Cognition 236 (2023) 105439

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

What sticks after statistical learning: The persistence of implicit versus explicit memory traces

Helen Liu[1], Tess Allegra Forest[1], Katherine Duncan[2], Amy S. Finn[2,*]

Department of Psychology, University of Toronto, 100 St. George Street, 4th floor, Sidney Smith Hall, Toronto, ON M5S 3G3, Canada

ARTICLE INFO

Keywords:
Statistical learning
Memory
Explicit memory
Implicit memory
Memory retention

ABSTRACT

Statistical learning is a powerful mechanism that extracts even subtle regularities from our information-dense worlds. Recent theories argue that statistical learning can occur through multiple mechanisms—both the conventionally assumed automatic process that precipitates unconscious learning, and an attention-dependent process that brings regularities into conscious awareness. While this view has gained popularity, there are few empirical dissociations of the hypothesized implicit and explicit forms of statistical learning. Here we provide strong evidence for this dissociation in two ways. First, we show in healthy adults ($N = 60$) that implicit and explicit traces have divergent consolidation trajectories, with implicit knowledge of structure strengthened over a 24-h period, while precise explicit representations tend to decay. Second, we demonstrate that repeated testing strengthens the retention of explicit representations but that implicit statistical learning is uninfluenced by testing. Together these dissociations provide much needed support for the reconceptualization of statistical learning as a multi-component construct.

1. Introduction

Statistical learning has been put forth as a powerful learning mechanism that supports the effortless extraction of meaningful regularities from our information-dense worlds (Saffran, Aslin, & Newport, 1996). It can enable language and word learning (Graf Estes, Evans, Alibali, & Saffran, 2007), support our perception of events (Buchsbaum, Griffiths, Plunkett, Gopnik, & Baldwin, 2015), and help us extract visual-spatial regularities (Fiser & Aslin, 2001). Throughout these varied contexts, statistical learning was thought to unfold outside of conscious awareness, with automatically extracted regularities implicitly shaping behaviour (Perruchet & Pacton, 2006; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997; Turk-Browne, Jungé, & Scholl, 2005). Accordingly, learners report little knowledge of learned patterns during debriefing (Turk-Browne et al., 2005), and learning is observed in indirect assessments that do not require conscious reflection (Kim, Seitz, Feenstra, & Shams, 2009). However, recent theoretical work proposes that statistical regularities in these same tasks are learned through multiple neurocognitive systems, generally conceived of as implicit and explicit (Arciuli, 2017; Batterink, Paller, & Reber, 2019; Conway, 2020; Daltrozzo & Conway, 2014; Savalia, Shukla, & Bapi, 2016). As yet, studies supporting the complementary contributions of both learning mechanisms are sparse, leaving questions about whether this reconceptualization is warranted.

While early studies found that learners cannot explicitly articulate their statistical learning (Saffran et al., 1997; Turk-Browne et al., 2005), others have shown that learners' confidence in statistical learning judgements tracks with their accuracy, suggesting that conscious access is possible (Batterink, Reber, Neville, & Paller, 2015; Bays, Turk-Browne, & Seitz, 2016). But occasional conscious access could arise without separate explicit and implicit statistical learning traces; learning within a single system may be privy to conscious access when sufficiently strong, but implicitly bias behaviour when weaker (Cleeremans, 2006). Stronger evidence for multiple traces comes from one study that measured learning in two ways: using a forced-choice recognition task which tracked explicit confidence, and implicit knowledge using a reaction-time-based target-detection task (Batterink et al., 2015). While performance on both tests reflected learning, performance across tasks did not correlate across participants. Though only powered to detect strong correlations, this provides initial evidence that explicit and

[*] Corresponding author.
E-mail address: finn@psych.utoronto.ca (A.S. Finn).
[1] These authors contributed equally as first authors
[2] These authors contributed equally as senior authors

https://doi.org/10.1016/j.cognition.2023.105439
Received 15 December 2020; Received in revised form 17 June 2022; Accepted 9 March 2023
Available online 17 March 2023
0010-0277/© 2023 Elsevier B.V. All rights reserved.

implicit forms of statistical learning could operate in parallel. Relatedly, Bays et al. (2016) found that reliable implicit learning (measured in a novel target detection task) was only observed for stimuli that were *not* later explicitly recognized. On the surface, this result suggests that separate implicit and explicit statistical learning processes not only exist, but that they also compete with each other. This interpretation should be tempered, though, because competitive patterns were also observed for implicit learning between their novel task and another implicit target detection task, similar to that employed by Batterink et al. (2015).

And, so, further empirical research is required to determine whether statistical learning results in dissociable memory traces. Here, we address the need to more carefully understand implicit and explicit traces of statistical memory by investigating two factors likely to dissociate these types of memory. First, we targeted their differential forgetting rates, an established gold standard for dissociating memory systems (Dosher & Rosedale, 1991). Indeed, explicit traces typically decay more quickly than implicit (Goshen-Gottstein & Kempinsky, 2001; Graf, Squire, & Mandler, 1984; Rappold & Hashtroudi, 1991). Building on this important body of work, we sought to determine whether parallel implicit and explicit statistical learning traces decay at different rates over the course of 24h. Of note, although the persistence of motor sequence learning is well studied (Galea, Albert, Ditye, & Miall, 2009; Kóbor, Janacsek, Takács, & Nemeth, 2017; Romano, Howard, & Howard, 2010; Sanchez, Gobel, & Reber, 2010; Willingham & Dumas, 1997), there are just three studies that have tested non-motor statistical learning across a 24h delay, with two showing preserved performance (Arciuli & Simpson, 2012; Kim et al., 2009) and the other showing improved performance following consolidation (Durrant, Taylor, Cairney, & Lewis, 2011). The relative contribution of explicit and implicit learning at each delay, however, remains unclear; while one study did measure memory in both ways (Kim et al., 2009), unlike other research (Batterink et al., 2015; Bays et al., 2016), explicit memory was not observed at even short delays, possibly because of rigorous retrieval demands.

In tandem with manipulating delay, we further asked whether repeated testing would dissociate implicit and explicit forms of statistical learning. Indeed, testing effects shape the long-term fate of explicit memories, with otherwise rapid forgetting rates (Rubin & Wenzel, 1996; Wixted, 2004) ameliorated by testing their content before a delay (Roediger & Karpicke, 2006). By contrast, testing effects are poorly understood for implicit memory, likely because it is standard to either retest all content across delays, or test different subsets at each delay. So, by including this factor, we not only test whether implicit and explicit forms of statistical learning are dissociable, we also gain insights into the generality of a well-documented memory enhancement strategy.

With these aims, we exposed participants to an auditory artificial language comprised of a continuous stream of structured triplets, in keeping with established auditory statistical learning paradigms (Finn, Kharitonova, Holtby, & Sheridan, 2019; Forest, Lichtenfeld, Alvarez, & Finn, 2019). We then *indirectly tested* learning with a target-detection task—designed to measure implicit knowledge—and *directly tested* learning with a recognition task and a confidence judgement—designed to measure explicit knowledge. Since the rates at which different types of test foils are falsely recognized as words provides insights into the nature of statistical learning traces (Endress & Mehler, 2009; Forest, Finn, & Schlichting, 2021), we included two types of foils in our explicit test. To chart consolidation trajectories, we tested participants' learning twice: immediately after exposure and 24-h after exposure. To preview, we observed increased performance in the indirect but not direct test across the delay. To ensure that this improvement indeed reflected consolidation and not exposure to the language, we included an additional Indirect Control Group who completed both tests in succession on the same day. Importantly, to determine the impact of repeated testing on both implicit and explicit knowledge, the immediate tests (both direct and indirect) probed the same half of the material while the delayed tests probed it exhaustively.

2. Method

2.1. Participants

Sixty individuals participated in the Experimental Group (39 female, 21 male; mean age = 19.9 years, *SD* = 2.4). A sample size of 60 was selected prior to data collection and provides over 80% power to detect moderate effect sizes (Cohen's d = 0.4). While no prior research has directly addressed the question at hand, statistical learning can be observed on direct and indirect assessments with effect sizes in the medium to large range (Cohen's d = 0.64–2.9; Batterink et al., 2015; Durrant et al., 2011; Kim et al., 2009). Sixty-three individuals participated in the Indirect Control Group, but there were errors during data collection for seven subjects (experiment code crashed, $N = 6$; withdrew from study half way through, $N = 1$), leaving 56 (36 female, 20 male; mean age = 19.7 years, *SD* = 2.0) for inclusion in analyses in the Indirect Control Group, to approximately match the sample size of the Experimental Group. All participants were fluent in English, had normal hearing, and had no history of psychiatric disorders. They were also given course credit or money for their participation and provided written informed consent. All research procedures were approved by the University of Toronto's research ethics board.

2.2. Stimuli

Participants were exposed to one of two artificial languages (A or B, available at https://osf.io/u4q2a/). Both languages were constructed from the same inventory of six vowels (a, e, i, o, u, ae) and consonants (t, d, b, p, k, g), paired to generate 18 syllables. These naturalistic syllables were produced by a formally trained opera singer and recorded using a Zoom H5 digital multitrack recorder. The syllable recordings were then edited to remove silences and normalized to a standard duration (450 ms) and average pitch (246 Hz). Syllables were then concatenated in a continuous auditory recording with no additional pauses between syllables. Different "words" – fixed syllable trigrams – comprised each language, such that trigrams that served as words in one language served as shifted non-word foils (described below) in the other (Fig. 1a and Supplemental Fig. 1). No consonants or vowels repeated within a word and no two words contained the same vowel or pair of consonants. Each word was repeated 75 times within a 10-min language in pseudorandomized order, with the restrictions that the same word could not appear twice in immediate succession and that each word was equally likely to follow every other word. This ensured an equal transitional probability of 0.2 between syllables at all word boundaries, as compared to a within word transitional probability of 1.

2.3. Procedure

After exposure to an artificial language, participants completed one test that directly assessed their explicit knowledge of the language's statistical structure, followed by another that indirectly assessed their implicit knowledge. These immediate (*first)* tests only assessed half of the words from the exposed language, counterbalanced across participants. To test the retention of both previously tested (*retested*) and untested (*unique*) words, participants then completed a direct test followed by an indirect test (both of which contained all exposed words) either 24-h later (Experimental Group) or immediately (Indirect Control Group).

Exposure. Languages were counterbalanced across participants and presented through Sony MDRX100 Series Stereo headphones at a volume adjusted by the participant for clarity and comfort. Participants were instructed to listen to the 10-min language and were told that they would be asked questions about it afterward. Following previous work (Finn & Hudson Kam, 2008; Finn & Hudson Kam, 2015), we asked them

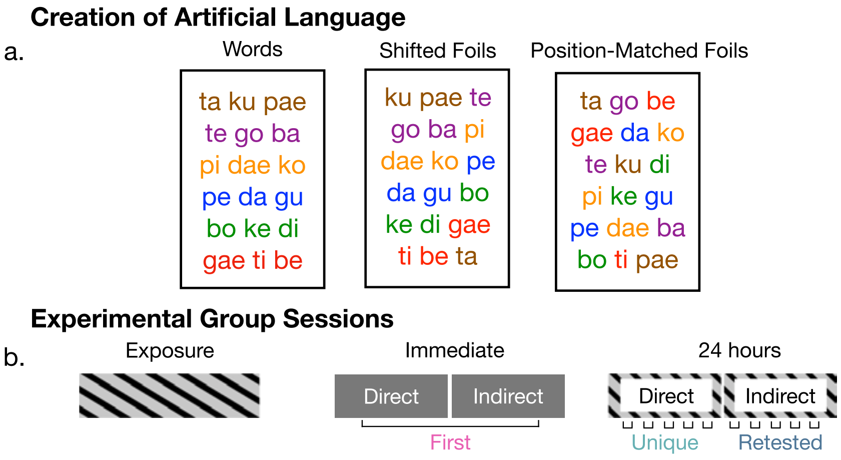

Fig. 1. a) The artificial languages (version A is depicted) were comprised of 6 trisyllabic words. Test items included these words and two foil types (6 each); shifted foils preserved one transitional probability, but broke the other; position foils broke all of the transitional probability information but maintained the position of the syllables relative to the words. Note that syllables are colored according to word to highlight the shifted foil structure. b) The experimental procedure consisted of an exposure phase in which participants listened to the 10-min language; immediately after, they completed a test which measured their learning of ½ of the words and foils on the direct test and the same ½ of the words on the indirect test. The direct test was always first. 24-h later the Experimental Group (visualized) completed a full test of all the words and foils on the direct test and all the words on the indirect test. The Indirect Control Group completed the same procedure but in immediate succession. Light and dark grey colour striping depicts that retested and uniquely tested words were randomly intermixed during their exposure and the delayed test.

to not tune out the language or subvocalize about other things. To encourage this, they were asked to complete a jigsaw puzzle during exposure.

Direct Test. We directly tested statistical learning using an old/new word recognition test with a confidence judgement; we chose this over the two-alternative-forced-choice (2AFC) test that is more frequently used in statistical learning because the 2AFC measure has been shown to reflect implicit memory traces (Voss, Baym, & Paller, 2008). Participants heard one trisyllabic stimulus at a time at the same speed as during exposure, 1/3 of which were intact “words” from the previously exposed languages, while 2/3 were “non-word” foils. Participants were asked to identify whether each was a word or non-word using a button press and then rate their confidence on a 5-point scale, with 5 being very confident and 1 being not confident.

We included two commonly used types of non-word foils: *shifted foils* and *position foils*. For Language A, each shifted foil began with the 2nd followed by the 3rd syllable of an exposed word, followed by the 1st syllable of a different word (Fig. 1a). Conversely, for Language B, each shifted foil began with the 3rd syllable of one word followed by the 1st and 2nd syllables of a different word. Thus, shifted foils contained one intact within-word syllable transition and one low-probability (0.2) word-boundary transition. This combination of forward and backward shifting also ensured that shifted foils for each language were words in the other, thus counterbalancing word vs. foil status across participants.

Each position foil was generated by recombining a different set of previously exposed syllables into a novel triplet while maintaining their original ordinal within-word position (e.g., syllables which occurred in the 1st position of a word during exposure also occurred in the 1st position of these foils). Syllables, consonants, and vowels were also never repeated within position foils, and no two syllables came from the same exposed word.

During the immediate test, participants heard nine different stimuli (three words, three shifted foils, three position foils), each repeated three times (27 trials). During the delayed test, participants heard 18 different stimuli (six words, six shifted foils, six position foils), each repeated three times (54 trials). Nine of these stimuli were tested during the immediate assessment while the other half had not been previously tested. Pseudorandomized trial sequences were generated for each testing session, such that the same word could not be presented on two consecutive trials and the ratio of word to foil trials was maintained in the first and second half of the test. Half of the participants were tested using one fixed sequence and the other half were tested with trials presented in the reversed order.

2.3.1. Indirect test

We used a syllable detection task to indirectly assess statistical learning, since previous work suggests that this kind of measure is more likely to reflect implicit representations (Batterink et al., 2015; Kim et al., 2009; Turk-Browne et al., 2005), and is reliable within individuals (Siegelman, Bogaerts, Kronenfeld, & Frost, 2018). At the beginning of each trial, participants were given a different target syllable. They heard the syllable 3 times and were asked to repeat it after each presentation to the experimenter who recorded their accuracy. One participant (in the Experimental Group) was not able to accurately repeat syllables and was excluded from the indirect test analyses; this participant also had the poorest target detection performance. After verifying that participants comprehended the target syllable, they were asked to press a key every time they heard it within a unique 24 s stream of the unsegmented artificial language. Each stream included 54 syllables, which corresponded to three presentations of each of the six trisyllabic words. Each tested syllable served as a target once per session and occurred exactly 3 times in its corresponding stream with the same presentation constraints as exposure (i.e., words could not repeat on consecutive trials). This resulted in 27 possible targets in the immediate test and 54 in the delayed test. Each target syllable was tested using a fixed, unique stream, generated for each testing condition (immediate and delayed for both languages A and B). Half of the participants were tested in one order while the other half were tested in the reverse order. Importantly, the immediate indirect test probed the same stimuli (half of the words) as the immediate direct test while the delayed tests (for both direct and indirect) probed all the material exhaustively.

Note that while the immediate indirect test only probed half the words, all words were included in test streams to maintain the statistics of the language. Accordingly, delay was correlated with language exposure in this test. To understand the possible impact of additional exposure during this test independent of the delay, we included an additional Indirect Control Group who received the exact same tests as the Experimental Group, just in immediate succession.

2.4. Data analysis

All analyses were performed using R, version 3.6.2 (R Core Team, 2013), with the exception of Bayes Factors, which were obtained to evaluate evidence for the null hypothesis using JASP, version 0.14, (JASP Team, 2020). All data and analysis scripts are publicly available at https://osf.io/u4q2a/. Repeated measures ANOVAs were used in analyses with one observation per participant per condition (e.g., dprime, consolidation scores), and linear mixed effects models were used in analyses with multiple observations per participant per condition (e.g., reaction time).

2.4.1. Direct recognition test

Participants’ ability to discriminate words from foils was quantified with dprime (Z(hit rate) - Z(false alarm rate)). To avoid hit and false alarm rates of 1 and 0, a loglinear correction was used (Stanislaw & Todorov, 1999). Specifically, 0.5 was added to the number of hit and false alarms and 1 was added to the total number of old and new trials before calculating hit and false alarm rates. For the delay test, only false alarms to uniquely tested foils (not foils that were presented in the immediate test) were used to calculate dprime to ensure that false alarms aren't driven by participants' exaggerated familiarity with lures that were presented in the earlier test.

To understand if participants were aware of their knowledge, word recognition was analyzed according to participants' confidence ratings. Five-point confidence ratings were z-scored within participant to account for differences in the use of the confidence scale. We then binarized these ratings as "high" when above 0 and "low" when below to have a sufficient number of trials per confidence bin when calculating dprime in each condition. Repeated measures ANOVAs were used to determine if dprime across conditions (first, retested, unique) and foil type differed according to confidence ratings (high, low). Participants with missing data (i.e., who did not use a particular confidence rating for a particular foil type or word condition) in at least one cell of any of the three foil-types (combined, shifted, position) were removed ANOVAs (N = 17 participants).

Our approach was to compare dprimes using a repeated measures ANOVA which included foil type, test session, and their interaction as within-participant predictors. To foreshadow, foil type impacted dprime (and false alarm rates) and thus, dprime was separately calculated using each foil type to gain insights into the content of learned word representations.

We then used repeated measures ANOVAs to determine if dprime, separately for each foil type, was influenced by testing condition (first, retested, unique). For all ANOVAs, Greenhouse-Geisser corrections were used for inference when Mauchly's test for sphericity surpassed an alpha of 0.05 (corrected p-values reported as p_{GG}). Planned paired t-tests were completed to unpack the pairwise differences contributing to a significant main effect. One-sample t-tests were used to ensure that performance was above chance in each condition, Bonferroni corrected for multiple comparisons.

2.4.2. Indirect syllable detection test

Learning was operationalized as faster response times (RT) to target syllables that occurred in later as compared to earlier positions within a word (*RT speeding*). Because it takes time to plan a motor response and responses made >750 ms after target onset bleed well into the onset of the next syllable, responses made within 200-750 ms of the target syllable onset were coded as hits (53%) and all other responses were coded as false alarms (3.7%) and excluded from subsequent RT analyses. RTs beyond 2 SD of the participants' mean were also excluded (2.4% of hit responses).

RT speeding was assessed using a linear mixed-effects model (Bates, Mächler, Bolker, & Walker, 2015). The model predicted RT with the linear effect of target syllable position (centered on position 2), testing condition (first, retested, unique; first served as the base level), and their interaction. Random slopes for these variables, grouped by participant, were also included in the model along with random intercepts.

To account for the possibility of online learning within each test stream of the syllable detection task (Batterink, 2017), we also included covariates coding for the number of times each target had been presented within that stream. Because there was one target per test stream, which was always presented three times, this value could only be 1, 2, or 3, and represented the 1st, 2nd, or 3rd time that target had been present in that test stream. However, to allow for the possibility that online learning would be more pronounced for later syllable positions, we modelled presentation number separately for each syllable position (terms "pos1_rep", "pos2_rep", and "pos3_rep" in the model syntax below). In other words, because practice detecting a particular target might lead to more learning for more predictable syllables, we modelled each syllable position (position 1, 2, and 3) with their own term. Random effects were modelled using an unstructured covariance matrix, allowing correlations between random effects:

$$\begin{aligned}&\text{lmer}\big(\text{rt} \sim \text{syllable\_pos}^{*}\text{test\_cond} + \text{pos1\_rep} + \text{pos2\_rep} + \text{pos3\_rep}\\&+ \left(\text{syllable\_pos}^{*}\text{test\_cond} \mid \text{subject}\right)\end{aligned}$$

Confidence intervals around unstandardized β estimates were generated using a bootstrapping procedure with 500 simulations. Individual participants' response speeding in each testing condition was estimated in separate linear regression models, which contained the same fixed-effects predictors and covariates as the group mixed effects model.

2.4.3. Correlation between direct and indirect tests

Pearson's product-moment correlations were computed to assess the relationship between the retention of learning measured through the indirect test (individual participant RT speeding) and the direct test (dprime). These correlations were performed separately for each testing condition (first, retested, unique). We note here that our indirect test included a maximum of 9 responses for each syllable position (1, 2 and 3), a lower number than has been used in previous studies (Batterink et al., 2015; Turk-Browne et al., 2005). Accordingly, our estimates of each participant's learning may have been less precise. However, this shortcoming is offset by our sample size being 2–3 times larger, which gave us the power to detect smaller effect sizes.

2.4.4. Consolidation

The dependent measures (dprime and RT speeding) obtained on each assessment have quite different scales. Thus, to directly compare the consolidation of learning assessed by each, we normalized the variables (using a z-score) to place them on the same scale. We then measured the difference between each participant's immediate and delayed performance (first compared to unique, first compared to retested) to derive a normalized consolidation score. For the consolidation scores from the direct measure, we again ran a repeated measures ANOVA which included foil type and delay test condition and their interaction as within-participant predictors. As in previous analyses on this direct measure, foil type impacted the consolidation scores and the two foil types were thus considered separately in subsequent analyses. To do this, we calculated separate consolidation indices by measuring the difference between z-scored dprimes on each participant's immediate and delayed performance (first compared to unique, first compared to retested) separately for shifted and position foils. Repeated measures ANOVAs were then conducted with the test type (direct, indirect) and condition (unique, retested) as within subject factors. In addition, follow-up paired t-tests were conducted to compare the consolidation observed on indirect vs. direct tests within test types.

3. Results

3.1. Direct test

Before assessing our central questions, we first confirmed that participants more accurately discriminated words from lures when expressing high as compared to low confidence in their judgements ($F(1,46) = 18.63$, $p < 0.001$, $\eta_G^2 = 0.05$; Fig. 2a), suggestive of the meta-awareness emblematic of explicit memory. This relationship between awareness and discrimination performance did not reliably differ across testing conditions ($F(2,92) = 1.95$, $p_{GG} = 0.15$, $\eta_G^2 = 0.01$), nor between lure types ($F(1,46) = 2.15$, $p = 0.15$, $\eta_G^2 = 0.003$). We also confirmed that recognition performance did not reliably differ across repetitions of test items ($F(2,118) = 1.83$, $p = 0.16$, $\eta_G^2 = 0.005$), and that repetition did not interact with test condition (first, unique, retested; $F(4,236) = 1.40$, $p = 0.23$, $\eta_G^2 = 0.007$; see also Supplementary Materials). We therefore do not consider item-repetition in further analyses.

Looking at dprime, we observed main effects of testing condition (F

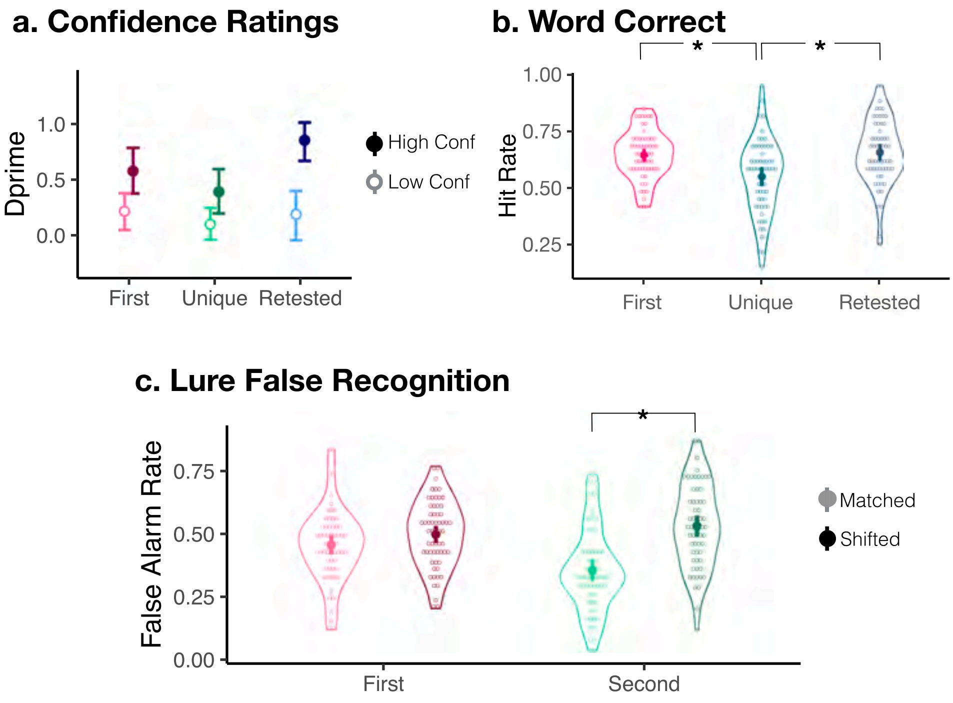

Fig. 2. a) Participants more successfully discriminated words from foils on high as compared to low confidence trials ($p < 0.001$) across all testing conditions (first, pink; retested, blue; unique, teal). In a, dots indicate the mean and bars reflect bootstrapped 95% confidence intervals, using a within-subject correction (Cousineau, 2005). As shown in the key, shading indicates confidence with darker shades reflecting high confidence and lighter reflecting low. b) Hit rates for words that were tested only (uniquely) during the second session were lower than for words that were tested in the first session ($p = 0.003$) and words that were retested during the second session ($p = 0.006$). c) While false alarms for matched and shifted foils were comparable during the first test ($p = 0.15$), they were more frequent for shifted as compared with matched foils during the second test ($p < 0.001$). In b and c, bold dots indicate the group mean, lines extending from the dots reflect bootstrapped 95% confidence intervals around the mean, open dots display individual participant means. In c, colour reflects foil type (per key). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

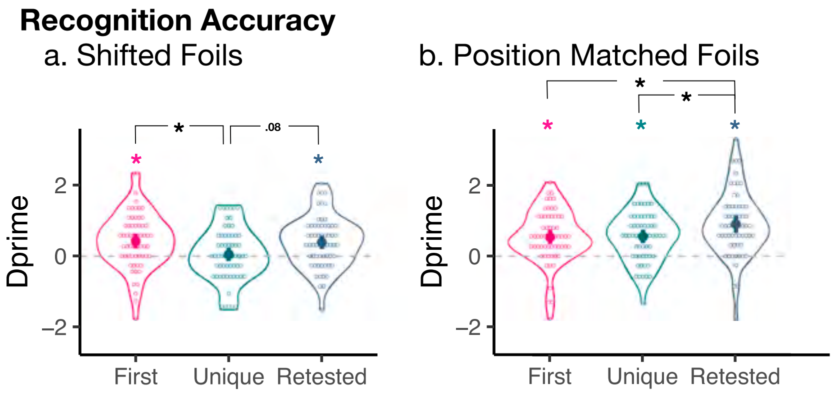

Fig. 3. a) In discriminating words from *shifted foils*, performance was reliably above chance when tested immediately ($p < 0.001$, pink) or retested after a delay ($p < 0.001$, blue), but not when uniquely tested after a delay ($p = 0.61$, teal). Discrimination depended on testing condition ($p_{GG} = 0.005$), with performance in the unique condition (teal) lower than that of the first ($p = 0.01$) and retested ($p = 0.004$) conditions. b) In discriminating words from *position foils*, performance was reliably above chance throughout all testing conditions (first $p < 0.001$; retested $p < 0.001$; unique $p < 0.001$). Performance also depended on testing condition ($p = 0.01$), with retesting leading to better performance compared to the first ($p = 0.008$) and uniquely tested conditions ($p = 0.004$). Bold dots indicate the group mean, lines extending from the dots reflect bootstrapped 95% confidence intervals around the mean, open dots display individual participant means, colored stars indicate a difference from chance (one-sample *t*-test; chance = dprime of 0). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(2,118) = 4.54, $p = 0.01$, $\eta^2_G = 0.03$) and foil type ($F(1,59) = 33.42$, $p < 0.001$, $\eta^2_G = 0.06$) that were qualified by an interaction ($F(2,118) = 9.33$, $p = 0.02$, $\eta^2_G = 0.01$). Correspondingly, hit rates also differed by testing condition ($F(2,118) = 7.00$, $p < 0.01$, $\eta^2_G = 0.06$), such that people more often recognized words on the first test (65%) and retested words on the second test (66%) than words that were only (uniquely) tested during the second session (55%; first vs. unique: $t(59) = 3.12$, $p = 0.003$; retested vs. unique: $t(59) = 2.88$, $p = 0.006$; Fig. 2b), consistent with the forgetting of untested material. False alarm rates, on the other hand, did not change across sessions ($F(1,59) = 2.56$, $p = 0.11$, $\eta^2_G = 0.008$, first: 48%; second: 45%). However, as with the dprime ANOVA, there was both a main effect of foil type ($F(1,59) = 28.75$, $p < 0.001$, $\eta^2_G = 0.08$), and an interaction between session and foil type ($F(1,59) = 9.72$, $p < 0.003$, $\eta^2_G = 0.03$; Fig. 2c). Specifically, participants falsely endorsed foils matched for syllable position and transition structure in the immediate test ($t(59) = -1.44$, $p = 0.15$), but found the position matched foils less alluring after a delay ($t(59) = -5.78$, $p < 0.001$; Fig. 2c). Thus, we probed the effect of testing condition for each foil type separately in subsequent analyses.

3.1.1. *Shifted foils*

First, we examined participants' ability to discriminate words from shifted foils, an ability which depends on fine-grained representation of syllable transitional probabilities (average transitional probability of 1 in words vs. 0.6 in shifted foils) and/or representation of syllable position within a word. When considering shifted foils, recognition performance was reliably above chance when tested immediately ($t(59) = 4.17$, $p < 0.001$, Cohen's $d = 0.54$; adjusted alpha = 0.017; Fig. 3a) and when retested after a delay ($t(59) = 4.23$, $p < 0.001$, Cohen's $d = 0.55$; adjusted alpha = 0.017; Fig. 3b), but not when uniquely tested after a delay ($t(59) = 0.51$, $p = 0.61$, Cohen's $d = 0.07$; adjusted alpha = 0.017; Fig. 3a). Indeed, this discrimination depended on testing condition ($F(2,118) = 6.05$, $p_{GG} = 0.005$, $\eta^2_G = 0.05$), with performance in the unique condition lower than that observed on both the first ($t(59) = 2.65$, $p = 0.01$, Cohen's $d = 0.34$) and retested ($t(59) = 2.96$, $p = 0.004$, Cohen's $d = 0.38$) conditions.

3.1.2. *Position foils*

We next examined participants' ability to discriminate words from position foils, an ability which can be supported by even a coarse-

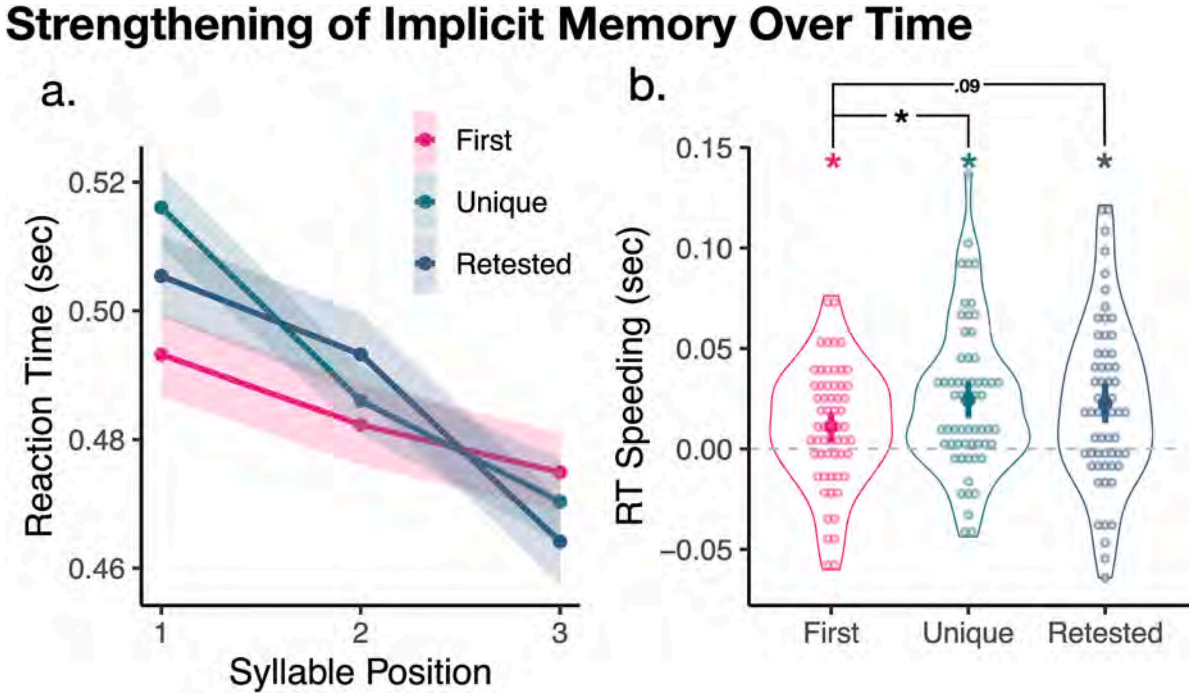

Fig. 4. a) Participants' RTs linearly decreased as a function of syllable position (response speeding) when tested initially ($p = 0.009$; pink), retested ($p < 0.001$; blue), and uniquely tested after a delay ($p < 0.001$; teal). Dots reflect the mean RT, and bands around connecting lines reflect the standard error of the mean, using a within-subject correction (Cousineau, 2005). b) RT speeding—the slope for which RT decreases across syllable position—is plotted for each test condition: the first test in pink, retested in blue and unique in teal. Compared to the first test session, response speeding was marginally more pronounced during the delayed test for repeated items ($p = 0.09$) and more pronounced for uniquely tested items ($p = 0.02$). Solid dots indicate the group mean, lines extending from the dots reflect the bootstrapped confidence interval around the mean, open dots display individual participant slopes, colored stars indicate a difference from chance (one-sample t-test; chance = 0). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

grained representation of transitional probabilities (average transitional probability of 1 in words vs. 0 in foils) but which would be hindered by reliance on representations of syllable position within a word. Unlike the discrimination of shifted foils, participants maintained a robust capacity to discriminate position foils from words throughout all testing conditions (first $t(59) = 5.54$, $p < 0.001$, Cohen's $d = 0.72$; unique $t(59) = 6.07$, $p < 0.001$, Cohen's $d = 0.78$; retested $t(59) = 7.60$; $p < 0.001$, Cohen's $d = 0.92$; adjusted alpha = 0.017; Fig. 3b). Performance, once again depended on testing condition ($F(2,118) = 4.83$, $p = 0.01$, $\eta_p^2 = 0.04$), with retesting leading to superior performance compared to the first ($t(59) = 2.71$, $p = 0.008$, Cohen's $d = 0.35$) and uniquely tested conditions ($t(59) = 2.96$, $p = 0.004$, Cohen's $d = 0.38$; Fig. 3b).

The qualitatively different patterns in forgetting observed across the two foil comparisons suggests that, if not initially tested, explicit representations of detailed syllable transition structure fade quickly, such that foils with some maintained transitions appear every bit as 'word-like' as fully intact words (shifted foils, unique testing condition). Initial testing, however, preserves these fine-grained representations, enabling participants to make the challenging discriminations with similar success as in immediate tests (shifted foils, retested condition).

3.2. *Indirect test*

The average time to detect target syllables is plotted in Fig. 4a and the distribution of response speeding (linear slope of RT across syllable positions) by testing condition is displayed in Fig. 4b. When immediately tested (first), participants' RTs linearly decreased as a function of syllable position ($\beta = 0.010$, $CI = [0.002–0.018]$, $t(73) = 2.68$, $p = 0.009$), consistent with prior research showing that learned statistical structure is associated with the faster detection of syllables occurring later in a word (including continued decreases in reaction time after the second syllable or shape (Batterink et al., 2015; Kim et al., 2009; Turk-Browne et al., 2005)). This response speeding was even more pronounced when tested during the second, delayed session (retested: $\beta = 0.020$, $t(52) = 4.15$, $p < 0.001$; unique: $\beta = 0.025$, $t(80) = 5.40$, $p < 0.001$), with statistically significant increases in speeding for uniquely tested syllables ($\beta = 0.014$, $CI = [0.001–0.027]$, $t(99) = 2.38$, $p = 0.02$), and marginal increases for repeatedly tested syllables ($\beta = 0.014$, $CI = [-0.002–0.021]$, $t(65) = 1.73$, $p = 0.09$). This suggests that implicit knowledge of the language's statistical structure was strengthened over the consolidation period, regardless of whether the material was initially tested ($p = 0.53$).

Notably, the opportunity for consolidation over the delay appeared to slow position 1 responses rather than speed responses to predictable position 2 and 3 syllables, akin to how others have observed position 1 slowing following learning (Batterink, 2017). However, as participants may have different baseline response speeds across these sessions, the slope across syllable positions is the only meaningful variable.

To ensure that this improvement was the consequence of our delay manipulation rather than increased experience with the language across testing sessions, we investigated response speeding in the Indirect Control Group, who were presented with both test phases in immediate succession (i.e., without the 24-h delay). While control participants still demonstrated significant response speeding across syllable positions in each testing condition (first test: $\beta = 0.013$, $CI = [0.009–0.018]$, $t(163) = 3.20$, $p = 0.002$; unique: $\beta = 0.019$, $CI = [0.015–0.025]$, $t(69) = 4.17$, $p < 0.001$; retested: $\beta = 0.02$, $CI = [0.017–0.026]$, $t(55) = 4.15$, $p < 0.001$), there was no improvement on the second test for unique ($\beta = 0.007$, $CI = [-0.000001–0.013]$, $t(72) = 1.00$, $p = 0.32$) or retested syllables ($\beta = 0.007$, $CI = [0.001–0.014]$, $t(177) = 1.22$, $p = 0.22$). Indeed, an ANOVA examining the impact of test condition (First, Retested, and Unique) on individual subjects' slopes indicated there was moderate evidence in favor of the null hypothesis ($BF_{01} = 9.09$), suggesting that test condition did not impact the slope of reaction time between syllables in the Indirect Control Group. These results therefore suggest that additional exposure between the first and second test is not responsible for the increase in response speeding, leaving consolidation of the implicit knowledge as the likely driver of our effect.

3.3. Correlations between direct and indirect test performance

Interestingly, participants' performance was not reliably correlated across indirect and direct assessments in any of the testing conditions (first: $r(57) = 0.13$, $\mathrm{BF}_{01} = 3.8$; retested: $r(57) = 0.16$, $\mathrm{BF}_{01} = 3.1$; unique: $r(57) = 0.19$, $\mathrm{BF}_{01} = 2.2$; all $ps > 0.14$; Fig. 5), with moderate evidence favoring the null relationship in two of the three correlations. This pattern further supports the idea that our direct and indirect tests assessed independent forms of statistical knowledge.

3.4. Comparing the consolidation of implicit and explicit memories

The above analyses suggest that implicit and explicit memories that are formed from statistical learning are different. In particular, implicit memories appear to be strengthened over the delay whereas explicit memories of detailed statistical structure (reflected in shifted foil discrimination) appear to be forgotten, particularly when not immediately tested. To directly compare the consolidation of the different types of statistical knowledge probed by each test, we normalized performance on indirect and direct tests. Because the foil type influenced the pattern of results observed on the direct tests, we first investigated whether foil type would also influence the consolidation pattern for direct knowledge. As expected, an ANOVA comparing consolidation scores of each foil type and test condition confirmed that there was a main effect of foil type ($F(1,58) = 8.67$ $p = 0.005$, $\eta^2_G = 0.03$) and an interaction between foil type and test condition ($F(1,58) = 7.57$, $p = 0.007$, $\eta^2_G < 0.01$), motivating us to separate subsequent comparisons according to foil type.

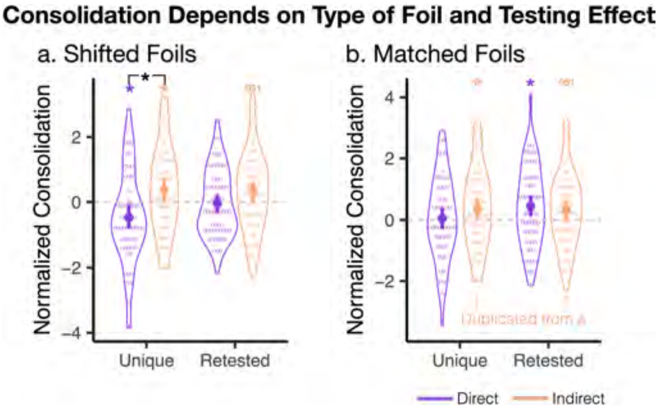

Fig. 6. For a) and b), dprime (for direct) and response speeding (for indirect) were z-scored to be placed on the same scale. The difference between each participant's immediate and delayed performance (first compared to unique, first compared to retested) was then calculated to create a normalized consolidation score, the y-axis in both a) and b). a) For shifted foils, consolidation of assessment type (direct vs. indirect) differed ($p = 0.003$) and assessment type interacted marginally with retesting ($p = 0.07$), with reliable test type differences within uniquely tested material ($p < 0.001$), but not repeatedly tested material ($p = 0.11$). b) For position foils, assessment type (direct vs. indirect) interacted marginally with retesting ($p = 0.08$, $\eta^2_p = 0.01$), but with no reliable differences in consolidation observed in either repeated or uniquely tested conditions ($ps > 0.17$). Note that a) and b) display the same indirect test consolidation scores for comparison, as indicated in text on panel b. Solid dots indicate the group mean, lines extending from the dots reflect bootstrapped confidence interval around the mean, open dots display individual participant consolidation scores, colored stars indicate a difference from chance (one-sample *t*-test; chance = 0).

Beginning with shifted foil discrimination on the direct test, an ANOVA showed weaker consolidation on this direct measure than on the indirect test (main effect of assessment type: $F(1,58) = 9.98$, $p = 0.003$, $\eta^2_G = 0.05$; Fig. 6a), with a marginal interaction between type of assessment and retesting ($F(1,58) = 3.52$, $p = 0.07$, $\eta^2_G = 0.01$; Fig. 6a). This interaction was driven by significantly stronger consolidation of uniquely tested material in the indirect as compared to direct test (*t(58)* = 3.58, $p < 0.001$; Fig. 6a), but not repeatedly tested material (*t(58)* = 1.60, $p = 0.11$; Fig. 6a). Conversely, when comparing indirect test performance to position foil discrimination, there was no effect of assessment type (direct vs. indirect; $F(1,58) = 0.23$, $p = 0.63$, $\eta^2_G = 0.001$; Fig. 6b), suggesting that more coarse explicit memories are retained (and strengthened with retesting) to a similar degree as implicit statistical knowledge. There was only a marginal interaction ($F(1,58) = 3.24$ $p = 0.08$, $\eta^2_p = 0.01$; Fig. 6b), with no reliable differences in consolidation observed in either repeated or uniquely tested conditions ($ps > 0.17$).

Finally, in support of the idea that a 24-h consolidation period drove the increase in consolidation score between the first and second indirect test, the indirect consolidation scores for subjects in our Indirect Control Group were no different than chance (uniquely tested: $t(54) = 0.63$, $p = 0.53$, $\mathrm{BF}_{01} = 5.61$; retested: $t(54) = 1.17$, $p = 0.24$, $\mathrm{BF}_{01} = 3.54$). Thus, without repeated testing, a 24-h delay has divergent effects on direct and indirect assessments of statistical knowledge, but only when direct tests require the fine-grained discrimination of transitional probabilities.

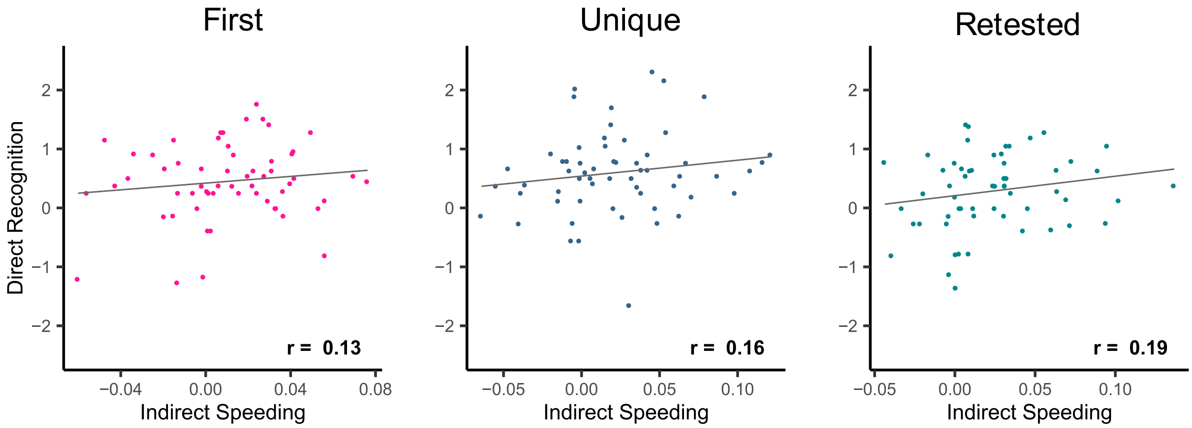

Fig. 5. Performance on direct and indirect tests was uncorrelated for the first ($r = 0.13$; pink), unique ($r = 0.19$; teal), and retested ($r = 0.16$; blue) conditions. In all correlation plots, dots depict individual participants, the line reflects the linear regression line and greyed area reflects 95% confidence interval around the regression line. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.5. Accounting for testing order

We chose to consistently perform the direct test before the indirect test in both sessions to minimize the influence of new learning. Specifically, the indirect test exposed participants to snippets of the original language, generating ample opportunity for relearning (Batterink et al., 2017); relearning that we could account for in our statistical modeling of the indirect test, but not factor out of performance on the direct test. In addition, the direct test contained twice as many foils as it did intact words and, thus, supplied a statistical structure that was markedly different from the original language. Combining across language A test items, transitional probabilities between the first and second syllables of true words dropped to 0.333 and to 0.667 between the second and third syllables. Analogously, the backward shifting of foils meant that language B participants experienced a transitional probability of 0.667 between the first and second syllables of true words across test items, and a 0.333 transitional probability between the second and third. Foils also introduced numerous non-word transitional probabilities of 0.333. Relearning this distorted transition structure would not optimally prepare participants to display the clear position-based speeding observed in the indirect test.

Even so, it is possible that learning during the direct test could have partially contributed to RT speeding on the indirect test. We therefore leveraged language differences to determine if this was the case and the degree to which it could have contributed to the robust demonstration of learning on the delayed indirect test. First, we determined that participants who were exposed to the different languages showed similar degrees of RT speeding ($\beta = -0.008$, $t(57) = 2.68$, $p = 0.13$) despite the different syllable transition distortions experienced when asked about the foils during their respective direct tests. Visualization of RT at each syllable position (Supplemental Fig. 3a), however, showed that participants tended to display steeper slopes for the syllable transitions that were more frequently experienced during their direct tests (syllable 2–3 transition for language A; syllable 1–2 transition for language B). To ensure that it was not these more frequent foil transitions that led to the delay-dependent improvements in the indirect test, we reconducted our primary analysis after dropping syllable 3 RTs from language A (leaving only the syllable 1–2 transition) and syllable 1 RTs from language B (leaving only the 2–3 syllable transition). We found that RT speeding was robustly stronger for delayed as compared to immediate tests in this subsample of trials (unique vs. first: $\beta = 0.03$, $t(93) = 3.47$, $p < 0.001$; retested vs. first: $\beta = 0.02$, $t(208) = 2.01$, $p = 0.05$; Supplemental Fig. 3b). This suggests, that while some learning may occur during the direct test, implicit representations of the original language experience are strengthened across the delay.

4. Discussion

Here, we empirically supported proposals for distinct statistical learning components which give rise to implicit and explicitly accessible knowledge (Arciuli, 2017; Batterink et al., 2019; Conway, 2020; Daltrozzo & Conway, 2014; Savalia et al., 2016). We interrogated each using respective indirect and direct assessments while manipulating retention delays alongside retesting. We dissociated explicit from implicit traces in two ways: showing how they differ in forgetting rates, and susceptibility to testing effects. Specifically, expression of implicit knowledge is *strengthened* over a consolidation period, regardless of whether it is immediately tested, while explicitly accessible precise knowledge is less durable, but clearly bolstered by initial testing. Moreover, performance on direct and indirect tests was not reliably correlated across participants. Combined, these results provide insights into *which* neurocognitive systems give rise to statistical learning.

Importantly, differential forgetting rates of implicit and explicit traces (Dosher & Rosedale, 1991) and test-retest (Roediger & Karpicke, 2006; Rubin & Wenzel, 1996; Wixted, 2004) are well-known effects in the broader learning and memory literatures that allow us to make important insights about statistical learning. Taking these in turn, the observed pattern of forgetting makes it implausible that direct and indirect performance reflected the same underlying memory representation. It is always possible that different measures—reaction time in a target detection task and old/new judgements in our experiment—simply reflect different ways of accessing the same learned representation. And, if we were to have simply observed greater forgetting in the direct as compared to indirect assessment, it may very well have been the case that weakened memories were harder to access through a deliberate old/new judgement than in a target detection task. But differences in access cannot explain why the consolidation of a single memory trace would result in stronger performance in the indirect assessment but weaker performance on test items that require precise representations in the direct assessment. Moreover, the words that resulted in the numerically strongest indirect test performance—those that were uniquely tested after a delay—were the very same words that resulted in the weakest performance on the direct test. With these opposing patterns of performance across testing conditions, the most parsimonious interpretation is that statistical learning results in dissociable memory representations.

What's more, these different representations appear to prioritize different aspects of the statistical structure. In particular, our use of two foil types also provided new insights into what is lost and retained in explicit representations. We found the sharpest decline in participants' sensitivity to (uniquely tested) detailed syllable transition structure, such that foils with some maintained transitions appear every bit as 'word-like' as fully intact words after a 24-h delay. By contrast, participants were less likely to perceive position matched foils as being words after a delay, suggesting that it is indeed the detailed transition information that slips from conscious access over time. Intriguingly, discriminating whole-words from shifted (but not position) foils parallels the indirect assessment: speeding in this task also depends on fine-grained differences in syllable transition, with participants faster to detect syllables following deterministic within-word transitions compared to lower probability word boundaries (1 vs. 0.2 probabilities). Yet, when assessed in this indirect procedure, participants demonstrated *stronger* learning across the delay, supported either by a slowing of responses to position 1 syllables or a speeding of later position syllables. Regardless, this further supports the conclusion that time may have the opposite impact on precise implicit and explicit traces.

Our observation that explicit representations become more general over time fits with work showing that consolidation during sleep is important for generalization and gist abstraction (Ellenbogen, Hu, Payne, Titone, & Walker, 2007). Although we measured neither sleep nor generalization, the reduced precision we observed may be useful for generalization. For example, extracting language-specific phoneme transition rules may require abstracting over individual words to endorse any allowable transition. From this perspective, it may be surprising that implicit traces remained so sensitive to gradations in transition structure.

But *why* do these explicit representations of syllable transitions become fuzzier over time? One intriguing possibility is that low frequency transitions become strengthened over the delay. Recall that the shifted foils contain one high frequency transition (TP = 1) and one low frequency transition (TP = 0.2). Modeling work suggests that certain hippocampal subfields represent low frequency sequential contingencies as strongly as high frequency ones (Schapiro, Turk-Browne, Botvinick, & Norman, 2017), so it is possible that known hippocampal replay mechanisms (Wilson & McNaughton, 1993) reinforce the lower frequency transitions after learning, boosting them to the higher frequency status. Alternatively, it may be that the single intact transition is what makes shifted foils appear more word-like over the delay. Future work using partially intact foils could distinguish these possibilities.

It is crucial here that we highlight, as others have (Batterink et al., 2019; Conway, 2020; Kalra, Gabrieli, & Finn, 2019; Keele, Ivry, Mayr, Hazeltine, & Heuer, 2003), the important distinction between consciously accessing a representation (that is, explicit knowledge) and whether learning takes place with an explicit intention to learn. Much work on statistical learning has debated whether the learning process itself requires attention (Finn, Lee, Kraus, & Hudson Kam, 2014; Musz, Weber, & Thompson-Schill, 2015; Pacton & Perruchet, 2008; Toro, Sinnett, & Soto-Faraco, 2005). Fewer studies address our question of whether learned representations are consciously accessible (Batterink et al., 2015; Bays et al., 2016; Turk-Browne et al., 2005). Still, the learning orientation and resultant memory representations may be related. Indeed, it has been shown that instructions to learn intentionally during statistical learning results in greater explicit knowledge (Bertels, Destrebecqz, & Franco, 2015). It has moreover been proposed that attention during statistical learning—while not required generally speaking—may be necessary for learning certain types of information (de Diego-Balaguer, Martinez-Alvarez, & Pons, 2016; Walk & Conway, 2016) and may result in more explicit representations. In keeping with prior work (Finn & Hudson Kam, 2008; Finn & Hudson Kam, 2015; Forest et al., 2019; Saffran et al., 1997), we kept the attentional demands during encoding quite low: participants were asked to listen, but not analyze the auditory stream while completing a puzzle. Future work manipulating attentional demands or learning instructions during encoding combined with the dependent measures used here over a delay would be most useful in better understanding the relationship between attentional processes during encoding and the resultant memory representations.

What's more, the dissociations reported here certainly suggest that multiple neurobiological mechanisms underlie statistical learning, but future work is required to determine how they map onto canonical memory systems (Squire, 2004). A suite of brain regions, including the frontal cortex, basal ganglia, hippocampus, and sensory regions have been implicated in statistical learning (Finn et al., 2019; Karuza et al., 2013; Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013; Schlichting, Guarino, Schapiro, Turk-Browne, & Preston, 2016), but how this circuitry maps onto different representations remains a question. Hippocampal contributions to implicit statistical learning (Schapiro, Gregory, Landau, McCloskey, & Turk-Browne, 2014; Schapiro, Turk-Browne, Norman, & Botvinick, 2015) exemplify the challenges in this pursuit; despite its classical links to the process of explicitly recalling a memory (Squire, 2004), good evidence now implicates this region in implicit forms of memory, when the implicit memory has a relational or associative structure (Cowell, Barense, & Sadil, 2019; Davachi, 2006; Henke, 2010; Shohamy & Turk-Browne, 2013). Thus, the contributions of the hippocampal memory system to statistical memory may be best understood through its contributions to relational processing, rather than the creation of consciously accessible memories.

Relatedly, there may be alternate ways to conceptualize the unique contributions of the neurobiological mechanisms which give rise to the forms of statistical learning representations dissociated here. For example, in addition to proposals that statistical learning systems can be differentiated in terms of their conscious access (Batterink et al., 2015), it has also been argued they might be more parsimoniously characterized by the cognitive demands required by learning (Conway, 2020). In particular, is has been suggested that one automatically acquires basic associations while the other relies on attention or working memory to learn more complex structures (especially non-adjacent structures) and bridge across modalities (Conway, 2020; Daltrozzo & Conway, 2014; Walk & Conway, 2016). It could be that the consciously accessible representation we are documenting here arises from this attentionally mediated system. However, this need not be the case since conscious accessibility can also arise from greater exposure alone (Keele et al., 2003). Thus, we think it is important to note here that explicit conscious accessibility is just one property that can distinguish the forms of statistical learning, but we find that this is nevertheless a diagnostic feature that is helpful in identifying dissociable representations that result from statistical learning. We hope that our findings inspire future work which could leverage similar dissociations in order to probe the relationships between processes during learning and the resultant representations.

Future work notwithstanding, the implications of our results are profound. We conclusively show that statistical learning results in multiple representations: one that is consciously accessible, malleable to later experience, and that becomes increasingly abstract; and another that shapes behaviour outside of conscious awareness, but retains precise details, and becomes stronger over time. Following this discovery are new pressing questions about why these multiple mechanisms are present, how they trade-off across learners, and how they contribute to learning different kinds of structure. Intriguingly, this discovery informs existing work on the development of statistical learning (Arciuli & Simpson, 2011; Finn et al., 2019; Janacsek, Fiser, & Nemeth, 2012; Raviv & Arnon, 2018; Schlichting et al., 2016; Shufaniya & Arnon, 2018), suggesting that statistical learning outcomes are likely to change across childhood *because* children's abilities to construct explicit traces grow while their ability to form implicit are present very early in life (Amso & Davidow, 2012; Finn et al., 2016). Thus, children could retain mostly implicit—but specific and less malleable—traces as compared with adults who retain both, thereby leaving a structured experience with only (or mostly) veridical, stable, but relatively inaccessible knowledge. Further work with children, or that manipulates the availability of certain neurocognitive systems in adults, could provide a unique window for understanding *why* both representations emerge from statistical learning and what behaviors they support.

Author contributions

H.L. K·D and A.S·F. contributed to the study design. Testing and data collection were performed by H.L. and T.A.F., K·D, T.A.F. and H.L performed the data analysis with input from A.S.F. H.L, A.S.F and T.A.F drafted the manuscript, and K·D provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

We thank Hannah Tarder-Stoll for her vocalization of stimuli; We thank Michaela Karam, Wujiamei Sun, Akshita Srivastava and Mishel Alexandrovsky for the help with data collection; And we thank the following funding sources: Natural Sciences and Engineering Research Council of Canada Discovery Grants (ASF:RGPIN-2016-05; KD:RGPIN/05582-201) and Canada Foundation for Innovation and Ontario Research Funds (ASF: 34947; KD:34479).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105439.

References

Amso, D., & Davidow, J. (2012). The development of implicit learning from infancy to adulthood: Item frequencies, relations, and cognitive flexibility. *Developmental Psychobiology, 54*, 664–673.

Arciuli, J. (2017). The multi-component nature of statistical learning. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 372*, 20160058.

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: The role of age and speed of stimulus presentation. *Developmental Science, 14*, 464–473.

Arciuli, J., & Simpson, I. C. (2012). Statistical learning is lasting and consistent over time. *Neuroscience Letters, 517*, 133–135.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 1*.

Batterink, L. J. (2017). Rapid statistical learning supporting word extraction from continuous speech. *Psychological Science, 28*, 921–928.

Batterink, L. J., Paller, K. A., & Reber, P. J. (2019). Understanding the neural bases of implicit and statistical learning. *Topics in Cognitive Science, 11*, 482–503.

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Mempry and Language, 83*, 62–78.

Bays, B. C., Turk-Browne, N. B., & Seitz, A. R. (2016). Dissociable behavioural outcomes of visual statistical learning. *Visual Cognition, 23*, 1072–1097.

Bertels, J., Destrebecqz, A., & Franco, A. (2015). Interacting effects of instructions and presentation rate on visual statistical learning. *Frontiers in Psychology, 6*.

Buchsbaum, D., Griffiths, T. L., Plunkett, D., Gopnik, A., & Baldwin, D. (2015). Inferring action structure and causal relationships in continuous sequences of human action. *Cognitive Psychology, 76*, 30–77.

Cleeremans, A. (2006). Conscious and unconscious cognition: A graded, dynamic perspective. In M. R. Jing, G. Rosenzweig, H. d'Ydewalle, H.-C. Zhang, Chen, & K. Zhang (Eds.), *Progress in psychological science around the world. I. Neural, cognitive, and developmental issues* (pp. 401–418). Hove, England: Psychology Press.

Conway, C. M. (2020). How does the brain learn environmental structure? Ten core principles for understanding the neurocognitive mechanisms of statistical learning. *Neuroscience & Biobehavioral Reviews, 112*, 279–299.

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. Tutorials in Quantitative Methods for. *Psychology, 1*, 42–45.

Cowell, R. A., Barense, M. D., & Sadil, P. S. (2019). A roadmap for understanding memory: Decomposing Cognitive Processes into Operations and Representations. *eneuro, 6*. ENEURO.0122-0119.2019.

Daltrozzo, J., & Conway, C. (2014). Neurocognitive mechanisms of statistical-sequential learning: What do event-related potentials tell us? *Frontiers in Human Neuroscience, 8*.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology, 16*, 693–700.

de Diego-Balaguer, R., Martinez-Alvarez, A., & Pons, F. (2016). Temporal attention as a scaffold for language development. *Frontiers in Psychology, 7*.

Dosher, B. A., & Rosedale, G. (1991). Judgments of semantic and episodic relatedness: Common time-course and failure of segregation. *Journal of Memory and Language, 30*, 125–160.

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia, 49*, 1322–1331.

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences, 104*, 7723–7728.

Endress, A. D., & Mehler, J. (2009). The surprising power of statistical learning: When fragment knowledge leads to false memories of unheard words. *Journal of Memory and Language, 60*, 351–367.

Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition, 108*, 477–499.

Finn, A. S., & Hudson Kam, C. L. (2015). Why segmentation matters: Experience-driven segmentation errors impair "morpheme" learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 41*, 1560–1569.

Finn, A. S., Kalra, P. B., Goetz, C., Leonard, J. A., Sheridan, M. A., & Gabrieli, J. D. E. (2016). Developmental dissociation between the maturation of procedural memory and declarative memory. *Journal of Experimental Child Psychology, 142*, 212–220.

Finn, A. S., Kharitonova, M., Holtby, N., & Sheridan, M. A. (2019). Prefrontal and hippocampal structure predict statistical learning ability in early childhood. *Journal of Cognitive Neuroscience, 31*, 126–137.

Finn, A. S., Lee, T., Kraus, A., & Hudson Kam, C. L. (2014). When it hurts (and helps) to try: The role of effort in language learning. *PLoS One, 9*, Article e101806.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12*, 499–504.

Forest, T. A., Finn, A. S., & Schlichting, M. L. (2021). *First general then specific memory representations emerge from structured experience*. JEP:General.

Forest, T. A., Lichtenfeld, A., Alvarez, B., & Finn, A. S. (2019). Superior learning in synesthetes: Consistent grapheme-color associations facilitate statistical learning. *Cognition, 186*, 72–81.

Galea, J. M., Albert, N. B., Ditye, T., & Miall, R. C. (2009). Disruption of the dorsolateral prefrontal cortex facilitates the consolidation of procedural skills. *Journal of Cognitive Neuroscience, 22*, 1158–1164.

Goshen-Gottstein, Y., & Kempinsky, H. (2001). Probing memory with conceptual cues at multiple retention intervals: A comparison of forgetting rates on implicit and explicit tests. *Psychonomic Bulletin & Review, 8*, 139–146.

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science, 18*, 254–260.

Graf, P., Squire, L. R., & Mandler, G. (1984). The information that amnesic patients do not forget. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 164–178.

Henke, K. (2010). A model for memory systems based on processing modes rather than consciousness. *Nature Reviews Neuroscience, 11*, 523–532.

Janacsek, K., Fiser, J., & Nemeth, D. (2012). The best time to acquire new skills: Age-related differences in implicit sequence learning across the human lifespan. *Developmental Science, 15*, 496–505.

JASP Team. (2020). *JASP (Version 0.14)*.

Kalra, P. B., Gabrieli, J. D. E., & Finn, A. S. (2019). Evidence of stable individual differences in implicit learning. *Cognition, 190*, 199–211.

Karuza, E. A., Newport, E. L., Aslin, R. N., Starling, S. J., Tivarus, M. E., & Bavelier, D. (2013). The neural correlates of statistical learning in a word segmentation task: An fMRI study. *Brain and Language, 127*, 46–54.

Keele, S. W., Ivry, R., Mayr, U., Hazeltine, E., & Heuer, H. (2003). The cognitive and neural architecture of sequence representation. *Psychological Review, 110*, 316–339.

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: Is it long-term and implicit? *Neuroscience Letters, 461*, 145–149.

Kóbor, A., Janacsek, K., Takács, Á., & Nemeth, D. (2017). Statistical learning leads to persistent memory: Evidence for one-year consolidation. *Scientific Reports, 7*, 760.

Musz, E., Weber, M. J., & Thompson-Schill, S. L. (2015). Visual statistical learning is not reliably modulated by selective attention to isolated events. *Attention, Perception, & Psychophysics, 77*, 78–96.

Pacton, S. B., & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 80–96.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Science, 10*, 233–238.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. from http://www.r-project.org/.

Rappold, V. A., & Hashtroudi, S. (1991). Does organization improve priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 103–114.

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science, 21*, Article e12593.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning:taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

Romano, J. C., Howard, J. H., & Howard, D. V. (2010). One-year retention of general and sequence-specific skills in a probabilistic, serial reaction time task. *Memory, 18*, 427–441.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*, 734–760.

Saffran, J. R., Aslin, R. N., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926–1928.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*, 101–105.

Sanchez, D. J., Gobel, E. W., & Reber, P. J. (2010). Performing the unexplainable: Implicit task performance reveals individually reliable sequence learning without explicit knowledge. *Psychonomic Bulletin & Review, 17*, 790–796.

Savalia, T., Shukla, A., & Bapi, R. S. (2016). A unified theoretical framework for cognitive sequencing. *Frontiers in Psychology, 7*.

Schapiro, A. C., Gregory, E., Landau, B., McCloskey, M., & Turk-Browne, N. B. (2014). The necessity of the medial temporal lobe for statistical learning. *Journal of Cognitive Neuroscience, 26*, 1736–1747.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience, 16*, 486–492.

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). *Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning* (p. 372). Philosophical Transactions of the Royal Society B: Biological Sciences.

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2015). Statistical learning of temporal community structure in the hippocampus. *Hippocampus, 26*, 3–8.

Schlichting, M. L., Guarino, K. F., Schapiro, A. C., Turk-Browne, N. B., & Preston, A. R. (2016). Hippocampal structure predicts statistical learning and associative inference abilities during development. *Journal of Cognitive Neuroscience, 29*, 37–51.

Shohamy, D., & Turk-Browne, N. B. (2013). Mechanisms for widespread hippocampal involvement in cognition. *Journal of Experimental Psychology: General, 142*, 1159–1170.

Shufaniya, A., & Arnon, I. (2018). Statistical learning is not age-invariant during childhood: Performance improves with age across modality. *Cognitive Science, 42*, 3100–3115.

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*, 692–727.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory, 82*, 171–177.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31*, 137–149.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition, 97*, B25–B34.

Turk-Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology, 134*, 552–564.

Voss, J. L., Baym, C. L., & Paller, K. A. (2008). Accurate forced-choice recognition without awareness of memory retrieval. *Learning & Memory, 15*, 454–459.

Walk, A. M., & Conway, C. M. (2016). Cross-domain statistical–Sequential dependencies are difficult to learn. *Frontiers in Psychology, 7*.

Willingham, D. B., & Dumas, J. A. (1997). Long-term retention of a motor skill: Implicit sequence knowledge is not retained after a one-year delay. *Psychological Research, 60*, 113–119.

Wilson, M., & McNaughton, B. (1993). Dynamics of the hippocampal ensemble code for space. *Science, 261*, 1055–1058.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55*, 235–269.