

As Trials Go By: Effects of 2-AFC Item Repetition on Statistical Learning Performance

Ana Paula Soares¹, Tiago França¹, Francisco-Javier Gutiérrez-Domínguez¹,
Inês Sousa², and Helena M. Oliveira¹

¹ Human Cognition Lab, CIPsi, School of Psychology, University of Minho

² Department of Mathematics and Applications, University of Minho

One of the most popular tasks used to test statistical learning (SL) involves asking participants to identify which of two stimuli, a triplet presented during the previous familiarization phase versus a new sequence made of the same stimuli never presented together, is more familiar based on the stream presented before, that is, to perform a two-alternative forced-choice (2-AFC) task. Despite the widespread use of this task, it has come under increasing criticism in current cognitive research due to psychometric flaws. A common practice to improve SL measurement involves increasing the number of 2-AFC trials by presenting the same items (triplets and foils) several times during the test phase. This work aimed to directly analyze the effect that this practice entails by examining how the proportion of correct discriminations of three-syllable nonsense words presented during the familiarization phase of an auditory triplet embedded task changed as the number of 2-AFC item repetitions increased. We also tested whether this effect was modulated by the predictability of the “words” embedded in the auditory streams (high and low) and the conditions under which they were presented to participants (implicit and explicit). Results showed that 2-AFC item repetitions had indeed detrimental effects on SL measurement, as indexed by a significant decrease in the proportion of correct discriminations as the number of items repetitions increased, both in the 2-AFC task performed under implicit and explicit conditions, although, in the first case, only for low-predictable “words.” These findings recommend caution when using this strategy to improve SL measurement.

Public Significance Statement

Despite the widespread use of the 2-AFC task to test SL, it has come under increasing criticism. A common practice to improve SL measurement involves increasing the number of 2-AFC trials by repeating the same items during the test phase. Here, we demonstrated this practice to introduce other sources of “noise” into the SL measurement, hence recommending caution when using this strategy as an attempt to improve SL measurement.

Keywords: statistical learning, 2-AFC, item repetition, predictability, explicit learning

Humans are experts at picking up regularities in the sensory environment, which allows them to structure the world in a regular and predictable way (e.g., Arciuli, 2017; Conway & Christiansen, 2005; Fiser & Aslin, 2001; Frost et al., 2015; Jiménez et al., 2020;

Reber, 1993; Siegelman & Frost, 2015; Thiessen & Erickson, 2013). The mechanism by which we extract those regularities in time and space, even without intention and/or awareness, is called implicit learning, statistical learning (SL), or implicit statistical learning, as recently some authors have proposed to mark the junction of two lines of research that, despite coming from different empirical and theoretical traditions, focused basically on the same phenomenon (see Perruchet & Pacton, 2006; Christiansen, 2019).

Although SL has been studied through the use of different tasks and paradigms, such as cross-situational learning, contextual cueing, Hebb repetition, or artificial grammar learning (see Arciuli & Conway, 2018; Frost et al., 2019, for reviews), since the late 1990s the triplet embedded paradigm is one of the most used in research. Following the seminal work of Saffran et al. (1996; see also Saffran et al., 1999), this task comprises two phases. The first is a familiarization phase, where participants are exposed to a continuous stream of stimuli (e.g., syllables, tones, visual shapes), which, unbeknownst to them, are grouped into triplets (e.g., “tokibu,” “gikoba,” “gopila,” “tipolu”) always appearing together in the stream with no pauses in between (“tokibugikobagopilatipolugikoba”), and without any information regarding the stimuli or the task (i.e., under incidental

This article was published Online First August 4, 2022.

Ana Paula Soares  <https://orcid.org/0000-0002-4047-3799>

Tiago França  <https://orcid.org/0000-0003-2928-4222>

Francisco-Javier Gutiérrez-Domínguez  <https://orcid.org/0000-0003-3690-5795>

Helena M. Oliveira  <https://orcid.org/0000-0003-0644-5302>

This study was conducted at the Psychology Research Centre (PSI/01662), University of Minho, and supported by Grant POCI-01-0145-FEDER-028212 from the Portuguese Foundation for Science and Technology and the Portuguese Ministry of Science, Technology, and Higher Education through national funds. This study was cofinanced by Fundo Europeu de Desenvolvimento Regional (FEDER) through COMPETE2020 under the PT2020 Partnership Agreement.

Correspondence concerning this article should be addressed to Ana Paula Soares, Human Cognition Lab, CIPsi, School of Psychology, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal. Email: asoares@psi.uminho.pt

or implicit conditions). At the end of the familiarization phase, participants are tested on SL through a two-alternative forced-choice (2-AFC) task in which pairs of stimuli, consisting of a triplet presented during the familiarization phase (e.g., “tokibu”) and a foil made of the same stimuli never presented together during exposure (e.g., “tokopi”), are presented. Here, participants are asked to select which one of the two options is more familiar based on the stream presented before. If performance exceeds the chance level, SL is assumed to have occurred. Note that an above-chance performance in the 2-AFC task is only possible if participants compute the statistical regularities embedded in the stream presented during the familiarization phase, typically the likelihood of two stimuli occurring together, a statistic known as transitional probability (TP), and that is, operationalized as the probability of a given item (Y) occurs another item (X) has already occurred [$P(Y|X)$]. TPs within triplets are usually of 1.0, meaning that, within a triplet, a given stimulus (e.g., syllable) is always preceded by another given stimulus (another syllable) in the input. Conversely, TPs across “word” boundaries are much lower, .33 in Saffran et al.’s (1996) work, which indicates that each of the four “words” used in their experiment could be followed by any of the remaining three “words” with the same level of probability. Keeping track of TPs embedded in the stream is, thus, assumed to be the mechanism driving correct “word” discrimination in the 2-AFC task.

However, despite the widespread use of this task to test SL, it has come under increasing criticism, particularly when used to differentiate performance across participants, a topic that is gaining attention in cognitive literature (e.g., Arciuli & Simpson, 2012; Bertels et al., 2012, 2015; Erickson et al., 2016; Franco et al., 2011; Frost et al., 2015; Grown et al., 2020; Kidd et al., 2020; Lages et al., 2022; Qi et al., 2019; Siegelman & Frost, 2015; Soares et al., 2018; Soares, Gutiérrez-Domínguez, et al., 2020; Soares, Silva, et al., 2021; Spencer et al., 2014; Turk-Browne et al., 2010). The use of a small number of 2-AFC trials (e.g., four in Spencer et al.’s, 2014 work) and of triplets with the same level of predictability (i.e., TPs of 1.0) certainly contribute to this state of affairs, as they do not allow for enough variance either within the sample or along with the ability to assess SL properly (see Arnon, 2020; Erickson et al., 2016; Frost et al., 2019; Siegelman, Bogaerts, Armstrong, & Frost, 2019; Siegelman, Bogaerts, Christiansen, & Frost, 2017; Siegelman, Bogaerts, & Frost, 2019; van Witteloostuijn et al., 2019, for similar arguments). To overcome such flaws, researchers have adopted different solutions, such as (a) increasing the number of 2-AFC trials by contrasting the triplets presented during exposure against different foils (e.g., 16 times in Arciuli & Simpson’s, 2012 study); (b) testing not only triplets but also pairs (bigrams) of stimuli (e.g., Siegelman et al., 2018; Siegelman, Bogaerts, & Frost, 2017); (c) using other variants of the 2-AFC task, such as a four-alternative forced-choice (4-AFC) pattern recognition task (e.g., Bertels et al., 2015; Siegelman et al., 2018; Siegelman, Bogaerts, & Frost, 2017) or a three-alternative forced-choice (3-AFC) completion task (e.g., Bertels et al., 2012, 2015; Siegelman, Bogaerts, & Frost, 2017; van Witteloostuijn et al., 2019); (d) using other experimental tasks, such as a target detection postlearning task (e.g., Batterink, Reber, & Paller, 2015; Batterink, Reber, Neville, & Paller, 2015; Siegelman et al., 2018; van Witteloostuijn et al., 2019), a statistically induced chunking recall task (e.g., Isbilen et al., 2020; Kidd et al., 2020), or a self-paced statistical learning task (e.g., Siegelman et al., 2018; Siegelman, Bogaerts, Armstrong, & Frost, 2019;

van Witteloostuijn et al., 2019); or even (e) collecting event-related potentials (ERPs) data during the test and/or familiarization phases (e.g., Abia et al., 2008; Batterink, Reber, & Paller, 2015; Batterink, Reber, Neville, & Paller, 2015; Cunillera et al., 2006; De Diego Balaguer et al., 2007; Gutiérrez-Domínguez et al., 2022; Sanders et al., 2002; Soares, Gutiérrez-Domínguez, et al., 2020; Soares, Gutiérrez-Domínguez, Lages, et al., 2022; Soares, Gutiérrez-Domínguez, Oliveira, et al., 2022; Soares, Lages, et al., 2022; Soares, Lages, Oliveira, & Gutiérrez-Domínguez, 2021).

This article focuses on exploring the effects that increasing the number of 2-AFC trials by repeating the same items (triplets and foils) during the test phase produces in SL measurement. Indeed, although some authors have already warned about the problems this solution might entail, namely the difficulty in disentangling the effects deriving from the learning, which occurred in the previous familiarization phase, from the effects that may arise from memory traces formed as a result of responses to previous 2-AFC trials (see Siegelman, Bogaerts, & Frost, 2017), the fact is that repeating the same trials several times during the test phase is one of the most common methodological options in the SL literature (see Siegelman, Bogaerts, & Frost, 2017, for simulations showing how increasing the number of 2-AFC trials reduces SL measurement error). Therefore, it is of paramount importance to directly examine this practice’s contributions to improving, or, conversely, worsening the assessment of SL.

Moreover, it is also worth noting that since SL is assumed to be an automatic mechanism that operates as a result of mere exposure (see Aslin, 2017; Saffran & Kirkham, 2018), the extraction of items TPs may occur, not only during the familiarization phase (as most SL studies assume), but also during the test phase, which might also contribute to hamper SL measurement. Indeed, if SL is still active in the test phase, as we assume it is, extending the number of 2-AFC trials by repeating the same items appearing again and again during the test phase may result not only in foils emerging as perceptual units as the number of item repetitions increase but also, importantly, the extraction of the foils’ TPs might also interfere with correct “word” discrimination as the 2-AFC task unfolds. Note that because foils are made of the same stimuli as the “words” presented during exposure, the continuous extraction of the TPs during the test phase may dramatically change the statistical properties of those “words,” making their recognition harder as the 2-AFC task advances. To illustrate, consider for instance that during exposure a specific syllable as “ki” is always preceded by the syllable “to,” as in the nonsense word “tokibu,” and that same syllable also appears in a foil such as “bukito.” This could make TPs drop from 1.0 in the familiarization phase to .50 in the test phase, or even less, according to the number of different arrangements using syllables from the “words” are combined in the foils. Thus, besides the confounding role that episodic memory might play in SL measurement, as pointed out by Siegelman, Bogaerts, and Frost (2017), we additionally claim that the continuous extraction of TPs during the test phase may also contribute to introducing additional “noise” in SL assessment, as foils may be gradually learned as (new) “words,” and the (old) “words” may be gradually “unlearned” as such.

Although testing the contribution of each of these factors in SL measurement is clearly beyond the scope of this article, the point we aim to stress here is that, altogether, they can contribute to worsening rather than improving SL assessment. This work is the first to provide the much-needed examination of the impact that increasing

the number of 2-AFC trials by presenting the same items (triplets and foils) several times during the test phase might have on the SL measurement, which will contribute not only to deepen our understanding of the effects of 2-AFC item repetitions on SL assessment but, importantly, to inform best practices in SL research. Specifically, in this study, we conducted a well-controlled experiment in which, after having been exposed to the repeated presentation of three-syllable nonsense words during the familiarization phase, as in the seminal work of Saffran et al. (1996), participants were successively confronted with one, two, three, up to eight repetitions of the same 2-AFC items (“words” and foils) while examining how the proportion of correct discriminations changed as the 2-AFC task unfolded or the number of 2-AFC item repetitions increased. In addition, we also analyzed whether the 2-AFC item repetition effect was modulated by the predictability of the “words” embedded in the speech streams (high and low), and the conditions under which they were presented to the participants (implicit and explicit) to further ascertain the conditions under which 2-AFC item repetitions were, or were not, impervious to interference. Despite the vast majority of SL studies conducted so far have used triplets presenting TPs of 1.0 under incidental (implicit) conditions (e.g., Arciuli & Simpson, 2012; Bertels et al., 2012, 2015; Frost et al., 2015; Kidd et al., 2020; Qi et al., 2019; Saffran et al., 1996, 1999; Spencer et al., 2014; Turk-Browne et al., 2010), an increasing number of studies have been using triplets with lower levels of predictability presented under implicit and also explicit (intentional) conditions (e.g., Arciuli et al., 2014; Batterink, Reber, Neville, & Paller, 2015; Batterink, Reber, & Paller, 2015; Bertels et al., 2015; Bogaerts et al., 2016; Grown et al., 2020; Gutiérrez-Domínguez et al., 2022; Lages et al., 2022; Siegelman et al., 2017; Soares, Gutiérrez-Domínguez, et al., 2020; Soares, Gutiérrez-Domínguez, Lages, et al., 2022). This is particularly relevant because these manipulations allow the analysis of how SL works in a wide range of situations, and to facilitate the generalization of the results obtained from SL experiments to “real” contexts, where the uncertainty is more the rule rather than the exception, and where individuals often use their previous knowledge to assist the acquisition of new one. Even though no previous study has attempted to directly address these issues with the collection of “real” data from “real” participants, we anticipate that increasing the number of 2-AFC repetitions will contribute to hamper “word” discrimination, as reflected by a significant decrease in the proportion of correct responses as the 2-AFC task unfolds or, in other words, as 2-AFC item repetition increases. Stronger item repetition effects are also expected for “words” with low- rather than high-TPs since “words” with higher levels of predictability are expected to produce stronger representations (e.g., Siegelman et al., 2018; Soares, Gutiérrez-Domínguez, et al., 2020; Soares, Gutiérrez-Domínguez, Lages, et al., 2022; Soares, Gutiérrez-Domínguez, Oliveira, et al., 2022). This might make these words less prone to interference, particularly when the 2-AFC task is performed under implicit conditions.

Indeed, even if implicit representations were classically assumed to be “immune” to interference (e.g., Graf & Schacter, 1987; Jacoby, 1983; Sloman et al., 1988), recent studies have challenged this view by showing not only that implicit representations are just as vulnerable to interference as explicit representations, but actually that they might be even more vulnerable to that kind of effects (e.g., Eakin & Smith, 2012; Ikier et al., 2008; Jiménez et al., 2006; Lustig & Hasher, 2001; Vaquero et al., 2020). For instance,

Jiménez et al. (2006) examined whether the learning obtained in a serial reaction time (SRT) task, performed under implicit or explicit conditions, was affected by the introduction of distractors in non-target (superficial changes) and target (structural changes) locations of the previously practiced sequence. This study revealed that the SRT performance of both learners with and without previous knowledge of the sequence embedded in the input, and that they would be tested later on their knowledge about the sequence (i.e., intentional vs. incidental learners, respectively), was affected by the introduction of those changes, even though each of them was selectively affected by different contingencies. Specifically, superficial changes impaired the SRT performance of incidental learners (but not intentional learners), whereas structural changes impaired the SRT performance of intentional learners (but not incidental learners). The fact that incidental learners were strongly affected by the introduction of distractors in nontarget locations, and intentional learners by the introduction of distractors in target locations was interpreted as reflecting the fact that learning under implicit conditions is more dependent on the reinstatement of the conditions, under which learning had occurred in the first place. Moreover, the authors also claimed that, under implicit conditions, the absence of relevant metaknowledge (conversely to its explicit counterpart) might also decrease the level of control exerted over the ongoing processes by using strategic operations directed to reduce the interference caused by nonrelevant information (see Cleeremans & Jiménez, 2002, for similar arguments; and also Vaquero et al., 2020, for a recent study testing the role of control processes on implicit sequence learning). These results led the authors to conclude that neither explicit nor implicit representations are impervious to interference, even though the between-subjects design adopted in that study, as well as in many other studies exploring the type of representations emerging from SL tasks (e.g., Arciuli et al., 2014; Batterink, Reber, Neville, & Paller, 2015; Batterink, Reber, & Paller, 2015), prevents us from drawing strong conclusions, since previous studies found substantial variability in the way individuals respond to SL tasks (see Siegelman, Bogaerts, Christiansen, & Frost, 2017; and also Siegelman, Bogaerts, & Frost, 2017), which might have confounded previous results. The use of within-subjects designs is thus strongly advisable (see Soares, Gutiérrez-Domínguez, et al., 2020; Soares, Gutiérrez-Domínguez, Lages, et al., 2022; Soares, Gutiérrez-Domínguez, Oliveira, et al., 2022, for recent examples), even if it necessarily implies not counterbalancing the order of the SL tasks across participants. Note that once an SL task is performed explicitly by participants, it can no longer be performed implicitly by the same participants, hence making the implicit followed by the explicit presentation of the SL task the only viable solution in that type of design.

Thus, following previous studies, we hypothesize that “words” learned under implicit conditions would be highly prone to 2-AFC item repetition effects. This hypothesis arises due to the implicit nature of the task, which does not allow participants to recruit other (strategic) processes directed to avoid the interference that the continuous extraction of TPs during the test phase might entail, particularly for the low-TP “words,” as they are assumed to generate weaker representations. Nevertheless, because the continuous extraction of the TPs might also change the structural (statistical) properties of the “words” previously presented, item repetition detrimental effects are also expected to be observed in the 2-AFC task performed under explicit conditions, although for the

high- and the low-TP “words” alike. Finally, a last point worth mentioning is that although we are aware that the within-subjects design adopted in the present study also prevents us from making direct comparisons between item repetition effects in the SL task performed under implicit versus explicit conditions, due to the absence of task counterbalancing; we would like to emphasize that in each SL task the stimuli (“words” and foils) were generated from different syllabaries to avoid carryover effects. Moreover, it is also important to highlight that since previous SL studies manipulating instructions in a between-subjects design have shown that the effect of explicit knowledge in SL performance is only noticeable when the instructions are specific enough (see Arciuli et al., 2014; Batterink, Reber, Neville, & Paller, 2015; Batterink, Reber, & Paller, 2015), performing the explicit version of the SL task after the implicit version of an analogous SL task may even be advantageous since participants could make use of their previous experience to respond to the second task more effectively, as intended with this manipulation. Still, acknowledging potential confoundings, the analyses were conducted separately for each type of SL task.

Method

Participants

Thirty-two undergraduate students (31 women, $M_{\text{age}} = 20.4$, $SD_{\text{age}} = 1.95$) from the University of Minho were recruited for this experiment. Since in a previous study using 32 participants, Soares, Gutiérrez-Domínguez, et al. (2020) obtained medium-to-large SL effects in a 2-AFC task that mimicked the one used here (even though effect sizes were not reported in that work, the computation of Cohen’s d based on the statistics provided, showed values ranging between .48 and 1.04), the same sample size was used. Moreover, a similar sample size ($N = 32$) was also obtained from power analysis conducted using G*Power (Faul et al., 2009), to achieve 95% power for detecting a medium effect at a significance criterion of $\alpha = .05$ ($1 - \beta = 0.95$; $\alpha = 0.05$) for an effect size of (U) = 0.40 ($\eta_p^2 \sim 0.082$). All participants were native speakers of European Portuguese with normal hearing and no reported history of learning or language disabilities and/or neurological problems. Written informed consent was obtained from each participant. The experiment was approved by the local Ethics Committee (University of Minho, Braga, Portugal, SECSH 028/2018).

Stimuli

The stimuli used were drawn from Soares, Gutiérrez-Domínguez, et al.’s (2020) work. Specifically, the materials included 32 unique consonant–vowel syllables evenly distributed across two different syllabaries of 16 consonant–vowel syllables each (e.g., “tu,” “ci,” “da,” “mi,” “ge,” “do” from syllabary A; and “ga,” “pa,” “be,” “me,” “gu,” “pi” from syllabary B), which were used in the implicit and explicit versions of the auditory SL task (aSL) to avoid carry-over effects. Syllables in each syllabary were organized into eight three-syllable nonsense words: four “words” presenting high levels of predictability (TP = 1.0) and four “words” presenting low levels of predictability (TP = .33). For instance, the nonsense word “tucida” from Syllabary A and the nonsense word “todidu” from Syllabary B correspond to high-TP “words,” as the syllables they contain appear only in those “words” and in that specific syllable position.

Conversely, the nonsense words “dotige” from Syllabary A and “pitegu” from Syllabary B correspond to low-TP “words,” as the syllables they contain appear in three different “words” in different syllable positions as “tidomi,” “migedo,” “gemiti” in Syllabary A and “tepime,” “megupi,” “and gumete” (see Table 1, for other examples).

Syllables were produced and recorded by a native speaker of European Portuguese with a duration of 300 ms each. The nonsense three-syllable words were concatenated in a continuous stream with the Audacity software (1999–2019) with a 50 ms interval between syllables (1,150 ms per “word”). Each nonsense word was repeated 60 times during the exposure phase. The stimuli were presented binaurally in random order with the restriction that the same nonsense word, or the same syllable in different “words,” was not presented consecutively in the stream. Hence, TPs across “word” boundaries were .14 for the high-TP “words,” as they could be preceded by any of the remaining seven “words” with the same level of probability, and .17 for low-TP “words,” since they could be preceded by any of the remaining nonsense “words” except for the one beginning with the same syllable as it ends (e.g., *gemiti**tidomi* was not allowed). Figure 1 depicts a graphic representation of a speech stream used in the Experiment with TPs both within and across “word” boundaries for the four high- and the four low-TP “words,” respectively. For simplicity, high and low-TP “words” are interspersed in Figure 1.

As in Soares, Gutiérrez-Domínguez, et al., (2020), the speech stream was edited to include a randomly superimposed chirp sound (a 0.1-s sawtooth wave sound from 450 to 1,450 Hz) to provide participants with a cover detection task, that ensures they would pay attention to the auditory stream during exposure. The target sound was programmed to appear in the interval between syllables in 30% of each “word” type, counterbalanced across syllables to prevent confounds. Results showed that participants paid appropriate and similar attention to the speech streams presented in both aSL tasks as intended—correct detections were 48.26 (± 2.1) in the implicit aSL task (98.1% of all responses, including false alarms) and 48.1 (± 1.7)

Table 1

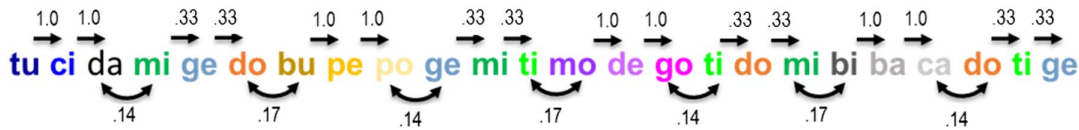
Three-Syllable Nonsense Words and Foils From Syllabary A and Syllabary B

TP of “word”		Syllabary	
		A	B
Nonsense words	High	tucida	todidu
		buepo	cegita
		modego	gapabe
	Low	bibaca	bomaco
		dotage	pitegu
		tidomi	tepime
Nonsense foils	High	migedo	megupi
		gemiti	gumete
		tumica	tomeco
		bugego	cegube
		modopo	gapita
	Low	bitida	botedu
		dobage	pimagu
		tidemi	tepame
		mipedo	megipi
		geciti	gudite

Note. TP = transitional probability.

Figure 1

Transitional Probabilities of the Syllables Within and Across “Word” Boundaries for the High-TP (1.0) and Low-TP (.33) “Words”



Note. TP = transitional probability. See the online article for the color version of this figure.

in the explicit aSL task (98.1%), not differing between each other ($p = .931$).

For the 2-AFC task, eight three-syllable foils were also created in each syllabary (see Table 1). The foils contained the same syllables used for the “words,” presented with the same frequency and syllable position as the syllables in the high- and low-TP “words.” For example, the most frequent syllables used during familiarization with Syllabary A (e.g., “do”) appeared three times in different “words” by the permutation of the same syllable across positions (“dotage,” “tidomi,” “migedo”), and also three times in the foils in the same positions (“dobage,” “modopo,” “mipedo”). Conversely, the less frequent syllables (e.g., “tu”) appeared only once, in a single “word” (e.g., “tucida”), and were also presented once in the foils in that same position (e.g., “tumica”)—see Table 1 for other examples. However, contrary to the syllables in the high and low-TP “words,” the syllables in the foils were never presented together during exposure (TPs = 0). Nevertheless, it is worth noting that due to constraints in the number of permutations allowed by the 16 syllables used in each syllabary, the foils created from the 12 syllables composing the high-TP “words,” contained two syllables from high-TP “words” and one syllable from low-TP “words.” Similarly, the foils created from the four syllables used in the low-TP “words” contained two syllables from low-TP “words” and one syllable from high-TP “words,” hence ensuring the same level of overlap across foil types. Four lists of materials were created to counterbalance syllables across positions and TP “word” (high and low) in each syllabary. Participants were randomly assigned to the lists with the constraint that the same number of participants would complete a given list (four participants per list). For clarity, Table 1 presents the stimuli used in List 1 from Syllabary A and List 1 from Syllabary B. The entire list of materials is available at <https://osf.io/65kd2/> (Soares, França, et al., 2022).

Procedure

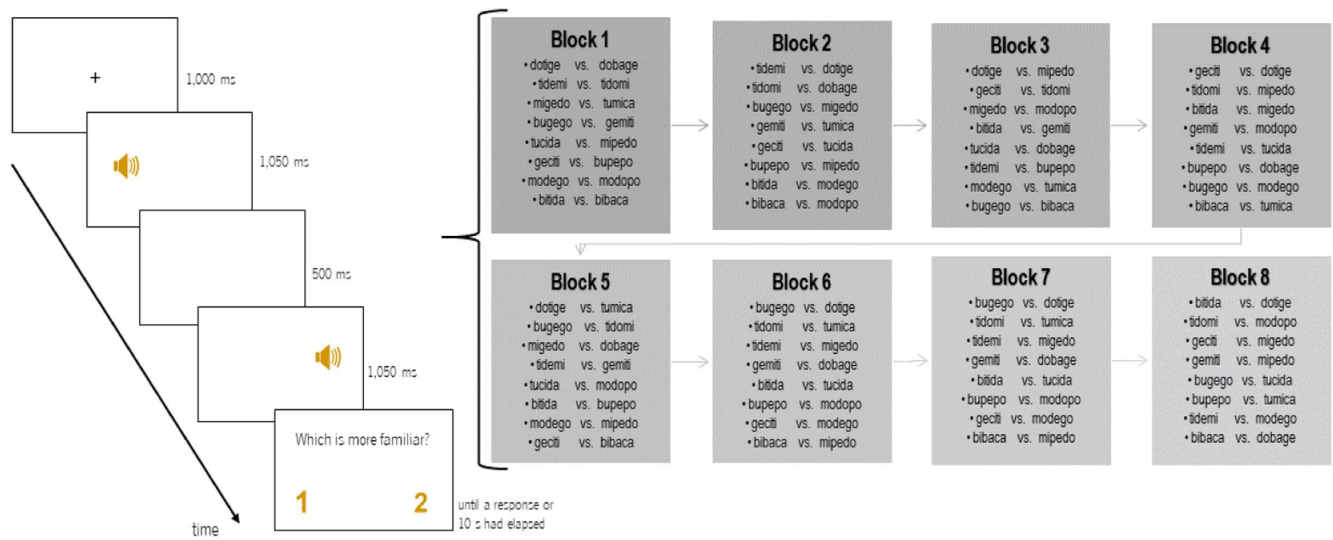
Data were collected in a soundproof booth at the Human Cognition Lab (School of Psychology, University of Minho). Participants were presented with the implicit version of the aSL task, and, subsequently, with the explicit version of the aSL task using two different syllabaries (A and B or B and A, respectively), as mentioned above. In the aSL implicit task, participants were instructed to pay attention to the auditory stream (presented at 60 dB sound pressure level via binaural headphones), because occasionally a deviant sound (i.e., a click) would appear, and their task would be to detect it as soon as possible by pressing the spacebar key. As mentioned above, such a request acted as a cover task to ensure that participants would pay attention to the stimuli during the

familiarization phase. Afterwards, participants underwent a test phase where they were asked to decide, as soon and as accurately as possible, which of two auditory stimuli (a “word” and a foil) “sounded more like” the stimuli presented before. This 2-AFC task comprised 64 trials where each of the eight “words” was paired exhaustively with each of the eight foils, a common practice in SL studies. Each trial began with the presentation of a fixation point for 1,000 ms, after which the first stimulus (“word”/foil) was presented, followed by the second stimulus, with a 500-ms interstimulus interval. The following trial began as soon as participants responded or 10 s had elapsed, as depicted in Figure 2.

The 64 2-AFC trials were presented in eight different blocks. Each block contained eight 2-AFC trials (see Figure 2). The blocks resulted from the implementation of a set of constraints to allow for an adequate tracking of how SL performance changes as the number of item repetitions increases throughout the 2-AFC task: (a) every “word” and every foil was presented once per block; (b) in each block, half of the high-TP “words” were paired with high-TP foils, and the other half with low-TP foils, guaranteeing the same distribution for the low-TP “words” within the block; and (c) the order (first or second) by which the stimuli (“word”/foil) were presented was also controlled, so that in each block half of the high-TP “words” and half of the low-TP “words” were presented first and, the other half, second. Thus, across blocks, each “word” and each foil were presented eight times, with half of the high-TP “words”/high-TP foils and half of the low-TP “words”/low-TP foils being presented first, and the other half, second. This was consolidated with an even distribution of the number of times each “word”/foil was paired with a foil/“word” from the same type—see Figure 2, for an example of the implementation with one list of materials. Trials in each block, as well as blocks, were presented randomly per participant.

After a brief interval, participants underwent the explicit version of the SL task. Again, they were exposed to another stream containing a different set of eight “words” (four high-TP “words” and four low-TP “words”) from the syllabary not used in the implicit version of the aSL task, counterbalanced across participants. The task followed the same procedure adopted in the implicit version, with the exception that, before exposure, participants were presented with the new “words” they would hear during the familiarization phase. Specifically, participants were told they would listen to “new words” from another foreign language. Each of the eight new “words” was then auditorily presented, and participants were asked to repeat each of them correctly before the familiarization phase began (see Batterink, Reber, & Paller, 2015; Gutiérrez-Domínguez et al., 2022; Lages et al., 2022; Soares, Gutiérrez-Domínguez, Lages, et al., 2022; Soares, Gutiérrez-Domínguez, Oliveira, et al., 2022;

Figure 2
Setup of the Experimental Procedure



Note. See the online article for the color version of this figure.

Soares, Lages, Oliveira, & Gutiérrez-Domínguez, 2021, for a similar procedure). Like in the implicit task, during the familiarization phase participants were asked to press a button whenever they heard the click sound, to ensure they paid attention to the auditory stimuli. After familiarization, participants performed another 2-AFC task, which mimicked the one used in the implicit version of the aSL task. The procedure lasted about 30-45 min per participant.

Results

To examine if SL performance was affected by 2-AFC item repetition, and the extent to which the effect was modulated by “word” predictability and/or by the conditions under which the to-be-learned regularities were presented to the participants, several analyses were conducted on the accuracy data (i.e., the proportion of correct identifications of “words” over foils). These analyses were conducted first for all the participants and, then, only for those who have shown evidence of learning at an individual level of analysis, to remove potential confounds that the inclusion of participants who have performed at, or below, the chance level might have introduced in the results. The analyses were conducted in the R software (R Core Team, 2021) using both analysis of variance (ANOVAs) and generalized linear mixed-effects models (glmm) with logistic link function and binomial variance. The first allowed us to test the hypotheses under study by directly controlling both the variability across participants and items at once, hence overcoming the drawbacks of the classical ANOVAs, namely in what refers to the need to conduct separate by-subject (F_1) and by-item (F_2) analyses (see Baayen et al., 2008; Judd et al., 2017; see also Soares, Lages, Oliveira, et al., 2019; Soares, Velho, & Oliveira, 2020; Soares, Lages, Velho, et al., 2021, for recent examples). In addition, the glmm allowed us to further estimate the impact of item repetition on SL performance each time an item was repeated (i.e., in each block), considering the variables under study as linear predictors containing random effects in addition to the usual fixed effects. Note that,

against the ANOVAs and also the traditional regression analyses, both of which assume complete independence between measurements, the glmm analyses allow us to assume that measurements of the same participant at different time points are correlated, hence providing additional depth to the previous analysis. The glmm models are able to decompose total variability in the data into within- and between-participants variability and, additionally, to make inferences on the mean effects taking into account those sources of variability. Therefore, the interpretation of the model’s parameters is the effect of the predictors on the odds of responding with accuracy (probability of producing a correct over an incorrect response).

The analyses were conducted for each aSL task (implicit and explicit) separately since the absence of task counterbalancing, across participants, due to the nature of the task, prevents us from making direct comparisons between item repetition effects in the SL task performed under implicit versus explicit conditions. The models were fitted using the lme4 R library and lmerTest library (Bates et al., 2011) to contrast simple effects with differences of least squares means. In a first step, the analyses were conducted by running a model with participants and items as random intercepts with the two repeated-measure factors: Block (B#1|B#2|B#3|B#4|B#5|B#6|B#7|B#8) and Type of “word” (High|Low) with random slope per subject, Block $\times$ Type of “word” + (Block + type of “word”|subject) + (1|item). However, due to strong collinearity observed between the intercepts and both the type of “word” and block, those factors were entered as fixed factors in the final models, Block $\times$ type of “word” + (1|subject) + (1|item). Note that, although Barr et al. (2013) suggested that glmm should be implemented with the most complex random-effect structure in factorial experiments, Bates et al. (2015; see also Matuschek et al., 2017) further demonstrated that this does not come without a cost, as the use of the most complex model also implies a loss in statistical power to detect the significance of fixed effects (i.e., Type II error rate inflation). Thus, the use of a more parsimonious model, particularly in situations

where there is collinearity between intercepts, should be considered even in situations where the model converges, such as in our case.

In the reported statistics, the p values were adjusted with Hochberg's method for all the post hoc comparisons equal to or below .05 (see Benjamini & Hochberg, 1995; Hochberg, 1988, for details). Responses were coded as 1 (*correct*) or 0 (*incorrect*) in each 2-AFC trial (maximum = 64 correct responses per aSL task, 8 per block). The raw data used in the statistical analyses reported here are available at <https://osf.io/65kd2/> (Soares, França, et al., 2022).

Table 2 presents the mean and the standard deviation (in brackets) of the proportion of correct responses obtained from all the participants ($N = 32$) in each aSL task (implicit and explicit) and for the high and low-TP “words” in each block (#1–#8).

The results of the ANOVA in the implicit aSL task revealed a main effect of type of “word,” $\chi^2(1) = 4.47, p = .034$, indicating that globally (i.e., regardless of the block), participants recognized significantly better low- (.64) as opposed to high-TP (.59) “words.” Neither the main effect of block nor the interaction between both factors reached statistical significance. In the explicit aSL task, no statistically significant effects were observed. Thus, when considering the SL performance of all participants included in the study as a whole, the results failed to reveal any significant effect of 2-AFC item repetition, both in the aSL task performed under implicit and explicit conditions, and for both kinds of “words.” Nevertheless, as mentioned above, it is possible that the inclusion of participants who have not shown signs of learning at an individual level of analysis might have contributed to these results.

Indeed, even though the results from the one-sample t tests against the chance level (.50), conducted on the overall SL performance (i.e., on mean accuracy considering the eight blocks as a whole), as reported in most studies, showed moderate but reliable signs of SL at a group level both in the aSL task performed under implicit, $M = .61, SD = .49, t(2047) = 10.38, p < .001$, and explicit conditions, $M = .66, SD = .47, t(2047) = 15.25, p < .001$; the results considering the overall SL performance at an individual level of analysis, showed substantial variability with approximately half of the participants in the implicit aSL task (43.8%) and one-third of the participants in the explicit aSL task (34.4%) performing at, or below, the chance level. Figure 3 depicts the distribution of the overall SL performance presented by each of the 32 participants in each of the aSL tasks.

Specifically, the results of the binomial distribution showed that only 18 participants in the implicit aSL task and 21 participants in the explicit aSL task reached an above-chance performance—note that only participants that responded correctly at least to 37 out of the 64 2-AFC trials (corresponding to 57.8% of accuracy) revealed evidence of learning ($p = .046$). Therefore, in the second set of analyses, we have only included the participants who have shown

evidence of SL in each task at individual levels, and for whom item repetition might have, indeed, produced any kind of effect. Note that, from a theoretical point of view, we can only talk about the effect of item repetition on SL performance if we assure that SL has actually occurred. Moreover, it is also worth noting that since the main purpose of this work was to examine the impact that 2-AFC item repetition had on SL performance as the number of item repetitions increased, the best solution was to consider the performance that each participant presented at the initial phase of the 2-AFC task (i.e., when items were presented for the first time) against performance each time the items were repeated (i.e., from Block 2 to Block 8), rather than taking the SL performance as a whole (i.e., regardless of the block), since the latter may already index the effect of item repetition on participants' SL performance. Thus, to discriminate “learners” from “nonlearners” in this second set of analyses we relied on the computation of SL performance participants revealed at the beginning of each 2-AFC task. Moreover, because considering only the results that participants showed in the first block (8 items) would lead us to exclude most of the sample—bear in mind that an above-chance performance with eight items for an $\alpha = .05$ requires participants to correctly recognize seven out of the eight items (87.5% accuracy), which was only achieved by three participants in the implicit aSL task and nine participants in the explicit aSL task, we opted to consider participants' SL performance in the first two blocks (16 items), that is, including those who correctly recognized at least 12 out of the 16 items (75% accuracy). Thirteen participants in the implicit aSL task and 15 participants in the explicit aSL task met that criterion. Figure 4 shows the proportion of correct recognitions presented by the participants who have shown evidence of learning at the beginning of each of the 2-AFC tasks per block (#1–#8) and TP “word” (high-TP and low-TP “words”).

The results of this second set of analyses, mimicking the ones conducted with all the participants, showed a significant main effect of block, $\chi^2(7) = 21.43, p = .003$ in the aSL task performed under implicit conditions, indicating that participants were more accurate recognizing “words” in the first (.77) than in all the other blocks except the second, even though the differences only reached statistical significance when the first block was compared with the fourth (.57; $p = .032$); and also more accurate recognizing “words” in the second block (.81) than in all the other blocks, even though the differences were only significant when compared with the fourth (.57; $p = .006$), fifth (.61; $p = .034$), sixth (.59; $p = .014$), and eighth blocks (.60; $p = .022$). Interestingly, the two-fold type of “Word” $\times$ Block interaction was also significant in the implicit aSL task, $\chi^2(7) = 22.78, p = .002$. The post hoc comparisons revealed that although for the high-TP “words” the differences across blocks failed to reach

Table 2

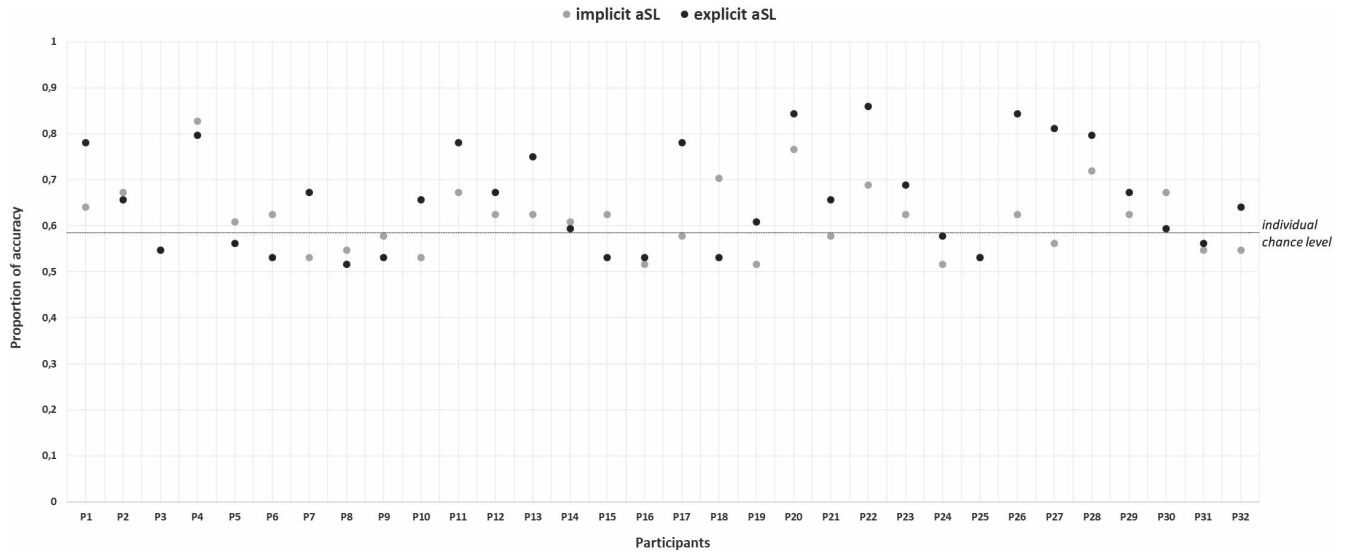
Mean (SDs) of Accuracy (Proportion of Correct Responses) by aSL Task, Type of “Word,” and Block for All the Participants ($N = 32$)

aSL task	Type of “word”	Block							
		B# ₁	B# ₂	B# ₃	B# ₄	B# ₅	B# ₆	B# ₇	B# ₈
Implicit	High-TP	.56 (.50)	.60 (.492)	.57 (.50)	.57 (.50)	.59 (.49)	.56 (.50)	.65 (.48)	.59 (.49)
	Low-TP	.73 (.45)	.72 (.45)	.63 (.48)	.59 (.49)	.64 (.48)	.63 (.48)	.58 (.50)	.59 (.49)
Explicit	High-TP	.70 (.46)	.70 (.46)	.67 (.47)	.67 (.471)	.64 (.48)	.66 (.48)	.61 (.49)	.65 (.48)
	Low-TP	.66 (.48)	.70 (.46)	.66 (.47)	.71 (.46)	.61 (.49)	.63 (.48)	.66 (.48)	.63 (.49)

Note. SD = standard deviation; aSL = auditory SL task; SL = statistical learning; TP = transitional probability.

Figure 3

Distribution of the Accuracy (Proportion of Correct Responses) in the Implicit and Explicit aSL Tasks Considering the Overall SL Performance Per Participant



Note. aSL = auditory SL task; SL = statistical learning. The dashed line shows the individual chance level.

statistical significance, for the low-TP “words” participants recognized significantly more “words” in the first (.87) and in the second (.85) blocks than in any other of the remaining blocks, even though the differences only reached statistical significance when the first block was compared to the fourth (.54; $p = .012$), fifth (.56; $p = .020$), sixth (.56; $p = .020$), seventh (.46; $p = .001$), and eighth (.60; $p = .053$) blocks—albeit in this last case, only at a marginally significant level; and the second block was compared with the fourth (.54; $p = .022$), fifth (.56; $p = .036$), sixth (.56; $p = .036$), and seventh (.46; $p = .002$) blocks. In the explicit aSL task, the results failed to show any statistically significant effect. Thus, when participants who did not show reliable signs of learning at the beginning of each 2-AFC task were excluded from the ANOVAs, the results clearly indicated that 2-AFC item repetitions had, indeed, detrimental effects on SL performance as expected, even though only for the low-TP “words” presented under implicit conditions.

To further estimate the effects of 2-AFC item repetitions on SL performance each time an item was repeated (i.e., at each block), in a third step, we conducted glmm analyses considering participants’ accuracy in each block (#1–#8) and for each type of “word” (high-TP/low-TP), and also the interactions between both terms as predictors in each of the aSL task, even though the ANOVA failed to show any statistically significant effect in the aSL task performed under explicit conditions. In each aSL task, we assumed a model that included a random intercept effect (participant-specific effect) and where the mean population for the odds ratio is described as follows:

$$\frac{p}{1-p} = \exp(\alpha + \beta X),$$

where, p is the probability of correct identifications, X is the design matrix for predictors, and α and β are the model’s parameters associated with the predictors. Therefore, the interpretation of the model’s parameters can be described as the effect of the predictors

(block, type of “word,” and the interaction between both) on the odds of responding accurately (i.e., the probability of producing a correct over an incorrect response) operationalized as follows:

$$\begin{aligned} \frac{p}{1-p} = \exp(\alpha + \beta_1 \times \text{Block} + \beta_2 \times \text{If}(\text{“word”} = \text{low-TP}) \\ + \beta_3 \times \text{Block} \times \text{If}(\text{“word”} = \text{low-TP})), \end{aligned}$$

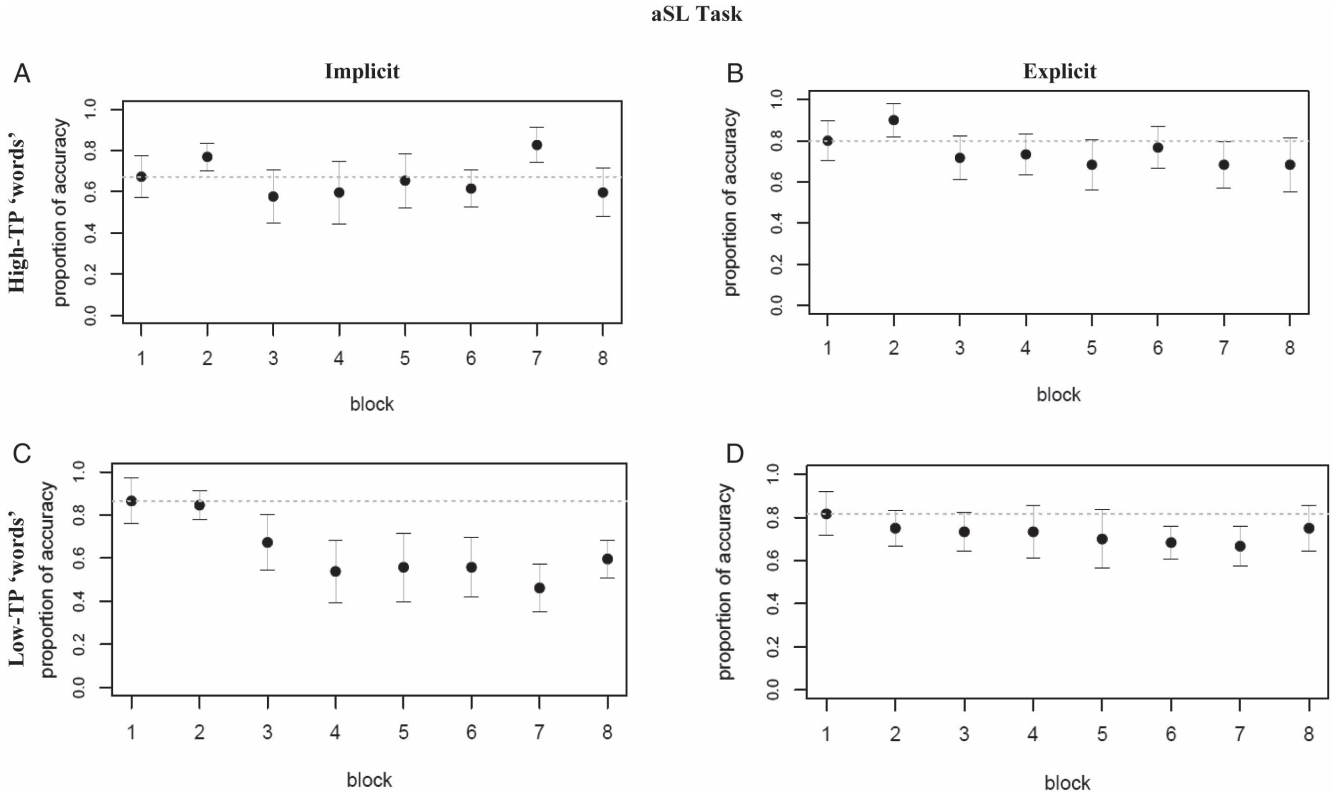
where, p is the probability of correct identifications, α and β are the model’s parameters associated with the predictors.

As shown in Table 3, despite the odds ratio of accuracy in the implicit aSL task not decreasing significantly as the 2-AFC task unfolded, the interaction between the two predictors reached statistical significance ($p < .001$). As illustrated in Figure 5, which depicts the mean estimated change in the proportion of accurate responses as a function of block (from #1 to #8) and type of “word” (high- vs. low-TP) in each aSL task and the correspondent 95% confidence intervals (bands) for the mean estimated change, the proportion of accurate responses decreased sharply as item repetitions (block) increased for the low-TP “words” with a decrease estimated in 19.8% each time a new block is presented or each time the same items were repeated, computed from $1 - \text{Exp}[\text{estimate}] \times 100$. This is not the case for the high-TP “words.” Moreover, in the implicit aSL task, the results also showed that, regardless of the block, the probability of responding accurately to the low-TP “words” is almost one and a half times higher (148%) than the probability of responding accurately to high-TP “words” ($p = .008$).

Furthermore, in the explicit aSL task, the results of glmm analyses revealed that, even though the probability of responding accurately to the low- versus high-TP “words” was virtual the same (note in Figure 5 that the confidence intervals for the mean estimated change of the proportion of accurate responses for the high- and low-TP “words” overlapped), item repetitions significantly impacted SL

Figure 4

Distribution of the Accuracy (Proportion of Correct Responses) in Each Block (#1–#8) and for Both Types of “Words” for Participants Who Have Learned in the Implicit ($n = 13$) and Explicit aSL Tasks ($n = 15$)



Note. aSL = auditory SL task; SL = statistical learning; TP = transitional probability. The error bars represent the standard error of the mean (SEMs). The dashed line shows the mean of the first block. (A) High-TP “words” in the implicit aSL task; (B) high-TP “words” in the explicit aSL task; (C) low-TP “words” in the implicit aSL task; and (D) low-TP “words” in the explicit aSL task.

performance as the 2AFC task unfolded with a decrease estimated in 11.7% each time a new block was presented or each time the same items were repeated. Thus, the results from the glmm analyses conducted showed that 2-AFC item repetitions had, indeed, detrimental effects on SL performance for the low-TP “words” in the aSL

task performed under implicit conditions and, importantly, in the aSL task performed under explicit conditions as well, regardless of the type of “word,” as hypothesized. Although the second set of ANOVAs failed to reveal a significant effect of item repetition (block) in the aSL task performed under explicit conditions, when

Table 3

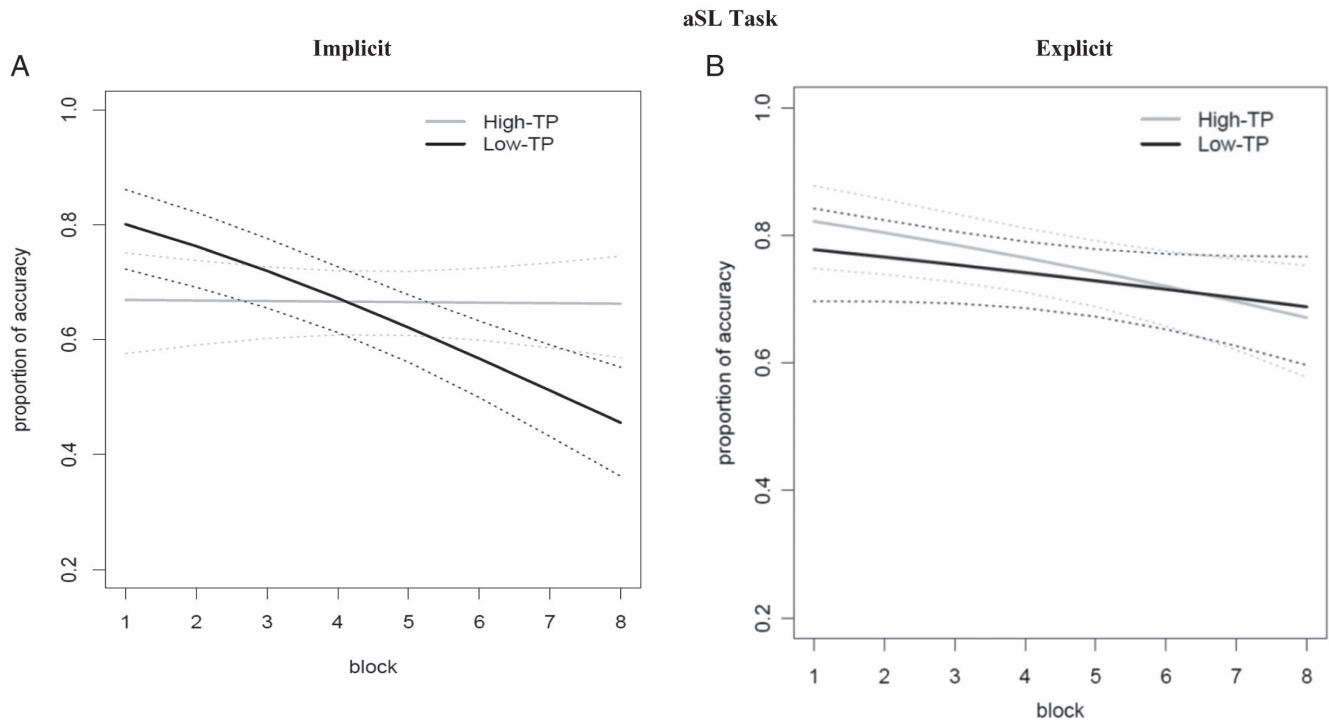
Parameter Estimates, Standard Errors, and p Values for Testing for No Effects on the Predictors for the Implicit aSL Task (Top Panel) and for the Explicit aSL Task (Bottom Panel)

Parameters	Estimate	Exp (estimate)	SE	p value
Implicit aSL task				
β_0	.708	2.030	.241	.003
β_1 (Block)	–.004	.996	.046	.927
β_2 (“word” = low-TP)	.907	2.476	.341	.008
β_3 (Block $\times$ (“word” = low-TP))	–.220	.803	.066	<.001
Explicit aSL task				
β_0	1.645	5.183	.264	<.001
β_1 (Block)	–.117	.8890	.047	.013
β_2 (“word” = low-TP)	–.328	.720	.342	.338
β_3 (Block $\times$ (“word” = low-TP))	.051	1.052	.065	.438

Note. aSL = auditory SL task; SL = statistical learning; TP = transitional probability.

Figure 5

Change in the Odds Ratio of Accuracy (Proportion of Correct Responses) Throughout the Blocks and for Both Types “Words” for Participants Who Have Learned in the Implicit ($n = 13$) and Explicit aSL Tasks ($n = 15$)



Note. aSL = auditory SL task; SL = statistical learning; TP = transitional probability. The error bars represent the standard error of the mean (SEMs). The dashed line shows the mean of the first block. (A) Implicit aSL task and (B) explicit aSL task.

the block was considered as a continuous measure, for example, assuming that SL performance at each time point [block] is not independent of other time points, and not as a discrete measure, as it was in the case of the ANOVAs, the results of the glmm analyses clearly indicate that item repetitions decreased SL performance, hence suggesting that the previous knowledge of the regularities embedded in the stream and the familiarization with the task were not strong enough to avoid 2-AFC item repetitions to interfere with SL performance, as discussed below.

Discussion

The present study examined how SL performance is affected by “word” and foil repetitions in a standard 2-AFC task aimed at testing SL. In addition, we also examined whether that effect was modulated by “word” predictability (TPs) and prior knowledge of the regularities (“words”) embedded in the speech streams, by asking participants to perform an aSL task in which “words” with high (1.0) and low (.33) TPs were embedded in the same stream, first, under incidental (implicit), and, subsequently, under intentional (explicit) conditions. With this within-subjects design, we aimed to minimize the role of individual differences in the results, even if it implied not counterbalancing the order of the aSL tasks across participants. Although previous studies have acknowledged the problems that increasing the number of 2-AFC trials by repeating the same items (“words” and foils) throughout the test phase might produce in SL

measurement, no previous study has attempted to directly quantify that effect while simultaneously examining the way “words” predictability and the conditions under which they are presented to participants affect the ability to correctly discriminate “words” from foils as the 2-AFC task unfolds or the number of repetitions increases. This article is, thus, the first to provide a much-needed examination of the impact this practice might have on the SL measurement with important methodological implications.

The results were straightforward. Although the ANOVAs conducted with all the participants failed to show clear effects of item repetition on SL performance in both aSL tasks, the second set of analyses, including only participants who have shown an above-chance performance at the beginning of each 2-AFC task (presumably the “purer” measure to assess if learning had occurred during the previous exposure phase, since the test items were presented only twice), led to a different scenario. Indeed, the results from the second set of ANOVAs indicated that item repetition had clear detrimental effects on SL performance, even though only for the low-TP “words” in the aSL task performed under implicit conditions. In addition, the results from the glmm analyses conducted in the third step, to further estimate the effects of 2-AFC item repetitions on SL performance each time an item was repeated (i.e., at each block), provided additional evidence that 2-AFC item repetitions had, indeed, detrimental effects on SL performance. It happened not only for the low-TP “words” presented under implicit conditions, with an estimated decrease of 19.8% each time a low-TP “word” is

repeated, but, importantly, also in the aSL task performed under explicit conditions, with an estimated decrease of 11.7% each time a low- or a high-TP “word” is repeated.

These findings are in line with our predictions and demonstrated not only that the representations generated from the incidental (implicit) and intentional (explicit) exposure to a continuous auditory stream comprising repetition of three-syllable nonsense “words” in a triplet embedded task are not impervious to interference, as shown by previous studies using other tasks and paradigms (e.g., Eakin & Smith, 2012; Ikier et al., 2008; Jiménez et al., 2006; Lustig & Hasher, 2001; Vaquero et al., 2020), but, importantly, that the common practice observed in SL research involving the increase of 2-AFC trials by repeating the same “words” several times, even if contrasted with different foils, introduces other sources of “noise” in the SL measurement, as the significant drop in SL performance as the 2-AFC task unfolded (or as the number of repetitions increases) suggests. Thus, besides the role that episodic memory traces might play herein (Siegelman, Bogaerts, & Frost, 2017), it is also important to consider that since the extraction of TPs is a continuous process throughout the aSL task (and not just restricted to the exposure phase as most SL studies assume), increasing the number of item repetitions throughout the 2-AFC task also creates new opportunities for learning to keep occurring and, thus, to interfere with the learning that presumably has happened in the previous familiarization phase, worsening SL measurement. Specifically, in that framework, it is possible to anticipate, as we have mentioned in the Introduction, the detrimental effects of 2-AFC item repetitions to arise from the action of two complementary mechanisms: (a) as the number of 2-AFC trials repetitions increased, foils could emerge as perceptual units competing strongly for recognition; and (b) because foils are made of the same stimuli (syllables) as the “words,” the continuous extraction of foil TPs might also make the “words,” presented during the exposure, to lose their statistical properties (TPs), contributing to hamper “word” recognition as the 2-AFC unfolds. This point is particularly relevant because, under implicit learning conditions, a successful “word” discrimination seems to be more dependent on the reinstatement of the conditions under which they were presented in the first place (i.e., during the familiarization phase), and also because the absence of relevant metaknowledge (conversely to its explicit counterpart), might also contribute to decreasing control over ongoing processes (i.e., the continuous extraction of the TPs during the test phase), hence creating the conditions for interference to occur.

Nevertheless, it is important to point out that item repetition interference effects were also observed in the 2-AFC task performed when participants had prior knowledge about the regularities (“words”) embedded in the streams and had the opportunity to previously perform an analogous task, albeit implicitly. These results showed that the metaknowledge generated from these explicit learning conditions was not sufficient to prevent interference from occurring, suggesting that the extraction of TPs seems to be, indeed, an automatic mechanism that operates independently of the participants’ intentions as a result of mere exposure, as claimed by several authors (see Aslin, 2017; Saffran & Kirkham, 2018). They also agree with previous studies (e.g., Jiménez et al., 2006; Vaquero et al., 2020), showing that implicit and explicit representations are vulnerable to interference, even though impaired by different contingencies. Although our study was not designed to determine the mechanisms that may underlie these results, it is

possible, as in Jiménez et al.’s (2006) study, that the structural changes produced by the continuous extraction of TPs during the test phase, have jeopardized the strategic processes that intentional learners might have set out to cope with the task’s requirements, which might even explain the absence of differences in the interference caused by low- and high-TP “words” under explicit conditions.

Another finding that deserves attention in our results is the fact that under implicit conditions the detrimental effects of 2-AFC item repetitions were observed for low, but not high-TP “words.” Indeed, although we have predicted low-TP “words” to be more prone to interference, since they were expected to generate weaker representations (e.g., Siegelman et al., 2018; Soares, Gutiérrez-Domínguez, et al., 2020; Soares, Gutiérrez-Domínguez, Lages, et al., 2022; Soares, Gutiérrez-Domínguez, Oliveira, et al., 2022), 2-AFC item repetitions effects were also expected to be observed for the high-TP “words.” The absence of detrimental effects for this type of “words” is not readily explainable. Even though, we can hypothesize that because high-TP “words” presented a relatively low level of learning (.56 when all the participants were considered and .67 when only those who have shown signs of learning at the beginning of the aSL task were taken into account), the interference caused by the continuous extraction of TPs during the test phase, as indexed by a significant decrease in SL performance as the 2-AFC task unfolded, is less likely to be noticed for the low- than for the high-TP “words,” which from the beginning of the 2-AFC task presented higher levels of learning (.73 when all the participants were considered and .87 when only those who have shown signs of learning at the beginning of the 2-AFC task were taken into account). Future studies using other manipulations and techniques, aimed to test whether the results observed here are consistent with a null hypothesis (i.e., a true absence of item repetition interference effects for this kind of “words”), such as Bayesian modelling (see Hout et al., 2017), should be conducted before strong conclusions can be drawn. In addition, it is also likely these results emerged due to the type of measures used in this work as indicative of the interference caused by item repetitions, since this effect is largely dependent on the level of learning presented in the first place (i.e., at the beginning of the 2-AFC task).

It is also important to note that, although the above mentioned argument might lead us to thinking that the interference effect observed for the low-TP “words” in the implicit aSL task has arisen from some sort of regression to the mean, a phenomenon showing that unusually large or unusually small results tend to be followed by results that are closer to the mean, this seems not to be the case. Note that the SL performance observed for the low-TP “words” in the first block seems to not be spurious, as a similar result was also observed in the second block when all participants or only those who have shown clear signs of learning at the beginning of the 2-AFC task were considered (see Table 2 and also Figure 4). Although by chance the same result is always possible, it is very unlikely, especially when they happened consecutively one after the other, as in this case. Moreover, it is also important to emphasize that observing a better performance for the low- rather than for the high-TP “words” is not new. Soares, Gutiérrez-Domínguez, Lages, et al. (2022), in a recent work analyzing the electrophysiological correlates of SL during the exposure phase to the same aSL task used in this work with another sample of adult participants, also showed better performance for the low- rather than for the high-TP “words” in a 2-AFC task using only 16 test trials. This unexpected result was

explained by the authors based on an inevitable consequence of the manipulation of words' predictability in their study, as well as in our study and in all the studies that used "words" with different TPs (e.g., Bogaerts, et al., 2016; Grown, et al., 2020; Gutiérrez-Domínguez et al., 2022; Lages et al., 2022; Siegelman, Bogaerts, & Frost, 2017; Soares, Gutiérrez-Domínguez, et al., 2020; Soares, Gutiérrez-Domínguez, Lages, et al., 2022; Soares, Gutiérrez-Domínguez, Oliveira, et al., 2022). Indeed, although high- and low-TP "words" were presented exactly the same number of times during exposure to control for "word" frequency effects, the fact is, that high-TP "words" are made of unique syllables (i.e., syllables that only occurred in a specific "word" in a given position), as opposed to low-TP "words," which contain syllables occurring in different "words" in different syllable positions (see Method section). This might make the recognition of the low-TP "words" in the 2-AFC task easier than the high-TP "words." Note that the learning of the low-TP "words" involved not only the encoding of a fewer number of different syllables than the high-TP "words" (4 vs. 12, respectively) but, importantly, syllables that occurred three times more during exposure than the syllables of the high-TP "words." Thus, when asked to decide which of the two stimuli ("word"/foil) "sounds more familiar" based on the stream presented before, the difference in the number of syllables and in the number of times each syllable appeared in the stream may have led participants to choose the stimuli with syllables that had occurred more frequently, and that certainly generated a stronger sense of familiarity when heard in the test phase.

This interpretation is also consistent with a vast amount of psycholinguistic studies showing that word frequency (i.e., the number of times a word appears in a language) and word familiarity (i.e., the degree to which individuals know and use words in their everyday life), exert a powerful influence on word processing, as words occurring more times and which individuals are more familiar with, are more easily acquired, recognized, and produced than words occurring less frequently in their daily lives (e.g., Balota et al., 2004; Brysbaert et al., 2011; Gernsbacher, 1984; Thompson & Desrochers, 2009; see also Soares et al., 2014, 2015, 2017; Soares, Lages, Silva, et al., 2019 for an extended discussion on the effects of lexical frequency/familiarity on word recognition). This is also in agreement with other works showing that, during exposure to a continuous stream, participants are sensitive to several regularities besides TPs, such as the frequency of syllable occurrence/co-occurrence, which could explain these results (see Thiessen, 2017 or Siegelman, Bogaerts, Armstrong, & Frost, 2019, for recent reviews; and also Soares, Lages, Oliveira, & Gutiérrez-Domínguez, 2021 for a recent study with electrophysiological evidence). Thus, the high accuracy rates observed for low- rather than for high-TP "words" in the implicit condition might have precisely stemmed from the use of more complex auditory streams, entailing not only a higher number of "words" than in previous SL works (e.g., Abula et al., 2008; Arciuli & Simpson, 2012; Batterink, Reber, & Paller, 2015; Bertels et al., 2012, 2015; Saffran et al., 1996, 1999; Spencer et al., 2014; Turk-Browne et al., 2010) but, importantly, "words" with different TPs. The unbalanced distribution in the number of syllables and in the number of times each syllable appeared in the stream may have also contributed to introducing important changes in the distributional properties of the input when compared to more heterogeneous streams, like the ones we have used in this work. In a more heterogeneous stream, syllables are much less predictable than in their homogeneous counterparts (streams entailing only words

with TPs of 1.0, for example), and, as a result, each syllable carries much more information per se, "forcing" the system to rely more on smaller (syllables) rather than larger (strings of syllables) units when extracting the regularities embedded in the input (see also Siegelman, Bogaerts, & Frost, 2019). Future studies should, thus, be conducted to further analyze if presenting less complex speech streams (made of a lower and/or a less diverse type of "words") would produce the same results.

Nonetheless, even if we assume the absence of significant 2-AFC item repetition interference effects for the high-TP "words" in the implicit condition, the important point to stress here is that, overall, the present findings suggest that increasing the number of 2-AFC trials by repeating the same trials throughout the test phase contributed to worsening, and not to improving, SL measurement. In this regard, it should also be noted that additional analyses conducted on the data, from each of the aSL tasks, comparing the error variance observed with 64 items (i.e., the full presentation of the eight blocks) against the ones observed with eight items (one block), 16 items (two blocks), 24 items (three blocks), 32 items (four blocks), 40 items (five blocks), 48 items (six blocks), and 56 items (seven blocks), showed remarkably similar values and, unsurprisingly, nonsignificant differences across them (implicit aSL: .238, .230, .227, .232, .236, .237, .237, respectively; explicit aSL: .224, .217, .214, 0.217, .216, .220, .222, .223, respectively). These findings challenge, thus, the assumption in SL research that increasing the number of 2-AFC items increases the resolution of the SL measure by reducing error variance, as argued by Siegelman, Bogaerts, and Frost (2017) based on computational simulations. This disparity in the results may be a consequence of the fact that Siegelman, Bogaerts, and Frost's (2017) results were based on simulated data and not on real data collected from real participants, as in our case, and, importantly, based on the assumption that responding to a given trial is independent from responding to another trial during the 2-AFC task, hence neglecting the interference effects caused by responding/being exposed to previous trials, as our data clearly indicate. Still, it is also important to highlight that in that work, Siegelman, Bogaerts, and Frost (2017) also recognized that extensive 2-AFC item repetitions could also interfere with the learning that had taken place during the previous exposure phase, hence recommending the use of a large number of 2-AFC trials, with a minimal number of repetitions, even though they did not provide any indication about the maximum number of repetitions recommended to avoid the occurrence of said interference. Our data suggest the use of up to a maximum of three repetitions per item, once after the third block, the decrease in SL performance relative to the first and the second blocks reached statistical significance, even though the estimated model indicates that the probability of responding accurately decreases 19.8% from one block to the following block when using low-TP "words" in the implicit condition, and 11.7% in the explicit condition regardless of the type of "word."

Conclusion

This is, to the best of our knowledge, the first study aimed to directly analyze the effect that 2-AFC item repetitions produce on SL measurement by examining how the proportion of correct discriminations of three-syllable nonsense words, presented during the familiarization phase of an auditory triplet embedded task, changed as the number of 2-AFC item repetitions increased.

We also tested whether this effect was modulated by the predictability of the “words” embedded in the streams (high and low) and the conditions under which they were presented to participants (implicit and explicit) to further ascertain the conditions under which 2-AFC item repetitions were, or were not, impervious to interference. Besides demonstrating for the first time that the representations generated, from the implicit and explicit exposure to speech streams in a triplet embedded task, were not impervious to interference, as shown by previous studies using other paradigms, our results clearly indicate that increasing the number of 2-AFC trials by repeating the same items (“words” and foils) several times throughout the 2-AFC task, even if contrasted with different foils, is not recommended, or at least recommended at a maximum of three repetitions per “word.” Although the results in the implicit version of the aSL task failed to reach significance for the most common “words” used in SL research (“words” with TPs of 1.0), we nevertheless recommend caution when using this strategy as a means of improving the sensitivity of the 2-AFC task to assess SL even with that kind of “words.” Note that this result may have arisen from the way item repetition interference was assessed in our work, that is, based on a significant decrease in SL performance across blocks, hence making the possibility of observing differences less likely when performance is slightly above chance, as in the case of the high-TP “words.” Future studies should, therefore, investigate whether these results could be observed when other types of item repetition interference measures are used (such as the simultaneous collection of electrophysiological data) and/or when streams involving fewer “words” and/or less diverse “words,” in their internal composition, are used instead. In any case, as a final message, the data obtained here recommends caution and encourages researchers to use other strategies to improve SL measurement. We acknowledge that this problem is not easy to solve, as the use of a low number of observations also raises important reliability problems, and the use of other postlearning tasks such as 3- or 4-AFC tasks, pattern completion tasks, target detection tasks, or even the induced chunking recall task, necessarily involves exposing participants to further stimulation, thereby allowing for more room for interference to occur. Researchers should, therefore, look for other strategies to provide a better proxy for SL, such as collecting electrophysiological data as the exposure to the auditory stream unfolds, a method being increasingly used in SL research (e.g., Abila et al., 2008; Cunillera et al., 2006; De Diego Balaguer et al., 2007; Gutiérrez-Domínguez et al., 2022; Soares, Gutiérrez-Domínguez, et al., 2020; Soares, Gutiérrez-Domínguez, Lages, et al., 2022; Soares, Gutiérrez-Domínguez, Oliveira, et al., 2022; Soares, Lages, et al., 2022; Soares, Lages, Oliveira, & Gutiérrez-Domínguez, 2021), or using behavioural tasks such as a self-paced statistical learning task (e.g., Siegelman et al., 2018; Siegelman, Bogaerts, Armstrong, & Frost, 2019), asking participants to press a key to proceed to the following stimulus during exposure, while using differences between responses times to predictable (second and third elements) versus unpredictable stimuli (first element) within a triplet as the exposure unfolds as a better proxy of SL.

Résumé

L’une des tâches les plus populaires utilisées pour tester l’apprentissage statistique (SL) consiste à demander aux participants

d’identifier lequel de deux stimuli, un triplet présenté lors de la phase de familiarisation précédente par opposition à une nouvelle séquence composée des mêmes stimuli jamais présentés ensemble, est plus familier sur la base du flux présenté auparavant, c’est-à-dire pour effectuer une tâche de choix forcé à deux alternatives (2-AFC). Malgré l’utilisation répandue de cette tâche, cette dernière a fait l’objet de critiques croissantes dans la recherche cognitive actuelle en raison de ses failles psychométriques. Une pratique courante pour améliorer la mesure de l’apprentissage statistique consiste à augmenter le nombre d’essais 2-AFC en présentant plusieurs fois les mêmes items (triplets et perturbateurs) pendant la phase de test. Ce travail visait à analyser directement l’effet que cette pratique entraîne en examinant comment la proportion de discriminations correctes de mots absurdes de trois syllabes présentés pendant la phase de familiarisation d’une tâche auditive de triplets enchâssés changeait lorsque le nombre de répétitions d’items 2-AFC augmentait. Nous avons également testé si cet effet était modulé par la prévisibilité des « mots » intégrés dans les flux auditifs (élevée et faible) et les conditions dans lesquelles ils étaient présentés aux participants (implicite et explicite). Les résultats ont montré que les répétitions d’items de 2-AFC avaient effectivement des effets néfastes sur la mesure de l’apprentissage statistique, comme l’indique une diminution significative de la proportion de discriminations correctes au fur et à mesure que le nombre de répétitions d’items augmente, à la fois dans la tâche 2-AFC réalisée dans des conditions implicites et explicites, bien que, dans le premier cas, seulement pour les « mots » de prévisibilité faible. Ces résultats recommandent la prudence dans l’utilisation de cette stratégie pour améliorer la mesure de l’apprentissage statistique.

Mots-clés : apprentissage statistique, 2-AFC, répétition d’items, prévisibilité, apprentissage explicite

References

- Abila, D., Katahira, K., & Okanoya, K. (2008). On-line assessment of statistical learning by event-related potentials. *Journal of Cognitive Neuroscience*, 20(6), 952–964. <https://doi.org/10.1162/jocn.2008.20058>
- Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 372(1711), Article 20160058. <https://doi.org/10.1098/rstb.2016.0058>
- Arciuli, J., & Conway, C. M. (2018). The promise and challenge of statistical learning for elucidating atypical language development. *Current Directions in Psychological Science*, 27(6), 492–500. <https://doi.org/10.1177/0963721418779977>
- Arciuli, J., & Simpson, I. C. (2012). Statistical learning is related to reading ability in children and adults. *Cognitive Science*, 36(2), 286–304. <https://doi.org/10.1111/j.1551-6709.2011.01200.x>
- Arciuli, J., von Koss Torkildsen, J., Stevens, D. J., & Simpson, I. C. (2014). Statistical learning under incidental versus intentional conditions. *Frontiers in Psychology*, 5, Article 747. <https://doi.org/10.3389/fpsyg.2014.00747.00747>
- Armon, I. (2020). Do current statistical learning tasks capture stable individual differences in children? An investigation of task reliability across modality. *Behavior Research Methods*, 52(1), 68–81. <https://doi.org/10.3758/s13428-019-01205-5>
- Aslin, R. N. (2017). Statistical learning: A powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(1–2), Article e1373. <https://doi.org/10.1002/wcs.1373>

- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412. <https://doi.org/10.1016/j.jml.2007.12.005>
- Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, 133(2), 283–316. <https://doi.org/10.1037/0096-3445.133.2.283>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Mächler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using Eigen and R syntax* [Computer software]. (R package Version 0.999375-42). <https://arxiv.org/abs/1406.5823>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory and Language*, 83, 62–78. <https://doi.org/10.1016/j.jml.2015.04.004>
- Batterink, L. J., Reber, P. J., & Paller, K. A. (2015). Functional differences between statistical learning with and without explicit training. *Learning & Memory*, 22(11), 544–556. <https://doi.org/10.1101/lm.037986.114>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Methodological*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bertels, J., Boursain, E., Destrebecqz, A., & Gaillard, V. (2015). Visual statistical learning in children and young adults: How implicit? *Frontiers in Psychology*, 5, Article 1541. <https://doi.org/10.3389/fpsyg.2014.01541>
- Bertels, J., Franco, A., & Destrebecqz, A. (2012). How implicit is visual statistical learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(5), 1425–1431. <https://doi.org/10.1037/a0027210>
- Bogaerts, L., Siegelman, N., & Frost, R. (2016). Splitting the variance of statistical learning performance: A parametric investigation of exposure duration and transitional probabilities. *Psychonomic Bulletin & Review*, 23(4), 1250–1256. <https://doi.org/10.3758/s13423-015-0996-z>
- Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58(5), 412–424. <https://doi.org/10.1027/1618-3169/a000123>
- Christiansen, M. H. (2019). Implicit-statistical learning: A tale of two literatures. *Topics in Cognitive Science*, 11(3), 468–481. <https://doi.org/10.1111/tops.12332>
- Cleeremans, A., & Jiménez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. In R. French & A. Cleeremans (Eds.), *Implicit learning and consciousness: An empirical, philosophical and computational consensus in the making* (pp. 1–40). Psychology Press.
- Conway, C. M., & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual, and auditory sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1), 24–39. <https://doi.org/10.1037/0278-7393.31.1.24>
- Cunillera, T., Toro, J. M., Sebastián-Gallés, N., & Rodríguez-Fornells, A. (2006). The effects of stress and statistical cues on continuous speech segmentation: An event-related brain potential study. *Brain Research*, 1123(1), 168–178. <https://doi.org/10.1016/j.brainres.2006.09.046>
- De Diego Balaguer, R., Toro, J. M., Rodríguez-Fornells, A., & Bachoud-Lévi, A. C. (2007). Different neurophysiological mechanisms underlying word and rule extraction from speech. *PLOS ONE*, 2(11), Article e1175. <https://doi.org/10.1371/journal.pone.0001175>
- Eakin, D. K., & Smith, R. (2012). Retroactive interference effects in implicit memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(5), 1419–1424. <https://doi.org/10.1037/a0027208>
- Erickson, L. C., Kaschak, M. P., Thiessen, E. D., & Berry, C. A. (2016). Individual differences in statistical learning: Conceptual and measurement issues. *Collabra*, 2(1), 2–14. <https://doi.org/10.1525/collabra.41>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12(6), 499–504. <https://doi.org/10.1111/1467-9280.00392>
- Franco, A., Cleeremans, A., & Destrebecqz, A. (2011). Statistical learning of two artificial languages presented successively: How conscious? *Frontiers in Psychology*, 2, Article 229. <https://doi.org/10.3389/fpsyg.2011.00229>
- Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, 145(12), 1128–1153. <https://doi.org/10.1037/bul0000210>
- Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, 19(3), 117–125. <https://doi.org/10.1016/j.tics.2014.12.010>
- Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 113(2), 256–281. <https://doi.org/10.1037/0096-3445.113.2.256>
- Graf, P., & Schacter, D. L. (1987). Selective effects of interference on implicit and explicit memory for new associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1), 45–53. <https://doi.org/10.1037/0278-7393.13.1.45>
- Growns, B., Siegelman, N., & Martire, K. A. (2020). The multi-faceted nature of visual statistical learning: Individual differences in learning conditional and distributional regularities across time and space. *Psychonomic Bulletin & Review*, 27(6), 1291–1299. <https://doi.org/10.3758/s13423-020-01781-0>
- Gutiérrez-Domínguez, F. G., Lages, A., Oliveira, H. M., & Soares, A. P. (2022). Neural signature of statistical learning: Proposed signs of typical/atypical language functioning from EEG time-frequency analysis. In G. Buela-Casal (Ed.), *International handbook of clinical psychology*. Thomson Reuters Editorial.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4), 800–802. <https://doi.org/10.1093/biomet/75.4.800>
- Hout, J. W., Heathcote, A., & Eidels, A. (2017). Bayesian analyses of cognitive architecture. *Psychological Methods*, 22(2), 288–303. <https://doi.org/10.1037/met0000117>
- Ikier, S., Yang, L., & Hasher, L. (2008). Implicit proactive interference, age, and automatic versus controlled retrieval strategies. *Psychological Science*, 19(5), 456–461. <https://doi.org/10.1111/j.1467-9280.2008.02109.x>
- Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically-induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science*, 44(7), Article e12848. <https://doi.org/10.1111/cogs.12848>
- Jacoby, L. L. (1983). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9(1), 21–38. <https://doi.org/10.1037/0278-7393.9.1.21>
- Jiménez, L., Oliveira, H. M., & Soares, A. P. (2020). Surface features can deeply affect artificial grammar learning. *Consciousness and Cognition*, 80, Article 102919. <https://doi.org/10.1016/j.concog.2020.102919>
- Jiménez, L., Vaquero, J. M. M., & Lupiáñez, J. (2006). Qualitative differences between implicit and explicit sequence learning. *Journal of*

- Experimental Psychology: Learning, Memory, and Cognition*, 32(3), 475–490. <https://doi.org/10.1037/0278-7393.32.3.475>
- Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, 68(1), 601–625. <https://doi.org/10.1146/annurev-psych-122414-033702>
- Kidd, E., Arciuli, J., Christiansen, M. H., Isbilen, E. S., Revius, K., & Smithson, M. (2020). Measuring children's auditory statistical learning via serial recall. *Journal of Experimental Child Psychology*, 200, Article 104964. <https://doi.org/10.1016/j.jecp.2020.104964>
- Lages, A., Oliveira, H. M., Arantes, J., Gutiérrez-Domínguez, F. G., & Soares, A. P. (2022). Drawing the links between statistical learning and children's spoken and written language skills. In G. Buela-Casal (Ed.), *International handbook of clinical psychology*. Thomson Reuters Editorial.
- Lustig, C., & Hasher, L. (2001). Implicit memory is not immune to interference. *Psychological Bulletin*, 127(5), 618–628. <https://doi.org/10.1037/0033-2909.127.5.618>
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315. <https://doi.org/10.1016/j.jml.2017.01.001>
- Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences*, 10(5), 233–238. <https://doi.org/10.1016/j.tics.2006.03.006>
- Qi, Z., Sanchez Araujo, Y., Georgan, W. C., Gabrieli, J. D., & Arciuli, J. (2019). Hearing matters more than seeing: A cross-modality study of statistical learning and reading ability. *Scientific Studies of Reading*, 23(1), 101–115. <https://doi.org/10.1080/10888438.2018.1485680>
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Reber, A. S. (1993). *Implicit learning and tacit knowledge. An essay on the cognitive unconscious*. Oxford University Press.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926–1928. <https://doi.org/10.1126/science.274.5294.1926>
- Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1), 27–52. [https://doi.org/10.1016/S0010-0277\(98\)00075-4](https://doi.org/10.1016/S0010-0277(98)00075-4)
- Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, 69(1), 181–203. <https://doi.org/10.1146/annurev-psych-122216-011805>
- Sanders, L. D., Newport, E. L., & Neville, H. J. (2002). Segmenting nonsense: An event-related potential index of perceived onsets in continuous speech. *Nature Neuroscience*, 5(7), 700–703. <https://doi.org/10.1038/nm873>
- Siegelman, N., Bogaerts, L., Armstrong, B. C., & Frost, R. (2019). What exactly is learned in visual statistical learning? Insights from Bayesian modeling. *Cognition*, 192, Article 104002. <https://doi.org/10.1016/j.cognition.2019.06.014>
- Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711), Article 20160059. <https://doi.org/10.1098/rstb.2016.0059>
- Siegelman, N., Bogaerts, L., & Frost, R. (2017). Measuring individual differences in statistical learning: Current pitfalls and possible solutions. *Behavior Research Methods*, 49(2), 418–432. <https://doi.org/10.3758/s13428-016-0719-z>
- Siegelman, N., Bogaerts, L., & Frost, R. (2019). What determines visual statistical learning performance? Insights from information theory. *Cognitive Science*, 43(12), Article e12803. <https://doi.org/10.1111/cogs.12803>
- Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining 'learning' in statistical learning: What does an online measure reveal about the assimilation of visual regularities? *Cognitive Science*, 42(S3), 692–727. <https://doi.org/10.1111/cogs.12556>
- Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language*, 81, 105–120. <https://doi.org/10.1016/j.jml.2015.02.001>
- Sloman, S. A., Hayman, C. A. G., Ohta, N., Law, J., & Tulving, E. (1988). Forgetting in primed fragment completion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(2), 223–239. <https://doi.org/10.1037/0278-7393.14.2.223>
- Soares, A. P., Costa, A. S., Machado, J., Comesaña, M., & Oliveira, H. M. (2017). The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. *Behavior Research Methods*, 49(3), 1065–1081. <https://doi.org/10.3758/s13428-016-0767-4>
- Soares, A. P., França, T., Gutiérrez-Domínguez, F. J., Sousa, I., & Oliveira, H. M. (2022, May 23). 2-AFC item repetition on statistical learning: Materials and data. <https://osf.io/65kd2/>
- Soares, A. P., Gutiérrez-Domínguez, F. J., Lages, A., Oliveira, H. M., Vasconcelos, M., & Jiménez, L. (2022). Learning words while listening to syllables: Electrophysiological correlates of statistical learning in children and adults. *Frontiers in Human Neuroscience*, 16, Article 805723. <https://doi.org/10.3389/fnhum.2022.805723>
- Soares, A. P., Gutiérrez-Domínguez, F., Oliveira, H. M., Lages, A., Guerra, N., Pereira, A. R., Tomé, D., & Lousada, M. (2022). Explicit instructions do not enhance auditory statistical learning in children with developmental language disorder: Evidence from ERPs. *Frontiers in Psychology*, 13, Article 905762. <https://doi.org/10.3389/fpsyg.2022.905762>
- Soares, A. P., Gutiérrez-Domínguez, F. J., Vasconcelos, M., Oliveira, H. M., Tomé, D., & Jiménez, L. (2020). Not all words are equally acquired: Transitional probabilities and instructions affect the electrophysiological correlates of statistical learning. *Frontiers in Human Neuroscience*, 14, Article 577991. <https://doi.org/10.3389/fnhum.2020.577991>
- Soares, A. P., Iriarte, A., Almeida, J. J., Simões, A., Costa, A., França, P., Machado, J., & Comesaña, M. (2014). Procura-PALavras (P-PAL): Uma nova medida de frequência lexical do Português Europeu contemporâneo [Procura-PALavras (P-PAL): A new measure of word frequency for contemporary European Portuguese]. *Psicologia: Reflexão e Crítica*, 27(1), 1–14. <https://doi.org/10.1590/S0102-79722014000100013>
- Soares, A. P., Lages, A., Oliveira, H., & Hernández, J. (2019). The mirror reflects more for d than for b: Right asymmetry bias on the visual recognition of words containing reversal letters. *Journal of Experimental Child Psychology*, 182, 18–37. <https://doi.org/10.1016/j.jecp.2019.01.008>
- Soares, A. P., Lages, A., Oliveira, H. M., & Gutiérrez-Domínguez, F. J. (2021). Extracting word-like units when two concurrent regularities collide: Electrophysiological evidence. In A. Botinis (Ed.), *Proceedings of 12th international conference of experimental linguistics* (pp. 215–218). ExLing Society. <https://doi.org/10.36505/ExLing-2021-12>
- Soares, A. P., Lages, A., Oliveira, H. M., & Gutiérrez-Domínguez, F. J. (2022). Can explicit instructions enhance auditory statistical learning in children with Developmental Language Disorder? In G. Buela-Casal (Ed.), *International handbook of clinical psychology*. Thomson Reuters Editorial.
- Soares, A. P., Lages, A., Silva, A., Comesaña, M., Sousa, I., Pinheiro, A. P., & Perea, M. (2019). Psycholinguistics variables in the visual-word recognition and pronunciation of European Portuguese words: A mega-study approach. *Language, Cognition and Neuroscience*, 34(6), 689–719. <https://doi.org/10.1080/23273798.2019.1578395>
- Soares, A. P., Lages, A., Velho, M., Oliveira, H. M., & Hernández, J. (2021). The mirror reflects more for genial than for casual: Right-asymmetry bias on the visual word recognition of words containing non-reversal letters. *Reading and Writing*, 34(6), 1467–1489. <https://doi.org/10.1007/s11145-020-10100-x>
- Soares, A. P., Machado, J., Costa, A., Iriarte, A., Simões, A., Almeida, J. J., Comesaña, M., & Perea, M. (2015). On the advantages of frequency measures extracted from subtitles: The case of Portuguese. *Quarterly Journal of Experimental Psychology*, 68, 680–696. <https://doi.org/10.1080/17470218.2014.964271>

- Soares, A. P., Nunes, A., Martins, P., & Lousada, M. (2018). Do children with Specific Language Impairment (SLI) present implicit learning (IL) deficits? Evidence from an Artificial Grammar Learning (AGL) paradigm. *BMC Health Services Research*, 18(2), Article 93. <https://doi.org/10.1186/s12913-018-3444-8>
- Soares, A. P., Silva, R., Faria, F., Oliveira, H. M., & Jiménez, L. (2021). Literacy effects on artificial grammar learning (AGL) with letters and colors: Evidence from preschool and primary school children. *Language and Cognition*, 13(4), 1–28. <https://doi.org/10.1017/langcog.2021.12>
- Soares, A. P., Velho, M., & Oliveira, H. M. (2020). The role of letter features on the consonant-bias effect: Evidence from masked priming. *Acta Psychologica*, 210, Article 103171. <https://doi.org/10.1016/j.actpsy.2020.103171>
- Spencer, M., Kaschak, M. P., Jones, J. L., & Lonigan, C. J. (2014). Statistical learning is related to early literacy-related skills. *Reading and Writing*, 28(4), 467–490. <https://doi.org/10.1007/s11145-014-9533-0>
- Thiessen, E. D. (2017). What's statistical about learning? Insights from modelling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1711), Article 20160056. <https://doi.org/10.1098/rstb.2016.0056>
- Thiessen, E. D., & Erickson, L. C. (2013). Beyond word segmentation: A two-process account of statistical learning. *Current Directions in Psychological Science*, 22(3), 239–243. <https://doi.org/10.1177/0963721413476035>
- Thompson, G. L., & Desrochers, A. (2009). Corroborating biased indicators: Global and local agreement among objective and subjective estimates of printed word frequency. *Behavior Research Methods*, 41(2), 452–471. <https://doi.org/10.3758/BRM.41.2.452>
- Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *The Journal of Neuroscience*, 30(33), 11177–11187. <https://doi.org/10.1523/JNEUROSCI.0858-10.2010>
- van Witteloostuijn, M., Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2019). Assessing visual statistical learning in early-school-aged children: The usefulness of an online reaction time measure. *Frontiers in Psychology*, 10, Article 2051. <https://doi.org/10.3389/fpsyg.2019.02051>
- Vaquero, J. M. M., Lupiáñez, J., & Jiménez, L. (2020). Asymmetrical effects of control on the expression of implicit sequence learning. *Psychological Research*, 84(8), 2157–2171. <https://doi.org/10.1007/s00426-019-01222-1>

Received August 25, 2021

Revision received May 23, 2022

Accepted June 9, 2022 ■