**Online and offline measures of statistical learning reflect sensitivity to pairwise relationships, but not chunking**

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

[**Introduction** 2](#_Toc51682728)

[**Results** 3](#_Toc51682729)

[**Experiment 1** 3](#_Toc51682730)

[**Experiment 2** 8](#_Toc51682731)

[**Feature Sensitivity in Online Target Detection** 10](#_Toc51682732)

[**Conclusion** 13](#_Toc51682733)

[**Methods** 13](#_Toc51682734)

[**Experiment 1** 13](#_Toc51682735)

[**Experiment 2** 16](#_Toc51682736)

[**Acknowledgements** 17](#_Toc51682737)

[**Contributions** 17](#_Toc51682738)

[**Competing Interests** 18](#_Toc51682739)

[**References** 18](#_Toc51682740)

[**Figure Legends** 21](#_Toc51682741)

[**Supplementary Materials** 22](#_Toc51682742)

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

\* *Corresponding Author.* Email: [ava.kiai@ae.mpg.de](mailto:ava.kiai@ae.mpg.de)

**Abstract**

Statistical learning allows individuals to rapidly extract regularities in the sensory environment. We replicated previous findings showing participants become sensitive to the implicit structure in a continuous speech stream of repeating tri-syllabic pseudowords, as measured by an online target detection task and offline word recognition task. Consistent with previous findings, we found only a weak correlation between these two measures of learning, leading us to question whether there is overlap between the information captured by these two tasks. Representational similarity analysis on reaction time data from the target detection task revealed that reaction time data reflects sensitivity to transitional probability, ordinal position, and pairwise groupings, but not triplet (pseudoword) groupings. Furthermore, individual performance on the word recognition task was only significantly predicted by their sensitivity to transitional probability. We conclude that these canonical SL tasks do not reliably measure chunking of embedded units (e.g. tri-syllabic pseudowords), but more simply a sensitivity to transitional probability and pairwise relationships of between stimulus items (e.g. syllables).

# **Introduction**

Statistical learning (SL) is a powerful mechanism by which the brain is able to detect subtle regularities in the sensory environment, which it can subsequently employ for cognitive operations, such as segmenting and representing units from continuous input, and predicting upcoming events. [1], [2] SL has been called a “domain-general mechanism,” as humans are capable of extracting low-level distributional properties from different sensory modalities, including sequences of auditory [3]–[6] and visual stimuli. [7]–[9]. Furthermore, SL is also known to be both fast and robust, typically requiring only a few minutes of exposure [3], [6], [10], leading to improved performance on behavioral tasks [11], with moderate test-retest reliability within individuals. [12]

The canonical measure of SL, as per [3], is administered after a learning phase and asks participants to discern which of two test items is more “familiar” (or, was previous observed during exposure). Success on the task consists of above-chance discrimination of the items that made up the stimulus stream (typically, those demarcated by low transitional probabilities).

More recently, studies have examined SL through a combination of tasks, including some from of “online” task. These online tasks have the potential to provide rich insight into the process of extracting regularities from continuous input, as they stand 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 whether it varies by stimulus). Typically, online SL tasks entail asking participants to detect a target stimulus via keypress while being presented with the exposure stream, and measuring reaction time (RT) to targets (“target detection tasks”). Studies employing this task in both visual [7], [13], [14] and auditory [15]–[17] SL studies all report that RTs to targets are modulated by the predictability of the target, suggesting sensitivity to statistical regularities facilitates the speed of detection. Targets with lower transitional probabilities elicit longer RTs than targets with higher transitional probabilities. This RT effect can be observed as early as the second presentation of a target stimulus. [10] Similarly, in Gómez et al., participants detected click sounds embedded in continuous speech, to the same effect: clicks occurring between embedded pseudowords were 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)

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), though the correlation coefficients tend to be rather small. (, Batterink & Paller, 2017; , Siegelman, Bogaerts, Kronenfeld, & Frost, 2018) Meanwhile, others report finding no correlation. [11], [13], [15], [18], [20]

The weak or nonexistent correlation between SL measures been largely discussed on a theoretical level. (Franco, Eberlen, et al., 2015; Gómez, Bion, & Mehler, 2010, but see Batterink, Reber, Neville, et al., 2015, for an empirical treatment of the subject) Some authors note that the discrepancy could be due to the simple fact that the RT tasks can be performed merely with implicit knowledge of the regularities, while the recognition task demands that this information be made explicit. [15] Alternatively, the weak relationship has also been ascribed to the different psychometric sensitivity of the two tasks: target detection tasks typically test all stimulus items in a larger number of trials over a longer test period, while word recognition tasks are often designed to test memory for only 4-8 items, and rarely exceed 36. (Siegelman et al., 2017)

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 then tested for a relationship between these two measures. In Experiment 2, we replicated our main target detection task results from Experiment 1, and added an additional control condition (where syllables were randomly ordered and which contained no statistical regularities) to confirm that our results were not artifacts of the stimuli or other confounds. Finally, we performed a representational similarity analysis on the combined target detection task data from these two experiments to gain insight into what structural features of the syllable stream (transitional probability, ordinal position, triplet grouping, duplet grouping) might be captured by reaction time data, in order to shed light on the empirical disparity between implicit, online and explicit, offline tests of SL.

# **Results**

## **Experiment 1**

To ensure that participants were able to perform the online target detection task, we computed mean detection accuracy across all trials (), which was above a 0.5 chance level (). We also calculated subject-wise mean accuracy for each ordinal position, word, and target syllable. Since we observed significant differences for these factors, we carried out two controls to ensure that differences in the number of observations for ordinal position and any unwanted variability in the stimuli did not drive our main findings. (See Supplementary Materials; **Table S1, Fig. S1**.)

#### **Ordinal Position Modulates Reaction Time**

Using a generalized linear mixed effects model (GLMM) with ordinal position as predictor, we found that reaction times are modulated by the ordinal position of the target syllable within the pseudoword (). (**Fig. 1a-b**) 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 (position 1) () were notably slower than those to word-medial (position 2) () and word-final syllables (position 3) (). The difference in mean RT between word-medial and word-final positions was smaller but still significant (). 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.

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

Although we found that the model that excluded the factor block was better able to represent our data, we wanted to directly test whether the graded RT effect emerged over the course of several blocks, or was present from the first block. We computed an ANOVA on our fuller model (as described in the previous section, see **Table S3**), and found a significant interaction between block and ordinal position (). Since each of our blocks are ~ 3 minutes long, we predicted that participants would have extracted the embedded regularities of the stream within the first two blocks. Indeed, previous studies found that only a few minutes of exposure [22] and even only a few occurrences of an embedded target syllable [10] were sufficient for participants to pick up the statistical structure, as measured by offline word recognition and RT, respectively. Furthermore, longer exposures have been shown to provide little benefit for learning. [6] Thus, we focused our follow-up analysis only on the first two blocks.

To specifically examine the change in the RT pattern between the first two blocks, we performed an ANOVA on a model identical in structure with that above, but using only data from blocks 1 and 2. Here, we observed no interaction, but significant main effects for both block () and ordinal position () factors. The effect of block was driven by a small, overall increase in RT between blocks 1 and 2 (), while contrasts between each position revealed a graded RT effect that was largest for position 1-2 () and 2-3 mean () differences (), but also significant for position 2-3 (). Since we did not find evidence that the RT effect was modulated by block (no interaction), we focused our next analysis on block 1. An ANOVA performed on a model using data only from block 1 and therefore only ordinal position as predictor revealed a main effect of ordinal position () and, between each level, the same graded RT effect (). These results suggest that the modulation of RT by ordinal position emerged during the first block, and remained relatively stable throughout the remainder of the experiment. (**Fig. 1c**) 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.

We also computed a measure of online statistical learning based upon the target detection task, borrowed from Siegelman and colleagues [19], as follows:

1. )

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 (). (**Fig. 1d**) This is an additional piece of evidence in support of early emergence of differentiated reaction times and stable behavior thereafter.

Lastly, we wanted to address a possible confound that could have helped generate the RT effect. As argued by Himberger et al. [20], the widely observed RT effect could be driven by a trivial overall speeding-up of RTs. They propose that since position of a target in a triplet and in the stream is confounded (both occur later for 3rd position targets than for 1st position targets), the graded RT effect could be spurious, driven by an overall decrease in RT and not sensitivity to stream statistics. For this critique to be valid, the mean difference in RT to targets in positions 1 - 2 and 2 – 3 must be equally large, and RTs must decrease over the course of exposure (both within trials and across blocks). To address this confound, we first performed a linear regression on mean RT differences between each position pairing (1-2, 2-3, and 1-3) with both block and pairs as predictors. We found no interaction between block and pairs, suggesting that the magnitude of the differences between positions, as well as the relationships between them, did not change over the course of the blocks (). We found only a main effect of pairs (). (**Fig. S2**) Tukey-corrected contrasts between pairs showed the mean differences between position 1-2 were significantly larger than those between positions 2-3 (), supporting the notion that the RT effect is not linear between the positions. Since the decrease in RT observed between positions 1 and 2 is indeed larger than between 2 and 3, the effect cannot be due to a simple confound of monotonically decreasing RTs over the course of the session. Additionally, there was no effect of block on overall RT () nor any difference between RTs when looking specifically at blocks 1 and 8 (). Finally, we also tested to see if RTs decrease monotonically over single trials (streams of 72 words). A linear model with block and target number (occurrences of each target within each trial) revealed no interaction of the factors on RT, but main effects of each. We therefore re-ran the model with only target number as predictor, which again reveled a main effect (). However, we observed that variation in RT between target numbers was not unidirectional, but wavered above and below the mean RT. (**Fig. S2**) Together, these results provide evidence against a claim that a generic decrease in RT spuriously induced the observed RT effect.

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

In the word recognition task, participants correctly distinguished the pseudoword from the part-word foils significantly above chance (chance level = 50%, or 8 out of 16 trials) (), indicating that participants were sensitive to the implicit regularities of the syllable stream and able to use this information to explicitly discriminate pseudowords from sequences of syllables that crossed word boundaries. **(Fig. 1e)**. 71% of participants (27 out of 38) completed the task with a mean accuracy greater than chance.

In an exploratory analysis, we also calculated the proportion correct responses for each pseudoword individually (out of 4 trials), 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 part-word foil could be sufficient to push a participant’s performance above chance level. We found that across participants, 3 out of the 4 pseudowords were discriminated above chance (2 out of 4 trials) () (**Fig. S3**)

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

Next, we calculated the measure of online SL (see equation 1 above) for each participant and correlated each individual’s online SL scores with their word recognition accuracy. We found a weak correlation between these two values at the 5% alpha level (). (**Fig. 1f**) Since we did not have complete data for all participants in both tasks, for this analysis we used data only from those participants with complete data sets (N = 32).

To compare this finding with previous literature which also found a correlation between these values, we repeated this analysis, but using the procedure from [17] to compute the “RT score”: . This analysis revealed an even weaker correlation, which did not reach statistical significance (). (**Fig. S4a**) Finally, we performed a third version of this analysis, where we considered the correlation between word recognition accuracy and the median difference between each position pairing (i.e. 1-2, 2-3, and 1-3), to see if certain pairs might better predict word recognition accuracy. To obtain “RT scores” that are comparable between participants, we z-normalized RT values for each participant, computed median RTs to each ordinal position, and computed the difference between the scaled median RTs for each position pairing for each participant. These values were correlated against the participant’s word recognition accuracy. We again found weak correlation between the two measures for all pairs (). (**Fig. S4b**)

#### **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). Likewise, our online word recognition task revealed results consistent with previous literature: targets in predictable locations (here, word-medial and word-final positions) elicited faster reaction times than targets in less predictable locations (word-initial positions). The rapid onset of this graded RT effect also corroborates previous findings stating statistical learning is a fast and robust mechanism.

However, it remains puzzling why these two measures of statistical learning are so weakly correlated. We followed procedures used by two earlier papers, as well as one of our own, to determine if a correlation exists between the online and offline measures, but only one of these methods revealed a significant relationship between performance in the two tasks.

We wished to investigate why graded response times in the online task failed to adequately predict accuracy in offline word recognition. On the basis of the literature reviewed, we reasoned that the lack of correlation may be due to the fact that the two tasks in fact measure learning of different features of the same input. Each of the syllables in the stream can be characterized by its membership to a specific group for each of the following features: ordinal position (1, 2, or 3), transitional probability (1 or 0.33), within-word duplet identity (*nu-ga, ga-di, ro-ki, ki-se*, etc.), and word or triplet identity (*nugadi, rokise*, etc.). “Success” on the online task (defined as: expressing a graded RT effect) minimally requires a tracking of transitional probability, but crucially can be achieved without any tracking of pseudoword triplets or the duplets that comprise them. Meanwhile, success in the offline task may necessitate some tracking of word identity and/or duplet identity. If indeed RTs do not capture any information related to word or duplet identity, this might explain the lack of correlation.

To empirically test this hypothesis, we ran a second experiment in which participants performed only the online detection task. We aimed to replicate the graded RT effect we found in the first experiment, and use the combined data set to determine whether we can uncover specific feature coding in RTs. Experiment 2 also consisted of an exposure phase/target detection task with a random stream (one created with the same stimuli but lacking any statistical regularities), which would allow us to confirm that our reported effects are driven primarily by the statistical regularities in the stream and not by unwanted variation in the stimuli acoustics.

## **Experiment 2**

As in Experiment 1, we first examined mean detection accuracy to ensure participants were engaged in the task. Overall detection accuracy () was significantly above a 0.5 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 (). We additionally tested for an effect of ordinal position, pseudoword, or syllable identity on accuracy. We found that accuracy varied by target syllable and following the same procedure as in Experiment 1 to ensure this source of variability did not spuriously induce the graded RT effect. (See Supplementary Materials, **Fig. S1**)

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

Our hypothesis stated that reaction times (in seconds) would be predicted by a combination of ordinal position and condition (structured, random). Using a GLMM with condition and target position as predictors, we observed an interaction between ordinal position and condition (). 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). As in Experiment 1, we observed slower RTs to word-initial syllables () than to word-medial syllables (), and word-final syllables (). The drop in mean RT between word-medial and word-final syllables was smaller but also significant (). (**Fig. 2a-b**).

Surprisingly, we observed the same general RT pattern in the random condition, although with far smaller marginal differences. Here, mean RTs to word-initial () were longer than those to word-medial () and to word-final () syllables as well. RTs to word-medial syllables were also longer than RTs to word-final syllables (). (**Fig. 2a-b**) 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 addition, we ran a linear model using condition and target number as predictors to confirm that a potential confound of monotonically decreasing RTs over the course of each trial did not drive this effect. We found an interaction between condition and target number (). However, RTs in fact followed a quadratically-shaped pattern over the course of trials for both conditions, generally increasing until occurrence 12-14, and then decreasing. (**Fig. S2**) We concluded that this confound was not present in our data.

In our second contrast, we evaluated the effect of condition for each level of ordinal position (i.e. how much condition affected RT to targets in each ordinal position). We observed that the presence of structure significantly decreased mean RT for word-medial () and word-final targets (). However, RTs to word-initial targets did not significantly vary between conditions. (). Together, these results demonstrate that the presence of an underlying structure in a continuous speech stream modulates RT to predictable target syllables above and beyond what might be observed due to random variation, whether that source is variability in participant performance or stimuli acoustics.

#### **Discussion**

In Experiment 2, we were able to replicate our main finding from Experiment 1. Namely, RTs to predictable targets in a continuous speech stream with embedded regularities, in the form of repeating tri-syllabic pseudowords, were significantly faster than those to less predictable targets. Notably, we established that this effect is not due to spurious variation in the stream or the differing detectability of target syllables, since the presence of structure magnifies the graded RT effect in comparison with that observed from exposure to a randomly ordered syllable stream.

Himberger and colleagues recently argued that the graded RT effect observed in numerous studies is an artifact unrelated to the regularities that experimenters expect participants to learn, but rather a consequence of general RT facilitation. [20] Several key features of our design make it unlikely that the RT effect we observed in Experiment 1 and in the structured condition of Experiment 2 was the result of such an artifact. First, our design did not confound the position of the target in the stream and in the word; streams began and ended with an amplitude ramp so that participants could not easily discern onsets and offsets, the first syllable could belong to any random position, and words were repeated in a random order a total of 72 times (in other words the permutation space was very large). Secondly, since our streams were longer than only a few triplet presentations and required participants to provide responses to 18-19 targets within a minute, it is highly unlikely that the speeded RT effect could have resulted from having a non-constant hazard rate (where the more time has passed before a target is heard, the faster the RT is). Third, RTs in both experiments did not trend towards monotonically faster responses; RTs in Experiment 1 hovered around the mean, while RTs in both conditions in Experiment 2 increased for a majority of the trial. Thus, faster responses to later stream positions could not explain our results.

## **Feature Sensitivity in Online Target Detection**

We used a representational similarity analysis to explore the reaction time data from the online target detection tasks further in order to determine if the observed patterns of RTs could reveal sensitivity to any specific features of the structured streams. Specifically, we hypothesized that one of four possible features could be encoded in the RT data: transitional probability, ordinal position, word identity, and duplet identity. We predicted that a failure to track one or more of these features (especially word and duplet identity) 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 similarity matrix (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 correlation values. We tested for the coding of each of the features mentioned above by running four ranked sum tests. For each test, we identified a *within* and an *across* group. Within groups consisted of the correlation values between all pairs of syllables characterized by that feature (e.g. correlation between two word-initial syllables). Across groups consisted of correlations between pairs where the pairing violates the feature (e.g. correlation between a word-initial and word-medial syllable) or represent of the opposite feature type (i.e. two levels of transitional probabilities). For each test, we performed a random sampling of correlation values from the respective cells, from all participants, to generate an array of *within* and array of *across* values. 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 paired ranked sum analyses 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 paired, two-sided Wilcoxon’s rank sum test on the resulting two arrays to determine whether similarity (as a proxy for feature coding) is higher within or across groups.

#### **RTs Track Ordinal Position, Transitional Probability, and Duplets**

For the test of ordinal position (N = 467), *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 pseudoword (e.g. nu-ga, ga-di, nu-di). Similarity within ordinal positions was higher than across ordinal positions ().

For the test of transitional probability (N = 239), *within* values included the correlation between all pairs of word-initial syllables (TP = 0.33). *Across* values consisted of the correlation between all pairs of word-medial and word-final syllables (TP = 1). Similarity was higher for the *within* as compared with the *across* group. ().

For the test of word identity (N = 467), *within* values included the correlation between all pairs of syllables within each pseudoword (e.g. nu-ga, ga-di, nu-di). *Across* values included “phantom” word pairs where each item in the pair is drawn from different pseudowords (e.g. nu-ki, nu-se). Here we observed no significant difference in the similarity between groups. ().

Finally, for the test of duplet identity (N = 156), *within* values were correlations between all pairs of consecutive syllables within pseudowords (e.g. nu-ga, ga-di), while *across* values were correlations between word-initial and word-final syllables within each psuedoword (e.g. nu-di). *Within* similarity was higher than *across* similarity. ().

Together, these analyses suggest that RT in the online detection task reflect sensitivity to categories of ordinal position, transitional probability, and duplet pairings; while failing to do so for word identity. (**Fig. 3**)

#### **Sensitivity to Transitional Probability Weakly Predicts Word Recognition Performance**

In the previous analysis, we found that RTs to targets in the online target detection task could reveal a sensitivity to three of the four features we had outlined: ordinal position, transitional probability, and duplet identity. Finally, we wished to test whether there exists a correlation between word recognition performance and sensitivity to any of these four features, as measured by within-across mean similarity, for each individual.

We first computed mean similarity for *within* and *across* groups for each of the four features for each participant. As above, we took each participant’s z-transformed correlation matrix of RTs to syllables, and subset correlation values using the same rubric as above to fill *within* and *across* groups. (However, here we did not perform a bootstrapping procedure.) We then subtracted the mean of *within* group similarity values from *across* group similarity values for each participant and each feature, to obtain four similarity measures for each individual. Similarity measures for each feature were correlated with word recognition performance. There was virtually no relationship between word recognition accuracy and ordinal position (), word identity (), and duplet identity (). However, within-across similarity for transitional probability was weakly correlated with word recognition accuracy (). (**Fig. 4**) (As we predicted the correlation would be positive for each feature, we performed one-sided tests for all features. However, our results did not change if we conducted the tests with a two-sided alternative (, ,)).

#### **Discussion**

Our representational similarity analysis revealed that RT can contain information about transitional probability, ordinal position, and duplet pairing of syllables in a structured stream, but not their word identity. This result appears intuitive on the basis of the results and discussion above, yet provides a few additional insights. First, it lends additional support for the claim that individuals are able to extract information about both ordinal position and transitional probability. Second, it suggests that participants are specifically sensitive to pairwise relationships between syllables, evidenced by the significant similarity for duplets. Finally, we found no similarity between RTs to syllables within versus across word boundaries. This fact may be key to why online and offline measures of SL correlate so poorly. Indeed, while online measures reflect 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. [1]

Saffran et al.’s original study concluded with the suggestion that infants exposed to the continuous syllable stream “succeeded in learning and remembering particular groupings of three-syllable strings.” [3] Since this seminal study, it has been proposed that explicit memory of the word chunk is what allows successful performance in the words vs. part-word discrimination task. [23], [24]

Yet, success in the offline task does not entail that an individual was able to generate a unified representation of the syllable triplets that formed the pseudowords. A sensitivity to transitional probabilities (mere segmentation) may be sufficient to achieve this. When a pseudoword is played in isolation, as in during the word recognition task, the transitional probabilities between the heard syllables sum to 2 (TP=1 between syllable 1 and 2, and also 1 between syllable 2 and 3). However, when hearing a part-word, the transitional probabilities sum to 1.33 (TP=0.33 between syllable 1 and 2, and 1 between syllable 2 and 3). Although we do not claim that participants actively sum transitional probabilities during the discrimination task, it is plausible that mere sensitivity to this feature could help an individual correctly select the pseudoword despite having no concept of the triplet grouping.

The notion that tracking transitional probabilities alone would allow above-chance performance on the pseudoword vs. part-word recognition task, but not be sufficient for “chaining” or “chunking” more than two items is supported by Endress & Mehler’s findings. [6] In their study, 5 minutes of exposure was sufficient for participants to discern words from part-words, but even after 40 minutes of exposure, they still could not discern words from phantom words (tri-syllabic sequences where syllables are drawn from different words, but ordered so as to maintain their original position in the word). Only with the introduction of linguistic cues (final syllable lengthening and pauses between words), did participants reject phantom words as often as they rejected part-words without additional cues. Another study found that correct judgments of whether a novel triplet belonged to an exposure stream of repeating visual shapes were improved by how closely the summed transitional probabilities within the triplet were to the original triplet structure. [25]

Nonetheless, it is uncontroversial to claim that tracking transitional probabilities is an ability that underlies the processing of any complex stream of information [26], and forms the basis of more complex processes, such as those involved in language acquisition. However, the tracking of transitional probabilities is only one of several mechanisms that humans can deploy to acquire artificial and real languages. [27]

# **Conclusion**

Online measures of statistical learning, such as the target detection task, can reveal subtle, dynamic properties of implicit learning. Meanwhile, offline tasks that require explicit discrimination between embedded pseudowords and foils show that participants can use implicitly acquired knowledge to make explicit decisions. We found evidence of learning using both tasks. In line with some earlier studies, we found a weak correlation between measures of learning in these two tasks. We used a representational similarity analysis to determine if this lack of correlation is due to the fact that reaction times to embedded targets provide orthogonal information about what was learned, compared with what is needed to accomplish the word recognition task. We found that RT patterns captured transitional probability, ordinal position, and duplet pairings of target syllables, but did not reflect pseudoword groupings. Furthermore, similarity within transitional probabilities was the only feature that could predict (albeit weakly) word recognition performance. We conclude that these canonical tasks of statistical learning afford little evidence that participants learned the higher-order units embedded in the stream, but rather an acquired sensitivity to co-occurring pairs of elements.

# **Methods**

## **Experiment 1**

#### **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 [3], 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 (26 female, mean age, 23.78 ± 9.31 sd). All participants reported having normal hearing, gave informed consent as approved by the Ethics Council of the Max-Planck Society, and were paid for their participation. 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 [17] 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 and 4.0.2 (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 ().

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;** lesser vs. fuller random slopes models deviance was and , respectively; ). Thus, we conducted further analysis on results of the lesser model.

## **Experiment 2**

#### **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 (16 female, mean age 24.50 ± 9.92 sd). All participants reported having normal hearing, gave informed consent as approved by the Ethics Council of the Max-Planck Society, and were paid for their participation. Inclusion criteria included the requirement to not have taken part in Experiment 1. 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 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 and 4.0.2 (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.

We performed a modelling procedure similar to that from Experiment 1. 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.)

# **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.

# **Competing Interests**

The authors declare no competing interests.

# **References**

[1] S. Dehaene, F. Meyniel, C. Wacongne, L. Wang, and C. Pallier, “The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees,” *Neuron*, vol. 88, no. 1, pp. 2–19, Oct. 2015.

[2] B. C. Armstrong, R. Frost, and M. H. Christiansen, “The long road of statistical learning research: past, present and future.,” *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 372, no. 1711, p. 20160047, Jan. 2017.

[3] J. R. Saffran, R. N. Aslin, and E. L. Newport, “Statistical Learning by 8-Month-Old Infants,” *Science (80-. ).*, vol. 274, no. December, pp. 1926–1928, Nov. 1996.

[4] B. Pelucchi, J. F. Hay, and J. R. Saffran, “Learning in reverse: Eight-month-old infants track backward transitional probabilities,” *Cognition*, vol. 113, no. 2, pp. 244–247, Nov. 2009.

[5] Richard N. Aslin, J. R. Saffran, and E. L. Newport, “Computation of Conditional Probabilities by Infants,” *Psychol. Sci.*, vol. 9, no. 4, pp. 321–324, 1998.

[6] A. D. Endress and J. Mehler, “The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words,” *J. Mem. Lang.*, vol. 60, 2009.

[7] N. B. Turk-Browne, J. A. Jungé, and B. J. Scholl, “The Automaticity of Visual Statistical Learning,” *J. Exp. Psychol. Gen.*, vol. 134, no. 4, pp. 552–564, 2005.

[8] N. B. Turk-Browne, B. J. Scholl, M. K. Johnson, and M. M. Chun, “Implicit Perceptual Anticipation Triggered by Statistical Learning,” *J. Neurosci.*, vol. 30, no. 33, pp. 11177–11187, 2010.

[9] S. Henin, N. B. Turk-Browne, D. Friedman, W. Doyle, O. Devinsky, and L. Melloni, “Learning hierarchical sequence representations across human cortex and hippocampus Keywords,” pp. 1–22, 2020.

[10] L. J. Batterink, “Rapid Statistical Learning Supporting Word Extraction From Continuous Speech,” *Psychol. Sci.*, pp. 1–8, 2017.

[11] N. Siegelman and R. Frost, “Statistical learning as an individual ability: Theoretical perspectives and empirical evidence,” *J. Mem. Lang.*, vol. 81, pp. 105–120, May 2015.

[12] N. Siegelman, L. Bogaerts, and R. Frost, “Measuring individual differences in statistical learning: Current pitfalls and possible solutions,” *Behav. Res. Methods*, vol. 49, no. 2, pp. 418–432, Apr. 2017.

[13] A. Franco, J. Eberlen, A. Destrebecqz, A. Cleeremans, and J. Bertels, “Rapid Serial Auditory Presentation,” *Exp. Psychol.*, vol. 62, no. 5, pp. 346–351, Nov. 2015.

[14] J. Bertels, A. Franco, and A. Destrebecqz, “How implicit is visual statistical learning?,” *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 38, no. 5, pp. 1425–1431, Sep. 2012.

[15] L. J. Batterink, P. J. Reber, H. J. Neville, and K. A. Paller, “Implicit and explicit contributions to statistical learning,” *J. Mem. Lang.*, vol. 83, pp. 62–78, 2015.

[16] L. J. Batterink, P. J. Reber, and K. A. Paller, “Functional differences between statistical learning with and without explicit training,” *Learn. Mem.*, vol. 22, pp. 544–555, 2015.

[17] L. J. Batterink and K. A. Paller, “Online neural monitoring of statistical learning,” *Cortex*, vol. 90, pp. 31–45, May 2017.

[18] A. Franco, V. Gaillard, A. Cleeremans, and A. Destrebecqz, “Assessing segmentation processes by click detection: online measure of statistical learning, or simple interference?,” *Behav Res*, vol. 47, pp. 1393–1403, 2015.

[19] N. Siegelman, L. Bogaerts, O. Kronenfeld, and R. Frost, “Redefining ‘Learning’ in Statistical Learning: What Does an Online Measure Reveal About the Assimilation of Visual Regularities?,” *Cogn. Sci.*, vol. 42, pp. 692–727, Jun. 2018.

[20] K. D. Himberger, A. S. Finn, and C. J. Honey, “Reconsidering the automaticity of visual statistical learning,” *PsyarXiv*, pp. 319–323, 2019.

[21] D. M. Gómez, R. A. H. Bion, and J. Mehler, “The word segmentation process as revealed by click detection,” *Lang. Cogn. Process.*, vol. 26, no. 2, pp. 212–223, 2010.

[22] J. R. Saffran, E. L. Newport, and R. N. Aslin, “Word Segmentation: The Role of Distributional Cues,” *J. Mem. Lang.*, vol. 35, no. 4, pp. 606–621, Aug. 1996.

[23] P. Perruchet and A. Vinter, “PARSER: A Model for Word Segmentation Pierre,” *J. Mem. Lang.*, vol. 39, pp. 246–263, 1998.

[24] P. Perruchet and S. Pacton, “Implicit learning and statistical learning: one phenomenon, two approaches,” *Trends Cogn. Sci.*, vol. 10, no. 5, pp. 233–238, 2006.

[25] T. A. Forest, A. S. Finn, and M. L. Schlichting, “What is Represented in Memory after Statistical Learning ?,” in *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 2020, pp. 1882–1888.

[26] F. Meyniel, M. Maheu, and S. Dehaene, “Human Inferences about Sequences: A Minimal Transition Probability Model,” *PLoS Comput. Biol.*, vol. 12, no. 12, p. 1005260, 2016.

[27] A. D. Endress and L. L. Bonatti, “Words, rules, and mechanisms of language acquisition,” *Wiley Interdiscip. Rev. Cogn. Sci.*, vol. 7, no. 1, pp. 19–35, 2016.

# **Figure Legends**

# **Supplementary Materials**

***Stimuli***

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

***Supplementary Information***

In Experiment 1, detection accuracy was modulated by ordinal position (, with each successive position having a higher mean accuracy (). Pairwise contrasts revealed mean differences between RTs to each position to be significant (). Accuracy also varied between the four words (). Specifically, between words *nugadi* and *zabetu* (), *rokise* and *zabetu* (), and *mipola* and *zabetu* (). (**Table S1a**.) This suggests that the accuracy was overall equal between all words except *zabetu*, for which detection was more difficult. Finally, we found that syllable identity also affected detection accuracy (). Certain CV syllable pairs may have been easier to detect than others, due to minor variations in stimuli acoustics. (**Table S1b., Fig. S1a**)

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 significant pairwise contrasts between all three levels of ordinal position (). Therefore, we concluded that slight variations in the acoustics of our stimuli did not significantly affect our results*.* In addition, to ensure that the difference in detection accuracy between ordinal positions did not affect the results, we ran the same model we used for our primary analyses with a subsample of RT data. Since accuracy varied between positions, the number of observations for each levels of factor ordinal position varied also. We subsampled our data set so that the observations for all ordinal positions were equal to the smallest group (). We observed the same main effect of ordinal position (). All pairwise contrasts between ordinal positions reached significance (). Given this result, and the fact that generalized mixed models are robust to unbalanced data sets, we concluded that the unequal number of observations between position levels did not skew our results.

In Experiment 2, we found no difference in accuracy as a function of ordinal position in the structured condition (). We also 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 conditions. We found no interaction between condition 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. Pairwise contrasts revealed a significant difference in mean accuracy between the two groups (). (Note however that the t-test on accuracy in random vs. structured, reported in the main text, did not reach significance, with 36.2 degrees of freedom) Meanwhile, when considering contrasts for accuracy to targets, collapsed over both conditions, only a few contrasts reached significance: *ga-di, di-be, ki-be, se-be, po-be* (all p < 0.05). This would suggest the effect is mainly driven by lower detectability for particular syllables, e.g. *be*. (**Fig. S1b**)

To ensure this variability in detection accuracy to syllables did not interfere with our main results, we again validated our 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 RT (in seconds) 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 significant pairwise contrasts between all three levels of ordinal position (). Therefore, we concluded that, once again, minute differences between stimuli did not affect our main results.

We additionally wished to confirm that there were no carry-over effects of condition, where the order in which participants saw the two conditions of the target detection task might influence the RT effect we were aiming to replicate. We ran a generalized linear mixed model with RT (in seconds) as outcome variable and condition, ordinal position, and condition order as predictors. We set the random effects term as a random intercept term with subject nested within condition order. We observed no three-way interaction between our three predictors (), suggesting that condition order did not significantly bias our results.

***Supplementary Tables***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | | | | | |
| **Table S1**  **A** | | | | | |
| 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 |
|  | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **B** |  |  |  |  |  |
| 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 S2** | | | | | | | |
| condition | target | N | mean | sd | se | ci |
| random | be | 316 | 0.737 | 0.441 | 0.025 | 0.049 |
|  | di | 333 | 0.883 | 0.322 | 0.018 | 0.035 |
|  | ga | 326 | 0.709 | 0.455 | 0.025 | 0.050 |
|  | ki | 318 | 0.852 | 0.355 | 0.020 | 0.039 |
|  | la | 314 | 0.828 | 0.378 | 0.021 | 0.042 |
|  | mi | 313 | 0.744 | 0.437 | 0.025 | 0.049 |
|  | nu | 322 | 0.807 | 0.395 | 0.022 | 0.043 |
|  | po | 317 | 0.817 | 0.387 | 0.022 | 0.043 |
|  | ro | 313 | 0.776 | 0.417 | 0.024 | 0.046 |
|  | se | 349 | 0.880 | 0.326 | 0.017 | 0.034 |
|  | tu | 324 | 0.722 | 0.449 | 0.025 | 0.049 |
|  | za | 305 | 0.777 | 0.417 | 0.024 | 0.047 |
| structure | be | 349 | 0.682 | 0.466 | 0.025 | 0.049 |
|  | di | 346 | 0.925 | 0.264 | 0.014 | 0.028 |
|  | ga | 345 | 0.791 | 0.407 | 0.022 | 0.043 |
|  | ki | 344 | 0.875 | 0.331 | 0.018 | 0.035 |
|  | la | 342 | 0.827 | 0.378 | 0.020 | 0.040 |
|  | mi | 343 | 0.872 | 0.335 | 0.018 | 0.036 |
|  | nu | 333 | 0.877 | 0.329 | 0.018 | 0.035 |
|  | po | 340 | 0.918 | 0.275 | 0.015 | 0.029 |
|  | ro | 331 | 0.801 | 0.400 | 0.022 | 0.043 |
|  | se | 347 | 0.827 | 0.379 | 0.020 | 0.040 |
|  | tu | 340 | 0.862 | 0.346 | 0.019 | 0.037 |
|  | za | 344 | 0.869 | 0.338 | 0.018 | 0.036 |
|  | | | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **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. |