681.

Dell, G. S., & Chang, F. (2013). The P-chain : relating sentence production and its disorders to comprehension and acquisition.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210. http://doi.org/10.1016/j.cognition.2008.07.008

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, *14*, 179–211.

Elman, J. L., Hare, M., & McRae, K. (1996). Cues, constraints, and competition in sentence processing. In M. Tomasello & D. I. Slobin (Eds.), *Beyond Nature-Nurture: Essays in Honor of Elizabeth Bates* (pp. 111–153). Psychology Press.

Farmer, T. A., Fine, A. B., & Jaeger, T. F. (2011). Implicit Context-Specific Learning Leads to Rapid Shifts in Syntactic Expectations. In L. In Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (CogSci11)* (pp. 2055–2060). Austin, TX: Cognitive Science Society.

Farmer, T. A., Fine, A. B., Yan, S., Cheimariou, S., & Jaeger, T. F. (2014). Error-Driven Adaptation of Higher-Level Expectations During Natural Reading. In P. In Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Meetings of the Cognitive Science Society* (pp. 2181–2186). Austin, TX: Cognitive Science Society.

Ferreira, F., & Clifton, C. J. (1986). The Independence of Syntactic Processing, *368*, 348–368.

Ferreira, V. S. (2003). The persistence of optional complementizer production: Why saying ‘“that”’ is not saying ‘“that”’ at all. *Journal of Memory and Language*, *48*, 379–398.

Fine, A. B., & Jaeger, T. F. (2011). Language comprehension is sensitive to changes in the reliability of lexical cues. In L. In Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society (CogSci11)* (pp. 925–930). Austin, TX: Cognitive Science Society.

Fine, A. B., & Jaeger, T. F. (2013). Evidence for Implicit Learning in Syntactic Comprehension. *Cognitive Science*, *37*(3), 578–591. http://doi.org/10.1111/cogs.12022

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

Fine, A. B., & Jaeger, T. F. (2016b). The Role of Verb Repetition in Cumulative Structural Priming in Comprehension. *Journal of Experimental Psychology : Learning, Memory, and Cognition*. http://doi.org/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. http://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 J. T. Hale (Ed.), *Proceedings of ACL: Workshop on Cognitive Modeling and Computational Linguistics* (pp. 18–26). Stroudsburg, PA: Association for Computational Linguistics.

Finegan, E., & Biber, D. (2001). Register variation and social dialect variation: The register axiom.

Frank, S. L., & Bod, R. (2011). Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, *22*(6), 829–834. http://doi.org/10.1177/0956797611409589

Fraundorf, S. H., & Jaeger, T. F. (2016). Readers generalize adaptation to newly-encountered dialectal structures to other unfamiliar structures q. *Journal of Memory and Language*, *91*, 28–58. http://doi.org/10.1016/j.jml.2016.05.006

Garnsey, S. M., Pearlmutter, N. J., Myers, E., & Lotocky, M. a. (1997). The Contributions of Verb Bias and Plausibility to the Comprehension of Temporarily Ambiguous Sentences. *Journal of Memory and Language*, *37*(1), 58–93. http://doi.org/10.1006/jmla.1997.2512

Garrod, S., & Pickering, M. J. (2009). Joint Action, Interactive Alignment, and Dialog. *Topics in Cognitive Science*, *1*(2), 292–304. http://doi.org/10.1111/j.1756-8765.2009.01020.x

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–64. http://doi.org/10.1016/j.tics.2010.05.004

Hale, J. T. (2001). A Probabilistic Earley Parser as a Psycholinguistic Model. In *NAACL '01*

Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (pp. 1–8).

Hare, M., Tanenhaus, M. K., & McRae, K. (2007). Understanding and producing the reduced relative construction: Evidence from ratings, editing and corpora. *Journal of Memory and Language*, *56*(3), 410–435. http://doi.org/10.1016/j.jml.2006.08.007

Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018). failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*(6), 864–877.

Jaeger, T. F., & Snider, N. (2008). Implicit learning and syntactic persistence : Surprisal and Cumulativity. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society (CogSci08)* (pp. 1061–1066). Austin, TX: Cognitive Science Society.

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: Syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition*, *127*(1), 57–83. http://doi.org/10.1016/j.cognition.2012.10.013

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition*, *124*(1), 66–71. http://doi.org/10.1016/j.cognition.2012.03.001

Kaschak, M. P. (2006). What this construction needs is generalized. *Memory & Cognition*, *34*(2), 368–379. http://doi.org/10.3758/BF03193414

Kaschak, M. P., & Glenberg, A. M. (2004). This construction needs learned. *Journal of Experimental Psychology. General*, *133*(3), 450–67. http://doi.org/10.1037/0096-3445.133.3.450

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Associationi*, *90*(430), 773–795.

Kleinschmidt, D. F., Fine, A. B., & Jaeger, T. F. (2012). A belief-updating model of adaptation and cue combination in syntactic comprehension. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society (CogSci12)* (pp. 599–604). Austin, TX: Cognitive Science Society.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception : Recognize the familiar , generalize to the similar , and adapt to the novel. *Psychological Review*, *122*(2), 148–203. http://doi.org/10.1037/a0038695

Kliegl, R., Masson, M. E. J., & Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition*, *18*(5), 655–681. http://doi.org/10.1080/13506280902986058

Kukona, A., Cho, P. W., Magnuson, J. S., & Tabor, W. (2014). Lexical interference effects in sentence processing: Evidence from the visual world paradigm and self-organizing models. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*, 326–47. http://doi.org/10.1037/a0034903

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language Cognition & Neuroscience*, *31*(1), 32–59. http://doi.org/10.1080/23273798.2015.1102299

Kuperman, V., Dambacher, M., Nuthmann, A., Kliegl, R., Kuperman, V., Dambacher, M., … Kuperman, V. (2010). The effect of word position on eye-movements in sentence and paragraph reading The effect of word position on eye-movements in sentence and paragraph reading, *0218*. http://doi.org/10.1080/17470211003602412

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

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–77. http://doi.org/10.1016/j.cognition.2007.05.006

Levy, R. (2011). Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics* (pp. 1–11).

Linzen, T., & Jaeger, T. F. (2015). Uncertainty and Expectation in Sentence Processing: Evidence

From Subcategorization Distributions. *Cognitive Science*. http://doi.org/10.1111/cogs.12274
Liu, L., Burchill, Z., Tanenhaus, M. K., & Jaeger, T. F. (2017). Failure to replicate talker-specific syntactic adaptation. In *Proceedings of CogSci*.
MacDonald, M. C. (1999). Distributional information in language comprehension, production, and acquisition: Three puzzles and a moral. In B. MacWhinney (Ed.), *The Emergence of Language* (pp. 177–196).
MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*(226), 1–16. http://doi.org/10.3389/fpsyg.2013.00226
MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology*, *24*(1), 56–98.
MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703. http://doi.org/10.1037//0033-295X.101.4.676
Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, *97*(2), 225–252. http://doi.org/10.1037/0033-295X.97.2.225
McDonald, S. A., & Shillcock, R. C. (2003). Low-level predictive inference in reading: the influence of transitional probabilities on eye movements. *Vision Research*, *43*(16), 1735–1751. http://doi.org/10.1016/S0042-6989(03)00237-2
McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. *Journal of Memory and Language*, *38*, 283–312.
Myslín, M., & Levy, R. (2016). Comprehension priming as rational expectation for repetition : Evidence from syntactic processing, *147*, 29–56. http://doi.org/10.1016/j.cognition.2015.10.021
Narayanan, S., & Jurafsky, D. (2004). A Bayesian Model of Human Sentence Processing. In *Proceedings of the twelfth annual meeting of the cognitive science society* (pp. 1–55).
Nicenboim, B., Logacev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in Psychology*, *7*(MAR). http://doi.org/10.3389/fpsyg.2016.00280
Pickering, M. J., McLean, J. F., & Branigan, H. P. (2013). Persistent structural priming and frequency effects during comprehension. *Journal of Experimental Psychology: Learning Memory and Cognition*, *39*(3), 890–897. http://doi.org/10.1037/a0029181
Qian, T., Jaeger, T. F., & Aslin, R. N. (2012). Learning to Represent a Multi-Context Environment: More than Detecting Changes. *Frontiers in Psychology*, *3*(July), 228. http://doi.org/10.3389/fpsyg.2012.00228
R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.
Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163.
Reitter, D., Keller, F., & Moore, J. D. (2011). A computational cognitive model of syntactic priming. *Cognitive Science*, *35*(4), 587–637. http://doi.org/10.1111/j.1551-6709.2010.01165.x
Roland, D., Dick, F., & Elman, J. L. (2007). Frequency of Basic English Grammatical Structures: A Corpus Analysis. *Journal of Memory and Language*, *57*(3), 348–379. http://doi.org/10.1016/j.jml.2007.03.002
Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika*, *70*(2), 377–381. http://doi.org/10.1007/s11336-005-1297-7
Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*(2), 195–223. http://doi.org/10.3758/BF03257252
Ryskin, R. A., Qi, Z., Duff, M. C., & Brown-Schmidt, S. (2017). Verb biases are shaped through lifelong learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(5), 781–794. http://doi.org/10.1037/xlm0000341

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–64. http://doi.org/10.3758/PBR.17.4.443

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–19. http://doi.org/10.1016/j.cognition.2013.02.013

Tabor, W., & Tanenhaus, M. K. (1999). Dynamical Models of Sentence Processing. *Cognitive Science*, *23*(4), 491–515. http://doi.org/10.1207/s15516709cog2304_5

Tabossi, P., Spivey-Knowlton, M. J., McRae, K., & Tanenhaus, M. K. (1994). Semantic effects on syntactic ambiguity resolution: Evidence for a constraint-based resolution process.

Thothathiri, M., & Snedeker, J. (2008). Give and take: syntactic priming during spoken language comprehension. *Cognition*, *108*(1), 51–68. http://doi.org/10.1016/j.cognition.2007.12.012

Tooley, K. M., Swaab, T. Y., Boudewyn, M. a., Zirnstein, M., & Traxler, M. J. (2013). Evidence for priming across intervening sentences during on-line sentence comprehension. *Language, Cognition and Neuroscience*, *29*(3), 289–311. http://doi.org/10.1080/01690965.2013.770892

Traxler, M. J. (2008). Lexically independent priming in online sentence comprehension. *Psychonomic Bulletin & Review*, *15*(1), 149–155. http://doi.org/10.3758/PBR.15.1.149

Trueswell, J. C., Tanenhaus, M. K., & Garnsey, S. M. (1994). Semantic Influences on Parsing: Use of Thematic Role Information in Syntactic Ambiguity Resolution. *Journal of Memory and Language*, *33*, 285–318.

Vasishth, S., & Gelman, A. (2017). The statistical significance filter leads to overconfident expectations of replicability, *103*(July), 151–175. http://doi.org/10.17605/OSF.IO/HBQCW

Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457–1475. http://doi.org/10.1037/a0036731

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E. J., & Brown, S. (2007). On the Linear Relation Between the Mean and the Standard Deviation of a Response Time Distribution. *Psychological Review*, *114*(3), 830–841. http://doi.org/10.1037/0033-295X.114.3.830

Weiner, E. J., & Labov, W. (1983). Constraints on the agentless passive. *Journal of Linguistics*, *19*(01), 29. http://doi.org/10.1017/S0022226700007441

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. http://doi.org/10.1016/j.cogpsych.2008.08.002

Wilson, M. P., & Garnsey, S. M. (2009). Making simple sentences hard: Verb bias effects in simple direct object sentences. *Journal of Memory and Language*, *60*(3), 368–392. http://doi.org/10.1016/j.jml.2008.09.005

Wood, S. N. (2016). mgcv: GAMs with GCV/AIC/REML smoothness estimation and GAMMs by PQL. *R Package Version 1.8-16.*

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall.

Yan, S., Farmer, T., & Jaeger, T. F. (2018). Adaptation of syntactic expectations affects re-analysis rather than initial surprisal.

Appendix A
Expectation adaptation as Bayesian belief-updating

Appendix A.1 presents the belief-updating model. Appendix A.2 summarizes the predictions of this model for block-level analyses like those conducted in the original study by Fine et al. (2013) and the replication by Harrington Stack et al. (2018). In the main text, we present trial-level analyses instead, as they make the link between the belief-updating model and predicted changes in surprisal more transparent.

A.1 The belief-updating model

The beta-binomial belief-updating model presented in Fine et al. (2010) starts from the assumption that comprehenders have prior expectations about the probability of two alternative syntactic structures, $p(MV) = \theta$ and $p(RC) = 1 - \theta$. Comprehenders are taken to have some degree of *uncertainty* about θ. This uncertainty over expectations is represented by the beta distribution with parameters α and β (here, representing MVs and RCs). Intuitively, these parameters can be thought of as pseudocounts of the number of times the comprehender has encountered each structure. As the sum of α and β gets larger, the uncertainty about $p(MV)$ and $p(RC)$ decreases—the beta distribution gets tighter and more peaked. Examples of the beta distribution and changes in that distribution are shown in Panel B of Figure 4 in the main text.

Under the assumption of a beta conjugate prior, the resulting uncertainty over $p(MV)$ at each point in the experiment is described as a function of α and β:

$$p(\theta) = Beta(\theta|\alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where both parameters can be understood as the sum of the *prior* pseudo counts for the respective structure (α_0 and β_0) and the number of observations of that structure within the experiment so far. For example, for a comprehender who has encountered 3 MVs and 10 RCs so far, the posterior beliefs about θ are:

$$p(\theta|\text{RC}) = Beta(\theta|\alpha + 3, \beta + 10) \propto \theta^{\alpha+2}(1-\theta)^{\beta+9}$$

In short, each time a comprehender encounters a structure, they update that structure's pseudocount. This posterior then becomes the prior for the next trial a comprehender encounters a relevant structure.

To approximate beliefs about all other structures that can occur in the same ambiguous environments as MVs and RCs, we modify the beta distribution by introducing a third parameter γ. Since comprehenders never encounter critical sentences in any structure other than RC and MV, this parameter never changes throughout the course of the experiment. We obtain expectations for $p(MV)$ and $p(RC)$ by marginalizing over the beta distribution, giving us:

$$p(MV) = {}^{\alpha}/_{\alpha+\beta+\gamma}, \text{ and } p(RC) = {}^{\beta}/_{\alpha+\beta+\gamma}$$

To further reduce the degrees of freedom of this model, we hold the relative proportion of α_0, β_0, and γ_0 constant at 0.008, 0.67, and 0.322, respectively (based on the relative frequency of MV and RC continuations in the ambiguous context in large databases of English, Roland et al., 2007). This allows us to have only a single degree of freedom, $\tau = \alpha_0 + \beta_0 + \gamma_0$.

A.2 Block-level predictions that correct for the differences in design

We can use the belief-updating model to calculate the average RC and MV surprisal for the three blocks of F13's and HS18's designs by aggregating over trial-level predictions (those shown

in Figure 2). From this, we then calculate the predicted differences in surprisal relevant to Predictions 1-3 for each of the two designs.

Panel A of Figure 13 shows the predicted mean surprisal for all three blocks of both the original F13, and the revised HS18, design. Consider, for example, the number of RCs in Block 1 for the RC-first group in Figure 2. In the revised design of HS18, there are now 16 more RCs in this block with increasingly smaller surprisal. As a consequence, the mean surprisal across all RCs in Block 1 of this revised design will be substantially smaller (3.21 bits for $\tau = 100$) than in the original experiment (4.84 bits). This in turn means that, for example, the predicted reduction in surprisal from Block 1 to Block 2 of the RC-first group differs between the original (predicted: 2.32 bits reduction in RC surprisal for $\tau = 100$) and the revised design (predicted: 1.42 bits reduction). Panel B of Figure 13 shows the predicted difference in surprisal relevant to the block-level assessment of Predictions 1-3. These are the differences that the block-level analyses in HS18 are aiming to replicate.

The predicted differences in Panel B always point in the same direction. They do so for all values of τ, and for both the original and the revised design. However, the size of the predicted difference in surprisal differs between the designs. For Predictions 1 and 3, which are both predictions about differences between groups, we find only small differences between the two design. With regard to Prediction 1, the revised design results in a minimally larger surprisal differences between groups than the original design, regardless of τ. With regard to Prediction 3, the original design results in minimally larger differences between groups for small τs, whereas the revised design results in minimally larger differences for large τs. For $\tau = 100$, the two designs do not differ much. For Prediction 2, which pertains to the difference in RC surprisal between Blocks 1 and 2, the original design in F13 is predicted to result in substantially larger surprisal differences, except for very large τs. Finally, Figure 13 shows that the predicted differences in surprisal tend to be larger for RCs than MVs, which should make it easier to detect changes in the magnitude of garden path effects for RCs (Predictions 2 and 3), compared to changes for MVs (Prediction 1).

*** Figure 13 APPROXIMATELY HERE ***

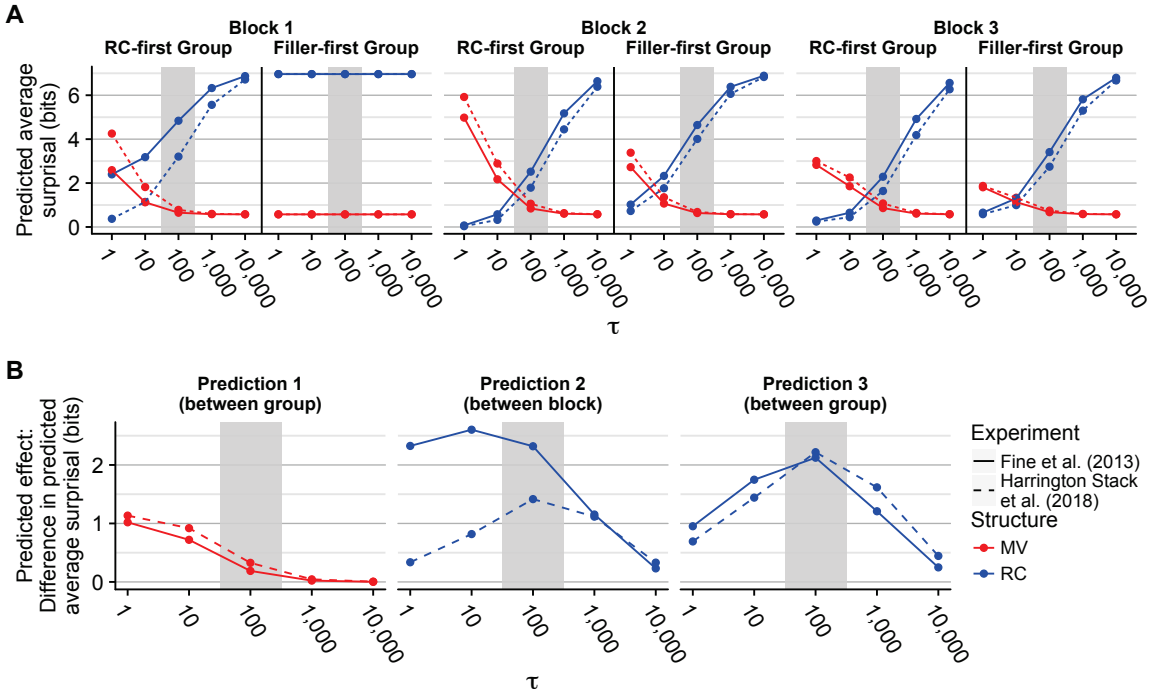

Figure 13 Predicted effects depending on the strength of prior beliefs (τ). In all panels we highlight $\tau = 100$, the value we committed to a priori in order to reduce researchers' degrees of freedom. Panel A: Predicted mean surprisal of RCs and MVs in Block 1-3 for both the original Experiment 2 in Fine et al. (2013) and Experiment 2 in Harrington Stack et al. (2018). Panel B: Predicted differences in the surprisal of the relevant structure for the three block-level analyses that assess Predictions 1-3, respectively.

We emphasize that the predicted effects in Panel B of Figure 13 are not to be equated with the statistical power to detect these effects. The statistical power also depends on the by-subject, by-item, and residual variance of RTs, and these variances differed substantially between both blocks and experiment (F13 vs. HS18). Specifically, the RTs in HS18 exhibited consistently larger variability (see Appendix B.2). For example, for Block 3 (relevant to Prediction 1) the residual variance of the corrected log-RTs in HS18 was 27% larger, compared to F13. Further, additional differences in stimuli design can affect the expected power. One such difference that we expect to substantially reduce the power for Prediction 1 in HS18's data is described in Appendix C.2. It is not trivial to weigh these various differences between F13 and HS18 against each other to obtain adequate power estimates for the two studies.

Appendix B
De-confounding task adaptation and the adaptation of syntactic expectation

This appendix provides supplementary information that motivates the GAMM-based procedure we used to correct log-transformed RTs for linear effects of word length effect and non-linear effects of trial position. We begin by motivating why our main analyses focus on (corrected) log-transformed RTs.

B.1 Reading times in both data sets are not normally distributed

The analyses employed by both F13 and HS18 assume that RTs are normally distributed. This simplifying assumption is widespread in the field (but see, e.g., Fine & Jaeger, 2016a; Nicenboim, Logacev, Gattei, & Vasishth, 2016), as reflected in the most common approaches to the analysis of RT data from self-paced reading experiments (e.g., the use of ANOVA or linear mixed models with raw or length-corrected RTs). However, this assumption is unwarranted. To begin with, the assumption of normality is unlikely to hold *prima facie* because of the inherent 'soft' floor for RTs. Indeed, RTs form self-paced reading experiments tend to have heavily skewed distributions (Figure 14), with strong correlations between means and standard deviations (Figure 15).

*** Figure 14 APPROXIMATELY HERE ***

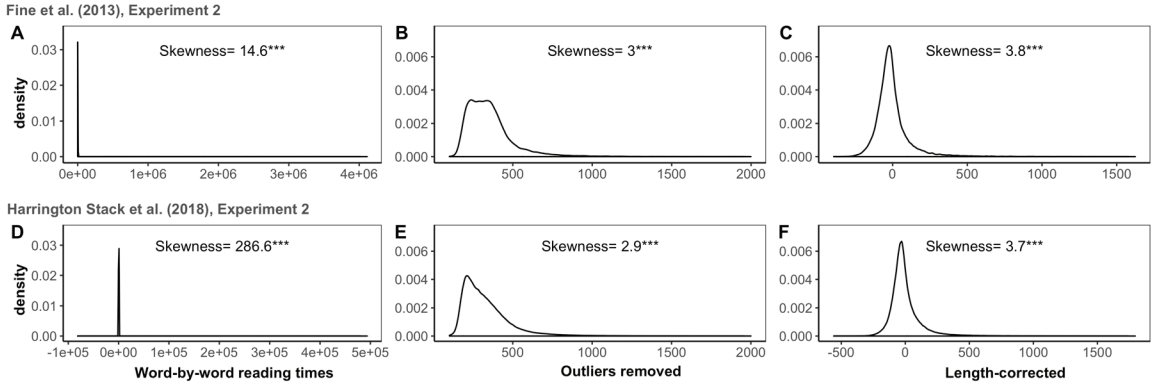

Figure 14: Marginal density of word-by-word reading times (RTs) using all data from Fine et al. (2013), Experiment 2 (Panel A-C) and Harrington Stack et al. (2018), Experiment 2 (Panel D-F). Panels A & D: Raw RTs. Panels B & E: Raw RTs after outlier exclusion as employed by both Fine et al. (2013) and Harrington Stack et al. (2018): 100 <= RT <= 2000 (< .5% data loss). Panels C & F: Length-corrected RTs, which also correct for individual differences in reading speeds between participants. The skewness

parameter is calculated as $E[((X-\mu)/\sigma)^3]$. *All distributions are significantly skewed based on D'Agostino test* (D'Agostino, 1970).

*** Figure 15 APPROXIMATELY HERE ***

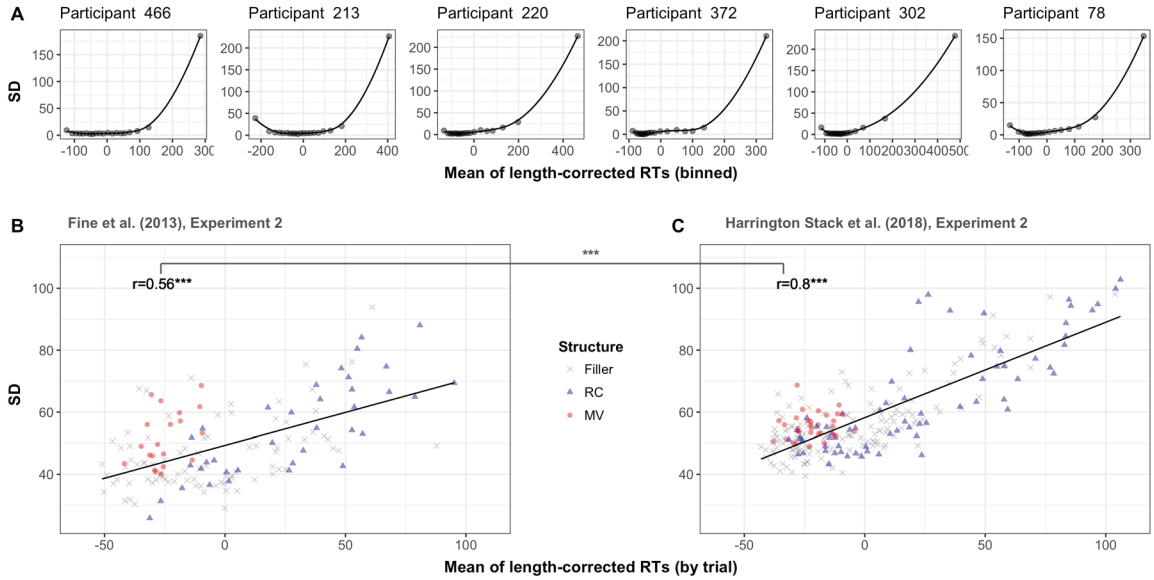

Figure 15: Correlation between means and standard deviations (SD) of outlier- and length-corrected word-by-word reading times (RTs, cf. the rightmost column in Figure 14). Panel A: Six randomly drawn participants from Harrington Stack et al. (2018), Experiment 2. RTs were binned into 20 quantiles with equally many data points each. Panel B: Correlation across participants in Fine et al. (2013), Experiment 2. Each data point represents one sentence (trial). Shape and color show the sentence structure. Panel C: Same as Panel B but for Harrington Stack et al. (2018), Experiment 2. Both for Panel B and C, RT means are a highly significant linear predictor of log-transformed SDs (p<.0001), and this relation is significantly stronger for Panel C, compared to Panel B (p<.001). Additionally, SDs were overall higher in Panel C (p<.0001), reducing the statistical power of the Harrington Stack et al.'s experiment.

If the self-paced RTs were to follow a normal distribution, the correlation between means and standard deviations should be zero (the mean and SD of a normal distribution are independent). Cleary, this is not the case. The RTs in F13 and HS18 (and other self-paced reading experiments we have examined) strongly violate the assumption of the homogeneity of variance. Rather, the mean and standard deviation of RTs exhibit strong correlations, as would be expected from, for example, log-normal or exponential Gaussian distributions (Wagenmakers & Brown, 2007).[8] It is largely an open question how much the assumption of normality and the resulting violation of the assumption of the homogeneity of variance affects analyses of reading times. Systematic evaluations of different types of models of RT distributions are to the best of our knowledge still lacking (unlike for simpler reaction time tasks, Kliegl, Masson, & Richter, 2010; Rouder, 2005; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Wagenmakers & Brown, 2007). In an ongoing large simulation project, we have found that log-normal analyses of self-paced reading data tend to have higher power and lower Type I error rates, than analyses based on the assumption of normality (Burchill, Bushong, & Jaeger, 2018) (see also Baayen & Milin, 2010; Nicenboim et al., 2016). This

[8] We note, however, that the quadratic relation sometimes observed between RT means and standard deviations is not expected under log-normal distributions. This might distinguish reading times from reaction times in simpler tasks (see references in text).

also was confirmed for both the F13 and HS18 data (not shown here). This motivates our choice to correct log-RTs, rather than untransformed RTs.[9]

B.2 Task adaptation: consequences of doubling the number of trials per block

Moving-window self-paced reading experiments like those employed by F13 and HS18 are subject to strong task adaptation effects. Unlike in everyday reading, participants only see one word at a time. The other words of a sentence are masked by dashes. Reading advances button press by button press. As participants get used to this unfamiliar task, reading times change drastically over the initial trials of an experiment. Figure 16 shows a decrease of almost 150ms per word from 450 to 300ms over the course of an experiment in HS18. Similar effects of task adaptation have been observed previously, and are the reason why previous analyses from our lab either statistically controlled for trial position (Craycraft, 2014; Fine & Jaeger, 2016a) or held it constant for critical comparisons (Fine et al., 2013; Fraundorf & Jaeger, 2016).

As can be seen in Figure 16, the difference in the number of sentence stimuli in HS18's vs. F13's design means that the two experiments are affected differently by task adaptation. And this in turn means that it is problematic to compare block-level results across the two experiments (this includes the type of argument Harrington Stack and colleagues present based on differences in block-level significance between the two experiments). This motivates the use of a GAMM-based residualization approach to correct (log-transformed) RTs for effects of word length and effects of trial position. GAMMs rather than linear mixed models were used to fully account for the effect of trial position, which is clearly non-linear (see also Figure 16).[10] We also fitted a more complex model in which both trial position and word length were allowed to have non-linear effects. However, both for the F13 and the HS18 data, word length effects were clearly linear. We thus decided to employ corrected (log) RTs from the simpler GAMM, in which only trial position was allowed to have non-linear effects.

*** Figure 16 APPROXIMATELY HERE ***

[9] We note that the variance of RTs was significantly larger in HS18's data. This is visible in Figure 15, and was confirmed in all mixed-effects analyses (e.g., the estimated residual variance of standard word length-corrected RTs was between 12-27% larger in SETA, compared to F13). This means that HS18's statistical power is likely significantly lower than their power analyses suggest, though likely still larger than F13's (because of the much larger number of participants).

[10] We note that task adaptation did not confound the original analyses of Predictions 1 and 3 in F13: *within* each experiment, adaptation to the task of self-paced reading cannot confound analyses for these predictions, as trial position is held constant for those comparisons. Within each experiment, only analyses of Prediction 2 were potentially confounded (as acknowledged in both Fine et al., 2013 and Harrington Stack et al., 2018). However, for comparisons *across* experiments (including comparisons of significance patterns), this is no longer the case because of the doubling of trials per block in the new design in HS18.

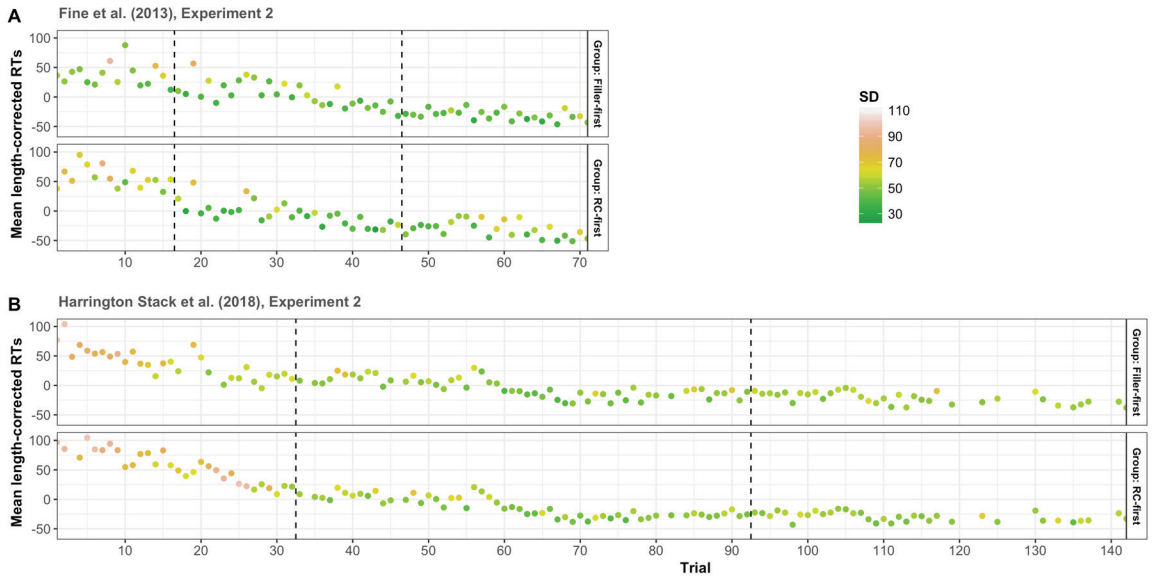

Figure 16: The standard deviation (SD) of length-corrected reading times (RTs) decreases as a function of trial order throughout the experiment, violating the assumption of homogenously distributed residuals if RTs are analyzed with standard Gaussian link functions. Panel A: Fine et al. (2013), Experiment 2. Panel B: Harrington Stack et al. (2018), Experiment 2. Data were obtained in the same way as in Figure 15. Block structure (Block 1-3) is indicated through dashed vertical lines.

B.3 Residualization models used to correct RTs

All models for RT correction were fit to only data from fillers, so as to avoid overfitting to the critical trials. Models were fit separately to the data from F13 and HS18 (the resulting fits were very similar). For the GAMM-corrected log-RTs, we regressed log(RT) ~ 1 + WordLength + s(TrialPosition) + s(Subject, bs = "re") + s(Subject, TrialPosition, bs = "re") + s(Subject, WordLength, bs = "re"), where WordLength and TrialPosition were first centered. For standard word length-corrected RTs, we regressed RT ~ 1 + WordLength + (1 + WordLength | Subject).

Corrected RTs for the disambiguation region of the critical items (RCs and MVs) were obtained by subtracting the RT predicted based on the respective correction model from the actual RT. Prior to analysis, we obtained average corrected per-word RTs for the disambiguation region for each unique combination of subject and item, so as to avoid Type I error inflation due to auto-correlations between the RTs of the three words in the disambiguation region. This was the case both for the analyses of Predictions 1-3 (for the Bayesian replication test under the assumption of identical predictions for F13 and HS18) and for the replication tests analyses that corrected for the differences in design between F13 and HS18.

Appendix C
Differences between the original and replication study and considerations for future studies

Table 1 provides a summary of differences between F13's and HS18's Experiment 2. that might be relevant to their interpretation. HS18 is intended as a close and high-powered replication of F13's Experiment 2. As summarized in the main text, some differences between the two studies are *predicted* to result in different results. Several other changes made to the design in HS18 are likely to work against the goal to maximize statistical power. We list these differences so that future work can take them into consideration.

We discuss two of these differences in some detail, because we consider them particularly relevant for future work. The first difference relates to differences in the structure of fillers (Appendix C.1). The second difference relates to differences in the way that MVs were disambiguated (Appendix C.2). Both differences are predicted to reduce—potentially by a

substantial amount—the power of HS18's Experiment 2 to detect the effects of Prediction 1. Finally, we discuss additional factors that are predicted to influence the degree of expectation adaptation in an experiment (Appendix C.3).

*** Table 1 APPROXIMATELY HERE ***

Change from F13 to HS18	Consequence	Expected to affect
Changes that are unambiguously improvements		
Doubled # of participants	• Increases number of observations that go into group-level RT estimates	• Increases power for effects of exposure group, i.e., in particular for Predictions 1 and 3.
Changes that change predicted effects		
Doubled # of RCs in Block 1	• Changes predicted differences in surprisal between exposure group	• Changes predictions and predicted effect sizes for Predictions 1, 2, & 3
Doubled # of RCs in Block 2	• Changes predicted differences in surprisal between exposure group	• Changes predictions and predicted effect sizes for Predictions 1, 2, & 3
Doubled # of MVs in Block 3	• Changes predicted differences in surprisal between exposure group	• Changes predictions and predicted effect sizes for Prediction 1
Doubled # of critical items	• Increases number of observations that go into each block-level RT estimate • Design predicts largest effects at *beginning* of each block. Longer blocks reduce predicted between-group differences in block-level averages.	• Changes power for block-level analyses (everything else being equal)
Doubled # of stimuli per block	• Changes how much, in particular, Block 1 and 2 are affected by task adaptation • Changes mean RTs per block (and relative differences in SD of residuals across blocks)	• Changes power (and possibly Type I error) of standard analyses like those employed in both papers for all predictions
	• Might increase proportion of attentional blinks or disengagement with task towards end of experiment	• Decreases power, in particular, for Prediction 1
Changes likely to affect ability to detect effect		
Differences in the ambiguous verb forms used in critical stimuli	• Changes the predicted magnitude of the *a priori* ambiguity effect	• Changes power for all predictions. E.g., reduction in ambiguity effect (Prediction 2) is only detectable for large *a priori* ambiguity effects
Differences in the disambiguation region used in critical stimuli	• Clarity of disambiguation signal seems to affect effect sizes (Craycraft, 2014 vs. Fine et al., 2013).	• Changes power for all predictions

	• 'Disambiguation' towards MV requires particular care.	• Changes power for Prediction 1
Different # of words *following* the disambiguation point	• Reduces disambiguation signal for MVs (the only signal is the distance to the end of the sentence).	• Decreases power for Prediction 1.
Changes reducing comparability, but *unlikely* to affect the results		
Experiment conducted over MTurk rather than in the lab	• Increased noise in RT measures (empirically confirmed, cf. Appendix B.2). We also found that the by-subject, by-item, and residual variances was often about 10-40% larger in HS18 compared to F13 (depending on the specific analysis).	• Decreases power, but likely not much, and large number of subjects should more than make up for it.
	• More heterogeneous subject population.	• Might increase generalizability of result
Other differences in the lexicalization of stimuli		
Repetition of ambiguous verbs within Block 1	• Could theoretically affect verb-specific adaptation, but Fine and Jaeger (2016) found no evidence for verb-specific adaptation.	
Repetition of ambiguous verbs within Block 2	• Could theoretically affect verb-specific adaptation, but Fine and Jaeger (2016) found no evidence for verb-specific adaptation.	
Repetition of ambiguous verbs between Block 2 (RCs) and 3 (MVs)	• Could theoretically affect verb-specific adaptation. Fine and Jaeger (2016) did not investigate consequences of observing same verb with *different* structures.	• If anything, this should arguably *in*crease the effect for Prediction 3

Table 1 Summary of differences between the original Experiment 2 in Fine et al. (2013) and Experiment 2 in Harrington Stack et al. (2018) intended as a replication of the original experiment.

C.1 Syntactic structure of fillers is predicted to affect expectation adaptation

Recall the wording of Prediction 3 that changes in the expectations for a structure "…should only depend on observations that speak to the relative frequency of [that structure] in the relevant contexts". In other words, any and all observations that comprehenders take to be informative about the probability of a structure in the current input are assumed to affect the expectations for that structure. One consequence of this is that it matters whether fillers contain environments in which the structures of interest (here: RCs or MVs) *could* occur. If a comprehender observes a syntactic environment in which the structures of interest could have occurred (here: an ambiguous verb form like *warned* that could be interpreted as either passive participle or simple past tense form), but the structure of interest was not observed, this means that a *competing* structure has been observed (here: other structures that could occur after ambiguous verb forms like "warned"; indicated with green lines in Figure 4, Panel B-C in the main text). Both HS18 and F13 avoided reusing in fillers any of the ambiguous verbs that occurred in critical items. Both HS18 and F13 also avoided fillers with the RC/MV ambiguity. However, F13 went one step further: "fillers never contained […] verb forms that (in their syntactic context) were ambiguous between a past tense and past participle interpretation" (Fine et al., 2013, p. 5). HS18 did not explicitly implement this constraint (Duane

Watson, p.c.). Future experiments should keep in mind that the inclusion of such instances—even in fillers—is predicted to affect participants' expectations, and thus the surprisal they will experience. For example, for the design of Experiment 2 in F13 and HS18, our model predicts that fillers of this type *reduce* the effect size for Prediction 1.

Prediction 3 also has a second, less immediately obvious, consequence. If comprehenders have implicit knowledge of statistical dependencies between the occurrence probabilities of different types of structures, the observation of one structure would allow comprehenders to adapt their beliefs about the probability of *all* structures whose occurrence this structure is informative about. An example from adaptation during speech perception serves to illustrate this point. The primary phonetic cue to voicing contrasts between syllable-initial plosives in English (/b/-/p/, /d/-/t/, /g/-/k/) *covary across talkers* (Chodroff & Wilson, 2017, 2018). And, critically, recent perception experiments suggest that listeners have implicit knowledge of this covariation, and take it into account during the recognition of speech categories: when listeners are exposed to an unfamiliar talkers pronunciations of /b/-/p/ and /g/-/k/, they can readily extrapolate to how that talker will pronounce /d/-/t/ (Chodroff, Golden, & Wilson, 2016).

We are not aware of similarly systematic experiments on, for example, the covariation of syntactic subcategorization preferences across talkers. There is, however, evidence that suggests that such covariation exists. For example, Finegan and Biber (2001) investigate a variety of reduction phenomena across different genres, registers, and topics. Finegan and Biber find a considerable amount of covariation between these phenomena across the different types of texts they investigate. Results like this suggest that the *objective* statistics of different syntactic structures can covary. They leave open whether comprehenders have implicit knowledge of this syntactic covariation, and whether they draw on this knowledge during language understanding (as they seem to do for phonetic covariation). For experiments on syntactic adaptation, this means experimenters should choose their fillers carefully—either minimizing their informativity about the target structures, or computationally modeling the potentially confounding effect of that informativity.

C.2 The importance of a clear disambiguation signal

The second consideration pertains to the strength of disambiguation signal, and the lexical material employed both in the ambiguous region and the disambiguation region. Expectation adaptation predicts changes in the size of ambiguity effects. Thus, predicted reductions in the ambiguity effect, for example, should be easier to detect for *a priori* larger ambiguity effects. One variable known to affect the size of the ambiguity effect is the subcategorization bias of the ambiguous verb form (Hare et al., 2007). Specifically, critical items with ambiguous verb forms that are strongly biased against the structure of interest should make it easier to detect expectation adaptation. Another mediator of the expected effect size is the clarity of the disambiguation signal. The less uncertainty about the correct syntactic parse comprehenders are left with after reading the first word of the disambiguation region, the easier it should be to detect a clear ambiguity effect.

In this context, we note an important asymmetry between the feasibility of testing changes in the size of ambiguity effects for MVs (Prediction 1) and RCs (Prediction 2): it more straightforward to obtain a clear disambiguation signal for RCs, than for MVs. And, as it turns out, this problem was exacerbated in HS18, compared to F13. As this might have contributed to HS18's failure to replicate Prediction 1, we elaborate.

Specifically, RC continuations as in (1b) provide comprehenders with a strong disambiguation signal—an observation of "conducted" following "The experienced soldiers warned about the dangers …" essentially requires readers to give up on the hypothesis that "warned" is the matrix verb (unless one considers "conducted" a typo or grammatical mistake). This is, of course, the very reason why sentences like (1b) enjoyed a stellar career as garden-path sentences. Critically, MV continuations as in (1a) do *not* provide an equally strong disambiguation signal. That is, even if we imagine a comprehender who strongly expects to be in an RC parse (e.g., because of having been exposed to many RCs and no MVs), the disambiguation region in (1a)—intended to indicate an MV parse—is actually perfectly compatible with a continued RC interpretation: "before the midnight raid" can be a modifier to the noun phrase "the dangers" or the verb phrase "warned about

the dangers" even under the RC parse.[11] This property of MV 'disambiguation' makes it much harder to detect ambiguity effects for MVs, compared to RCs. And this, in turn, makes it difficult to reliably test Prediction 1, which are about *changes* in the size of ambiguity effect. In fact, the only unambiguous disambiguation signal is provided by the period at the end of the sentence (and thus outside the three-word disambiguation region typically employed in studies on the RC/MV ambiguity). This problem is not limited to sentence (1). All stimuli of previous studies on the RC/MV ambiguity that we reviewed have this property (which was not critical to the goals of studies on the *existence* of garden paths but matters when we test predictions about changes in garden path effects). Indeed, it seems to be difficult, if not impossible, to create stimuli that avoid this asymmetry for the RC/MV ambiguity (but see Craycraft, 2014, for improvements over the stimuli used in earlier studies).

It is primarily due to a combination of a specific property of the self-paced reading paradigm employed by F13 as well as a specific property of their stimuli that *any* signal for Prediction 1 is expected at all during the three-word disambiguation region. The moving-window word-by-word self-paced reading F13 employed masked all words both *before* and after they were shown. As is common for this paradigm, words were masked with "-" and spaces between the words were visible to readers. This means that participants know how many words are yet to be revealed before the sentence ends. And, critically, the RC and MV stimuli in F13's experiment all had exactly four words remaining in the sentence starting at the first word of the disambiguation region (as in (1): "… conducted/before the midnight raid"). Since a hypothetical strongly RC-biased comprehender who expects the sentence onset "The experienced soldiers warned about the dangers …" to have an RC parse still requires a matrix verb for a well-formed sentence, this means that each of the remaining four words further increases the probability that the sequences of words seen so far must be interpreted as an MV. This is similar to the case of spoken language understanding, where prosodic information can be used to anticipate clause or sentence boundaries. The fact that F13 analyzed the first three of these four words (following earlier garden-path experiments on the RC/MV ambiguity MacDonald et al., 1994) is thus likely to have given them about as much statistical power as possible in detecting an ambiguity effect.[12]

HS18 employ the same word-by-word moving window masked procedure. Their stimuli, however, did not share the second property of F13's experiment: across ambiguous MV stimuli, between 3-8 words followed the region intended to be ambiguous (mean = 4.9, SD = 1.48; cf. mean = 4, SD = 0 for F13). Seven (37%) of the 19 MVs in HS18 (1 was removed by HS18 for other reasons) had more than 4 words following the ambiguous region. Additionally, one (5%) of the ambiguous MVs had only 3 words following the ambiguous region, which meant that the large sentence-final reading time increases were part of the disambiguation region. In short, 8 out of the 10 additional MVs designed by HS18 had properties that made it hard or impossible to detect an ambiguity effect on them.[13] The true statistical power of HS18's experiment with regard to Prediction 1 is thus likely to be substantially lower than suggested by their power analysis. Perhaps most important is that design changes like this can actually reduce the predicted effect size even compared to the original experiment with half as many items. Consider, for the sake of this

[11] This becomes apparent if we consider an unambiguous RC version of the sentence in (1a): "The experienced soldiers who were warned about the dangers before the midnight raid …". This sentence can be completed while leaving the RC interpretation intact ("… conducted another raid.").

[12] We note that inclusion of the fourth and final word is not necessarily expected to further increase the statistical power to detect an effect: self-paced reading exhibits very large increases in reading times at the end of sentences (Kuperman et al., 2010), as also evidenced in F13's experiments (Fine et al., 2013, Figure 4). Self-paced reading experiments tend to be intentionally designed such that critical regions do not involve the last word, as was also the case for Fine et al. (2013).

[13] Additionally, we found that words constituting the three-word disambiguation region differed *within*-item for at least three (35%) of the 10 new MV items in HS18 (Item 158, 160, 164). The final region differed in these items, and in one additional item (Item 171). In all cases, each of the two different versions of each item were seen by approximately equally many participants. We communicated these problems to HS18 and the revised data and scripts on their OSF site now exclude these items.

argument, that the 8 problematic MVs had no detectable ambiguity effect, and the remaining 12 MVs had the predicted ambiguity effect, and thus a predicted between-group difference in the predicted surprisal of .326 bits (left-most panel in Figure 13-B). This would leave 40% fewer items to detect the predicted between-group difference in MV surprisal.

In short, several properties of the RC/MV ambiguity might limit researchers' ability to detect expectation adaptation on MVs (see also Figure 12 in the main text). Future studies on expectation adaptation might consider alternative garden path structures with clearer disambiguation points.

C. 3 Additional factors that are predicted to affect the speed of expectation adaptation

We briefly illustrate two aspects that might affect the degree to which expectation adaptation is observed in an experiment, that are not captured by the simple belief-updating model we have employed here.

The first point concerns the interpretation of τ. Above we have sometimes referred to τ as the strength of prior beliefs. This follows the most common nomenclature for this type of parameter but is potentially confusing. Rather, we propose that τ captures how *relevant* comprehenders take previous experience to be to the current environment. This perspective highlights that the relevance of prior expectations to expectations for the present environment—or even the relevance of *some* prior expectations based on *specific* previous experience that is considered similar to the present environment—itself needs to be inferred. For example, if comprehenders are capable of learning environment-specific expectations and remembering them, a full model of expectation adaptation might require inferences about relevant similar experiences.

Additionally, adaptation to a new unfamiliar environment (such as an experiment) might be faster in the presence of additional cues that prior expectations might not transfer to this environment (cf. wonky world experiments or artificial language learning). Conversely, the more likely the current environment is to be inferred to follow prior expectations, the more expectation adaptation should be constrained by those prior expectations. Indeed, there is abundant evidence that some expectations are very hard or impossible to overcome even with prolonged exposure (e.g., second language acquisition in adults, as recently reviewed in Pajak, Fine, Kleinschmidt, & Jaeger, 2016). While limitations of this type are often attributed to cognitive limitations, it is worth pointing out that it can be rational to not adapt to quickly—avoiding reduced efficiency due to volatile expectations that mismatch the statistics of the input. Rational adaptation should itself rely on prior knowledge about the volatility of the relevant statistics and the factors that predict changes in those statistics (for relevant discussion, see Kleinschmidt & Jaeger, 2015; Qian, Jaeger, & Aslin, 2012; Yu and Cohen, 2008).

Future comparisons of best-fitting τs across different types syntactic structures should thus take into account that expectation adaptation might be more rapid for statistics for which comprehenders' previous experience suggests that they vary across linguistic environments (REF). For example, if the frequency of passive structures varies strongly between environments but the frequency of different ditransitive structures does not, adaptation to changes in passive frequency should be faster.