**Online and offline measures of statistical learning reflect sensitivity to distinct features in a continuous speech sequence**

Ava Kiai1 & Lucia Melloni1,2

1Neuroscience Department, Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany

2Department of Neurology, New York University Langone School of Medicine, New York, USA

Contents

[**Abstract** 1](#_Toc49355881)

[**Introduction** 2](#_Toc49355882)

[**Experiment 1** 3](#_Toc49355883)

[**Method** 3](#_Toc49355884)

[**Results** 5](#_Toc49355885)

[**Discussion** 10](#_Toc49355886)

[**Experiment 2** 11](#_Toc49355887)

[**Method** 11](#_Toc49355888)

[**Results** 12](#_Toc49355889)

[**Discussion** 14](#_Toc49355890)

[**Coding of Structural Features in Online Target Detection** 15](#_Toc49355891)

[**Conclusion** 16](#_Toc49355892)

[**Acknowledgements** 16](#_Toc49355893)

[**Contributions** 16](#_Toc49355894)

[**References** 16](#_Toc49355895)

[**Supplementary Materials** 19](#_Toc49355896)

**Keywords**: statistical learning, implicit vs. explicit learning, sequence representation

## **Abstract**

Here, we replicated the online target detection task and found further evidence for transitional probabilities modulating RT. We also found significant word recognition performance at the group level in the classic 2AFC task. We failed to uncover a correlation between these two measures. We extended this replication by running a second experiment in which participants performed the online target detection task for both structured and randomly ordered sequences of continuous speech syllables. Data from this task in both experiments was combined to address the question of why the target detection task fails to correlate with word recognition performance.

## **Introduction**

Statistical learning (SL) is a powerful mechanism by which the brain is able to detect implicit regularities in the sensory environment. (Dehaene et al. 2015) Regardless of input modality, humans are sensitive to subtle distributional properties of incoming streams of information, and able to employ this information to complete subsequent tasks. (Armstrong, Frost, and Christiansen 2017; Frost et al. 2015)

Numerous experiments have shown individuals are capable of extracting transitional probabilities embedded in streams of auditory (Endress and Mehler 2009; Pelucchi, Hay, and Saffran 2009; Saffran, Aslin, and Newport 1996; Saffran, Newport, and Aslin 1996) as well as visual sequences of stimuli. (Turk-Browne, Jungé, and Scholl 2005). The ability to track regularities in sensory streams not only leads to improvement in performance on behavioral tasks, but can also be observed in neural responses to the sequence. This neural correlates of this process can be observed whether the items are within (Turk-Browne et al. 2010) or outside (Batterink and Paller 2019; Turk-Browne et al. 2008) the focus of attention (for a review, see Batterink, Paller, & Reber, 2019). Complex statistics, such as non-adjacent relationships, can also be detected alongside adjacent relationships (Bonatti et al. 2005).

Yet, despite the considerable amount of evidence in favor of SL as a robust ability, the behavioral measures used to evaluate the degree of learning are psychometrically weak and have yielded highly variable results among participants (both within and across studies). (Siegelman, Bogaerts, and Frost 2017) Since the original study using continuous speech to test for SL in infants (Saffran, Aslin, and Newport 1996), this ability has been canonically measured using an explicit recognition task, often with little or no modification from the original, after participants have been exposed to the stimuli sequence in a learning phase. (Siegelman, Bogaerts, and Frost 2017) The task typically consists of pairing each of the units that make up the sequence (e.g. tri-syllabic pseudowords or triplets of visual shapes) with a foil, a test item of equal length but which either violates the transitional probability structure of the learned units, or is made up of random elements from the stimulus set (syllables, shapes). Participants then report which of the pair is more familiar, more similar to the exposure stream, or explicitly asked which appeared in the previous part of the experiment.

Meanwhile, several tasks have been developed to measure statistical learning “online.” These tasks aim to capture the dynamic aspects of SL (e.g. how early it occurs, how robust or stable it is over the course of the exposure phase) and provide more insight into the contribution of individual stimulus items in generating the ultimate effect. Typically, online SL tasks entail asking participants to detect a particular stimulus while they are being exposed to the stream embedded with the statistical regularities they are supposed to learn. A common variant of the task asks participants to detect a target stimulus that is part of the pattern being learned (e.g. syllable, shape). These “target detection tasks” have been used in both visual (Bertels, Franco, and Destrebecqz 2012; Franco et al. 2015; Turk-Browne, Jungé, and Scholl 2005) and auditory (Batterink et al. 2015; Batterink and Paller 2017; Batterink, Reber, and Paller 2015) SL studies. They provide an additional advantage in that they also double as a cover task to keep participants engaged and attentive, while preserving the statistical structure of the sequence fully intact. Similar tasks involve having participants detect a secondary stimulus embedded in the sequence (e.g. click sounds embedded in continuous speech, Gómez, Bion, and Mehler 2010).

These online tasks have the potential to provide rich insight into the process of extracting regularities from continuous input. The tasks mentioned above all report that reaction times (RT) to more predictable targets become faster than those to less predictable targets, suggesting an acquired sensitivity to the statistical regularities of the stream, which facilitates the detection speed. This RT effect can be observed as early as with the second presentation of the target item. (Batterink 2017) Gómez et al. similarly claim that clicks sounds occurring between embedded pseudowords become faster than clicks placed within pseudowords, suggesting that the stronger predictions generated by the learned word units interfere with click detection and thereby incur longer RTs. (but see Franco et al. 2014 for a non-replication)

This same logic underlies the serial reaction time task, whereby participants view or listen to a sequence of items which they advance at their own pace by keypress. This method has been showed to reveal the same effect, where RTs in advancing predictable shapes are faster than RTs in advanced unpredictable or less predictable shapes. (Karuza et al. 2014; Siegelman et al. 2018)

Intriguingly, there is conflicting empirical evidence as to whether the speeded RTs to predictable items indeed correlate with standard measures of SL, administered after the familiarization phase.

Several studies report significant correlations between online (target detection or analogous tasks) and offline measures (2AFC familiarity or recognition tasks). (Batterink and Paller 2017; Siegelman et al. 2018) Meanwhile, a comparative number of studies report no correlation. (Batterink et al. 2015; Franco et al. 2014; Siegelman and Frost 2015)

[What if I include a table with the rho and p and short index of the two studies for each?]

This lack of correlation between SL measures been largely discussed on a theoretical level, with occasional papers addressing the disparity empirically. (Batterink et al. 2015)

Here, we addressed the question of why these two measures might be uncorrelated or weakly correlated, despite strong evidence that both tasks are indeed sensitive to the learning of embedded regularities. In Experiment 1, participants performed an online target detection task during exposure to a continuous stream of speech syllables, followed by a standard 2AFC pseudoword vs. part-word recognition task. We examined the relationship between these two measures. In Experiment 2, participants performed the online target detection task for two continuous speech streams, one in which the embedded regularities from Experiment 1 were preserved, and another in which syllables were randomly ordered and which contained no statistical regularities. We used the combined data from these two Experiments to examine what features (statistical, theoretical) of the stream are captured by reaction time data, in order to shed light on the empirical disparity between online and explicit, offline tests of SL.

# **Experiment 1**

## **Method**

#### **Stimuli**

Speech stimuli consisted of 12 consonant-vowel (CV) pairs. We selected 5 unique vowels that are maximally separated in their manner and place of articulation. We ensured that none of these vowels typically occurred in unstressed syllables in spoken German. We then selected 12 unique consonants, in order to render each syllable phonetically distinct from the others. We used the CELEX database to calculate the frequency of occurrence of each of our syllables in spoken German, as well as the frequency of co-occurrence between each pair of syllables. We eliminated high-frequency CV pairings from our list of possible syllables and formed the final words by combining three syllables (each with distinct vowels) for which no transitions were frequent in spoken German. Final syllabes were: be, di, ga, ki, la, mi, nu, po, ro, se, tu, za.

A male native speaker of German was recorded pronouncing each syllable in our set separately and with a flat intonation. Each syllable was repeated several times to ensure we obtained a quality token. The token which most closely followed the IPA pronunciation was selected as the final syllable. The syllables were then high-pass filtered at 50 Hz and silences before and after syllable were removed using a custom script in Matlab 2017b. The 12 syllables were normalized for pitch and intensity using Praat to ensure relative homogeneity between tokens. Finally, syllables were temporally compressed to 240 ms in duration and a 10 ms silence was added at the end of each syllable, for a total duration of 250 ms.

Syllables were combined into 4 tri-syllabic pseudowords such that each word featured no repeating consonants or vowels and similarity between any possible succeeding pairs of syllables was minimized. We also ensured that no pairs were phonotactically illegal or shared a resemblance with existing words in German. Pseudowords for our study were: nugadi, rokise, mipola, zabetu. Part-words, used in the word recognition task, were of the form C’AB (word-final syllable from one word followed by word-initial and word-medial syllables from another): dizabe, semipo, lanuga, turoki.

Continuous speech sequences (24) were created in Matlab by concatenating syllables comprising the four pseudowords such that no words repeated consecutively. Each stream was comprised of 216 syllables (72 words) and was 54 seconds long. As per the design in (Saffran, Aslin, and Newport 1996), standard in SL studies, the only cue to segmenting the sequence lay in the transitional probabilities between syllables. The transitional probability of word-medial and word-final syllables (relative to the preceding syllable) was 1, while the transitional probability of word-initial syllables was 0.33. The first syllable in each stream could be a word-initial, word-medial, or word-final syllable. If stream began with the word-medial (word-final) syllable of a word, the word-initial (word-initial and word-medial) syllable of that word would be the last (two) syllable(s). Speech streams were ramped up and down in amplitude using a linear slope over a period of 1.5 seconds (6 syllables) so that onset and offset syllables were not clearly distinguishable and could not serve as cues to word boundaries.

#### **Procedure**

41 individuals participated in the study (27 female, mean age, 27.44 ± 5.78 sd). Two participants were removed from the data pool due to technical failure. Of the 39 remaining datasets, 33 were used in analyzing the target detection task (one participant failed to follow instructions, and technical issues caused partial data loss for the other five). Since the design of our experiment was modular, technical failure in one task did not necessarily affect data in another. Of the 39 datasets, we were able to use 38 for analyzing the word recognition task (data from one participant in this task was overwritten). For the correlation analysis comparing target detection task performance with word recognition performance, we included only participants for whom we had data for both tasks (32).

A previous study by Batterink and colleagues (Batterink and Paller 2017) using similar online and offline tasks as us had observed a significant correlation coefficient of 0.51 with 24 participants. A power analysis revealed this analysis to have a power of 0.74, suggesting that this effect size is rather large based on Cohen’s effect sizes for r values of 0.1, 0.3, and 0.5, respectively representing small, medium, and large effects. We calculated that in order to obtain a test with at least 80%, we would need 27 participants, and for 90% 36 participants. Our sample of 33 then was theoretically sufficient to observe a correlation effect as large as Batterink et al. reported.

Participants were seated in a dimly-lit, sound-attenuated booth, approximately 52 cm from the monitor and listened to the stimuli via headphones connected to a headphone amplifier (Beyerdynamics-DT-770 80 Ohm; Lakepeople G103P1262). Stimulus intensity level was approximately 57 dB (LAF: min 44 dB, max 76 dB), as measured by a NTi Audio device connected to an artificial ear on which the experiment headphones were mounted.

The experiment was designed using Presentation® (Version 20.1 Build 12.04.17) and delivered on two versions of the software (Version 20.0 Build 07.26.17 and Version 21.1 Build 09.05.19, Neurobehavioral Systems, Inc., Berkeley, CA, [www.neurobs.com](http://www.neurobs.com)). The experiment was conducted on a 64-bit Windows machine (Fujitsu Celsius M740B) running Windows 10.

The experiment consisted of an exposure phase, during participants performed the target detection task, followed by the word recognition task. Our experiment also included an additional task, designed to measure perceived speed of the speech stream before versus after the exposure phase. Results from this task will not be discussed here.

During the exposure phase, participants listened to a total of approximately 24 minutes of continuous speech. Participants were told they would hear brief sequences of sounds from an alien language. Audio was presented binaurally. Before the start of each stream, one of the 12 syllables was displayed orthographically on the screen and played aurally twice. Participants were instructed to press the spacebar as fast as they could during the subsequent stream whenever they heard this target syllable. Each of the 12 syllables served as a target syllable twice. The presentation order of syllables was pseudo-randomly shuffled for each participant with the constrain that a syllable from each ordinal position in the pseudoword (1st/word-initial, 2nd/word-medial, or 3rd/word-final) was tested before any were repeated. The 24 streams were organized into 8 blocks, where each block consisted of 3 streams with one target syllable from each ordinal position tested. Within each stream, target syllables appeared 17-18 times. Participants could take self-paced breaks between blocks.

In the word recognition task, participants completed 16 trials of a two-alternative forced-choice task. In each trial, a pseudoword and a part-word were presented (counterbalanced across trials), and participants were prompted to determine which of the pair was a word in the alien language they had just heard in the previous section. The inter-stimulus-interval between words was 400 ms, while inter-trial-interval was 1.2 seconds. Each pseudoword was paired with each part-word once (4 x 4 trials).

#### **Analyses**

All analyses were performed in RStudio (version 1.2.1335; RStudio Team 2018) using R version 3.6.1 (R Core Team 2019). Raw data from Presentation® was transformed into appropriate formats using Matlab R2017b (version 9.3.0.713579). Presentation® scenario files, data wrangling scripts, and R scripts with used packages are available publicly on Github.

For the target detection task analyses, we considered only those responses that occurred within a boundary of ± 3 times the median absolute deviation over all RT values. This procedure ensures that RT cutoffs would be based on the distribution of the raw data and not arbitrary limits. At the same time, the use of the median as the centrality metric is arguably more appropriate, given that the mean can be a biased estimator of RT data, which typically follows a gamma, lognormal, or ex-Gaussian distribution. This procedure eliminated only 0.034% of the data and resulted in RT that ranged from 0 to 943 ms (versus the original 0 to 1298 ms). This procedure did not significantly change the overall mean accuracy ().

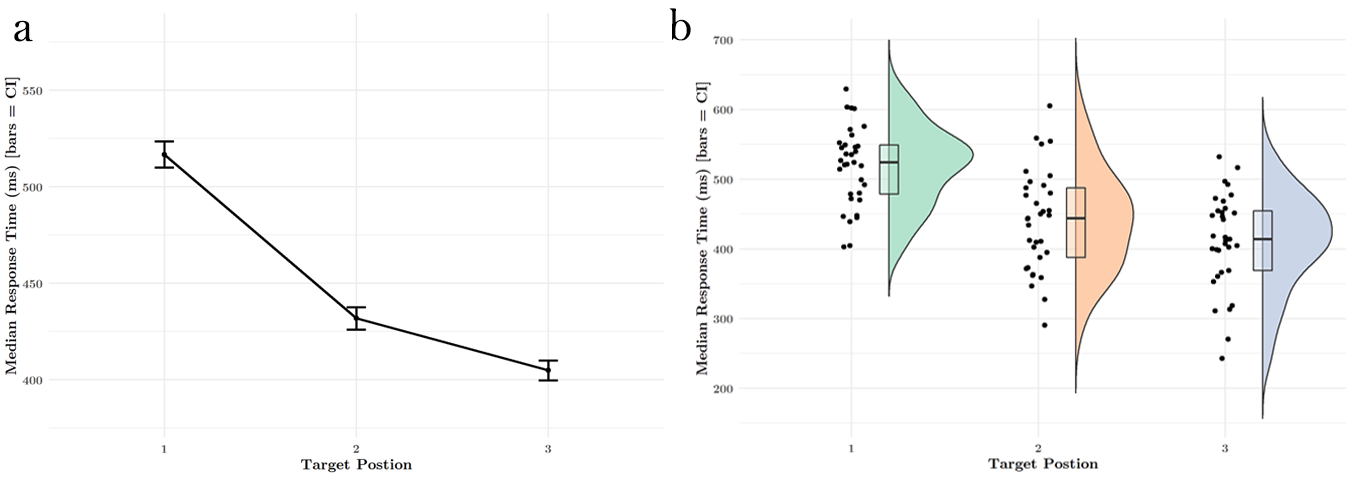
## **Results**

To ensure that participants were able to perform the task, we computed mean detection accuracy across all trials (), which was above a 0.5 chance level (). We also calculated mean accuracy for each ordinal position, word, and target syllable. (See Supplementary Materials.) We observed that certain syllables were detected less often than others, suggesting some unwanted variability in the stimuli (notably for syllables *ro, za, be, and mi).* (**Table S1**, **Fig. S1**) In the following section, we describe a control carried out to ensure that stimulus-specific effects did not drive our main finding.

***Ordinal Position Modulates Reaction Time***

To replicate findings that showed graded reaction times in response to syllables in different ordinal positions, we ran a generalized linear model with reaction time (in seconds) as outcome variable, fitted with a gamma function and log link function. Our full model included both ordinal position and block as fixed effects factors, and subject as a random intercept-random effects factor. This model was compared with a lesser model in which only ordinal position was used as a fixed effect. The lesser model provided a better fit of the data, with a lower AIC () value and significantly lower deviance (). (See **Table S3** for regression results.) We also compared both the fuller and the lesser models with random slopes for levels of ordinal position in the random effects term, but the lesser model with only varying random intercepts in the random effects term still proved a better fit for observed data (see **Table S3.** Model output for the fuller random slopes model not shown; lesser vs. fuller random slopes models deviance was and , respectively; ). Thus, we conducted further analysis on results of the lesser model.

We found that reaction times are modulated by ordinal position (main effect of ordinal position, ). (**Fig. 1**) We conducted pairwise comparisons on estimated marginal means for levels of the factor position with Tukey adjustment, to explore the drop in reaction times between each ordinal position. Specifically, RTs to word-initial syllables () are notably slower than those to word-medial () and word-final syllables (). The estimated drop in mean RT between positions 1 and 2 was 78 ms, while the drop between positions 1 and 3 was roughly 110 ms. The difference in mean RT between positions 2 and 3 was smaller, at about 32 ms (). These results suggest both ordinal position and transitional probability contributed to the graded RT effect. Indeed, if participants were sensitive only to transitional probability, we would expect to find a significant difference in RT to syllables in word-initial positions versus word-medial and word-final positions, but no difference between word-medial and word-final syllables, since the latter two have the same transitional probability (TP = 1). Rather, we find that RTs to word-final syllables are also significantly faster than RTs to word-medial syllables, suggesting that the position of the syllable in the pseudoword structure also speeds up RT.



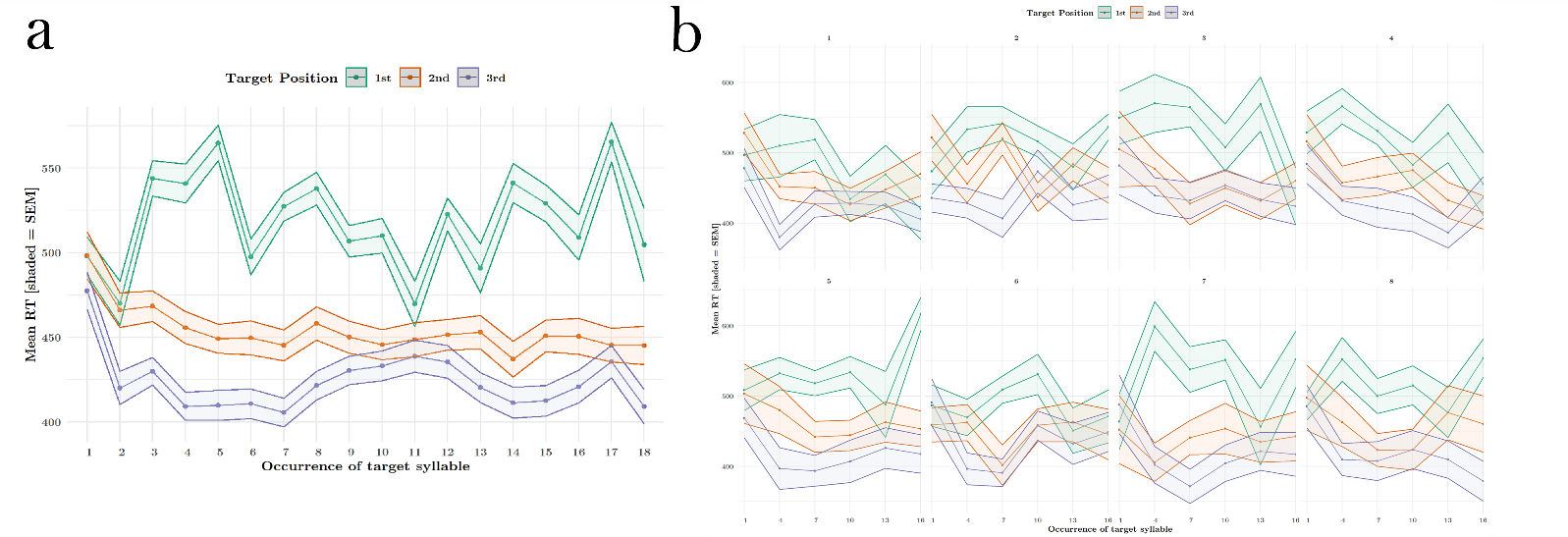
**Figure 1. A. Reaction times to target syllables are modulated by ordinal position in pseudowords. Participants responded more slowly to syllables in the word-initial (1st) position (TP = 0.33) than to syllables in the word-medial (2nd) or word-final (3rd) position. This effect was present in block 1 and remained stable throughout subsequent blocks. B. Median RTs for each position collapsed over blocks. Dots are individual participant medians. Box plots indicate group median and CI’s, half-violins describe distribution.**

Given our observation that detection accuracy varied as a function of syllable identity, sought to validate the above results by regressing out the effect of individual syllable as a function of ordinal position. We ran a generalized mixed model with ordinal position and target syllable as fixed effects factors and subject as random effect factor, with the RT (in ms) as outcome variable. We then subtracted the resulting residual values for each data point from the raw RT, and re-ran the lesser model as specified above using the adjusted RT values. We still observed the main effect of ordinal position (*)* and therefore concluded that slight variations in the acoustics of our stimuli did not significantly affect our results*.*

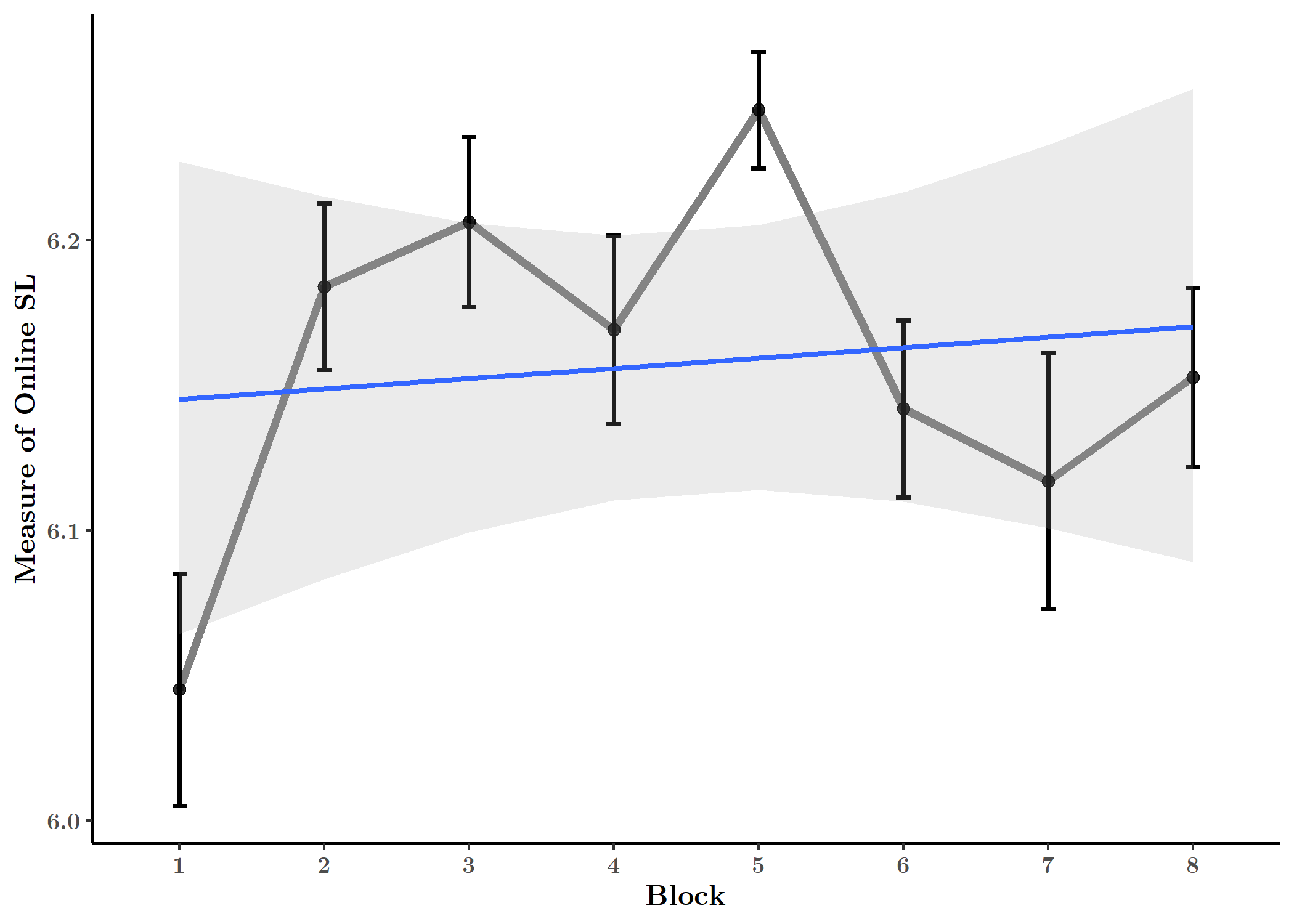
***Rapid Onset of Graded RT to Predictable Syllables***

Although we found that model that excluded the factor block was better able to represent our data, we wanted to directly test whether the observed reaction time difference emerged over the course of several blocks, or was present from the first block.

We computed pairwise contrasts on median RTs between ordinal positions for each block and found that the RT difference between positions 1 and 2, as well as 1 and 3, were significant in all 8 blocks (). The RT difference between positions 2 and 3 was significant in all blocks except blocks 2, 3 and 6 (). We additionally ran three linear models to determine whether the difference in mean RT between positions 1 and 2, 1 and 3, and 2 and 3, respectively, varied between blocks. This analysis would reveal whether the magnitude of the RT effect is modulated by block. We found no main effect as well as no significant contrasts between any blocks for any of the three analyses. These results suggest that the RT modulation by position emerged during the first block, and remained relatively stable throughout the remainder of the experiment. (**Fig. 2**) This might account for why the factor of block did not contribute significantly to model fit (): reaction times differentiate early on and persist across the whole data set, (**Fig. 2a**) despite some variation observable within individual blocks (**Fig. 2b**),



**Figure 2. Moving Average of RTs to target syllables, by ordinal position. Moving averages were calculated by taking the average of every three trials (each block contained ~18 occurrences of each syllable).**



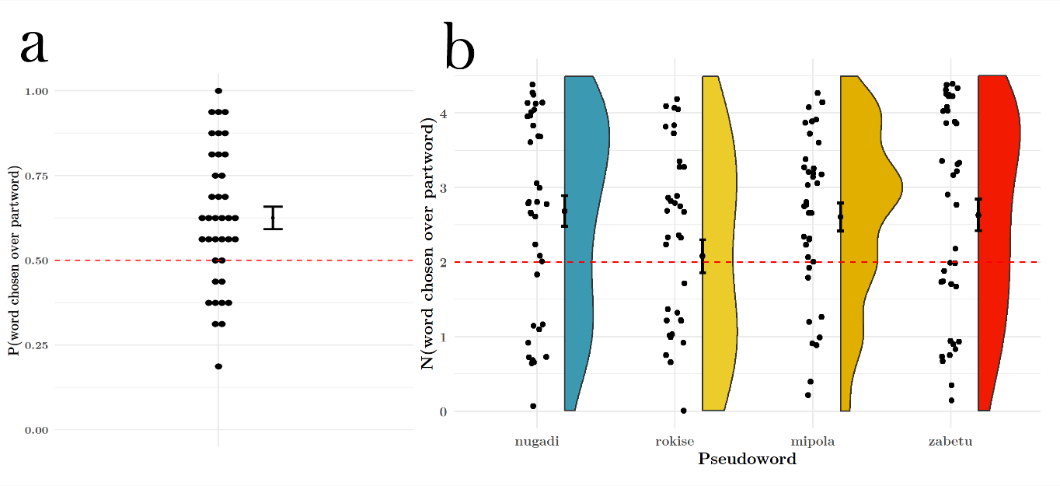
**Figure 3. Online measure of SL per block. As per Siegelman and colleagues, we used this composite measure to evaluate the magnitude of the reaction time effect (specifically, the difference between the log average RT to first – (second and third) position syllables), as a proxy for statistical learning. We observed no significant effect of block on this metric, suggesting that the modulation of RT as a function of target syllable position in the pseudowords had already occurred in the first block.**

Finally, we computed a measure of online statistical learning from the target detection task, borrowed from Siegelman and colleagues (Siegelman et al. 2018), as follows:

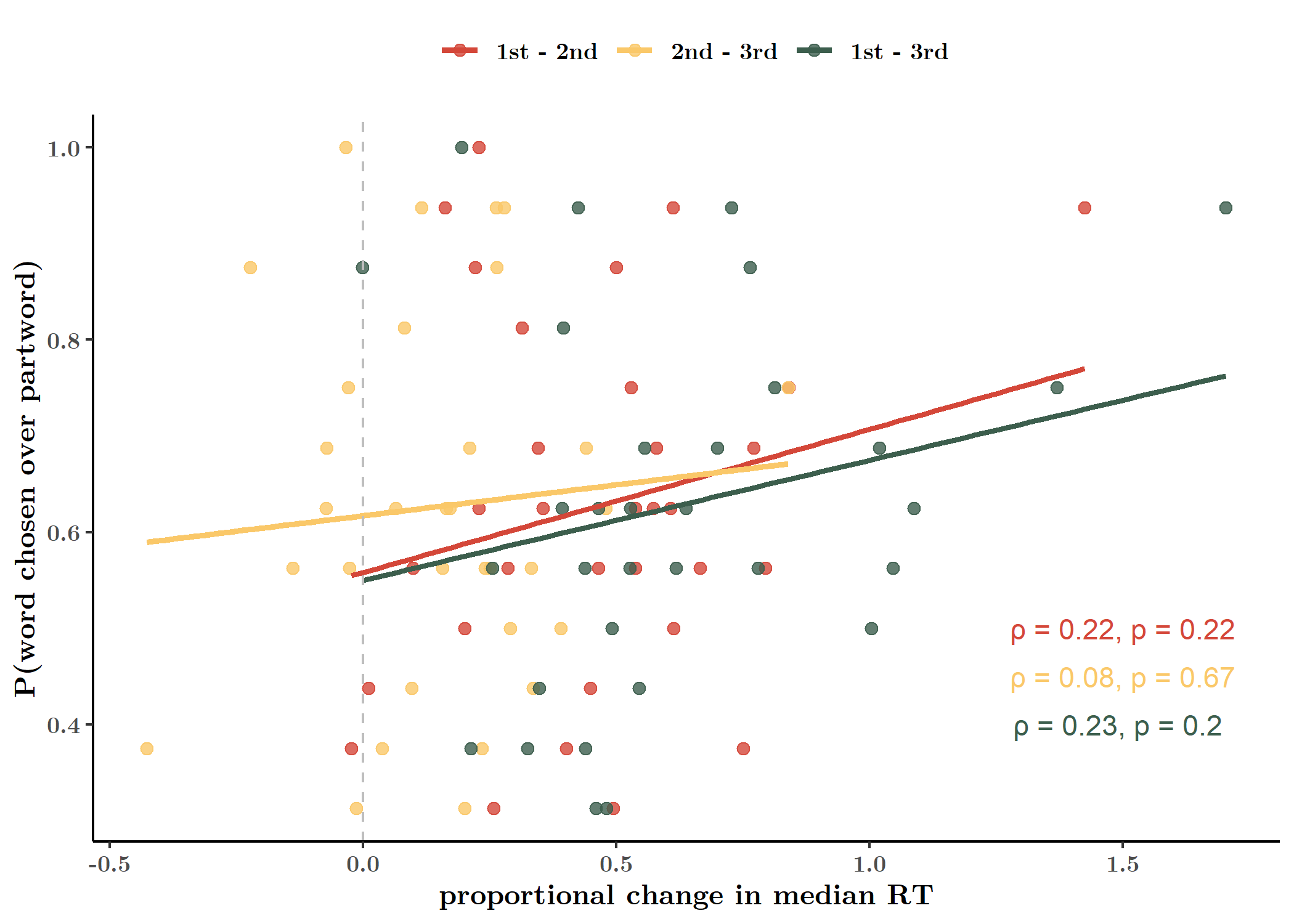
)

We computed this measure of SL for each participant and for each block. A linear model with the SL measure as outcome variable and block as predictor failed to reveal any effect of block on this composite measure (). This is an additional piece of evidence in support of early emergence of differentiated reaction times and stable behavior thereafter.

***Pseudowords Can Be Distinguished From Part-words***

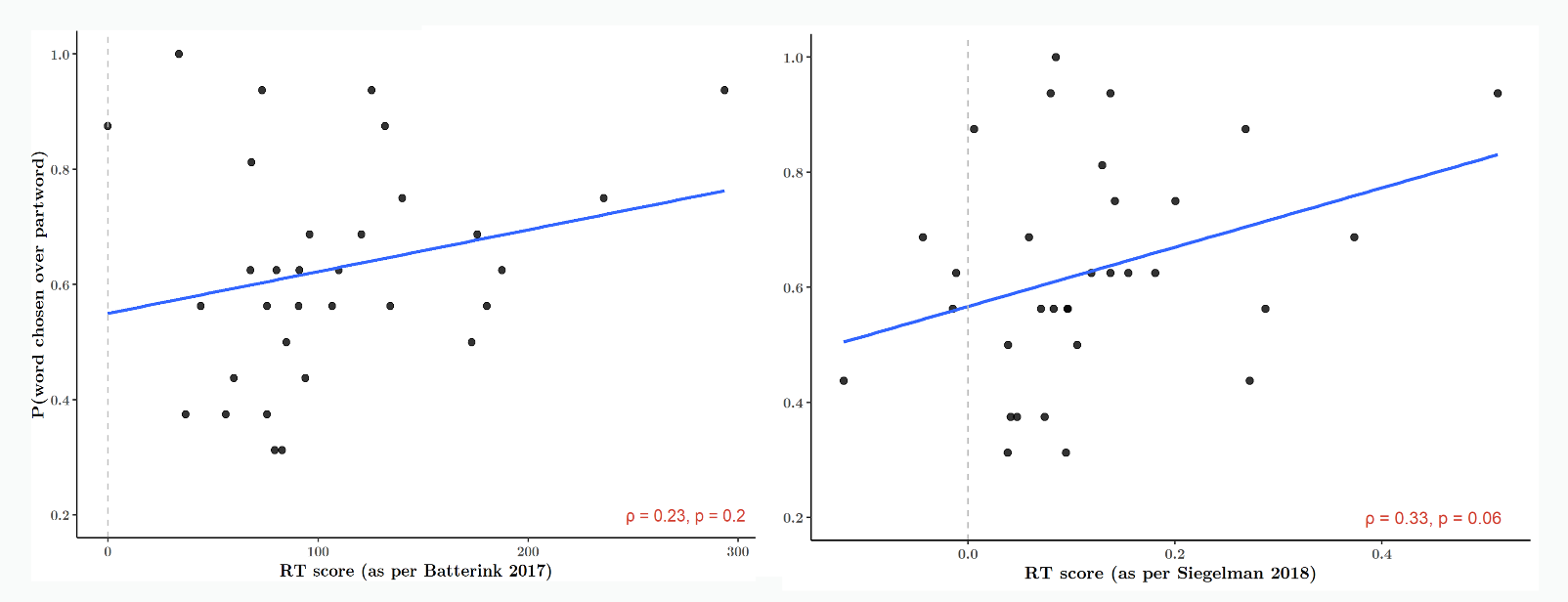
**Figure 4. Pseudoword vs. part-word discrimination. A. Word recognition performance was above 50% chance level. B. Recognition for 3 out of the 4 pseudowords was above chance.**

In the explicit word recognition task, The proportion of trials on which participants correctly distinguished the pseudoword from the part-word was significantly above a chance level of 0.5 (), indicating that participants were sensitive to the implicit regularities of the syllable stream. **(Fig. 4a)**. 71% of participants (27 out of 38) had proportion correct trials greater than the 0.5 chance level. In an exploratory analysis, we also calculated the proportion correct responses for each pseudoword individually, in order to determine if any particular word was driving the overall word recognition effect. This is because distinguishing only a single word from its paired part-word could be sufficient to push a participant’s performance above chance level. We found that across participants, 3 out of 4 words were discriminated above a chance level of 2 correct discriminations out of a possible 4. () (**Fig. 4b**)

***Online and Offline Measures Are Weakly Correlated***

**Figure 5. Pearson correlation of online and offline learning measures (median).**

Finally, also asked whether the online and offline measures of statistical learning were correlated, i.e. whether sensitivity to transitional probability online predicts explicit word recognition. Since we had an unequal number of data sets for the two tasks, we used data only from participants with complete data from both tasks (N = 32). For this analysis, we z-normalized RT values for the data set, computed median RTs for each participant and for each ordinal position, and computed the difference between the scaled median RTs for each position pairing and for each participant. These values were correlated against the participant’s proportion correct word recognition performance out of all 16 2AFC trials. Surprisingly, word recognition performance (prop. correct responses) and response time change was weakly correlated (Test for association between paired samples of Pearson's product moment correlation coefficient). (**Fig. 5**)



**Figure 6. Online SL measures as per two previous studies also failed to correlate with offline SL performance.**

We also repeated this correlation analysis using two alternate methods of calculating the online measure of SL, or “RT score”, reported in (Batterink et al. 2015; Batterink and Paller 2017) and in (Siegelman et al. 2018). The procedure as per Batterink & Paller 2017 entails a simple subtraction of the median RT to position 3 from position 1. Thus, the RT score values are on the millisecond scale (). (**Fig. 6a**) Siegelman et al. 2018 subtracted the mean of the log RTs to positions 2 and 3 from the mean log RT to position 1 (). (See previous section; **Fig. 6b**) Neither of these methods revealed a significantly stronger correlation between the measures of SL (Pearson’s test).

## **Discussion**

Our study replicated two tasks that measure statistical learning in distinct ways. Our offline word recognition task revealed a well-established effect of statistical learning, which is the ability to explicitly discriminate a properly formed pseudoword from a sequence of syllables that was heard but which span a word boundary (a part-word). This pseudoword vs. part-word test is more conservative than its sister version, the word vs. non-word test, in which words are tested against random combinations of syllables which never occurred in that particular order during the learning phase. (Batterink et al. 2015)

The online target detection task asks participants to response as soon as they heard a target syllable. Reaction times appeared to be modulated primarily by their transitional probability, which is equivalent to their predictability, such that highly predictable syllables occurring in word-medial and word-final positions elicited much faster responses than less predictable syllables occurring in word-initial positions. The rapid differentiation of response times, in the first block, confirms previous findings that statistical learning is a very fast and robust mechanism.

However, it remains puzzling why these two measures of statistical learning are uncorrelated. Previous studies reporting correlation analyses between similar online and offline SL tasks reported mixed results. (See Introduction.) In an exploratory analysis, we followed the procedure used by two earlier papers to determine if a correlation exists between the online and offline measures. None of the methods revealed a significant relationship.

Multiple explanations may account for this finding. First, the two tasks vary in their psychometric sensitivity. The target detection task in our experiment comprised a total of roughly 432 “trials” (occurrences of a target to be detected; ~18 per stream x 24 streams) and 35 trials per test item (syllable), while the word recognition task consisted of only 16 trials and 8 test items (4 pseudowords and 4 part-word foils). The correlation therefore may be weak (or the computed coefficient unstable) due to a lack of power stemming from the experiment design. Nonetheless, a vast majority of statistical learning studies have used this task with little or no modification. Test trials in the 2AFC recognition task are rarely greater than 36, typically testing only 4-8 pseudowords or visual triplets (in the case of visual statistical learning). (Siegelman, Bogaerts, and Frost 2017)

Second, it had been previously noted that explicit tasks such as the word recognition task are different in nature than implicit detection tasks (Batterink et al. 2015), which may more closely resemble serial response time tasks while yielding similar behavioral results. (Karuza et al. 2014, 2016; Schapiro et al. 2013) Online tasks have the potential to capture the dynamic progress of statistical learning and therefore reveal more subtle learning effects. (Siegelman et al. 2018). Theoretically, it is unclear whether the tracking of statistical regularities or transitional probabilities necessarily entails the chunking or explicit representation of larger units within a sensory stream. (Dehaene et al. 2015) These regularities may result in the generation of an event boundary (Schapiro et al. 2013; Zacks and Swallow 2007), without enabling explicit recall of the event itself.

We wished to investigate why graded response times in the online task failed to adequately predict offline pseudo-word recognition. The syllables in the stream are each characterized by several features: ordinal position, transitional probability, within-word duplet pairing, and within-word triplet pairing. Success on the online tasks requires only a tracking of transitional probability, while success in the offline task requires a representation of the within-word triplet pairing (in other words, the representation of the tri-syllabic pseudoword as a single unit). We decided to investigate whether we could determine which of these features may be reflected in the target detection task. Indeed, a lack of correlation between online and offline tasks may be due to the fact that they test for sensitivity to different properties of the stream.

To do this, we first aimed to replicate our online target detection task findings. We ran a correlation analysis on the combined data set in order to explore what features of the stream were being captured by the online task, using a larger dataset. By comparing the current (structured stream) task with a condition in which participants are exposed to the same stimuli with the same detection task, but without any embedded regularities or adjacency relationships (pseudo-random syllable stream), we could additionally demonstrate that the reported effect is primarily driven by statistical regularities and not uncontrolled variation in our stimuli acoustics.

# **Experiment 2**

## **Method**

***Stimuli***

The syllable stimuli used in Experiment 2 were identical to those used in Experiment 1. For this experiment we synthetized 12 “structured” streams and 12 “random” streams in Matlab. For structured streams, the procedure was identical to that mentioned above. For random streams, the 12 syllables were pseudo-randomly permuted out to the same length as the structured stream (216 syllables), with the sole constraint that a syllable could not be repeated consecutively. Thus, transitional probabilities between adjacent syllables were roughly 0.083. Speech streams were ramped up and down in amplitude over a period of 1.5 seconds so that onset and offset syllables were not clearly distinguishable and could not serve as cues to word boundaries. Within each stream, target syllables appeared 17-18 times. Participants could take self-paced breaks between streams presentations.

***Procedure***

21 individuals participated in the study (15 female, mean age, 28.08 ± 6.82). The same inclusion criteria were used as in Experiment 1, in addition to requiring that they not have taken part in the previous experiment. One participant was excluded due to technical failure. Technical failure caused data loss in the random condition for one other participant, leaving data from 20 participants (19 in the random condition, 20 in the structured condition).

Participants were seated approximately 52 cm from the monitor and listened to the stimuli via headphones connected to the PC server. Stimulus intensity level was again measured by a NTi Audio device connected to an artificial ear. Volume levels were in the range reported for Experiment 1.

The experiment was designed using Presentation® (Version 20.1 Build 12.04.17) and delivered on two versions of the software (Version 20.0 Build 07.26.17 and Version 21.1 Build 09.05.19, Neurobehavioral Systems, Inc., Berkeley, CA, [www.neurobs.com](http://www.neurobs.com)). The experiment was conducted on a 64-bit Windows machine (Fujitsu Celsius M740B) running Windows 7.

Participants completed two exposure phases, one with a continuous stream of random syllables and one with a continuous structured stream. During both phases, participants completed the target detection task. Each phase consisted of a total of approximately 12 minutes of continuous speech, divided into ~1 minute long streams. Participants could take self-paced breaks between streams. The instructions and task procedure for each phase was identical to that in Experiment 1, with the exception that participants only performed the task once for each syllable instead of twice. Each stream featured ~18 occurrences of the target syllable. Random and structured exposure orders were counterbalanced across participants.

Our experiment also included an additional non-SL task, which was completed after each exposure phase. Results from this task will not be discussed here.

***Analysis***

All analyses were performed in RStudio (version 1.2.1335; RStudio Team 2018) using R version 3.6.1 (R Core Team 2019). Raw data from Presentation® was transformed into appropriate formats using Matlab R2017b (version 9.3.0.713579). Presentation® scenario files, data wrangling scripts, and R scripts with used packages are available publicly on Github.

For the target detection task analyses, we used the same criterion to eliminate outliers as in Experiment 1 (± 3 times the median absolute deviation). This procedure eliminated only 1.93% of the data and resulted in RT that ranged from 119 to 941 ms (originally, 0 to 1997 ms). This procedure did not affect the overall detection accuracy ()

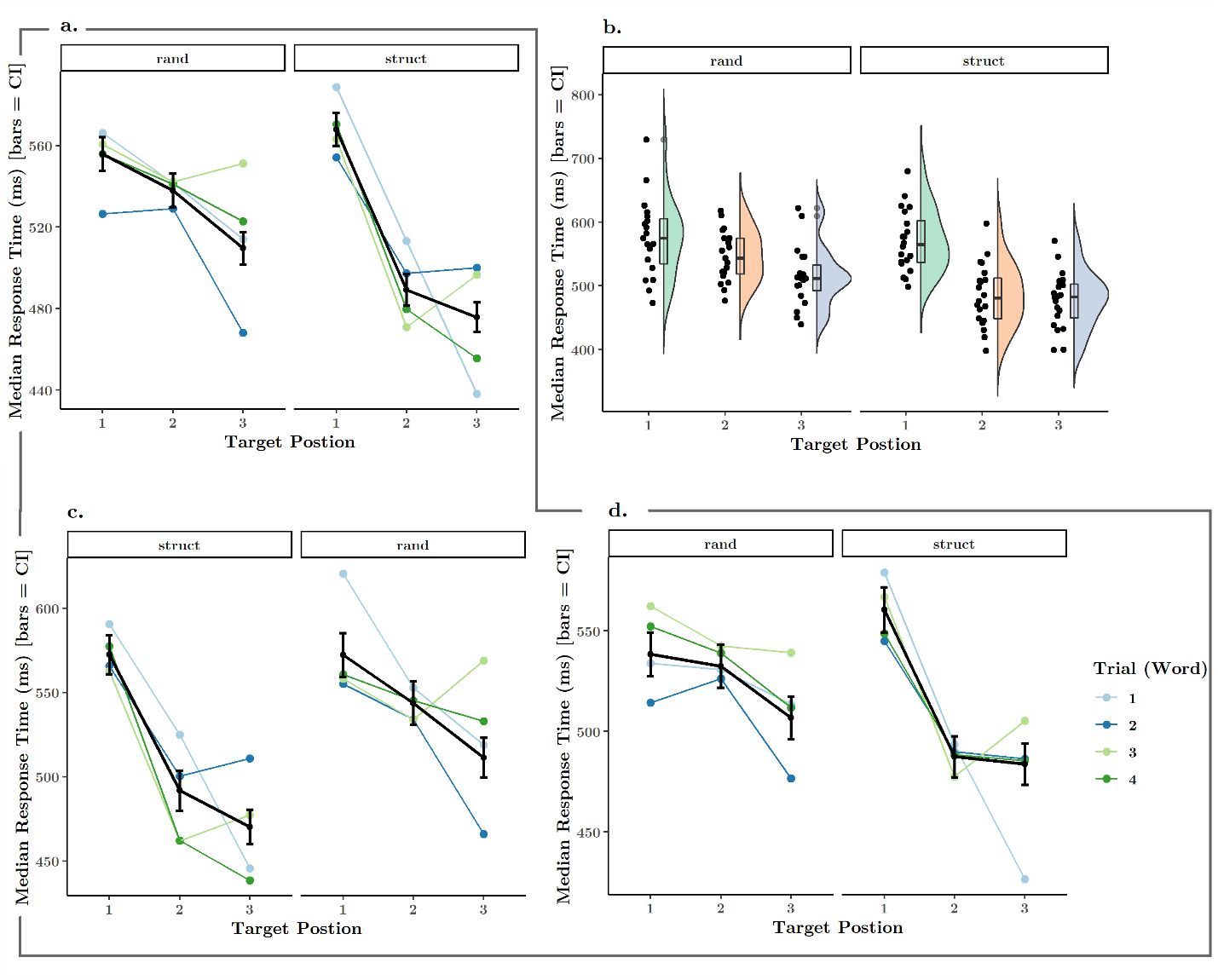
To compare the reaction times between structured and random conditions, we dummy-coded the random streams with the same ordinal positions as the structured streams. Thus, if the syllables in the structured stream 1 followed the order: 3,1,2,3,1,2,3…, we applied the same position coding to random stream 1, even though these position codes correspond to no meaningful property in the random stream. This procedure however, allowed us to compare RTs for the same variable (ordinal position) between the two conditions.

## **Results**

As in Experiment 1, we first examined mean detection accuracy to ensure participants were engaged in the task. Overall detection accuracy was 0.82 (), and significantly above a 50% chance level (). Detection accuracy was higher in the structured condition () than in the random condition (), but this difference was not significant when tested through a one-sided test predicting the mean for the structured condition to be greater than random (). (**Fig. S2, Table S2**) We additionally checked to see if ordinal position and word yielded difference effects on accuracy in the structured condition, and if syllable identity did so in both conditions. (See Supplementary Materials.)

***Ordinal Position in Structured Stream Modulates Reaction Time***

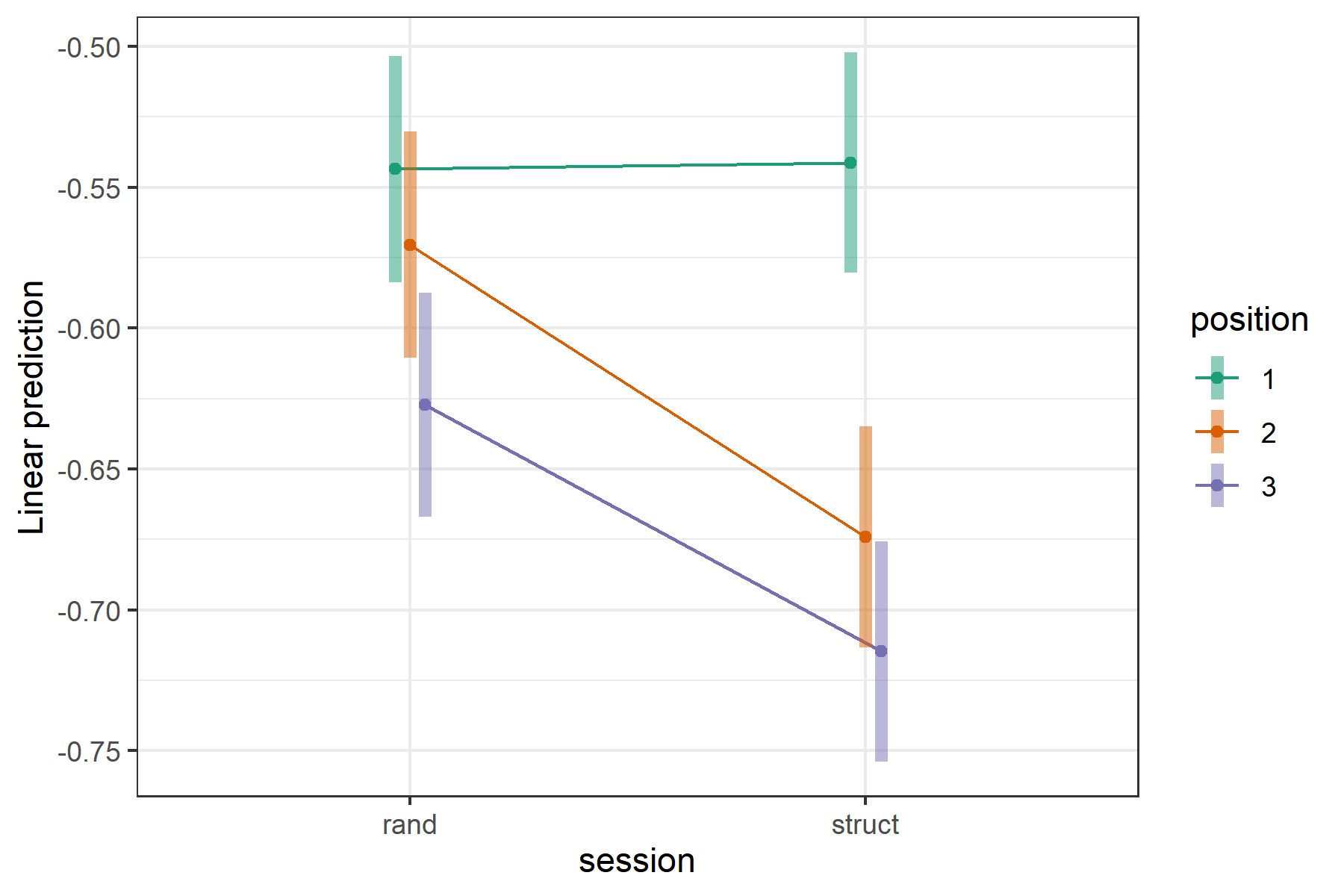
We performed a modelling procedure similar to that from Experiment 1. Our hypothesis stated that reaction times (in seconds) would be predicted by a combination of ordinal position and condition (structured, random). We included subject as a nested effect within condition order (whether participants completed the structured condition before the random condition, or vice versa), as condition order was our between-subjects variable. We further specified the random effects term by allowing random intercepts and uncorrelated random effects for each level of condition. This structure allows the graded RT curve for each participant to vary between conditions, as well as their baseline RT (intercept). (See **Table S4** for regression results.)

We observed main effects of both condition and ordinal position, as well as an interaction between them (). We then performed two planned contrasts. First, we evaluated the effect of ordinal position within each level of condition (i.e. to determine the modulation of reaction times for each condition). We observed a significant drop in estimated means between positions 1 () and 2 (; ), as well as between 1 and 3 () in the structured condition. There was a smaller, but also statistically significant decrease in means between positions 2 – 3 (). (**Fig. 7a**). This result replicates the graded RT effect we observed in Experiment 1.

**Figure 7 Reaction time to target syllables modulated by ordinal position. A. RTs to targets in 2nd and 3rd ordinal positions were faster than RTs to 1st position targets for both random and structured conditions. However, the structured condition saw a much more pronounced RT effect. B. Distribution of median RTs to each target position for each participant. C. Median RTs to each ordinal position for participants who performed the structured condition first, random condition second. C. Median RTs to each ordinal position for participants who performed the random condition first, structured condition second. Legend: Colored lines connect overall median RTs to targets within the same pseudoword**

In the random condition, we were surprised to observe a similar RT pattern, where differences in estimated means for certain pairs of positions reached still significance, with a small but significant change occurring between positions 1 () and 3 (), as well as between position 2 () and 3. (**Fig. 7a**) Given that there were no regularities in the random stream that could bias reaction times to certain tokens more than others, we hypothesized that the modulation observed here is due to variations in the acoustic features of the stimuli, as also noted above for Experiment 1.

In our second contrast, we evaluated the effect of condition for each level of ordinal position (i.e. how much condition affected reaction times to targets in each ordinal position). We observed that the presence of structure significantly decreased mean reaction times for 2nd () and 3rd position targets (). Meanwhile, RTs to 1st position targets remained roughly the same (reaction times for 1st position targets were in fact slightly higher in the structured vs. the random condition, but this change was not statistically significant; ). (**Fig. 7c-d, Fig. 8**)



**Figure 8 Linear predictions from GLM modelling RT as a function of ordinal position and condition. RTs to targets in the second and third positions were significantly shortened in the structured condition compared with the random condition. RTs to targets in the first position were unchanged.**

We additionally tested a model that included condition order as a fixed effect, to determine whether the there were any carry-over effects of condition (i.e. confounding effect of exposure to random then structured streams, or vice versa). We did not detect a three-way interaction between condition order, ordinal position, and condition in modulating response times (). (**Fig. S3**) However, we note that this null result may be due to lack of statistical power, and some confounding effect of viewing structured or random streams first may have occurred.

## **Discussion**

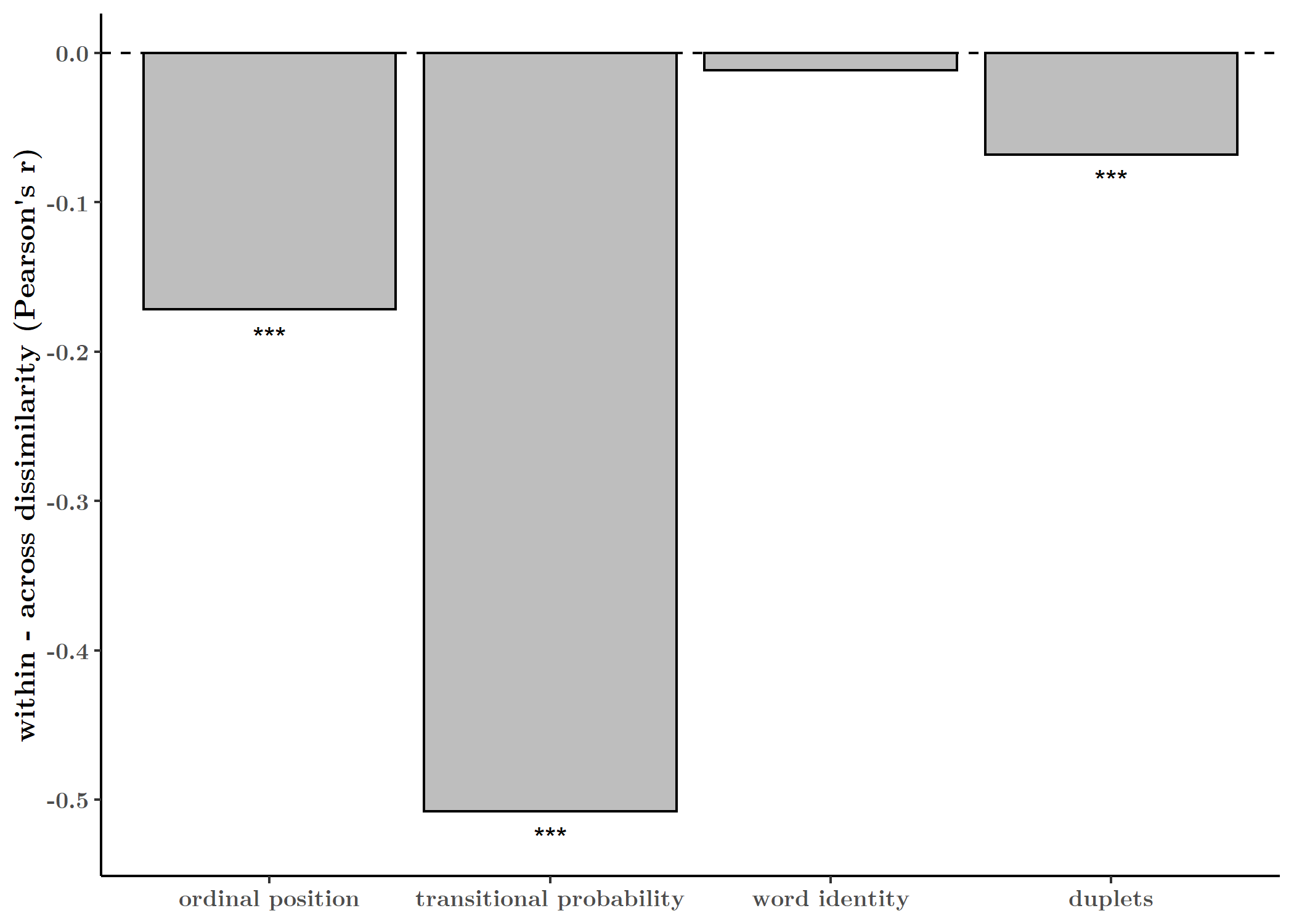
Here we were able to replicate our findings from Experiment 1, in which the presence of implicit statistical structure modulates reaction times to items embedded in the stream. Notably, we could establish that this modulation is most extreme for those items which become strictly predictable (those with a transitional probability of 1, in word-medial and word-final positions), as compared with the same items tested in a stream of randomly ordered syllables.

# **Coding of Structural Features in Online Target Detection**

In a final analysis, we aimed to explore the reaction time data from the online target detection task further, in order to determine if the observed patterns of RTs could reveal any specific information about the implicit structure of the stream heard during the exposure. Specifically, we hypothesized that one of four possible features could be encoded in reaction time data: transitional probability, ordinal position, word identity, and duplets. We predicted that a failure to track one or more of these features may explain why online RT scores fail to correlate with offline word recognition performance.

We first combined the data from Experiment 1 (N = 33) with the data from structured condition in Experiment 2 (N = 20) for a combined data set with greater power (N = 53). For each participant, we computed a dissimilarity matrix (1- Pearson correlation) on RTs between each pair of syllables, thus generating a 12-x-12 matrix of correlation values for each participant. We then applied a Fisher’s z-transformation to each matrix to normalize dissimilarity values. For each of the four analyses mentioned above, we identified a *within* and an *across* group. Within groups consisted of those cells in the correlation matrix that correspond to the feature being tested (e.g. ordinal position), while across groups features cells that represented a violation of that group. For each test, we performed a random sampling of values from the respective cells from all participants. The sampling procedure was repeated 200 times with replacement, with the N for each test being equal to 4/5 times the length of the shorter of the two arrays being compared. (Since arrays for this analysis must have the same length and not all tests entailed the same number of comparisons, we effectively subsampled from both arrays out to a common length.) We then computed a Wilcoxon’s rank sum test on resulting two arrays.

## **Reaction Times to Targets Reveal Sensitivity to Ordinal Position, Transitional Probability, and Duplet Pairings**



**Figure 9 Within versus across group similarity for four features of the stimulus stream. For each feature, we determined within and across groups; pairs of syllables that share the property versus pairs that violate the property, respectively. We sampled 200 times with replacement from the 12-x-12 syllable-wise dissimilarity matrices created for each participant.**

For the test of transitional probability (N = 239), within values included the correlation between all pairs of word-initial syllables (TP = 0.33) versus the correlation between all pairs of word-medial and word-final syllables (TP = 1). Across values included pairs with crossed probabilities (e.g. TP = 1 vs. TP = 0.33). We observed a significant shift in the means of these two groups. (). For the test of ordinal position (N = 468), within values included the correlation between all pairs of word-initial syllables (e.g. nu-ro), all pairs of word-medial syllables (e.g. ga-ki), and all pairs of word-final syllables (e.g. di-se). Across values included correlations between syllables within each word (e.g. nu-ga, ga-di, nu-di). Here we also observed a significant shift in means between the two groups. (). For the test of word identity (N = 468), within values included the correlation between syllables within each word (e.g. nu-ga, ga-di, nu-di). Across values included “phantom” word pairs where each item in the pair is drawn from two different words (e.g. nu-ki, nu-se). Here we observed no significant difference. (). Finally, to test duplet identity (N = 156), we compared values from all pairs of duplets within words (e.g. nu-ga, ga-di) versus pairs of word-initial and word-final syllables within words (e.g. nu-di). Our Wilcoxon test suggested the two groups differ in their means. ().

Together, these analyses suggest that RT in the online detection task track transitional probability, ordinal position, and duplet groupings; meanwhile, they fail to track word identity. (**Fig. 9**) This result appears intuitive on the basis of the previous results above, however it provides support for the claim that individuals are sensitive to both ordinal position and transitional probability. In addition, we found evidence that duplet identity is also tracked by implicit responses. However, we found no similarity between RTs to syllables within versus across word boundaries. This fact, also suggested by the other findings reported here, may be key to why online and offline measures of SL fail to correlate. Indeed, while online measures track low-level statistics of the stimulus stream, they do not entail a “chunking” of the sensory stream, i.e. an explicit representation of word or event boundaries. (Dehaene et al. 2015)

## **Correlating Individual RT Patterns with Offline Word Recognition Performance**

# **Conclusion**

Online detection task reveals sensitivity to transitional probability, ordinal position, and duplet pairs, but not the pseudo-word units. We see a weak correlation not only because the tasks potentially tap into different representations (implicit vs. explicit), but also because the target detection task doesn’t inherently reflect word-level chunking, but rather sensitivity to pairwise relationships.

# **Acknowledgements**

We thank R. Muralikrishnan for help with programming the experiment. We also thank Yue Sun for help with stimulus creation. Many thanks to Valeria Peviani, Martina Vilas, and Valerie Pu for helpful comments. This research was supported by the MPIEA and a grant to L.M. (XXX).

# **Contributions**

L.M. & A.K. designed the studies. A.K. programmed the study, collected and analyzed the data. A.K. wrote the paper. L.M. made edits to the paper.

# **References**

Armstrong, Blair C., Ram Frost, and Morten H. Christiansen. 2017. “The Long Road of Statistical Learning Research: Past, Present and Future.” *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 372(1711): 20160047. http://www.ncbi.nlm.nih.gov/pubmed/27872366 (December 17, 2018).

Batterink, Laura J. 2017. “Rapid Statistical Learning Supporting Word Extraction From Continuous Speech.” *Psychological Science*: 1–8. https://doi.org/10.1177/0956797617698226 (August 30, 2019).

Batterink, Laura J., and Ken A. Paller. 2017. “Online Neural Monitoring of Statistical Learning.” *Cortex* 90: 31–45. https://www.sciencedirect.com/science/article/pii/S0010945217300540?via%3Dihub#fig1 (July 25, 2019).

———. 2019. “Statistical Learning of Speech Regularities Can Occur Outside the Focus of Attention.” *Cortex* 115: 56–71. www.sciencedirect.com (August 30, 2019).

Batterink, Laura J., Ken A. Paller, and Paul J. Reber. 2019. “Understanding the Neural Bases of Implicit and Statistical Learning.” *Topics in Cognitive Science*: 1–22. https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12420 (June 17, 2019).

Batterink, Laura J., Paul J. Reber, Helen J. Neville, and Ken A. Paller. 2015. “Implicit and Explicit Contributions to Statistical Learning.” *Journal of Memory and Language* 83: 62–78. http://dx.doi.org/10.1016/j.jml.2015.04.004 (September 3, 2019).

Batterink, Laura J., Paul J. Reber, and Ken A. Paller. 2015. “Functional Differences between Statistical Learning with and without Explicit Training.” *Learning and Memory* 22: 544–55. http://www.learnmem.org/cgi/doi/10.1101/lm.037986.114. (September 3, 2019).

Bertels, Julie, Ana Franco, and Arnaud Destrebecqz. 2012. “How Implicit Is Visual Statistical Learning?” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 38(5): 1425–31. http://doi.apa.org/getdoi.cfm?doi=10.1037/a0027210 (September 10, 2019).

Bonatti, Luca L, Marcela Peña, Marina Nespor, and Jacques Mehler. 2005. “Linguistic Constraints on Statistical Computations The Role of Consonants and Vowels in Continuous Speech Processing.” *Psychological Science* 16(6). https://journals.sagepub.com/doi/pdf/10.1111/j.0956-7976.2005.01556.x (July 10, 2019).

Dehaene, Stanislas et al. 2015. “The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees.” *Neuron* 88(1): 2–19. http://dx.doi.org/10.1016/j.neuron.2015.09.019 (September 26, 2018).

Endress, Ansgar D, and Jacques Mehler. 2009. “The Surprising Power of Statistical Learning: When Fragment Knowledge Leads to False Memories of Unheard Words.” *Journal of Memory and Language* 60. https://ac.els-cdn.com/S0749596X08000995/1-s2.0-S0749596X08000995-main.pdf?\_tid=ec4b4780-2f43-43cd-b52b-edae1927b0ee&acdnat=1551180035\_7b6a1ae732e95a1222fed8e16168b830 (February 26, 2019).

Franco, Ana et al. 2015. “Rapid Serial Auditory Presentation.” *Experimental Psychology* 62(5): 346–51. https://econtent.hogrefe.com/doi/10.1027/1618-3169/a000295 (September 10, 2019).

Franco, Ana, Vinciane Gaillard, Axel Cleeremans, and Arnaud Destrebecqz. 2014. “Assessing Segmentation Processes by Click Detection: Online Measure of Statistical Learning, or Simple Interference?” https://link.springer.com/content/pdf/10.3758%2Fs13428-014-0548-x.pdf (February 13, 2019).

Frost, Ram, Blair C. Armstrong, Noam Siegelman, and Morten H. Christiansen. 2015. “Domain Generality versus Modality Specificity: The Paradox of Statistical Learning.” *Trends in Cognitive Sciences* 19(3): 117–25. https://www.sciencedirect.com/science/article/pii/S1364661314002770 (December 17, 2018).

Gómez, David M, Ricardo A H Bion, and Jacques Mehler. 2010. “The Word Segmentation Process as Revealed by Click Detection.” *Language and Cognitive Processes* 26(2): 212–23. https://www.tandfonline.com/action/journalInformation?journalCode=plcp21 (February 13, 2019).

Karuza, Elisabeth A. et al. 2016. “Sampling over Nonuniform Distributions: A Neural Efficiency Account of the Primacy Effect in Statistical Learning.” *Journal of Cognitive Neuroscience* 28(10): 1484–1500.

Karuza, Elisabeth A, Thomas A Farmer, Alex B Fine, and Florian Jaeger. 2014. “On-Line Measures of Prediction in a Self-Paced Statistical Learning Task.” *Proceedings of the Annual Meeting of the Cognitive Science Society* 36(36): 726–30.

Pelucchi, Bruna, Jessica F. Hay, and Jenny R. Saffran. 2009. “Learning in Reverse: Eight-Month-Old Infants Track Backward Transitional Probabilities.” *Cognition* 113(2): 244–47. https://www.sciencedirect.com/science/article/pii/S0010027709001826 (February 10, 2019).

Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport. 1996. “Statistical Learning by 8-Month-Old Infants.” *Science* 274(December): 1926–28. http://www.ncbi.nlm.nih.gov/pubmed/918660 (September 24, 2018).

Saffran, Jenny R., Elissa L. Newport, and Richard N. Aslin. 1996. “Word Segmentation: The Role of Distributional Cues.” *Journal of Memory and Language* 35(4): 606–21. https://www.sciencedirect.com/science/article/pii/S0749596X96900327?via%3Dihub (May 13, 2019).

Schapiro, Anna C et al. 2013. “Neural Representations of Events Arise from Temporal Community Structure.” *Nature Neuroscience* 16(4): 486–92. http://www.nature.com/articles/nn.3331 (February 15, 2019).

Siegelman, Noam, Louisa Bogaerts, and Ram Frost. 2017. “Measuring Individual Differences in Statistical Learning: Current Pitfalls and Possible Solutions.” *Behavior Research Methods* 49(2): 418–32. http://link.springer.com/10.3758/s13428-016-0719-z (December 17, 2018).

Siegelman, Noam, Louisa Bogaerts, Ofer Kronenfeld, and Ram Frost. 2018. “Redefining ‘Learning’ in Statistical Learning: What Does an Online Measure Reveal About the Assimilation of Visual Regularities?” *Cognitive Science* 42: 692–727. http://doi.wiley.com/10.1111/cogs.12556 (December 17, 2018).

Siegelman, Noam, and Ram Frost. 2015. “Statistical Learning as an Individual Ability: Theoretical Perspectives and Empirical Evidence.” *Journal of Memory and Language* 81: 105–20. https://www.sciencedirect.com/science/article/pii/S0749596X15000200?via%3Dihub (February 26, 2019).

Turk-Browne, Nicholas B., Justin A. Jungé, and Brian J. Scholl. 2005. “The Automaticity of Visual Statistical Learning.” *Journal of Experimental Psychology: General* 134(4): 552–64. http://doi.apa.org/getdoi.cfm?doi=10.1037/0096-3445.134.4.552 (February 26, 2019).

Turk-Browne, Nicholas B., Brian J Scholl, Marvin M Chun, and Marcia K Johnson. 2008. “Neural Evidence of Statistical Learning: Efficient Detection of Visual Regularities Without Awareness.” *Journal of Cognitive Neuroscience* 21(10): 1934–45. www.omniglot.com. (February 26, 2019).

Turk-Browne, Nicholas B, Brian J Scholl, Marcia K Johnson, and Marvin M Chun. 2010. “Implicit Perceptual Anticipation Triggered by Statistical Learning.” *The Journal of Neuroscience* 30(33): 11177–87. https://www.jneurosci.org/content/jneuro/30/33/11177.full.pdf (September 3, 2019).

Zacks, Jeffrey M, and Khena M Swallow. 2007. “Event Segmentation.” *Current Directions in Psychological Science*: 80–84. https://journals.sagepub.com/doi/pdf/10.1111/j.1467-8721.2007.00480.x (February 26, 2019).

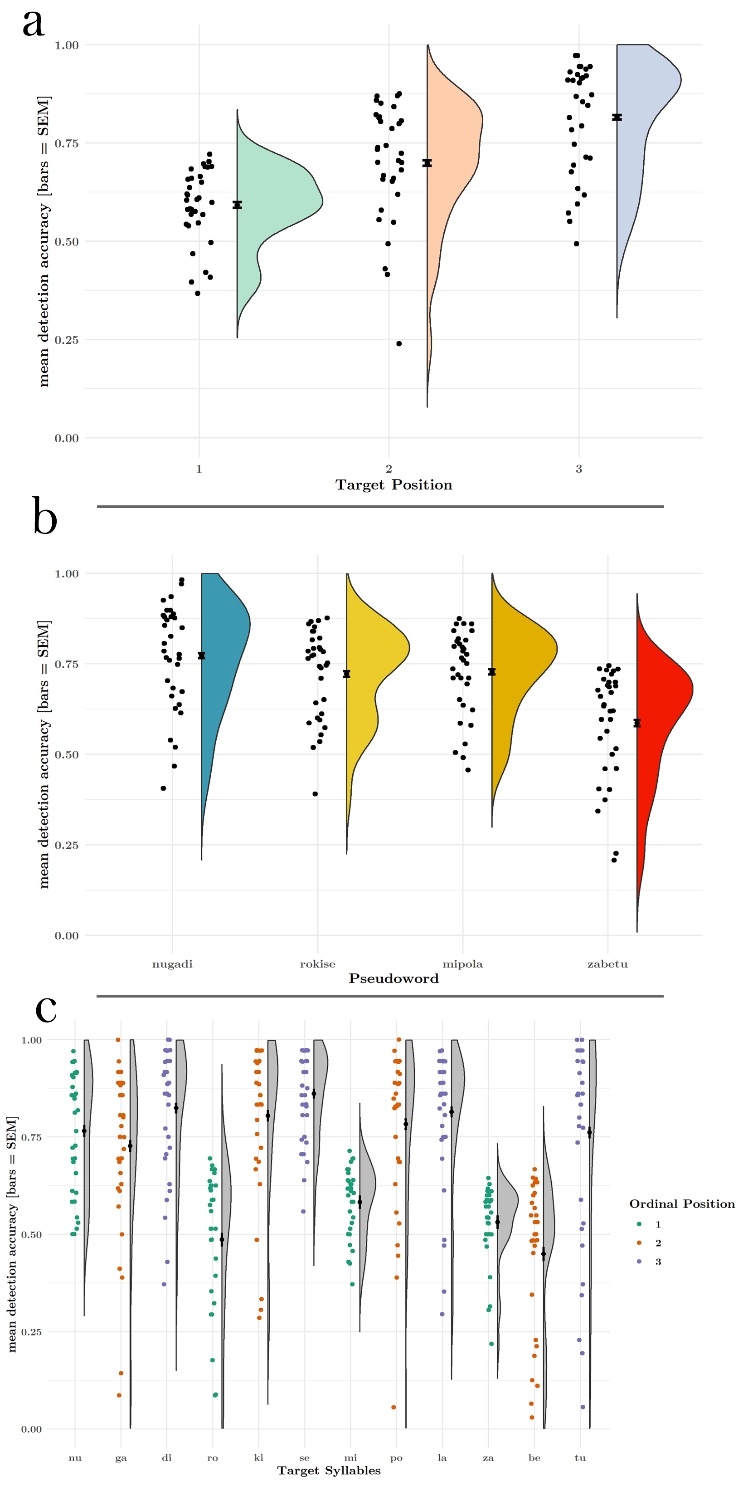
# **Supplementary Materials**

***Stimuli***

Stimuli and code can be found on Github: <https://github.com/avakiai/statistical-learning>.

***Supplementary Figures, Tables***

**Fig. S1. Mean detection accuracy for each ordinal position, syllable and pseudoword.**

****The hit rate was modulated by ordinal position, with each successive position having a higher mean accuracy (). **(Fig. S1a**)

Accuracy also varied between the four words (). (**Fig. S1b, Table S1a**) Specifically, between words nugadi and rokise (), nugadi and zabetu (), rokise and zabetu (), and mipola and zabetu ().

Syllable identity additionally affected detection accuracy (). (**Fig. S1c, Table S1c**) Certain CV syllable pairs may have been easier to detect than others, due to minor variations in stimuli acoustics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | | | |
| **Table S1a** | | | | | |
| word | N | mean | sd | se | ci |
| nugadi | 3,425 | 0.772 | 0.419 | 0.007 | 0.014 |
| rokise | 3,378 | 0.722 | 0.448 | 0.008 | 0.015 |
| mipola | 3,452 | 0.727 | 0.445 | 0.008 | 0.015 |
| zabetu | 3,306 | 0.586 | 0.493 | 0.009 | 0.017 |
|  | | | | | |

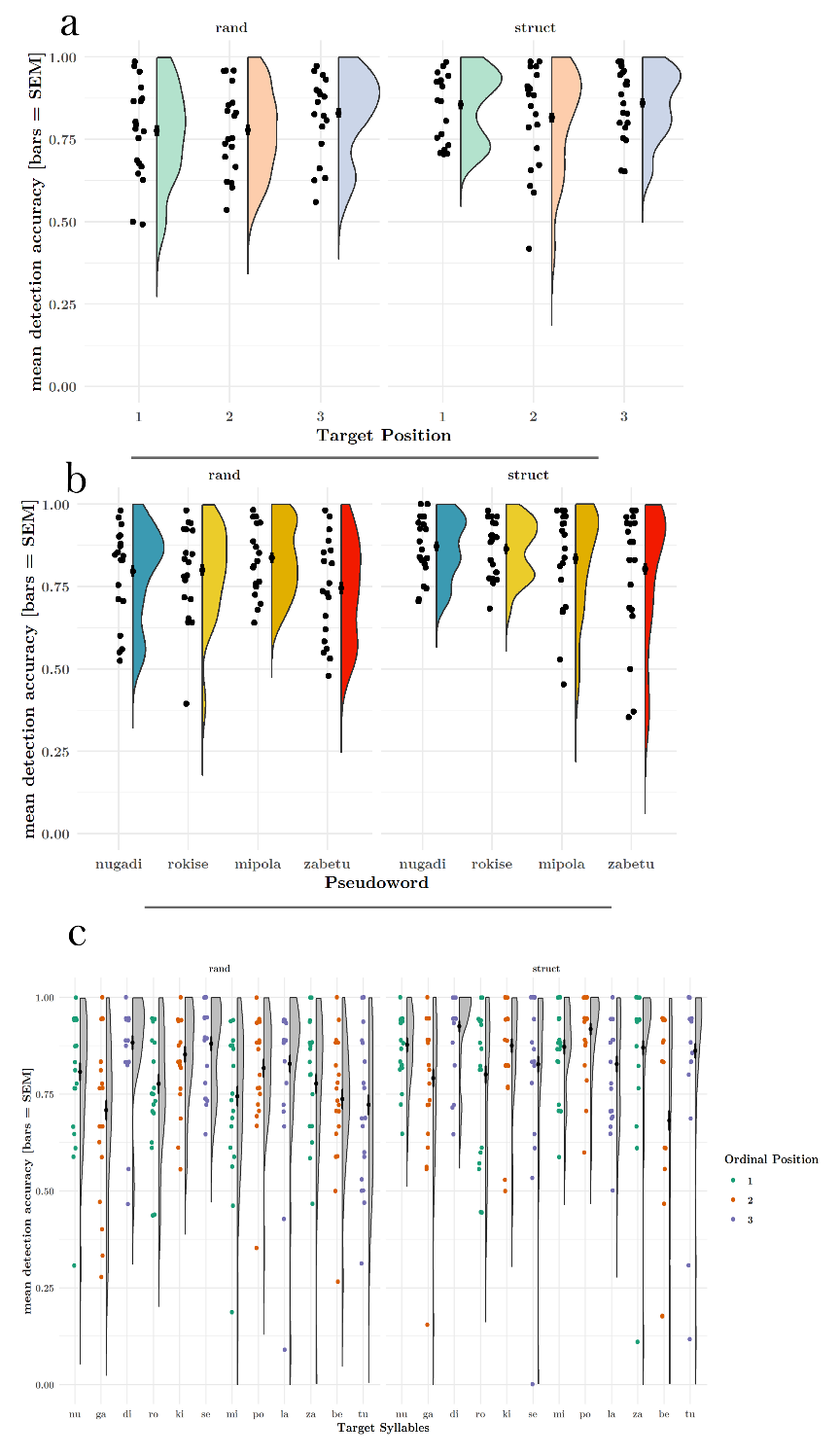
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table S1b** |  |  |  |  |  |
| target | N | mean | sd | se | ci |
| be | 1,015 | 0.449 | 0.498 | 0.016 | 0.031 |
| di | 1,154 | 0.824 | 0.381 | 0.011 | 0.022 |
| ga | 1,149 | 0.727 | 0.446 | 0.013 | 0.026 |
| ki | 1,150 | 0.804 | 0.397 | 0.012 | 0.023 |
| la | 1,152 | 0.814 | 0.389 | 0.011 | 0.022 |
| mi | 1,136 | 0.583 | 0.493 | 0.015 | 0.029 |
| nu | 1,122 | 0.766 | 0.424 | 0.013 | 0.025 |
| po | 1,164 | 0.783 | 0.413 | 0.012 | 0.024 |
| ro | 1,082 | 0.486 | 0.500 | 0.015 | 0.030 |
| se | 1,146 | 0.861 | 0.346 | 0.010 | 0.020 |
| tu | 1,148 | 0.761 | 0.426 | 0.013 | 0.025 |
| za | 1,143 | 0.531 | 0.499 | 0.015 | 0.029 |

***Table 1. Estimated marginal mean contrasts for reaction times to targets in each ordinal position. All contrasts reached statistical significance at the 5% alpha threshold. Contrasts were performed on model output from the generalized linear model described above.***

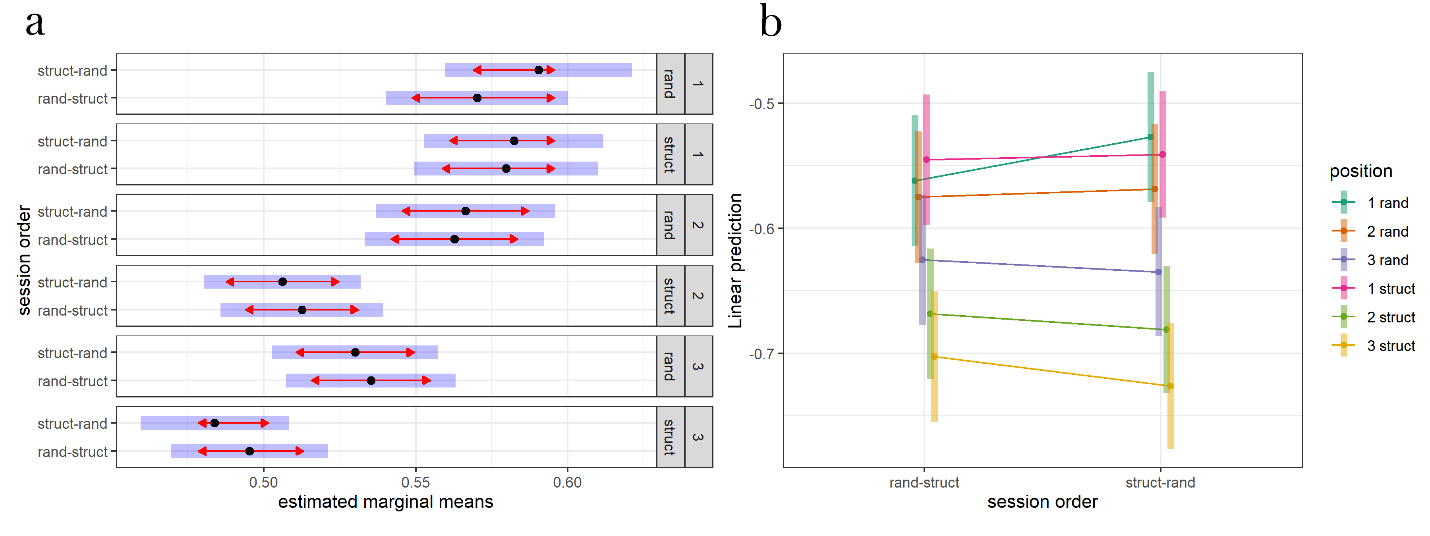
In Experiment 2, accuracy was again modulated by ordinal position in the structured condition (). **(Fig. S2a**) [If we do the lm/anova on per-subject accuracy,] We found no difference in accuracy as a function of odinal position in the structured condition ().

We also found that accuracy varied by pseudoword in the structured condition (**Fig. S2b, Table S2a**, ). [If we do the lm/anova on per-subject accuracy,] We found no difference in accuracy as a function of pseudoword in the structured condition ().

Lastly, we checked for an effect of syllable identity in both sessions. We found an interaction between these two factors (), with certain syllables having higher accuracy values than others in both conditions. (**Fig. S2c, Table S2b**) [If we do the lm/anova on per-subject accuracy,] We found no interaction between session and target identity, but main effects of both on detection accuracy (). As reported above, detection accuracy in the structured condition was higher than in the random condition. When this contrast was performed (as in the main text) on subject-wise accuracy with a df of 37.5, here, we averaged over detection accuracy per syllable for each subject. Thus, here we observe a significant contrast (). Meanwhile, when considering contrasts for accuracy to targets, collapsed over both conditions, only a few contrasts reached significance: ga-di, po-be, se-be, ki-be (all p < 0.05). This would suggest primarily, lower detectability for a particular syllable, *be*.

**Fig. S2. Mean accuracy for Experiment 2.

**Table 2. Estimated marginal mean contrasts for levels of position within each condition. In the structured condition, all contrasts reached significant, replicating the results from Exp. 1. In the random condition, contrasts between positions 1 and 3 as well as 2 and 3 reached significance. S-values reflect degree of surprisal. Infinite s-values indicate p-value is too small to accurate calculate s-value.**

****

**Fig. S3. No interaction of condition order with condition and ordinal position**

|  |  |  |  |
| --- | --- | --- | --- |
| **Table S3. GLM Results** | | | |
|  | | | |
|  | *Dependent variable:* | | |
|  |  | | |
|  | reaction time (s) | | |
|  | lesser | lesser (random slopes) | fuller |
|  | (1) | (2) | (3) |
|  | | | |
| Intercept(Pos 1) | -0.63\*\*\* (-0.68, -0.57) | -0.64\*\*\* (-0.68, -0.61) | -0.69\*\*\* (-0.76, -0.62) |
| Pos 2 | -0.16\*\*\* (-0.18, -0.14) | -0.14\*\*\* (-0.19, -0.10) | -0.11\*\*\* (-0.17, -0.05) |
| Pos 3 | -0.23\*\*\* (-0.25, -0.21) | -0.22\*\*\* (-0.28, -0.15) | -0.18\*\*\* (-0.24, -0.12) |
| Block 2 |  |  | 0.07\*\* (0.01, 0.13) |
| Block 3 |  |  | 0.08\*\*\* (0.02, 0.15) |
| Block 4 |  |  | 0.08\*\* (0.02, 0.14) |
| Block 5 |  |  | 0.11\*\*\* (0.05, 0.17) |
| Block 6 |  |  | 0.03 (-0.03, 0.09) |
| Block 7 |  |  | 0.05 (-0.01, 0.11) |
| Block 8 |  |  | 0.06\* (-0.005, 0.12) |
| Pos 2:Block 2 |  |  | -0.04 (-0.12, 0.04) |
| Pos 3:Block 2 |  |  | -0.03 (-0.11, 0.05) |
| Pos 2:Block 3 |  |  | -0.07\* (-0.16, 0.01) |
| Pos 3:Block 3 |  |  | -0.03 (-0.11, 0.05) |
| Pos 2:Block 4 |  |  | -0.05 (-0.13, 0.04) |
| Pos 3:Block 4 |  |  | -0.07 (-0.15, 0.01) |
| Pos 2:Block 5 |  |  | -0.06 (-0.14, 0.03) |
| Pos 3:Block 5 |  |  | -0.14\*\*\* (-0.22, -0.06) |
| Pos 2:Block 6 |  |  | -0.01 (-0.10, 0.07) |
| Pos 3:Block 6 |  |  | 0.01 (-0.06, 0.09) |
| Pos 2:Block 7 |  |  | -0.06 (-0.15, 0.02) |
| Pos 3:Block 7 |  |  | -0.09\*\* (-0.17, -0.004) |
| Pos 2:Block 8 |  |  | -0.06 (-0.14, 0.03) |
| Pos 3:Block 8 |  |  | -0.06 (-0.14, 0.02) |
|  | | | |
| Fixed Effects | Subject | Position | Subject | Subject |
| Fixed Effects Struct. | Rand. Int. | Rand. Int., Slope | Rand Int. |
| Observations | 9,531 | 9,531 | 9,531 |
| Log Likelihood | 3,199.61 | 3,312.13 | 3,224.15 |
| Akaike Inf. Crit. | -6,389.23 | -6,604.26 | -6,396.29 |
| Bayesian Inf. Crit. | -6,353.42 | -6,532.64 | -6,210.07 |
|  | | | |
| *Note:* | \*p\*\*p\*\*\*p<0.01 | | |
|  | Fitted using Gamma distribution and log link function. | | |

|  |  |
| --- | --- |
| **Table S4. GLM Results** | |
|  | |
|  | *Dependent variable:* |
|  |  |
|  | reaction time (s) |
|  | |
| Intercept(Pos 1/Rand) | -0.54\*\*\* (-0.58, -0.50) |
| Struct | 0.002 (-0.04, 0.04) |
| Pos 2 | -0.03\*\* (-0.05, -0.01) |
| Pos 3 | -0.08\*\*\* (-0.10, -0.06) |
| Struct:Pos 2 | -0.11\*\*\* (-0.14, -0.08) |
| Struct:Pos 3 | -0.09\*\*\* (-0.12, -0.06) |
|  | |
| Fixed Effects | Condition Order/Subject + Condition |
| Fixed Effects Struct. | Rand. Int. + Rand. Slope |
| Observations | 6,524 |
| Log Likelihood | 4,140.28 |
| Akaike Inf. Crit. | -8,250.56 |
| Bayesian Inf. Crit. | -8,148.81 |
|  | |
| *Note:* | \*p\*\*p\*\*\*p<0.01 |
|  | Fitted using Gamma distribution and log link function. |