Silva, S., Inácio, F., Folia, V., & Petersson, K.M. (2017). Eye-movements in Implicit Artificial Grammar Learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49* (3), 1387-1402. DOI: 10.1037/xlm0000350.

Eye-movements in Implicit Artificial Grammar Learning

Susana Silva [1, 4], Filomena Inácio [1], Vasiliki Folia [2, 3], & Karl Magnus Petersson [1, 2, 3]

[1] Cognitive Neuroscience Research Group, Centre for Biomedical Research (CBMR), Universidade do Algarve, Portugal

[2] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

[3] Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, the Netherlands

[4] Neurocognition and Language Group, Center for Psychology of the University of Porto, Portugal

Corresponding author

Susana Silva

Center for Psychology at the University of Porto, Porto, Portugal

E: zanasilva@gmail.com

ABSTRACT

Artificial grammar learning (AGL) has been probed with forced-choice behavioral tests (active tests). Recent attempts to probe the outcomes of learning (implicitly acquired knowledge) with eye-movement responses (passive tests) have shown null results. However, these latter studies have not tested for sensitivity effects, for example, increased eye movements on a printed violation. In this study, we tested for sensitivity effects in AGL tests with (Experiment 1) and without (Experiment 2) concurrent active tests (preference- and

grammaticality classification) in an eye-tracking experiment. Eye movements discriminated between sequence types in passive tests and more so in active tests. The eye-movement profile did not differ between preference and grammaticality classification, and it resembled sensitivity effects commonly observed in natural syntax processing. Our findings show that the outcomes of implicit structured sequence learning can be characterized in eye tracking. More specifically, whole trial measures (dwell time, number of fixations) showed robust AGL effects, whereas first-pass measures (first-fixation duration) did not. Furthermore, our findings strengthen the link between artificial and natural syntax processing, and they shed light on the factors that determine performance differences in preference and grammaticality classification tests.

Keywords: Eye-tracking, implicit learning, artificial grammar learning, syntactic processing, preference classification

INTRODUCTION

The artificial grammar learning (AGL) paradigm probes implicit sequence learning (Forkstam & Petersson, 2005; Reber, 1967; Seger, 1994; Stadler & Frensch, 1998; van den Bos & Poletiek, 2008) and models aspects of the acquisition of structural knowledge such as linguistic syntax (Christiansen, Conway, & Onnis, 2012; Christiansen, Louise Kelly, Shillcock, & Greenfield, 2010; Conway, Karpicke, & Pisoni, 2007; Lelekov-Boissard & Dominey, 2002; Silva, Folia, Hagoort, & Petersson, 2016; Tabullo, Sevilla, Segura, Zanutto, & Wainselboim, 2013; Zimmerer, Cowell, & Varley, 2014). The paradigm involves exposure and test phases. In the *exposure* phase, participants are given positive examples of a grammar, often letter sequences. In implicit versions of AGL, participants are kept unaware that the sequences are constructed according to rules (Figure 1) and may thus be referred to as *grammatical* sequences. In the *test* phase, novel grammatical sequences are presented together with sequences containing at least one violation of grammar rules (i.e., *non-grammatical sequences*). Participants are asked to make grammaticality judgments under forced-choice conditions, and any implicitly acquired knowledge is inferred from the accuracy of those judgments—that is, from *behavioral discrimination* between grammatical and non-grammatical sequences.

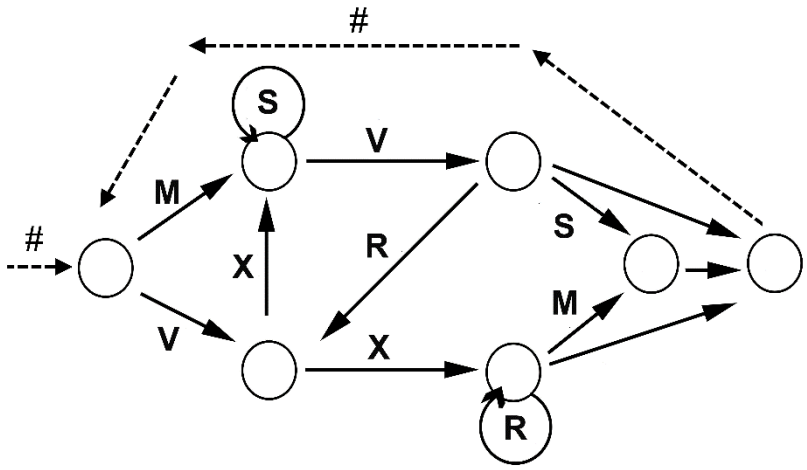

FIGURE 1. The artificial grammar used in this study. Grammatical sequences are generated by traversing the transition graph along the indicated directions (e.g., MSVRXVS). An example of a non-grammatical counterpart would be MSXRXVS, with X being the violating target letter and V a legal target letter.

The importance of keeping participants unaware of the learning targets has generated some discussion on grammaticality judgment tasks because the test-instructions highlight the existence of rules and might therefore lead to explicit processing (Buchner, 1994; Manza & Bornstein, 1995). Indirect accuracy-free judgments, such as preference classification (like/dislike), have been proposed as an alternative (Forkstam, Elwér, Ingvar, & Petersson, 2008; Gordon & Holyoak, 1983; Manza & Bornstein, 1995), with the advantage of allowing for a baseline (pre-exposure) measure of accuracy underlying a proper-learning design (Petersson, Elfgren, & Ingvar, 1999b, 1999a). Even though preference judgments are sensitive (Folia et al., 2008; Folia & Petersson, 2014; Forkstam et al., 2008; Silva et al., 2016; Uddén, Ingvar, Hagoort, & Petersson, 2012), an involuntary index of learning would be even more akin to the implicit character of the process, and it would afford expanding AGL research to populations such as infants and animals. Eye movements are not always involuntary (Hayhoe & Ballard, 2011), but the probability of being so is high, in the context of viewing AGL test sequences. In addition, eye-tracking measures reflect acquired knowledge when learning is implicit (Giesbrecht, Sy, & Guerin, 2013; Jiang, Won, & Swallow, 2014, but see Coomans, Deroost, Vandenbossche, Van Den Bussche, & Soetens, 2012, for the potential role of covert attention). In this study, we investigate the suitability of eye-tracking measures in characterizing the outcomes of AGL (implicitly acquired knowledge), focusing on the possibility that some form of ocular discrimination of sequence types parallels the behavioral discrimination that is observed in successful implicit AGL.

Eye-tracking measures have been extensively used in spatial implicit learning, where space is the learning target. Paradigms measuring the anticipation of the spatial position of a target have relied on saccade latency (Amso & Davidow, 2012) and saccade length (Jiang et al., 2014). Visual search paradigms relating to contextual cuing effects (implicit learning of spatial context) have measured the number of saccades (Hout & Goldinger, 2012) or fixations (Manelis, & Reder, 2012) required to scan a scene before the target is found. Scan-path measures, defining the exploration overlap of scenes, have been used to index implicit memory (Ryals, Wang, Polnaszek, & Voss, 2015).

In AGL there are no spatial targets and different approaches are required. To our knowledge, only three studies have probed the outcomes of AGL with eye-tracking methodologies. Heaver (2012) tested participants for pupillary responses to grammatical and non-grammatical sequences at the test phase, and found no discrimination of sequence types based on pupil size. Wilson and colleagues (Wilson et al., 2013; Wilson, Smith, & Petkov, 2015) delivered auditory stimuli through speakers and analyzed the time participants gazed at the speaker area as a function of the grammatical status of the sequence. The paradigm worked for primates (Wilson et al., 2013, 2015), who showed longer gaze times for non-grammatical sequences, but it did not show any effects in humans (Wilson et al., 2015). However, a behavioral forced-choice (grammaticality classification) did work in humans, and it was suggested that this might be due to increased levels of attention in the active (forced-choice) compared with the passive (eye-tracking only) task. A slightly different, yet related explanation for why eye-tracking measures alone might fail to capture AG knowledge relates to the processes that may or may not be recruited depending on the behavioral task (e.g., Leeser, Brandl, & Weissglass, 2011). Given that AGL involves syntax-like processing (e.g., Christiansen et al., 2010) - and hence a focus on dependencies between sequence elements—the required type of analysis may not be recruited unless there is an active and suitably syntax-oriented task. The results on implicit AGL with preference classification— apparently a nonsyntax-oriented task—contribute to argue against this possibility (Folia et al., 2008; Folia & Petersson, 2014; Forkstam et al., 2008; Silva et al., 2016; Uddén et al., 2012), but it may nevertheless be considered.

Whether passive tests fail in facilitating attention in general, or in eliciting syntactic analysis in particular, one may expect that the eye-tracking signatures of AGL resemble the so-called *sensitivity effects*. Sensitivity effects have been described in the literature on natural syntax processing, and they refer to the fact that readers fixate longer or regress more frequently from a violating word compared with its syntactically correct counterpart (Godfroid et al., 2015; Keating, 2009; Lim & Christianson, 2014; Sagarra & Ellis, 2013). The reason why

sensitivity effects may be expected is not that AGL materials resemble written words: AGL sequences are meaningless and unpronounceable, and they are presented one at a time, so interword regressions do not exist. Instead, sensitivity effects may be expected on the grounds that AGL models the acquisition and the processing of natural syntax (Christiansen et al., 2012, 2010; Conway et al., 2007; Lelekov-Boissard & Dominey, 2002; Silva et al., 2016; Tabullo et al., 2013; Zimmerer et al., 2014), and so the processing of dependencies among sequence items (letters, in this case) is likely to mirror the processing of dependencies among words (sentence subunits) in natural language. Moreover, sensitivity effects have been obtained in natural language without readers being specifically asked to do syntactic judgments, so it is possible that they emerge in passive eye-tracking tests, when no additional task is requested. However, natural language is different in one fundamental aspect. Unlike AGL stimuli, natural language sentences have both lexical and sentence-level meaning. The presence of semantic content may be sufficient to increase the levels of attention or to drive syntactic analysis. From this viewpoint, it is less certain that sensitivity effects emerge in AGL, which is semantic-free. As already noted above, Wilson and colleagues (2015) suggested that AGL effects do not show up in eye-tracking measures. However, Wilson and colleagues (2015) did not probe sensitivity effects (increased eye movements on the target letter or event, the one violating the grammar) and so the possibility of observing sensitivity effects in implicit AGL remains untested.

The first objective of our study was to test for sensitivity effects in a proper-learning implicit AGL paradigm (pretest-posttest design, with pre-exposure and post-exposure measures of knowledge) with and without a concurrent forced-choice, active test. In the first experiment (see Table 1), we used active tests and participants were also tested in a baseline (pre-exposure) preference classification task. We compared this with a final (post-exposure) preference classification as well as with a grammaticality classification test. In Experiment 2, we started with passive tests and added a final active test (grammaticality classification) for within-subject comparisons. We predicted that sensitivity effects would be weaker with passive, eye-tracking only tests (Experiment 2) than with active ones (Experiment 1), and that the introduction of an active test would boost ocular discrimination in Experiment 2. An issue of interest was the comparison between ocular discrimination in final preference versus grammaticality classification in Experiment 1. Several AGL studies have shown quantitative differences in behavioral performance for final preference versus grammaticality classification (Folia et al., 2008; Folia & Petersson, 2014; Forkstam et al., 2008; Silva et al., 2016; Uddén et al., 2012). Behavioral tests completely depend on offline (final) decision processes, which are highly susceptible to the self-monitoring of performance (e.g., 'Should I say I like it?' in preference, vs.

'Should I say it is correct?' in grammaticality). Differences between preference and grammaticality decisions concerning the processes engaged may be responsible for the quantitative differences observed so far in behavioral tests. In contrast, eye-tracking measures are online measures that capture the whole judgment process. This may include final decision processes and influences of self-monitoring, but it also includes the whole processing time before a specific response is planned, making eye-tracking measures less susceptible to decision-related influences than behavioral ones. Thus, if differences between preference and grammaticality classification show up in behavioral tests but not in concurrent eye-tracking measures, this would suggest that final decision processes are critically involved in behavioral differences.

Table 1

Design of the Two Experiments

Phase	Day 1	Day 2	Day 3	Day 4	Day 5
Experiment 1					
Exposure (G)	Yes	Yes	Yes	Yes	Yes
Active test (G-NG)	**Baseline preference**				**Final preference Grammaticality**
Experiment 2					
Exposure (G)	Yes	Yes	Yes	Yes	Yes
Passive test (G-NG)	**Passive baseline Passive Test 1**[a]	**Passive Test 2**	**Passive Test 3**	**Passive Test 4**	**Passive Test 5**
Active test (G-NG)					**Grammaticality**

Note. G and NG refer to sequence types (G = grammatical; NG = non-grammatical). Text in bold indicates eye-tracking recordings.

[a] Passive 1 was run after exposure on Day 1.

The second objective of this study was to determine the type of sensitivity effect associated with implicitly acquired knowledge. Despite claims that there is no one-to-one mapping between eye movements and awareness (Godfroid & Schmidtke, 2013) and that triangulation with verbal data is required to determine whether learning was implicit or not (Godfroid & Winke, 2015), it has been proposed that regressions (movements from right to left)

are associated with explicit knowledge (Godfroid et al., 2015). This claim was based on the assumptions that regressions are controlled processes (Reichle, Warren, & McConnell, 2009), and that implicit knowledge is accessed by automatic rather than controlled processing. In our study, we tested for the more general concept of *second-pass reading*, including regressions (right to left movements) as well as progressions (left to right) to the violating (target) letter after the first-fixation on it. For this reason, we used measures related to whole-trial time (dwell time, number of fixations), considering first-pass (first-fixation duration) and second pass measures (dwell-to-first-fixation ratio) separately.

In the two experiments, we controlled for the effects of local subsequence familiarity, measured as associative chunk strength (ACS, Knowlton & Squire, 1996; Meulemans & Linden, 1997), to rule out the possibility that learning is based on overt, surface features of the sequences (Shanks & John, 1994) instead of structural features of the underlying grammar (Folia et al., 2008; Folia & Petersson, 2014; Forkstam et al., 2008; Silva et al., 2016; Uddén et al., 2012). As in our previous studies, we used a multiday paradigm to allow abstraction and consolidation processes to take place (e.g., Nieuwenhuis, Folia, Forkstam, Jensen, & Petersson, 2013).

EXPERIMENT 1: EYE MOVEMENTS IN ACTIVE TESTS

In the first experiment, we tested whether eye movements concurrent with active, forced-choice classification tests reveal artificial grammar learning (AGL). We used a proper-learning paradigm (Folia et al., 2008; Folia & Petersson, 2014; Petersson et al., 1999b, 1999a), where the focus is on changes in discrimination between sequence types (grammatical vs. non-grammatical) after exposure.

METHOD

PARTICIPANTS

Thirty-three healthy adults with normal or corrected-to-normal vision volunteered to take part in the experiment. Due to excessive eye-tracking artifacts, three participants were excluded from further analysis. From the remaining 30 participants, 13 were female (M age ± SD = 26 ± 5). All participants were prescreened for medication use, history of drug abuse, head trauma, neurological or psychiatric illness, and family history of neurological or psychiatric

illness. Written informed consent was obtained from all according to the protocol of the Declaration of Helsinki.

STIMULUS MATERIAL

Sequences were generated from the Reber grammar represented in Figure 1 (5 to 12 consonants long, from the alphabet [M, S, V, R, X], see the Appendix 1). For a detailed description of the procedure to generate the stimulus material, see Forkstam, Hagoort, Fernandez, Ingvar, & Petersson, 2006). For the exposure phase (see Table 1), we generated one acquisition set with 100 grammatical sequences (G). To engage participants in same/different judgments (cf. Procedure section), we paired 50 of these sequences with themselves ("same") and the remaining 50 with another string from the set ("different"). We created five different pairings for presentation in each of the 5 days of exposure, using the same 50/50 proportion. For the test phase, we generated three additional classification sets, each with 60 novel grammatical (G) and 60 non-grammatical (NG) sequence pairs that were matched for associative chunk strength (ACS). In sum, each classification set consisted of 30 sequences of each sequence type: high ACS grammatical (HG), low ACS grammatical (LG), high ACS non-grammatical (HNG), and low ACS non-grammatical (LNG). HG sequences were paired with HNG, and LG with LNG, such that each pair differed in one letter, named the *target letter* (legal in G vs. violating in NG). The target letter appeared in random, nonterminal positions.

PROCEDURE

Participants were exposed to implicit acquisition sessions over 5 days (see Table 1). The sessions were constructed as short-term memory tasks of visually presented grammatical sequences. Each sequence from the 100-sequence set was presented during 4 s on a computer screen, followed by a fixation cross for 1 s. After the cross, either the same or a different sequence was presented for 4s. The participant responded whether the sequences were either the same or different, in a self-paced manner and without performance feedback. Each session lasted approximately 30 min. In the test sessions, participants performed a forced-choice classification task. On the first day, before the first acquisition session, participants classified 120 sequences according to whether they liked it or not, based on their immediate intuitive impression, or "gut feeling" (i.e., *baseline preference* classification). They did the same with novel sequences on the fifth day, after the last acquisition session (i.e., *final preference*

classification). Then we informed participants about the existence of an underlying complex set of rules generating the acquisition sequences, and they performed the third and last classification session. They classified sequences in the new set as grammatical or not (*grammaticality* classification) on the basis of their immediate intuitive impression ("gut feeling"). The three classification sets were disjoint (no overlap) and balanced across participants. Each sequence was presented for four seconds, after which the participant responded with a button press. At the end of the experimental procedure, participants filled in a questionnaire to assess potential explicit knowledge of the grammar. They were asked whether they had noticed any regularity in the stimuli. They were also asked about any technique they might have used for classification, including any combination of letters and/or the location or pattern of letters within the sequences. Finally, they were invited to generate 10 grammatical sequences.

EYE –TRACKING DATA RECORDING AND PREPROCESSING

Eye movements from test sessions were recorded with an EyeLink 1000 eye-tracking system (http://sr-research.com). Sequences were presented centrally on the computer screen, and they were preceded by fixation crosses aligned with the first (left-most) letter. The monitor, 55.8 cm wide, was placed 70 cm away from the participant. At this distance, each letter (font size 36) encompassed approximately 1° of the horizontal visual angle. Before each classification session, a five-point calibration procedure was implemented, and calibration was repeated after tracking errors larger than 0.5°. Participants placed their head on a chin rest. They were asked to stand still, relax, and blink as little as possible during sequence presentation. The raw signal was inspected, such that participants with high levels of artifacts (blinks and signal loss) were excluded from the analysis (n = 3). The analysis was based on the number and duration of events (fixations and saccades). Each letter sequence and target letter was surrounded by rectangular areas of interest, such that four target-letter-related eye-movement features would be computed: the *dwell-time proportion* (fixation and saccade times on the letter, relative to dwell time on the whole sequence), the *proportion of fixations* (number of fixations on letter relative to those on sequence), the (absolute) *duration of the first-fixation*, and the *ratio between dwell time* on the target letter and the *first-fixation* on it (*dwell/first-fixation*). The first two features provide an overall picture of the processing of the target letter. First-fixation duration indicates the first-pass response to the violation, whereas the ratio between dwell and first-fixation signals the amount of second-pass responses in relation to first-fixation duration, which may

vary across participants/trials and thus becomes normalized. We preferred this relative measure of second-pass over an absolute one because it seemed to better capture how much the participant needed to expand her/his first (variable) contact with the target. Data were inspected for outliers (±3 SD > M), and outlier trials were removed from the analysis. Null values for first-fixation duration and dwell-to-first-fixation ratio were classified as missing values (no fixation on the critical letter). The data points that entered the analysis (out of 7200 potential data points—30 participants x 120 items x 2 tests) are quantified in Tables 2 and 3.

STATISTICAL ANALYSIS

Behavioral and eye-tracking data were analyzed with linear mixed-effects models as implemented in the lme4 package (Bates, 2010; Bates, Maechler, Bolker, & Walker, 2014) for R (http://www.R-project.org/). We focused on changes in the effects of grammatical status (gram, G vs. NG) and/or ACS (high vs. low) across tests. We compared baseline preference with final preference to check for learning (increased discrimination between G and NG), and then we compared the two active tests (final preference and grammaticality). The primary interaction of interest was Test x Gram, defining grammar-based learning. Conversely, Test x ACS tested for learning based on the knowledge of surface features. The Test x Gram x ACS interaction defined the extent to which grammaticality or ACS effects depended on each other.

The full model had test (baseline preference vs. final preference or final preference vs. grammaticality), grammatical status (gram, G vs. NG), and ACS (high vs. Low) as fixed factors, together with random intercepts for participants. The model was fitted using the ML criterion so as to allow significance testing, which was achieved by comparing the full model with models without the interactions whose significance was being tested. Namely, we first tested the Test x Gram x ACS interaction by comparing the full model with a second one (Model 2, without the third-order interaction), testing for (Test x Gram) + (Test x ACS). Then we tested Test x Gram and Test x ACS by respectively comparing Model 2 with Model 3a (without Test x Gram), defined by (Test x ACS) + Gram, and Model 2 with Model 3b (without Test x ACS), defined by (Test x Gram) + ACS. Additionally, and given the large sample size, absolute *t* values larger than 2 were taken as indicators that the fixed-effects parameters were significant at the 5% level (Baayen, Davidson, & Bates, 2008). When significant, Test x Gram x ACS interactions were broken down (Test x Gram in high ACS vs. low ACS). For significant Test x Gram interactions, we ran post hoc tests of grammatical status effects on pre-exposure and post-exposure tests separately. Ideally, there should be no pre-exposure grammatical effects (no grammar knowledge), but these do

not contradict learning evidence as long as significant Test x Gram interactions exist, and this is why a proper-learning design is important. Concerning post-exposure grammatical effects, these should be observed as evidence that effective sensitivity to grammatical status resulted from exposure.

We used a similar approach to analyze behavioral data. Here, the dependent variable was the participant's endorsement rate, defining the proportion of items that were classified as grammatical (endorsed G items are correct responses, whereas endorsed NG items are incorrect). We complemented the analysis of behavioral data with estimates of accuracy and *d'* against chance levels by means of one-sample *t* tests.

Post-experimental data (questionnaires) were analyzed for indices of structural explicit knowledge: Verbal reports concerning awareness of rules were checked for consistency with the grammar (full consistence would indicate awareness), and the accuracy in generating grammatical sequences was computed (proportion of valid sequences, among the 10 sequences requested). Valid (grammatical) sequences were then analyzed one-by-one, so as to exclude generated sequences that had been presented during the acquisition or classification tasks. Our assumption was that the generation (recall) of sequences that were previously seen by participants is not a valid expression of structural knowledge because it may simply reflect participants' memory for concrete exemplars (see, e.g., Pothos, 2007). Memory for concrete exemplars is highly unlikely to account for eye-tracking sensitivity effects (response to violation letters) and is thus irrelevant for understanding our results. After excluding non-novel sequences, we were left with *generator participants* (those generating novel grammatical sequences) and *nongenerators* (generated none). Generators may be considered potential explicit learners but it may also not be the case: a small number of novel grammatical sequences may result from chunk memory (i.e., memory of frequent fragments, which may be concatenated as legal sequences by chance; see Pothos, 2007), and chunk memory is also irrelevant for understanding ocular responses to a violating letter. Still, we wanted to grant that the whole group's pattern of results did not reflect the influence of generators (*potential explicit learners*). To that end, we did a control analysis in which we considered the behavioral and eye-tracking data of nongenerators (*strict implicit learners*) separately. If nongenerators replicated the pattern of the whole group and survive the exclusion of potential explicit learners, this would be evidence that our pattern of findings reflects implicitly acquired knowledge.

RESULTS

BEHAVIORAL RESULTS

Accuracy was at chance levels in baseline preference (M = 49%), t(29) = -0.539, p > .59, and above chance levels after exposure (final preference: M = 59%, t[29] = 4.32, p < .001; grammaticality: M = 63%, t[29] = 4.85, p < .001). Discrimination between G and NG sequences (difference between endorsement rates) increased after exposure (see Figure 2), as shown by a significant Test x Gram interaction for baseline preference against final preference (see Table 2). The non-significant Test x Gram x ACS interaction indicated that increased discrimination did not depend on ACS. The Test x ACS interaction was non-significant, ruling out ACS-based learning. Comparisons between final preference and grammaticality classification showed increased discrimination in the latter (see Table 3), and again there were no significant effects involving

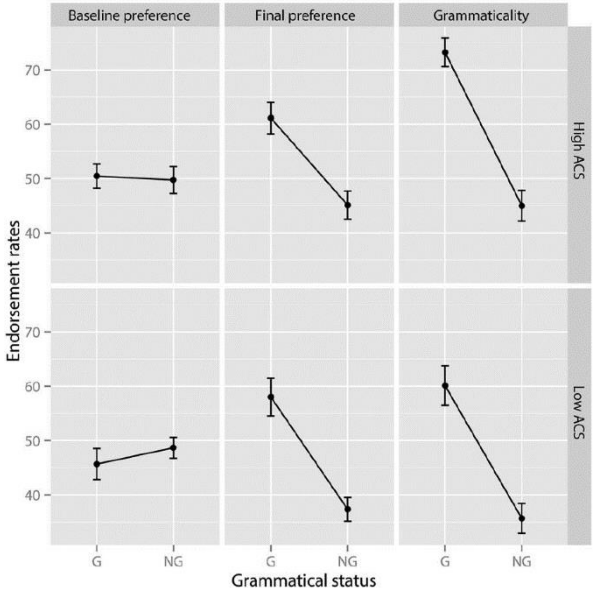

FIGURE 2. Mean endorsement rates (classification as grammatical) in Experiment 1 as a function of test, grammatical status (G = grammatical; NG = non-grammatical) and associative chunk strength (ACS). Error bars indicate the standard error of the mean.

ACS. In line with this, *d'* did not differ significantly from zero in baseline preference (M = -0.045), t(29) = -0.56, p > .57, but it did so in final preference (M = 0.544), t[29] = 3.99, p < .001, and grammaticality (M = 0.878), t(29) = 4.75, p < .001. In summary, the results showed that the exposure to grammatical examples induced the acquisition of knowledge based on grammatical status and not on ACS, entirely consistent with previous findings (Folia et al., 2008; Folia & Petersson, 2014; Forkstam et al., 2008; Silva et al., 2016; Uddén et al., 2008).

Table 2

Experiment 1: Comparison between Baseline Preference and Final Preference

Effect	Behavioral (endorsement rates)	Eye-tracking			
		First-fixation duration	Dwell time (proportion)	Fixation (proportion)	Dwell/first-fixation
Fixed effect					
Test x Gram x ACS	$X^2(2) = 1.63, p = .44$	$X^2(2) = 1.17, p = .56$	$X^2(2) = 7.46, p < .05$	$X^2(2) = 14.0, p < .001$	$X^2(2) = 0.48, p = .78$
Test x Gram	$X^2(1) = 33.4, p < .001$	$X^2(1) = 1.18, p = .28$	$X^2(1) = 18.7, p < .001$	$X^2(1) = 19.1, p < .001$	$X^2(1) = 15.8, p < .001$
Test x ACS	$X^2(1) = 0.58, p = .44$	$X^2(1) = 0.03, p = .87$	$X^2(1) = 0.14, p = .70$	$X^2(1) = 0.14, p = .71$	$X^2(1) = 1.89, p = .17$
Random effect	Var (SD)	Var (SD)	Var (SD)	Var (SD)	Var (SD)
Participant (intercept)	77.2 (8.79)	651.8 (25.5)	0.0003 (0.0173)	0.0002 (0.0159)	0.0240 (0.1551)
Residual	326.6 (18.1)	12060 (109.8)	0.0044 (0.0662)	0.0056 (0.07514)	1.0529 (1.0261)
Number of observations	480	4188	6095	6240	4246

Note. N = 30. Test = Baseline Preference vs. Final Preference; Gram = Grammatical status (grammatical vs. non-grammatical); ACS = Associative Chunk Strength (high vs. low); Var = variance.

Post-experimental verbal reports showed no evidence of explicit learning or awareness of the underlying grammar. Some participants reported decision criteria other than gut-feeling (e.g., terminal letters), but these were never fully consistent with the grammar. In the sequence generation task, some participants generated valid (grammatical) sequences. However, only a few of these were novel relative to the acquisition and classification sets, suggesting that most sequences were memorized exemplars. Novel sequences were generated by 13 participants (17 generated none), and the mean accuracy level for the whole group was 7%. A closer inspection

showed that the structure of the successfully generated novel sequences (as well as that of unsuccessfully generated ones) was based on the concatenation of frequent chunks (e.g., MS + VRX), indicating that the generation of novel sequences was based on memory for chunks rather than structural knowledge. Altogether, these facts strongly suggest that structural explicit knowledge did not take place. Nevertheless, we analyzed the behavioral accuracy levels for the nongenerators (17 participants with successful generation = 0) separately, so as to make sure that the global indices of knowledge were not expressing the performance of generators (generation > 0), who might be considered potential explicit learners under utmost skepticism. In line with our expectations, the accuracy of nongenerators (strict implicit learners) was at chance levels in baseline preference (*M* = 51%), *t*(16) = .298, *p* > .76, and above chance levels after exposure (final preference: *M* = 59%, *t*[16] = 4.07, *p* = .001; grammaticality: *M* = 62%, *t*[16] = 3.94, *p* = .001). Therefore, the grammar-based learning pattern observed in the whole group did not result from the influence of potential explicit learners. We repeated this control analysis for eye-tracking data, as shown subsequently.

EYE-TRACKING RESULTS

The comparison between baseline preference and final preference showed increased post-exposure discrimination (significant Test x Gram interactions; see Figure 3 and Table 2) in all eye-tracking measures but first-fixation duration. Consistent with this, post hoc comparisons revealed significant differences between G and NG sequences in final preference for dwell time, X^2(1) = 77.8, *p* < .001, fixations, X^2(1) = 72.1, *p* < .001, and dwell/first-fixation, X^2(1) = 51.1, *p* < .001, but not for first-fixation duration (*p* > .14). At baseline preference, there were grammatical effects on dwell, X^2(1) = 10.8, *p* < .001, and fixations, X^2(1) = 7.33, *p* < .01, but not on dwell/first-fixation (*p* > .18) or first-fixation (*p* > .91). Comparisons between final preference and grammaticality (see Table 3) showed no changes. In both comparisons (baseline preference vs. final preference, final preference vs. grammaticality), there were significant Test x Gram x ACS interactions, but they were merely quantitative and did not affect the learning pattern. From baseline preference to final preference, discrimination increased for both High ACS (dwell: X^2[1] = 16.7, *p* < .001; fixations: X^2[1] = 14.9, *p* < .001) and Low ACS sequences (dwell: X^2[1] = 5.06 *p* < .05; fixations: X^2[1] = 6.43, *p* < .05), and from final preference to grammaticality it remained constant in both ACS levels (High ACS: dwell: X^2[1] = 0.84, *p* = .36; fixations: X^2[1] = 0.16, *p* = .69; Low ACS: dwell: X^2[1] = 0.12, *p* = .73; fixations: X^2[1] = 0.42, *p* = .51). There was no evidence of ACS-based change (Test x ACS) in eye movements.

The ocular patterns of nongenerators (participants generating no valid sequences, n = 17) were similar to those of the whole group (see Figure 4). In the comparison between baseline preference and final preference, there were significant Test x Gram interactions for dwell time, X^2 (1) = 4.37, p = .036, number of fixations, $X^2(1)$ = 4.92, p = .026, a marginal interaction for dwell/first fixation, $X^2(1)$ = 2.81, p = .093, and no interaction for first fixation duration, $X^2(1)$ = 1.73, p = .18. Interactions among test, grammaticality, and ACS were non-significant (all ps > .13), and so were Test x ACS interactions (all ps > .30). Comparisons between final preference and grammaticality classification showed non-significant effects.

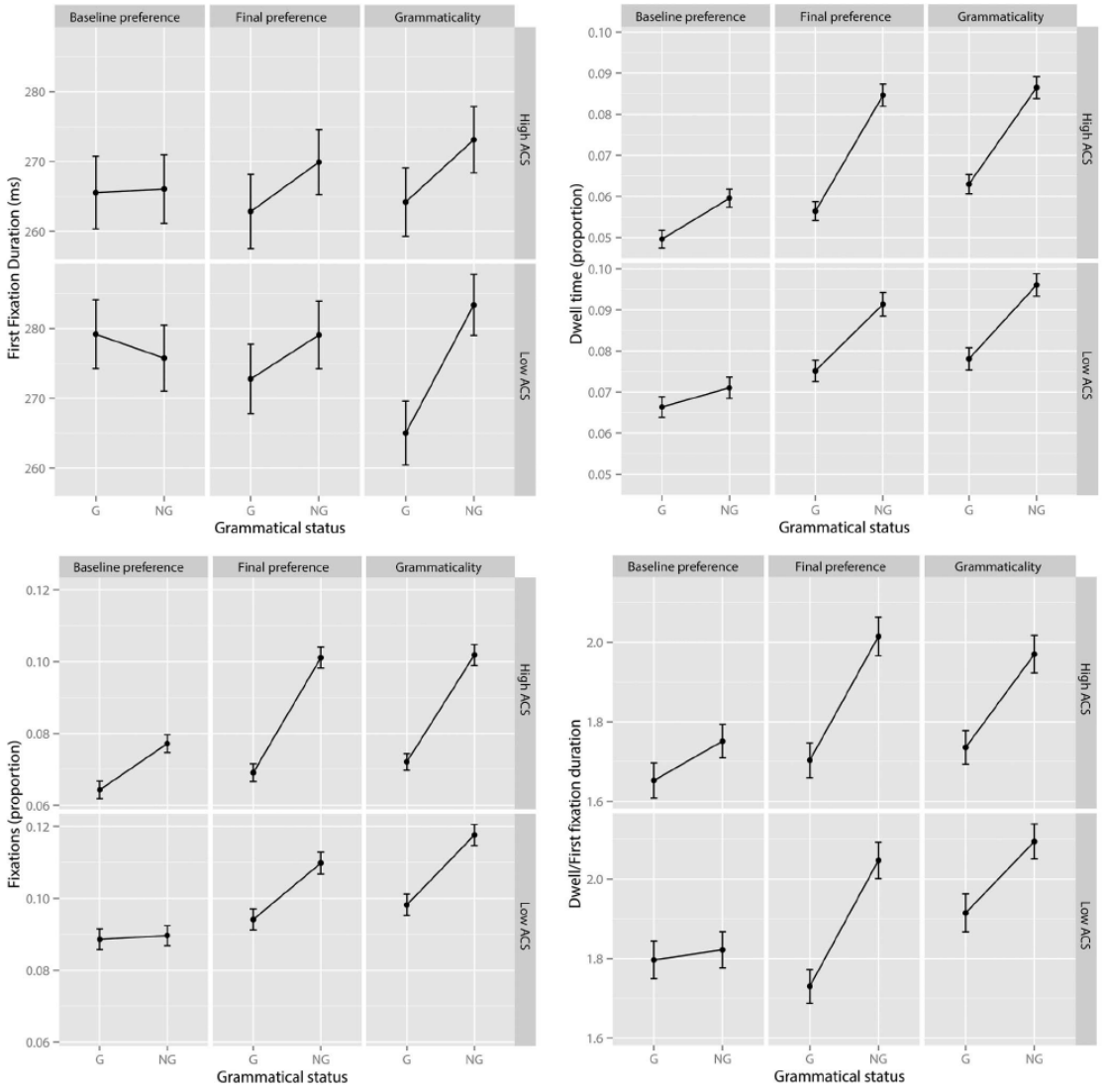

FIGURE 3. Mean eye-tracking measures for the target letter in Experiment 1 as a function of test, grammatical status (G = grammatical; NG = non-grammatical) and associative chunk strength (ACS). Error bars indicate the standard error of the mean.

Table 3

Experiment 1: Comparison between Final Preference and Grammaticality Classification

Effect	Behavioral (endorsement rates)	Eye-tracking			
		First-fixation duration	Dwell time (proportion)	Fixation (proportion)	Dwell/first-fixation
Fixed effect					
Test x Gram x ACS	$X^2(2) = 1.26, p = .53$	$X^2(2) = 1.17, p = .56$	$X^2(2) = 7.46, p < .05$	$X^2(2) = 12.6, p < .01$	$X^2(2) = 0.48, p = .78$
Test x Gram	$X^2(1) = 4.45, p < .05$	$X^2(1) = 1.20, p = .27$	$X^2(1) = 0.13, p = .72$	$X^2(1) = 0.06, p = .81$	$X^2(1) = 2.78, p = .10$
Test x ACS	$X^2(1) = 2.32, p = .13$	$X^2(1) = 0.58, p = .45$	$X^2(1) = 0.14, p = .70$	$X^2(1) = 1.11, p = .29$	$X^2(1) = 3.80, p = .05$
Random effect	Var (SD)	Var (SD)	Var (SD)	Var (SD)	Var (SD)
Participant (intercept)	70.2 (8.38)	580.7 (24.1)	0.0003 (0.0183)	0.0003 (0.0172)	0.0278 (0.1666)
Residual	428.6 (20.7)	12023 (110)	0.0048 (0.0649)	0.0059 (0.0769)	1.1184 (1.057)
Number of observations	480	4425	6098	6264	4246

Note. N = 30. Test = Final Preference vs. Grammaticality Classification; Gram = Grammatical status (grammatical vs. non-grammatical); ACS = Associative Chunk Strength (high vs. low); Var = variance.

DISCUSSION

With the exception of first-fixation duration, all eye-tracking measures paralleled behavioral findings and showed increased discrimination between grammatical and non-grammatical sequences after exposure. Thus, eye-tracking measures showed sensitivity effects in our active forced-choice test. First-fixation duration did not show any significant sensitivity effects, an issue we return to in the General Discussion. Unlike behavioral measures, eye movements revealed no differences between preference and grammaticality classification, suggesting that previous evidence of quantitative differences in the sensitivity of both tests (e.g., Folia et al., 2008) may reflect decision-related processes (see General Discussion). Neither behavioral nor eye-tracking results indicated learning based on surface features (ACS). The

observed pattern of eye-tracking results remained after the exclusion of potential explicit learners. In summary, this experiment showed that eye movements capture the outcomes of implicit AGL when participants are engaged in an active, forced choice task. In Experiment 2, we test whether this is or is not the case during passive testing, where no instruction is provided.

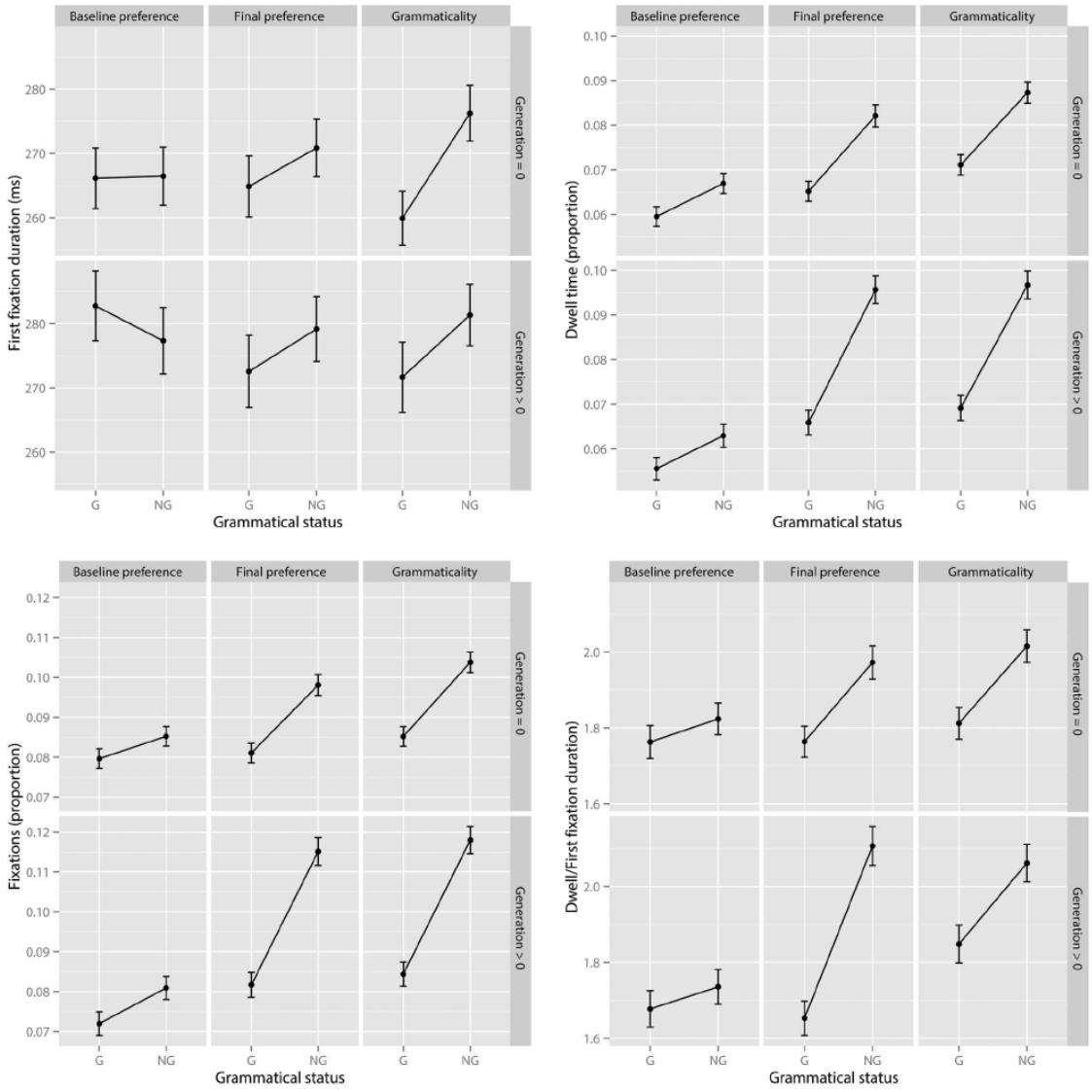

FIGURE 4. Mean eye-tracking measures for the target letter in Experiment 1 as a function of test, grammatical status (G = grammatical; NG = non-grammatical) and performance in the sequence generation task (between-subjects factor: nongenerators [generation = 0, n = 17] vs. generators [generation = 0, n = 13]). Error bars indicate the standard error of the mean.

EXPERIMENT 2: EYE MOVEMENT IN PASSIVE TESTS

As in Experiment 1, we approached AGL with a proper-learning paradigm using passive tests (see Table 1). A group of participants different from that of Experiment 1 was exposed to the artificial grammar, and eye movements were recorded before and after exposure, under no instruction other than to look at the sequences. To reach a within-subjects comparison of test effects (passive vs. active), we added an active test upon completion of the experiment (see Table 1). If discriminative eye movements are facilitated by active tests, ocular discrimination should be less apparent in the present experiment compared with the previous one, and the introduction of an active test in the present experiment should boost discrimination.

METHOD

PARTICIPANTS

Twenty-nine participants took part in the experiment, and 1 was excluded for excess of artifacts. The remaining 28 (*M* age ± *SD* = 25 ± 8; 23 female) complied with the selection criteria of Experiment 1.

STIMULUS MATERIALS

The grammar from Experiment 1 was used to generate one acquisition set (64 items) and seven test sets (16 x 4 = 64 items each). The structure of the stimulus material was identical to Experiment 1.

PROCEDURE

Participants were exposed to five acquisition sessions (see Table 1), on five different days. Sessions were approximately 20 min long. As in Experiment 1, they did same/different judgments on paired sequences (32 same/32 different, five different pairings across the five sessions). Before the first session, they underwent a passive baseline test, where eye-tracking measures were collected in response to 32 G and 32 NG sequences (16 high and 16 low ACS in each group). At the end of each acquisition session, a passive test was run (Passive Tests 1 through 5). In all passive tests, participants were instructed to look at the sequences. On Day 5, the passive test was followed by a grammaticality classification (active) test similar to Experiment 1.

EYE-TRACKING DATA RECORDING AND PREPROCESSING

Data recording and preprocessing followed the steps described for Experiment 1. Artifact inspection led to the exclusion of 1 participant. The data points that entered the analyses (out of 10752 potential data points—28 participants x 64 items x 6 tests, for the first comparison; out of 3584 data points—28 participants x 64 items x 2 tests, for the other comparison) are quantified in Table 4 and Table 5, respectively.

STATISTICAL ANALYSIS

The analysis was similar to that in Experiment 1. We focused on two different comparisons: across all passive tests (six levels for test factor), and between the last passive test and the active grammaticality test (two levels). In this experiment, behavioral data could not be analyzed with a proper learning approach because no active baseline was included. Therefore, we analyzed endorsement rates, accuracy and *d'* in the (single) active test of this experiment.

RESULTS

BEHAVIORAL RESULTS

Accuracy was significantly above chance levels (M = 65%), $t(27) = 4.99$, $p < .001$. Participants discriminated between grammatical and non-grammatical sequences in grammaticality classification (see Figure 5; gram: $X^2[2] = 48.1$, $p < .001$), and this was independent from ACS (Gram x ACS: $X^2[1] = 66.2$, $p = .18$). The *d'* was significantly different from zero (M = 0.90), $t(29) = 4.92$, $p < .001$.

Post-experimental data paralleled that of Experiment 1. Participants showed no evidence of explicit knowledge of the artificial grammar in their verbal reports, although some participants generated valid sequences. As in Experiment 1, only a few sequences were novel (M = 7% novel, correct sequences provided by 11 participants), and these were made up of frequent chunks. The accuracy level of nongenerators (n = 17) in the grammaticality classification task was above chance (M = 59%), $t(16) = 2.52$, $p = .023$. As in Experiment 1, we analyzed separately the ocular patterns of these 17 nongenerators for control (see subsequent text).

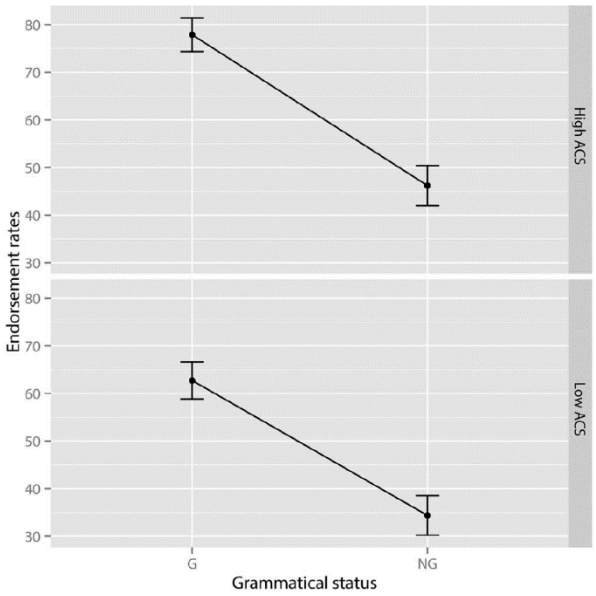

FIGURE 5. Mean endorsement rates (classification as grammatical) in Experiment 2 as a function of test, grammatical status (G = grammatical; NG = non-grammatical) and associative chunk strength (ACS). Error bars indicate the standard error of the mean.

EYE-TRACKING RESULTS

Discrimination based on grammatical status increased across passive tests (baseline plus five subsequent tests) for the proportion of dwell time and dwell-to-first-fixation ratio (see Figure 6 and Table 4). There were also marginal changes for the proportion of fixations. Nevertheless, individual comparisons between baseline and each subsequent test indicated significant differences in only one case, namely for dwell time on Day 4 against baseline (*b* = 0.0105, *SE* = 0.00519, *t* = 2.02).

Nongenerators alone (participants generating zero valid sequences, *n* = 17) were not able to fully provide the pattern of Test x Gram interactions seen for the whole group (see Figure 7): The interaction was marginal for dwell time, $X^2(1) = 10.33$, *p* = .066, and non-significant for fixations (*p* > .14) as well as dwell/first-fixation time (*p* > .46). For dwell time and number of fixations, this seemed to be due to loss of statistical power because the group of generators (participants generating valid sequences, *n* = 11) showed even fewer significant interactions (dwell: *p* > .50; fixations: *p* > .61). Thus, the ocular pattern of generators (potential explicit learners) does not seem to have been responsible for the results of the whole group. A different scenario showed up for dwell/first-fixation, where the Test x Gram interaction was significant for generators, $X^2(1) = 13.38$, *p* = .023, and non-significant for nongenerators (*p* > .46). Still, the

interaction among test, grammaticality, and generation (generators vs. nongenerators) was non-significant, $X^2(1) = 6.38$, $p > .38$. For nongenerators, the interaction among test, grammaticality, and ACS was never significant (all *p*s > .40), and so was the interaction between test and ACS (all *p*s > .09).

Table 4

Experiment 2: Comparison across Passive Tests (Passive Baseline and Passive Tests 1 Through 5)

Effect	First-fixation duration	Fixation (proportion)	Dwell time (proportion)	Dwell/first-fixation
Fixed effect				
Test x Gram x ACS	$X^2(6) = 2.41, p = .88$	$X^2(6) = 3.73, p = .71$	$X^2(6) = 4.05, p = .67$	$X^2(6) = 3.53, p = .74$
Test x Gram	$X^2(5) = 5.72, p = .33$	$X^2(5) = 9.35, p = .10$	$X^2(5) = 14.1, p < .05$	$X^2(5) = 11.2, p < .05$
Test x ACS	$X^2(5) = 4.07, p = .54$	$X^2(5) = 5.89, p = .32$	$X^2(5) = 6.19, p = .29$	$X^2(5) = 3.24, p = .66$
Random effect	Var (SD)	Var (SD)	Var (SD)	Var (SD)
Participant (intercept)	1245 (35.3)	0.00016 (0.01246)	0.00012 (0.01082)	22020 (0.04849)
Residual	11597 (107.7)	0.00382 (0.06180)	0.00260 (0.05095)	1.45024 (1.20430)
Number of observations	7034	9032	8820	6869

Note. N = 28. Test = Passive Baseline vs. Passive Tests 1 through 5; Gram = Grammatical status (grammatical vs. non-grammatical); ACS = Associative Chunk Strength (high vs. low); Var = variance.

Comparisons between Passive Test 5 and the active grammaticality test that was performed immediately after (see Table 5) revealed significant increases in discrimination for first-fixation duration and proportion of dwell time. There was a marginal increase for proportion of fixations. Consistent with the learning profile signaled by interactions, passive

baseline did not show any grammaticality effects (*p*s > .31), Passive Tests 1 through 5 (collapsed) showed significant grammaticality effects on dwell time, $X^2(1) = 24.9$, $p < .001$, fixations, $X^2(1) = 34.9$, $p < .001$, and dwell/first-fixation, $X^2(1) = 24.8$, $p < .001$, but not on first-fixation duration ($p > .44$), and the active grammaticality test showed significant grammaticality effects on all measures (first-fixation: $X^2[1] = 14.4$, $p < .001$; dwell: $X^2[1] = 24.4$, $p < .001$; fixations: $X^2[1] = 21.2$, $p < .001$; dwell/first-fixation: $X^2[1] = 7.82$, $p < .001$).

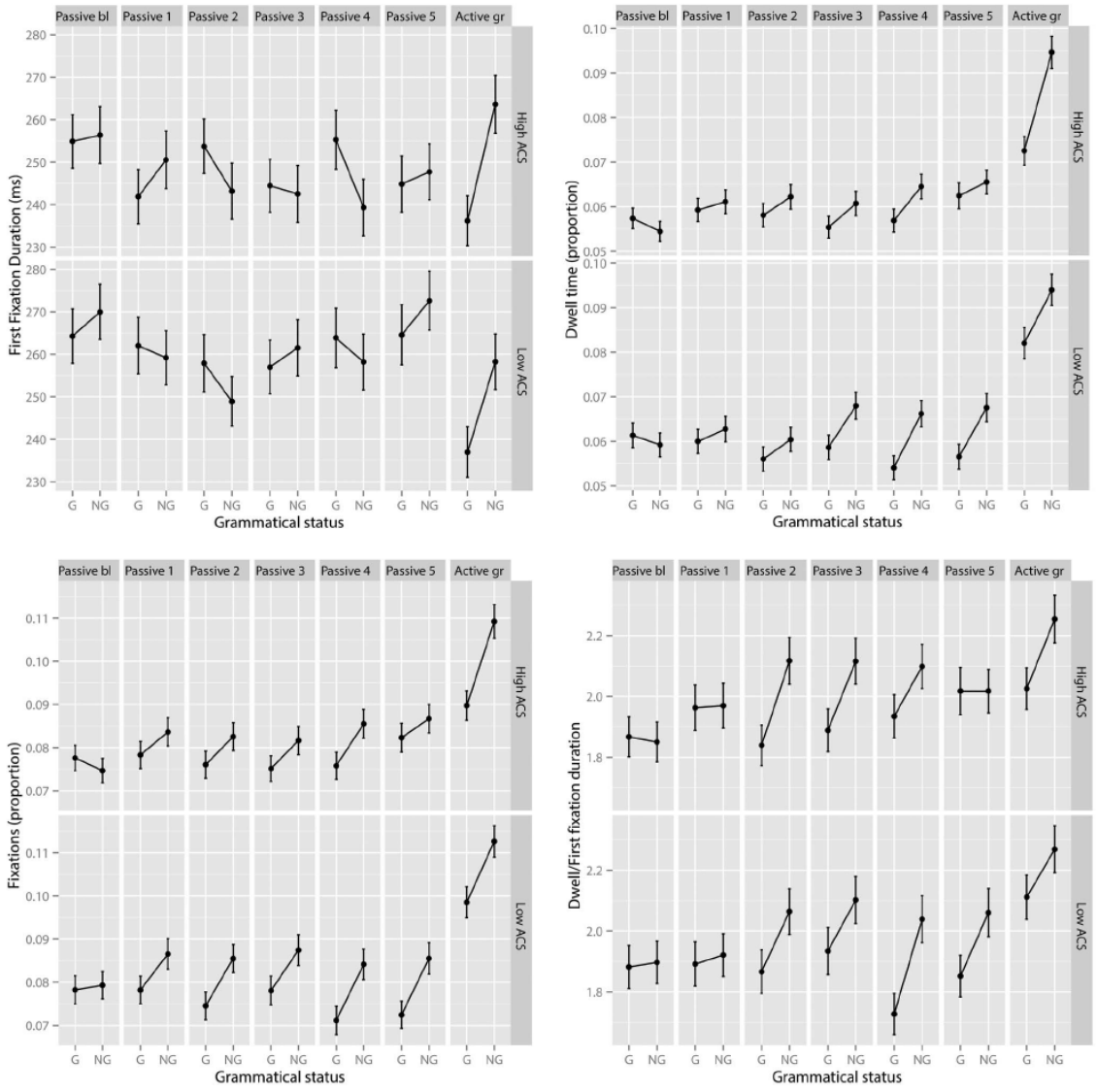

FIGURE 6. Mean eye-tracking measures for the target letter in Experiment 2 as a function of test (Passive bl = passive baseline; Passive 1–5 = Passive Tests 1 through 5; Active gr = active grammaticality classification), grammatical status (G = grammatical; NG = non-grammatical) and associative chunk strength (ACS). Error bars indicate the standard error of the mean.

Nongenerators alone did not show the grammaticality-related changes of the whole group (dwell: $p > .22$; fixations: $p > .16$; first-fixation: $p > .12$), but generators alone did not show it either (dwell: $p > .11$; fixations: $p > .33$; first-fixation: $p > .16$). So, once again, the global pattern of results was not due to the influence of generators. Nongenerators showed no Test x Gram x ACS interactions ($p > .05$), and they showed a significant Test x ACS interaction for first-fixation duration ($p > .05$).

Table 5

Experiment 2: Comparison between Passive Test 5 and Grammaticality Classification

Effect	First-fixation duration	Fixation (proportion)	Dwell time (proportion)	Dwell/first-fixation
Fixed effect				
Test x Gram x ACS	$X^2(2) = 0.59, p = .74$	$X^2(2) = 2.56, p = .28$	$X^2(2) = 4.86, p = .09$	$X^2(2) = 2.27, p = .32$
Test x Gram	$X^2(1) = 4.30, p < .50$	$X^2(1) = 2.81, p = .09$	$X^2(1) = 5.07, p < .05$	$X^2(1) = 0.77, p = .38$
Test x ACS	$X^2(1) = 5.45, p < .05$	$X^2(1) = 5.22, p < .05$	$X^2(1) = 1.87, p = .17$	$X^2(1) = 0.99, p = .32$
Random effect	Var (SD)	Var (SD)	Var (SD)	Var (SD)
Participant (intercept)	1502 (38.8)	0.00020 (0.01421)	0.00019 (0.01380)	0.03600 (0.18970)
Residual	11462 (107.1)	0.00443 (0.06657)	0.00344 (0.05864)	1.57400 (1.25440)
Number of observations	2368	3020	2892	2329

Note. N = 28. Test = Passive Test 5 vs. Grammaticality; Gram = Grammatical status (grammatical vs. non-grammatical); ACS = Associative Chunk Strength (high vs. low); Var = variance.

DISCUSSION

As predicted, the absence of an active test weakened ocular discrimination. Compared with Experiment 1 (eye-tracking coupled with an active task), the Test x Grammatical status interactions— which once again excluded first-pass measures—were less significant for the passive tests in Experiment 2. For proportion of fixations, the effect went from significant to marginally significant. Critically, introducing an active test immediately after the last passive test boosted ocular discrimination in three of the four measures (first-fixation duration, proportion of dwell time, and proportion of fixations). Therefore, an active test seems to facilitate the ocular expression of artificial grammar learning. Similar to Experiment 1, the eye-tracking pattern observed in the whole group did not result from the influence of potential explicit learners, with a possible exception from dwell/first-fixation. We return to this issue in the General Discussion.

GENERAL DISCUSSION

In this study, we wanted to determine whether eye-tracking captures the implicitly acquired knowledge of an artificial grammar and shed light on some restrictions to this possibility. Our first goal was to test the hypothesis that an eye-tracking AGL test shows more robust discrimination between grammatical and non-grammatical sequences when it is coupled to an active test than when this is not the case. In line with our hypothesis, eye movements were significantly sensitive to the outcomes of implicit AGL during both the active final preference classification (Experiment 1) and the active grammaticality classification (Experiments 1 and 2), but less during passive tests, when no instructions were provided other than looking at the sequences (Experiment 2). In addition, eye movements reflected the knowledge of participants who showed no awareness of the grammar by all standards (verbal reports, sequence generation, performance in preference, implicit tests). Thus, we showed that eye-tracking measures alone are able to capture the outcomes of implicit artificial grammar learning and that the sensitivity of eye-tracking measures to implicit knowledge is boosted in the presence of an active forced-choice task.

The most important contribution of our study was to show that implicitly acquired AG knowledge may be captured with eye-tracking. Capturing implicit AGL outcomes in humans with eye-tracking measures has failed in previous studies. Wilson and colleagues (2015) found null results when using an auditory paradigm probing ocular responses to the whole sequence, and it was suggested that eye-tracking-only, passive tests are unable to capture AG knowledge in humans. In line with this, Heaver (2012) probed pupillary responses to visual (whole) AG

sequences and also found null results. In both studies, behavioral discrimination was observed after exposure, suggesting that knowledge had been acquired but it was not being properly captured by eye-tracking measures. Drawing on sensitivity effects, which rely on responses to the violating event rather than the whole sequence, we captured eye-tracking signatures of implicitly acquired AG knowledge.

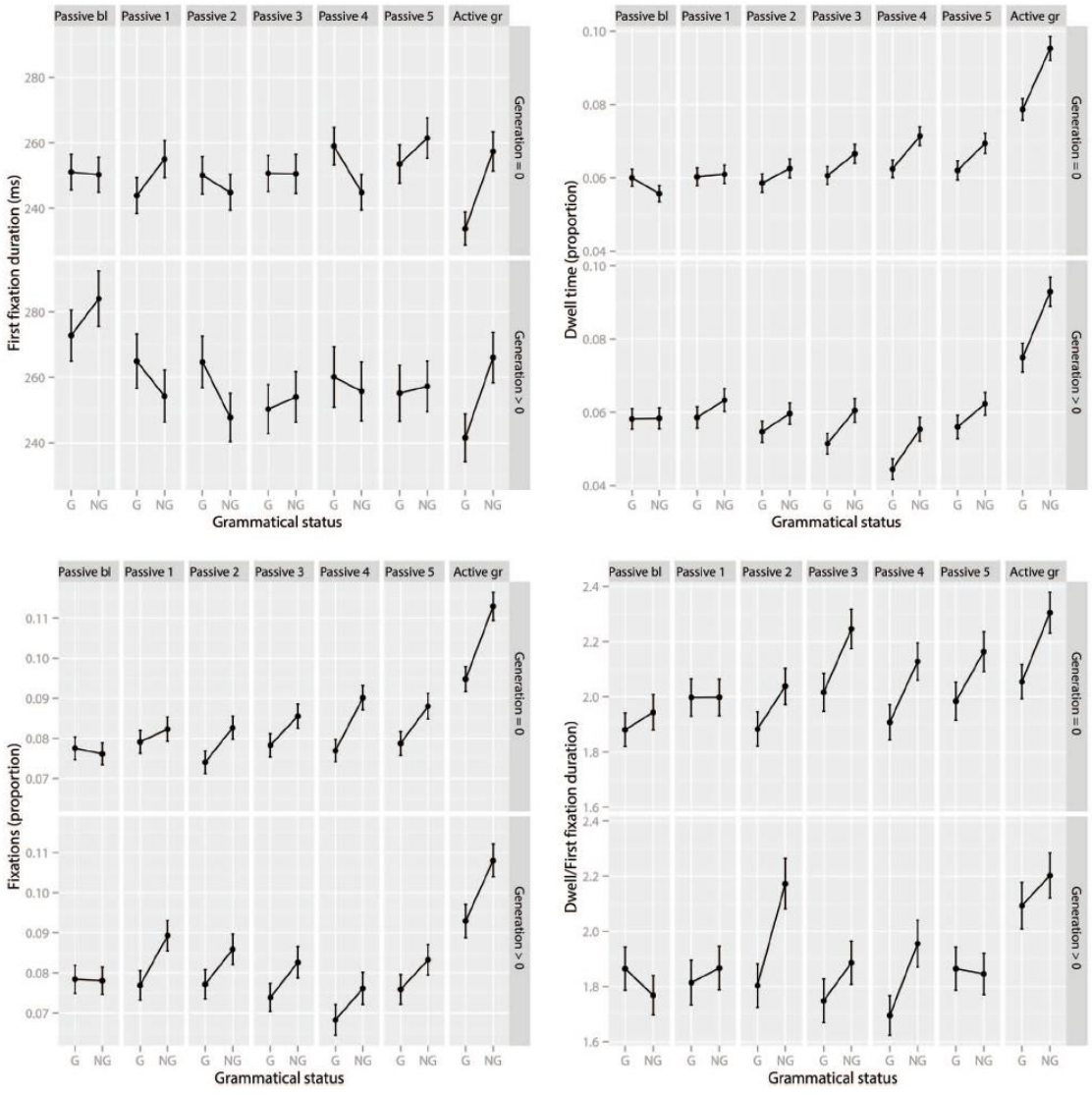

FIGURE 7. Mean eye-tracking measures for the target letter in Experiment 2 as a function of test, grammatical status (G = grammatical; NG = non-grammatical), and performance in the sequence generation task (between-subjects factor: nongenerators [generation = 0, n = 17] vs. generators [generation = 0, n = 11]). Error bars indicate the standard error of the mean.

The sensitivity of eye-tracking measures to implicit artificial grammar learning occurred in the expected direction, that is, as post-exposure increases in proportion of dwell time, proportion of fixations and dwell-to-first fixation ratio for non-grammatical target letters. The presence of sensitivity effects in AGL tests, paralleling the ones observed in tests of natural syntax knowledge, is consistent with the idea that the outcome of AGL is structural, syntax-like knowledge (Christiansen et al., 2012, 2010; Conway et al., 2007; Lelekov-Boissard & Dominey, 2002; Silva et al., 2016; Tabullo et al., 2013; Zimmerer et al., 2014).

Eye-tracking measures were not sensitive to the learning of subsequences (ACS). ACS effects on eye movements were not expected from the behavioral results of Experiment 1 because these showed no ACS-based learning (no Test x ACS interactions), in line with previous studies of ours (Folia et al., 2008; Folia & Petersson, 2014; Forkstam et al., 2008; Silva et al., 2016; Uddén et al., 2012). However, even if behavioral ACS effects on endorsement rates had been observed, it is unclear whether ocular effects on a single violating letter would also be observed. The ACS of a letter sequence presented at the final test phase quantifies how often the bigrams and trigrams of that sequence appeared at the exposure phase, and thus it concerns units larger than one single letter. Therefore, there might be a lack of sensitivity in this respect. Nevertheless, this lack of local subsequence familiarity (ACS) effect is consistent with previous and current behavioral results.

Our second goal was to determine specific eye-tracking signatures of implicitly acquired knowledge. Previous literature has suggested that implicit knowledge on structured sequences, including natural syntax, is better expressed in first-pass eye-tracking measures compared with second-pass measures. Going against this expectation, whole-trial measures (dwell time and number of fixations) revealed AG knowledge in both the active and passive conditions (Experiment 1 and 2) of our study, whereas first-pass measures (first-fixation duration) did not. Critically, we ruled out the possibility that this eye-tracking pattern resulted from explicit learning. Concerning dwell/first-fixation (second-pass measure), we saw sensitivity to acquired knowledge, but our results were not clear as to whether it reflected knowledge that may be considered implicit beyond any doubt: In Experiment 2, unsuccessful generators (strict implicit learners) did not show learning effects on dwell/first fixation, whereas successful generators (potential explicit learners) did so. Moreover, in Experiment 1, the significant interaction for the whole group became marginal after the exclusion of potential explicit learners. Therefore, for second-pass measures (dwell/first-fixation), two different scenarios seem possible: Either our potential explicit learners were effectively explicit and dwell/first-fixation reflects mostly explicit knowledge as suggested in the literature, or these learners were actually implicit and second-

pass measures may express implicitly acquired knowledge. As we stressed throughout this article, the first scenario is unlikely: Potential explicit learners performed above chance levels in the preference classification test (an implicit behavioral test), they did not show awareness of the grammar in their verbal reports, they generated only a small amount of novel grammatical sequences, and these novel sequences could be explained by memory for chunks rather than structural knowledge. Therefore, the most likely scenario is that all participants— even those who generated new strings—acquired implicit knowledge, that dwell/first-fixation patterns reflect implicit knowledge, and some reason other than explicit learning made successful generators more responsive in terms of second-pass eye signatures. In this view, the assumption of a strong association between implicit knowledge and first-pass reading (Godfroid et al., 2015) may be premature, either because second-pass reading is not always a reflection of controlled (vs. automatic) processing or because cognitive control is not incompatible with access to implicitly acquired knowledge (Schott et al., 2005).

Finally, concerning the reasons why an active test boosts ocular discrimination, these remain unspecified. One could think that repeated testing throughout the learning phase (alternate learn-test design, Experiment 2) would introduce noise by forcing participants to process a repeated proportion of non-grammatical sequences, thus leading to weaker learning outcomes. Alternate designs have been shown to elicit weaker learning results when compared with continuous learning designs (Citron, Oberecker, Friederici, & Mueller, 2011) as the one we used in Experiment 1 (but see Forkstam et al., 2006). However, the behavioral and the eye-tracking results of the active test (immediately following passive tests in Experiment 2) provided evidence that knowledge was being concealed - rather than impeded - by passive tests. Earlier in this article, we raised two possible explanations for why passive tests may conceal acquired knowledge: either passive, eye-tracking-only tests are generally unable to provide optimal levels of attention because there is no goal other than looking at the sequences, or passive tests do not specifically elicit the syntactic (structure-related) analysis of AGL sequences needed for expressing knowledge. Further work on this issue should compare eye-tracking sensitivity to AGL classification instructions that activate syntactic analysis to different degrees (e.g., instructions focusing on the visual properties of letters may weaken syntactic analysis).

CONCLUSION

Our results are novel in showing that eye-tracking measures alone are able to express the implicit knowledge resulting from learning an artificial grammar, even though adding an active, forced-choice test boosts ocular discrimination. The possibility of using instruction-free

settings such as eye-tracking to measure the outcomes of implicit structured sequence learning opens new avenues in research. When using eye-tracking concurrently with two different forced-choice active tests, preference and grammaticality classification, we also found highly similar eye-movement profiles. This overcomes behavioral differences observed so far and indicates that differences observed in behavioral testing may result from processes related to final decisions, namely participants' self-monitoring of response direction. Finally, our findings suggest that whole-trial measures may be relevant, and even crucial, to capture the outcomes of implicit structured sequence learning.

REFERENCES

Amso, D., & Davidow, J. (2012). The development of implicit learning from infancy to adulthood: Item frequencies, relations, and cognitive flexibility. *Developmental Psychobiology*, *54*(6), 664–673. http://doi.org/10.1002/dev.20587

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. http://doi.org/10.1016/j.jml.2007.12.005

Bates, D. (2010). *lme4: Mixed-effects modeling with R*. New York: Springer. Retrieved from http://lme4.rforge.r-project.org/book/

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. (R package, Version 1.1–7). Retrieved from http://cran.r-project.org/package%03lme4

Buchner, A. (1994). Indirect effects of synthetic grammar learning in an identification task. *Journal of Experimental Psychology : Learning, Memory, and Cognition*, *20*(3), 550–566.

Christiansen, M. H., Conway, C. M., & Onnis, L. (2012). Similar neural correlates for language and sequential learning: Evidence from event-related brain potentials. *Language and Cognitive Processes*, *27*(2), 231–256. http://doi.org/10.1080/01690965.2011.606666

Christiansen, M. H., Louise Kelly, M., Shillcock, R. C., & Greenfield, K. (2010). Impaired artificial grammar learning in agrammatism. *Cognition*, *116*(3), 382–393. http://doi.org/10.1016/j.cognition.2010.05.015

Citron, F. M. M., Oberecker, R., Friederici, A. D., & Mueller, J. L. (2011). Mass counts: ERP correlates of non-adjacent dependency learning under different exposure conditions. *Neuroscience Letters*, *487*(3), 282–286. http://doi.org/10.1016/j.neulet.2010.10.038

Conway, C. M., Karpicke, J., & Pisoni, D. B. (2007). Contribution of implicit sequence learning to spoken language processing: Some preliminary findings with hearing adults. *Journal of Deaf Studies and Deaf Education*, *12*(3), 317–334. http://doi.org/10.1093/deafed/enm019

Coomans, D., Deroost, N., Vandenbossche, J., Van Den Bussche, E., & Soetens, E. (2012). Visuospatial perceptual sequence learning and eye movements. *Experimental Psychology*, *59*(5), 279–285. http://doi.org/10.1027/1618-3169/a000155

Folia, V., & Petersson, K. M. (2014). Implicit structured sequence learning: An fMRI study of the structural mere-exposure effect. *Frontiers in Psychology*, *5*(FEB), 1–13. http://doi.org/10.3389/fpsyg.2014.00041

Folia, V., Uddén, J., Forkstam, C., Ingvar, M., Hagoort, P., & Petersson, K. M. (2008). Implicit learning and dyslexia. *Annals of the New York Academy of Sciences*, *1145*(2008), 132–50. http://doi.org/10.1196/annals.1416.012

Forkstam, C., Elwér, A., Ingvar, M., & Petersson, K. M. (2008). Instruction effects in implicit artificial grammar learning: a preference for grammaticality. *Brain Research*, *1221*, 80–92. http://doi.org/10.1016/j.brainres.2008.05.005

Forkstam, C., Hagoort, P., Fernandez, G., Ingvar, M., & Petersson, K. M. (2006). Neural correlates of artificial syntactic structure classification. *NeuroImage*, *32*(2), 956–67. http://doi.org/10.1016/j.neuroimage.2006.03.057

Forkstam, C., & Petersson, K. M. (2005). Towards an explicit account of implicit learning. *Current Opinion in Neurology*, *18*(4), 435–41. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16003121

Giesbrecht, B., Sy, J. L., & Guerin, S. A. (2013). Both memory and attention systems contribute to visual search for targets cued by implicitly learned context. *Vision Research*, *85*, 80–87. http://doi.org/10.1016/j.visres.2012.10.006

Godfroid, A., Loewen, S., Jung, S., Park, J. H., Gass, S., & Ellis, R. (2015). Timed and untimed grammaticality judgments measure distinct types of knowledge: Evidence from eye-movement patterns. *Studies in Second Language Acquisition*, *37*(2), 269–297. http://doi.org/10.1017/S0272263114000850

Godfroid, A., & Schmidtke, J. (2013). What do eye movements tell us about awareness ? A triangulation of eye-movement data, verbal reports, and vocabulary learning scores. *Noticing and Second Language Acquisition: Studies in Honor of Richard Schmidt*, (January 2013), 183–205. Retrieved from http://sls.msu.edu/files/5213/8229/7769/Godfroid__Schmidtke_2013.pdf

Godfroid, A., & Winke, P. (2015). Investigating implicit and explicit processing using L2 learners' eye-movement data, (May), 325–348. http://doi.org/10.1075/sibil.48.14god

Gordon, P. C., & Holyoak, K. J. (1983). Implicit learning and generalization of the "mere exposure" effect. *Journal of Personality and Social Psychology*, *45*(3), 492–500. http://doi.org/10.1037/0022-3514.45.3.492

Hayhoe, M. M., & Ballard, D. H. (2011). Mechanisms of gaze control in natural vision. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The Oxford handbook of eye movements* (pp. 607–617). New York: NY: Oxford University Press. http://doi.org/http://dx.doi.org/10.1093/oxfordhb/ 9780199539789.013.0034

Heaver, B. (2012). Psychophysiological Indices of Recognition Memory. *Diss. University of Sussex,*.

Hout, M. C., & Goldinger, S. D. (2012). Incidental learning speeds visual search by lowering response thresholds, not by improving efficiency: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(1), 90–112. http://doi.org/10.1037/a0023894

Jiang, Y. V., Won, B.-Y., & Swallow, K. M. (2014). First Saccadic Eye Movement Reveals Persistent Attentional Guidance by Implicit Leaming. *Journal of Experimental Psychology:*

Human Perception and Performance, *40*(3), 1161–1173.

Keating, G. (2009). Sensitivity to Violation of Gender Agreement in Native and Nonnative Spanish. *Language Learning*, *59*(September), 503–535.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(1), 169–81. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8648284

Leeser, M. J., Brandl, A., & Weissglass, C. (2011). NoTask effects in second language sentence processing research. In P. Trofimovich & K. Donough (Eds.), *Applying priming methods to L2 learning, teaching and research: Insights from psycholinguistics* (pp. 179–198). Amsterdam: Benjamins.

Lelekov-Boissard, T., & Dominey, P. F. (2002). Human brain potentials reveal similar processing of non-linguistic abstract structure and linguistic syntactic structure. *Neurophysiologie Clinique*, *32*(1), 72–84. http://doi.org/10.1016/S0987-7053(01)00291-X

Lim, J. H., & Christianson, K. (2014). Second language sensitivity to agreement errors: Evidence from eye movements during comprehension and translation. *Applied Psycholinguistics*, *36*(6), 1283–1315. http://doi.org/10.1017/S0142716414000290

Manelis, A., Reder, L. M., Manelis, A., & Reder, L. M. (2012). Procedural learning and associative memory mechanisms contribute to contextual cueing : Evidence from fMRI and Procedural learning and associative memory mechanisms contribute to contextual cueing : Evidence from fMRI and eye-tracking, 527–534. http://doi.org/10.1101/lm.025973.112

Manza, L., & Bornstein, R. F. (1995). Affective discrimination and the implicit learning process. *Consciousness and Cognition*, *4*, 399–409.

Meulemans, T., & Linden, M. Van Der. (1997). Associative Chunk Strength in Artificial Grammar Learning. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *23*(4), 1007–1028.

Nieuwenhuis, I. L. C., Folia, V., Forkstam, C., Jensen, O., & Petersson, K. M. (2013). Sleep Promotes the Extraction of Grammatical Rules. *PLoS ONE*, *8*(6). http://doi.org/10.1371/journal.pone.0065046

Petersson, K. M., Elfgren, C., & Ingvar, M. (1999a). Dynamic changes in the functional anatomy of the human brain during recall of abstract designs related to practice. *Neuropsychologia*, *37*(5), 567–587. http://doi.org/10.1016/S0028-3932(98)00152-3

Petersson, K. M., Elfgren, C., & Ingvar, M. (1999b). Learning-related effects and functional neuroimaging. *Human Brain Mapping*, *7*(4), 234–243. http://doi.org/10.1002/(SICI)1097-0193(1999)7:4<234::AID-HBM2>3.0.CO;2-O

Pothos, E. M. (2007). Theories of artificial grammar learning. *Psychological Bulletin*, *133*(2), 227–44. http://doi.org/10.1037/0033-2909.133.2.227

Reber, A. S. (1967). Implicit Learning of Artificial Grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 855–863.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin and Review*, *16*(1), 1–21. http://doi.org/10.3758/PBR.16.1.1

Ryals, A. J., Wang, J. X., Polnaszek, K. L., & Voss, J. L. (2015). Hippocampal contribution to implicit configuration memory expressed via eye movements during scene exploration. *Hippocampus*, *25*(9), 1028–1041. http://doi.org/10.1002/hipo.22425

Sagarra, N., & Ellis, N. C. (2013). From seeing adverbs to seeing verbal morphology. *Studies in Second Language Acquisition*, *35*(2), 261–290. http://doi.org/10.1017/S0272263112000885

Schott, B. H., Henson, R. N., Richardson-Klavehn, A., Becker, C., Thoma, V., Heinze, H.-J., & Duzel, E. (2005). Redefining implicit and explicit memory: The functional neuroanatomy of priming, remembering, and control of retrieval. *Proceedings of the National Academy of Sciences*, *102*(4), 1257–1262. http://doi.org/10.1073/pnas.0409070102

Seger, C. A. (1994). Implicit learning. *Psychological Bulletin*, *115*(2), 163–196.

Shanks, D. R., & John, M. F. S. (1994). Characteristics of dissociable learning systems. *Behavioral and Brain Sciences*, *17*(February 2010), 367–395. http://doi.org/10.1017/S0140525X00035032

Silva, S., Folia, V., Hagoort, P., & Petersson, K. M. (2016). The P600 in Implicit Artificial Grammar Learning. *Cognitive Science*, *41*(1), 137–157. http://doi.org/10.1111/cogs.12343

Stadler, M. A., & Frensch, P. A. (1998). *Handbook of implicit learning*. Thousand Oaks, CA: Sage Publications Inc.

Tabullo, Á., Sevilla, Y., Segura, E., Zanutto, S., & Wainselboim, A. (2013). An ERP study of structural anomalies in native and semantic free artificial grammar: Evidence for shared processing mechanisms. *Brain Research*, *1527*, 149–160. http://doi.org/10.1016/j.brainres.2013.05.022

Uddén, J., Folia, V., Forkstam, C., Ingvar, M., Fernandez, G., Overeem, S., … Petersson, K. M. (2008). The inferior frontal cortex in artificial syntax processing: an rTMS study. *Brain Research*, *1224*, 69–78. http://doi.org/10.1016/j.brainres.2008.05.070

Uddén, J., Ingvar, M., Hagoort, P., & Petersson, K. M. (2012). Implicit Acquisition of Grammars With Crossed and Nested Non-Adjacent Dependencies: Investigating the Push-Down Stack Model. *Cognitive Science*, *36*(6), 1078–1101. http://doi.org/10.1111/j.1551-6709.2012.01235.x

van den Bos, E., & Poletiek, F. H. (2008). Intentional artificial grammar learning: When does it work? *European Journal of Cognitive Psychology*, *20*(4), 793–806. http://doi.org/10.1080/09541440701554474

Wilson, B., Slater, H., Kikuchi, Y., Milne, A. E., Marslen-Wilson, W. D., Smith, K., & Petkov, C. I. (2013). Auditory Artificial Grammar Learning in Macaque and Marmoset Monkeys. *Journal of Neuroscience*, *33*(48), 18825–18835. http://doi.org/10.1523/JNEUROSCI.2414-13.2013

Wilson, B., Smith, K., & Petkov, C. I. (2015). Mixed-complexity artificial grammar learning in humans and macaque monkeys: Evaluating learning strategies. *European Journal of Neuroscience*, *41*(5), 568–578. http://doi.org/10.1111/ejn.12834

Zimmerer, V. C., Cowell, P. E., & Varley, R. A. (2014). Artificial grammar learning in individuals with severe aphasia. *Neuropsychologia*, *53*(1), 25–38. http://doi.org/10.1016/j.neuropsychologia.2013.10.014

ACKNOWLEDGEMENT

This work was supported by Max Planck Institute for Psycholinguistics, Donders Institute for Brain, Cognition and Behaviour, Fundação para a Ciência e a Tecnologia (PTDC/PSI-PCO/110734/2009; SFRH/BD/85439/2012; UID/BIM/04773/2013 CBMR 1334; PEst-OE/EQB/LA0023/2013 and UID/PSI/00050/2013), Vetenskapsrådet, The Swedish Dyslexia Foundation. We thank Christian Forkstam for help with the experimental set-up and some data acquisition.