LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Implicit Learning in Production: Productive Generalization of New Form–Meaning Connections in the Absence of Awareness

Giulia Bovolenta [a] and John N. Williams [b]

[a]University of York [b]University of Cambridge

Abstract: Second language implicit learning research has shown that a variety of linguistic features can be acquired without awareness. However, this research overwhelmingly uses comprehension tests to measure implicit learning. It remains unclear whether newly acquired implicit knowledge can also be recruited for production. To address this question, we developed a novel paradigm based on elicited recall and false memory that was used to both train participants and test their implicit knowledge in production, including generating new instances of the rule. Participants learned a semiartificial language containing a rule based on one in a natural language (the alternation between Czech spatial prepositions *v* and *na*). Participants who remained unaware of the rule, as assessed by verbal report, nevertheless were able to use it in a production test involving novel items, while believing that they were performing a cued recall test. Even without extensive training, newly acquired implicit knowledge can immediately be evident in production.

Keywords implicit learning; language; production; elicited recall; false memory

---

A one-page Accessible Summary of this article in non-technical language is freely available in the Supporting Information online and at https://oasis-database.org

The authors have no known conflict of interest to disclose. This research was supported by the Economic and Social Research Council, UK (doctoral studentship to Giulia Bovolenta, award ref. 1368466).

Correspondence concerning this article should be addressed to Giulia Bovolenta, Department of Education, University of York, Heslington YO10 5DD, United Kingdom. Email: giulia.bovolenta@york.ac.uk

The handling editor for this article was Sarah Grey.

DOI: 10.1111/lang.12551

Introduction

Research on implicit learning in the field of second language (L2) acquisition has revealed that adult L2 learners can acquire knowledge of novel language rules while remaining unaware of what they have learned (Williams, 2005). However, this strand of research in adults tends to rely on comprehension tasks as measures of implicit knowledge. There is limited evidence that it is possible to acquire productive language skills through implicit learning (Dell et al., 2000; Denhovska & Serratrice, 2017). This is partly for methodological reasons: Eliciting production is difficult, and even more so if the aim is to avoid drawing attention to the rule(s) participants need to rely on to respond accurately in the test. This study was designed to investigate the following question: Can learners develop and correctly use implicit knowledge, without awareness, during production? To address this question, we developed a novel paradigm, based on elicited recall, as a way to both train participants and test their implicit knowledge in production, including generating new instances of the rule. We trained participants on a semiartificial language based on a rule naturally found in Czech, specifically, the usage rule for a pair of spatial prepositions (*v* and *na*) that alternate depending on the distinction between open and enclosed spaces.

Background Literature

Implicit Learning of Language

In cognitive psychology, the distinction between explicit and implicit knowledge in its basic form hinges on conscious mental access (Williams, 2009). Explicit knowledge is knowledge that we know that we know (Dienes & Perner, 1999) and can therefore use deliberately; implicit knowledge, by contrast, can guide our behaviour even if we are not aware that we have it (Cleeremans et al., 1998). In the field of L2 acquisition, it is generally agreed that the development of implicit knowledge is necessary for the attainment of L2 proficiency (DeKeyser, 2003; N. C. Ellis, 2011; R. Ellis, 2012; Hulstijn, 2005; Krashen, 1982). Implicit knowledge or, possibly, explicit knowledge that has become highly automatized through practice (Li & DeKeyser, 2017) is the basis for fluent, automatic L2 processing in both comprehension and production.

For our purposes, implicit learning will be defined as a type of incidental learning in which learners spontaneously "pick up" some target regularity in the absence of relevant instruction and without engaging in conscious hypothesis testing, or, more broadly, without an intention to learn the regularity in question. Given the difficulty of ascertaining that learning is implicit in the moment it is happening without interfering with it (Rebuschat et al., 2015),

researchers often (but not always) use measures of awareness following exposure as a means of inferring whether learning was implicit, on the assumption that explicit learning processes, if they had occurred, would have led to conscious knowledge. Under this definition, the implicitness of the learning process is separate from that of the learning product: Implicit learning could potentially lead to explicit knowledge if the person becomes aware of what they have learned following exposure, for example, through spontaneous insight. Conversely, in the long term, explicit knowledge acquired through explicit learning could contribute to the development of implicit knowledge (e.g., through practice). Hence the validity of the inference from unawareness of product to implicitness of process has to be evaluated in the context of each particular study.

Whereas implicit knowledge is studied in L2 acquisition as one of the components of language proficiency, the study of implicit learning as a field is characterized by strict control over the learning process and rigorous methods for determining the nature of resulting knowledge. The application of implicit learning paradigms to L2 acquisition research has provided evidence for implicit learning of various different aspects of language, such as word segmentation (Saffran et al., 1996), orthography (Pacton et al., 2001), phonotactics (Chambers et al., 2003), syntactic structure (Francis et al., 2009; Rebuschat & Williams, 2012; Williams & Rebuschat, 2012), and morphology (Marsden et al., 2013; Rogers et al., 2016). With regard to form–meaning connections (particularly relevant to the present study), there is evidence that participants can implicitly learn the correlation between novel grammatical morphemes and thematic roles (Fukuta & Yamashita, 2021; Leung & Williams, 2011), and semantic-based selectional restrictions of novel verbs (Paciorek & Williams, 2015a, 2015b). A number of studies have focused on correlations between novel determiners and noun animacy. Some studies have found effects in participants who report no awareness of the rules (Batterink et al., 2016; Chen et al., 2011; Kerz et al., 2017; Leung & Williams, 2014; Williams, 2005), whereas others have found no effect (Faretta-Stutenberg et al., 2011; Hama & Leow, 2010), and still others have found an effect only when using trial-by-trial subjective measures to isolate decisions based on implicit knowledge (Rebuschat et al., 2015; Zhao et al., 2021). It is not our intention here to address particular reasons for the disparities in the above results, but the present study does add to this literature by examining a similar issue in the form of the correlation between a preposition and the spatial properties of the noun it accompanies, using an awareness measure based on verbal report.

Production in Implicit Learning

In L2 research, a strand of literature has examined the effects of implicit instruction on the development of L2 knowledge, with the inclusion of production measures to test learning outcomes (Goo et al., 2015; Norris & Ortega, 2000; Spada & Tomita, 2010). However, these studies looked at the effect of implicit or incidental instruction on L2 learners with prior knowledge of the language and did not presuppose a lack of awareness of the rules being tested. By contrast, implicit learning studies—studies targeting the implicit learning of novel regularities under controlled conditions and testing the resulting knowledge in unaware participants—have mostly trained and tested participants using comprehension tasks, rather than production ones. Even after quite brief training, participants may show sensitivity to the regularity in question by performing at above-chance levels on, for example, grammaticality judgment or forced-choice tasks; however, it remains unclear whether the implicit knowledge acquired under these circumstances would be evident in language production.

Can we learn to produce language following a new rule system without being aware of it? Some preliminary evidence comes from work by Dell et al. (2000) on implicit learning of phonotactics, where a novel phonotactic constraint was shown to influence the probability of speech errors. However, the effect in that study was observed after extensive training over days. Can this phenomenon be shown after only brief exposure to a novel linguistic system under experimental conditions, and for grammar rather than phonotactics? In a study by Denhovska and Serratrice (2017), participants who learned novel morphology (Russian gender inflection) under incidental learning conditions, and who remained unaware of the rule, showed evidence of learning in a grammaticality judgment task, but not in a written production test (a fill-in-the-blank task). The authors therefore suggested that production of newly acquired rules may rely on explicit knowledge, at least at the early stages of learning. However, it is also possible that the production test used in the study was not sensitive enough to pick up the level of implicit knowledge that may develop after limited exposure.

Whereas implicit knowledge underlies automaticity in L2 processing, including production, brief exposure presumably results in weak and fragmentary representations with insufficient strength to support automatic processing (Cleeremans & Jiménez, 2002), at least in the sense of speed of access. Therefore, we focus on a situation in which there is no time pressure on the production process. If weakly encoded representations can still be deployed without conscious control (due to lack of awareness, which is our criterion for

implicitness), then it is possible that they will influence production. On the other hand, even without time pressure, speech production poses additional demands relative to comprehension (e.g., planning and word retrieval) and cannot rely on additional semantic cues, which can decrease reliance on structural information in comprehension (Clahsen & Felser, 2006). Accordingly, production abilities often lag behind comprehension skills in L2 acquisition.

Thus, it may be that stronger linguistic representations are needed for rule knowledge to be detected in production, relative to comprehension. If that is the case, detecting implicit knowledge in production may prove problematic: The linguistic representations acquired by participants in implicit learning studies tend to be quite weak, if conceptualized in terms of accuracy. Typical accuracy in the two-alternative forced-choice tests is around 60%, against a chance level of 50% (Williams, 2005). When representations become stronger, however, so does the likelihood that participants will become aware of their knowledge (Cleeremans & Jiménez, 2002). Since typical research on implicit learning relies on lack of awareness at the time of testing in order to conclude that implicit learning took place, such participants may not provide the data needed to establish implicit learning. What is needed is a test that is sensitive enough to detect knowledge that may be weak or fragmentary, without raising awareness of the rule being tested.

In L2 research, the elicited oral imitation task (Erlam, 2009; Slobin & Welsh, 1973) is often used to detect knowledge that is used automatically and unconsciously in production. The task relies on the principle that verbatim recall is supported by linguistic knowledge. For instance, English speakers are better at remembering lists of English words than of Aymara words of the same length (Erlam, 2009). The facilitatory effect of existing representations is thought to reflect a process of regeneration. According to the "regeneration hypothesis" (Lee & Williams, 1997; Lombardi & Potter, 1992; Potter & Lombardi, 1990), a sentence in a known language can be decoded for meaning and then reencoded in form upon recall, whereas one in an unknown language can only be stored in short-term phonological memory as a string of phonemes. In the elicited imitation task, participants hear a sentence in the target language and have to repeat it. Ungrammatical sentences may also be included among the items. When asked to repeat ungrammatical sentences, proficient speakers tend to correct any mistakes in the sentences without realizing (Erlam, 2009; Slobin & Welsh, 1973), which lends support to the regeneration hypothesis. In general, increasing the time between listening to the sentence and having to repeat it is thought to increase the extent to which participants reconstruct the sentence, rather than rely on rote memory (Spada et al., 2015).

A similar principle has been exploited in artificial grammar learning research. In a study by Reber (1967), rule learning was assessed by string recall in a between-subjects design. Participants were instructed to memorize sets of letter strings and reproduce them in writing. In the experimental group, the strings were generated according to a set of rules, whereas in the control group, strings were composed of the same letters, but arranged randomly. After the first two sets, recall performance became better in the experimental group than in the control group, indicating some degree of acquisition of the rules underlying the letter strings. More recently, Isbilen et al. (2018, 2020) used recall of rule-based strings as a measure of learning in a series of statistical learning studies. In their 2020 study, participants were first exposed to a (meaningless) artificial language composed of trisyllabic nonwords concatenated together in a continuous stream, following a study by Saffran et al. (1996). The participants were then given six-syllable strings, which they were asked to recall immediately. These longer strings could be composed either of two words from the language (experimental items) or of the same syllables arranged in a random order. Participants were significantly more accurate in their recall of experimental items compared to random ones, indicating that they had acquired knowledge of the trisyllabic sequences in the language. Similar results were obtained by Isbilen et al. (2018) with an artificial language including nonadjacent dependencies. In the test phase of this study, Isbilen et al. also included generalization strings, that is, strings with trained dependencies but novel intervening syllables. Participants were still significantly better at recalling these novel grammatical strings than the random ones, showing that their grammatical knowledge could generalize to new instances. In sum, these studies demonstrate the usefulness of short-term recall as a way of measuring learning effects. We used a similar principle in our study, as detailed below.

The Present Study

Our research question was the following: Can learners develop and correctly use implicit knowledge during production? Participants were exposed to a semiartificial language consisting of English lexis combined with four novel spatial prepositions (*gi*, *ro*, *wa*, *ne*). The distribution of two of the novel prepositions followed a rule found in a natural language, Czech, where spatial prepositions *v* and *na* alternate based on the physical properties of the place in question (open vs. enclosed space). Although *v* and *na* are the actual forms of the prepositions used in Czech, we did not use the forms *v* and *na* as stimuli in our study. Instead, we used the forms *gi*, *ro*, *wa*, and *ne*, which were randomly assigned to belong either to the system pair (which reproduced the *v*/*na*

distinction) or the random pair, which alternated randomly. For example, for a particular participant, *wa* would be used with enclosed spaces, *gi* with open spaces ("system pair"), and *ro* and *ne* equally with open and enclosed spaces ("random pair"). We based our stimuli on previous studies on the implicit learning of form–meaning connections, which have used a similar set of forms, namely, *gi*, *ro*, *ul*, and *ne* (Leung & Williams, 2012, 2014; Williams, 2005). We decided to use *wa* instead of *ul*, because the sound *ul* was harder to discern when produced by the speech synthesizer that we used to create auditory stimuli.

We used a recall paradigm to test participants' rule knowledge by measuring its effects on the recall of rule-based (system pair) prepositions in comparison to random (random pair) ones. We also tested participants' ability to generalize the rule in production by using a modified ("false memory") version of the recall paradigm, which avoided raising participants' attention to the rule we were testing. As a first test of rule knowledge, following exposure, participants performed a cued recall task in which they saw a picture from the training phase and attempted to recall the sentence that went with it. We compared recall accuracy for sentences using the rule-based prepositions (system items) and those using the randomly distributed prepositions (random items).

Although recall was not immediate in our study, it exploited the same principle as in the study by Isbilen et al. (2020). We expected participants to be able to rely on their rule knowledge to "reconstruct" (either consciously or unconsciously) the rule-based items during recall, whereas they could rely only on verbatim item memory to correctly recall the random ones. Therefore, if participants had acquired knowledge of the rule, we would expect them to have better memory for system items than for random ones. Crucially, for the effect of implicit knowledge to emerge in recall, it is necessary to avoid ceiling effects, which would occur if the string could be easily recalled by memory alone. In the recall task used by Isbilen et al. (2020), participants were presented with six-syllable words, that is, items exceeding the average working memory span by one syllable, and were asked to recall them immediately. In our case, because the novel elements of the sentence were mostly (but not entirely, as described below) limited to the prepositions, the amount of novel phonological material that participants had to recall was fairly small (the remainder of the sentence consisted of English lexis that either was kept constant or could be inferred from the picture). Therefore, we had a longer gap than did Isbilen et al. between exposure and recall. Participants were asked to recall sentences from the training phase in a subsequent cued recall (production) test.

As a second test of rule knowledge, we tested generalization of the rule to new instances. In the elicited imitation task, testing generalization is sometimes achieved through the insertion of ungrammatical sentences, which participants may automatically correct when recalling them, if they have implicit knowledge of the rule. However, as discussed earlier, weak representations characterize the knowledge developed by participants in implicit learning studies; these would be unlikely to support the kind of automatic correction that may be observed in proficient L2 learners. At the same time, any test employed should avoid raising attention to the rule, for the purposes of demonstrating implicit learning. One way to do so would be to ask participants to recall new sentences where the rule is used in novel contexts, as previously done by Isbilen et al. (2018). Another way, which could provide even stronger evidence for rule knowledge in production, is to test participants by exploiting another aspect of human memory, namely, the false memory phenomenon (Deese, 1959; Roediger & McDermott, 1995). With false memory, participants tend to erroneously believe that they have seen an item if it conforms to an experienced pattern. For instance, participants who had to learn a list of words (e.g., *bed*, *rest*, *awake*) would also produce associated (but unpresented) words (e.g., *sleep*) in immediate recall 40% of the time (Roediger & McDermott, 1995).

In this study, we took advantage of the false memory phenomenon to test participants' ability to generalize the rule they had learned to new instances. In addition to recall of trained sentences, we induced participants to produce new sentences without directly asking them to do so. During the recall task, participants were presented with a series of pictures that looked familiar, because they contained images from the items used in the training phase. However, the images in each picture were combined in novel ways, meaning that the sentence that would correctly describe the picture as a whole had not been heard during training. Our intention was that participants would be under the impression that they were still doing a recall test, even though those items were novel, and would produce a novel sentence while believing that they were recalling a sentence they had heard during training.

We were also interested in knowing whether participants would show evidence of rule knowledge in comprehension, so we ran an additional comprehension test following the recall test. However, this was a separate research question that does not form part of the current study, so we will briefly report the comprehension test in the Procedure section but will not discuss its findings (which can be found in Appendix S2 in the Supporting Information online). After the recall and comprehension tests, we used a questionnaire to assess participants' awareness of the rule.

Our dependent variables were (a) recall of trained rule-based items, (b) generation of novel rule-based items, and (c) awareness of the rule. Because we were interested in implicit learning, our main experimental hypotheses concerned participants who would not become aware of the rule as assessed by a debriefing questionnaire administered immediately after testing (variable c). With regard to dependent variables a and b, we hypothesized that, if participants had acquired any (implicit) knowledge of the rule, they should show evidence of it in a production test following the training phase. Specifically, we expected them to show higher accuracy when recalling rule-based (system) items compared to non-rule-based (random) ones and above-chance accuracy in producing new rule-based items. We did not have any specific hypotheses for participants who became aware of the rule, since they were not the focus of the study. However, based on evidence from previous implicit learning studies, we may expect them to perform better than unaware participants on the tests of trained item recall and generalization, because they could rely on explicit knowledge of the rule.

Method

Participants

Forty-two native English speakers (29 females, $M = 20.5$ years, $SD = 2.05$) from the University of Cambridge and surrounding community took part in the experiment, receiving £6 as compensation. We based our sample size on previous studies on implicit learning of novel morphosyntactic rules in semiartificial languages, which found learning effects in generalization (significant difference from 50% chance-level accuracy): Leung and Williams (2012; $N = 25$, 20 unaware) and Williams (2005; Experiment 1: $N = 41$, 33 unaware; Experiment 2: $N = 24$, 11 unaware). A prior (unpublished) experiment carried out as part of this research project and additional piloting suggested that we could expect approximately 57% of participants to remain unaware of the rule, which would result in approximately 24 unaware participants in this study. Of the 42 subjects who took part in our study, two reported some knowledge of a Slavic language (Russian), which encodes a similar distinction to the one we used in our rule (distinct spatial prepositions for open vs. enclosed spaces). Other foreign languages spoken by participants were French (13), German (10), Spanish (6), Italian (4), Mandarin Chinese, Irish, Greek, Hindi, Urdu, and Arabic (1 each).

Materials

All materials used in the experiment, including the debriefing questionnaire, are available via the OSF (https://doi.org/10.17605/OSF.IO/86b7c) and IRIS

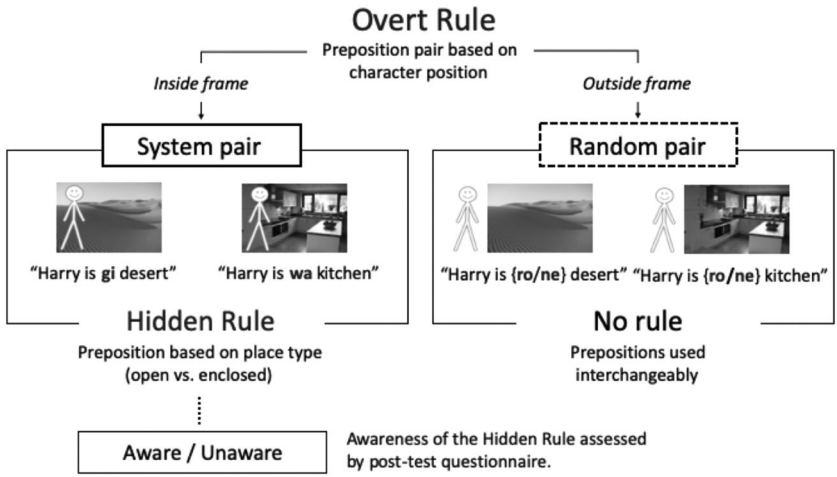

Figure 1 Description of rules used in the experiment.

(Bovolenta & Williams, 2022). We based our materials on the pair of spatial prepositions *v* and *na*, found in Czech, which alternate based on the type of location they refer to (enclosed vs. open, as in *Žena je* ***v*** *kanceláři* "The woman is in the office," but *Žena je* ***na*** *hřišti* "The woman is in the playground"). We selected 80 place nouns, 40 for each type of location. Four pseudowords ("gi," "ro," "wa," and "ne") were used as spatial prepositions. Place nouns were embedded in simple sentences, all with the following structure: subject – "is" – preposition – place noun (e.g., "Harry is gi desert"). The subject was always one of two characters, "Harry" or "Lucy" (which corresponded to a stick figure drawing of either a male or a female), randomly assigned at each trial. Sentences were presented auditorily. During training, each sentence was accompanied by a visual representation, composed of a drawing representing either Harry or Lucy, and a photograph depicting the place noun (see Appendix S1 in the Supporting Information online for more details of how the stimuli were created).

Overt and Hidden Rules

There was a semisystematic relationship between the preposition (nonword) used in a sentence, the position of the character relative to the place photograph in the picture, and the type of place depicted in the photograph, which was governed by two rules: an overt and a hidden one (see Figure 1 and explanation below). The place photograph always occupied the centre of the picture, but

Table 1 Sample preposition assignment to experimental conditions

Distinction	System items	Random items
Open places	*Gi*	*Ro* / *Ne*
Enclosed places	*Wa*	*Ro* / *Ne*

Note. Prepositions used in system items alternate based on place type; prepositions used in random items alternate at random, ignoring the open/enclosed distinction.

the position of the character relative to it varied. The character could either be superimposed on the left half of the photograph (system condition) or appear beside it (random condition). There was a systematic relationship between the character's position and the preposition used in a sentence. Out of four possible prepositions ("gi," "ro," "wa," "ne"), two were used only in system sentences (e.g., "gi" and "wa" in the left-hand panel of Figure 1) and the other two only in random sentences (e.g., "ro" and "ne" in the right-hand panel of Figure 1). This was the overt rule, which participants were encouraged to discover. They were told to pay attention to the position of the character on screen, because it would help them to remember the sentences.

Besides the character's position (overt rule), there was another difference between system and random sentences. In system sentences there was an additional systematic relationship governing preposition use. The preposition used for each system sentence, out of the two possible ones associated with this condition, was determined by the nature of the place depicted in the picture as either an open or enclosed space, as in the desert or kitchen photographs in Figure 1, where "gi" would map onto open spaces and "wa" would map onto enclosed spaces. This was the hidden rule, and participants were not given any indications that this rule may exist. By contrast, in random sentences there was no systematic relationship between prepositions and places. The two possible prepositions were used interchangeably, irrespectively of place type (Figure 1 and Table 1).

Therefore, the overt and hidden rules were not orthogonal. Rather, the hidden rule was nested within the overt rule, as it only applied in one of the two conditions (system), with the random condition serving as a control condition. The hidden rule was the focus of our implicit learning experiment. The reason for having the overt rule was to direct participants' conscious attention away from the hidden rule, in order to reduce the chances that they would process it consciously. Hence the training task was intended to encourage incidental learning of the hidden rule, but explicit learning of the overt rule.

Table 2 Sample of items used in training and testing (production task)

	Training		Testing (trained items)		Testing (generalization)	
Images	System	Random	System	Random	System	Random
Place photos 1–32	×	×	×	×		
Place photos 33–56	×					×
Place photos 57–80		×			×	

The assignment of specific prepositions to the different conditions was randomized for each participant. We chose to associate the system condition only with pictures where the character was superimposed onto the place photograph, and not with those where the character was next to the photograph, to maintain ecological validity and ensure that the rule would be learnable. The Czech *v*/*na* pair that we based the hidden rule on means "in/at/on," but there is no equivalent preposition pair meaning "near/by/next to" that would be sensitive to place type. Therefore, associating system sentences with the character being outside the photograph would have risked introducing additional confounds.

Training and Testing Items

All training and testing items were derived from the 80 place photographs selected, which could be displayed in either system or random condition pictures (for a total of 160 possible pictures). Given that we intended to test for hidden rule knowledge by comparing recall of rule-based (system) items to that of control (random) items, we wanted to ensure that the items would be comparable, to avoid confounds. Therefore, the same set of 64 "matched" items was included both in the training phase, and in the testing phase as trained items. These items were made up of the same set of 32 place photographs, appearing once in the system and once in the random condition over the course of the training task, and later, of the recall test (Table 2). This allowed us to then test participants' recall by comparing two sets of items (system vs. random) that all contained the same places but differed in the prepositions used (and whether these prepositions were rule-based). Therefore, any advantage in recall accuracy for system items could be attributed to the presence of rule-based prepositions.

Whereas we used long-term recall to assess learning, the training phase task also involved recall, but over a shorter interval. Participants only had to commit two sentences to memory at a time and then recall them. A short

recall interval was employed to ensure a high level of accuracy following the principle of "errorless learning" (see Wilson et al., 1994), but we also wanted to ensure that the memory load was sufficiently high to encourage rule extraction following research showing the effect of item complexity on rule extraction (see Radulescu et al., 2020). We increased the amount of novel material that participants had to recall by introducing pseudowords for place names in 50% of the matched items, which participants had to recall together with the rest of the sentence (e.g., "Harry is *gi dreet*" instead of "Harry is *gi* kitchen"). Pseudoword items were evenly divided among the system and random conditions (counterbalanced across participants). If a matched item in the system condition had a pseudoword place name, its random counterpart would retain the English place name (and vice versa), thereby ensuring that all places were seen at least once with the appropriate English name.[1]

Participants were trained on 112 sentences. Of these, 64 were the matched items we have just described, and the remaining 48 were created using 48 unique place photographs, which were evenly split between system and random condition pictures (Table 2). Participants were then tested on 88 sentences, which were a mix of old and new items. Of these sentences, 64 were the matched items from training, which were used to test recall (trained items). The remaining 24 were a subset of the 48 unique items from training, now presented in the opposite condition (generalization items). Using place photographs already seen in training ensured that generalization items would seem familiar to participants. However, because the photographs had not been seen in those specific pictures (with the character in that position) during training, the sentences that participants would need to "recall" in response to those items were in fact novel. The assignment of unique items to different conditions in training and testing was counterbalanced across participants.

Procedure

The experiment was carried out using PsychoPy2 (Peirce et al., 2019); see Appendix S1 in the Supporting Information online for details of the laboratory setting. The procedure consisted of a training phase, a recall testing phase, and a comprehension testing phase (see below for rationale), delivered in this order, followed by a short debriefing questionnaire. Each of the experimental tasks was preceded by a set of written instructions, which participants read at their own pace (for the specific instructions, see Bovolenta & Williams, 2022, and the OSF at https://doi.org/10.17605/OSF.IO/86b7c). At the start, participants were informed that they would hear a series of sentences, each accompanied by a drawing of the scene it described. Every sentence would

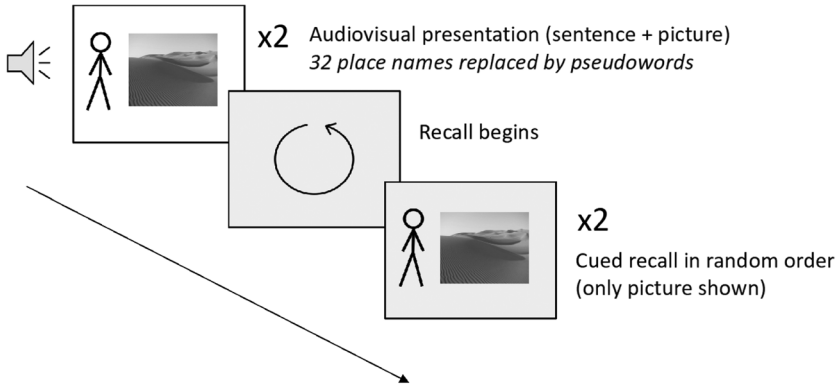

Figure 2 Training phase procedure.

be almost entirely in English apart from one word, which would be a foreign word. Participants were encouraged to discover the overt rule. They were told that different words were associated with different character positions on screen, and that they would be tested on the association later. After instructions, participants did a practice block for each task and then began the task. The whole session lasted about 45 minutes.

Training Phase

The training phase consisted of 112 sentences, divided into seven blocks of 16. Within each block, sentences with corresponding pictures were presented in sets of two trials, followed by two recall trials (showing the picture only, with no audio), one for each sentence (Figure 2; for more details, see Appendix S1 in the Supporting Information online). Trial order within each pair of recall trials was randomized, to ensure that participants would form an association between each sentence and the relevant picture, rather than simply memorizing the prepositions in the order in which they had heard them (since prepositions were the only part of the sentence that could not be reconstructed from the picture).

Recall Test

The recall test consisted of 88 sentences, divided into 11 blocks of eight items each. It consisted entirely of trained item recall and generalization trials, with no further exposure to the stimulus sentences. Participants were told that they were being asked to recall sentences from the training phase, using the

pictures as a cue. At this stage, participants were not explicitly told to repeat the pseudoword matching the place photograph, but only to recall the sentence including the preposition. There were eight blocks of trained items (64 matched items) and three blocks of generalization items (24 items). Trained and generalization item blocks were randomly intermixed, and the items within individual blocks were presented in random order. Participants were not told about the presence of generalization items and believed that all the pictures they saw had already been presented during training. None of the participants reported becoming aware of this manipulation at any stage during or after the experiment.

Comprehension Test

Given the novelty of the production testing measure we used, we also included a comprehension test as a potentially more sensitive measure of implicit knowledge. Participants completed the comprehension test immediately after the recall test. We included this task so that if no effects were detected in production, we might have additional evidence to help us determine whether this was due to a lack of knowledge or to low task sensitivity for the production task. Given that the results of the production test are informative enough with regard to our central research questions, we will not report findings from the comprehension test here; but information on test design, analysis, and results can be found in Appendix S2 in the Supporting Information online.

Debriefing Questionnaire

At the end of the session, participants were asked to fill in a debriefing questionnaire designed to assess awareness of the overt and hidden rules, before being informed of the nature of the experiment. Awareness of the overt and hidden rules was assessed from answers to the following questions: (a) Did you think the use of the words *ro*, *gi*, *wa*, and *ne* was governed by any rules? (b) Did you think it depended on whether the character was inside or outside/near the place pictured? (c) Did you think there were any other rules? and (d) Could you give a rough translation of the words *ro*, *gi*, *wa*, and *ne*? The full questionnaire can be found at https://doi.org/10.17605/OSF.IO/86b7c and in Bovolenta and Williams (2022).

Data Analysis

Recall Test Scoring

Each trial in the production task was marked as either correct or incorrect depending on whether participants successfully reproduced the correct preposition as part of the sentence; failure to produce the correct character name or

place name was not marked as an error (see Appendix S3 in the Supporting Information online for the exact scoring criteria used). When scoring the task, we extracted two measures of accuracy: overt rule and hidden rule accuracy. Overt rule accuracy was the proportion of all recall trials in which participants correctly used the overt rule, that is, when out of four possible prepositions, they used one of the two that were appropriate for that condition (system or random, as cued by the character's position on screen relative to the place photograph). To calculate hidden rule accuracy, we retained only the trials in which participants had used the overt rule correctly, and calculated the proportion of those trials for which participants correctly recalled the exact preposition used in the item during training. Given that the choice was between two possible prepositions, chance level for hidden rule accuracy was 50%.

Debriefing Questionnaire Scoring

Participants were classed as aware of the hidden rule if their answers to the awareness questions made reference to the relevant distinction, as "open" versus "enclosed" spaces or "indoor" versus "outdoor" spaces, for example, or in any other recognizable way, even indirectly (see Appendix S3 in the Supporting Information online for further information).

Planned Analyses

Given that the focus of the experiment was on implicit learning and therefore on unaware participants, we ran two separate models for aware and unaware participants. The dependent variable in the recall test was hidden rule accuracy (i.e., accuracy on trials where participants had responded correctly to the overt rule, by producing one of the two prepositions that were associated with the character's position on screen). This meant that only a subset of trials was retained for analysis. The exact number of trials retained in each analysis can be found in the summary tables for individual models (see below).

We analyzed production data using mixed-effect modelling, implemented in R version 4.0.3 (R Core Team, 2020) using the lme4 package (Bates et al., 2015). We used generalized linear mixed-effect models for binomial data to analyze accuracy scores from the recall test and created separate models for the aware and unaware groups, and for different types of items. For trained items, the independent variables were condition (levels: system, random) and block number (scaled and centered), entered as a continuous variable. For generalization items, we used intercept-only models to assess deviation from chance performance, following Wang (2020). Alpha was set at .05 (see Appendix S3 in the Supporting Information online for further details).

Results

We excluded participants who did not become aware of the overt rule ($n = 6$). This was taken as a sign of lack of engagement with the task, given that participants were told about the existence of the rule in the instructions. A further three participants were excluded from the analysis for failing to perform all the tasks in the experiment. A total of 33 participants were included in the analysis (22 females, $M = 20$ years, $SD = 2.09$). Based on their awareness of the hidden rule as assessed by the debriefing questionnaire, participants were either classed as aware ($n = 16$) or unaware ($n = 17$).

Debriefing Questionnaire

The debriefing questionnaire was scored independently by two researchers involved in the project. Interrater agreement on overt rule awareness was 100%. Interrater agreement on hidden rule awareness was initially 97%, and complete agreement was achieved following discussion. Of the participants included in the analysis, 10 made reference to the distinction between indoor and outdoor places (mentioning both dimensions directly or indirectly), and were therefore classed as being aware, following the criteria described in the Data Analysis section. A further six participants did not mention the indoor/outdoor distinction, but mapped at least a subset of the prepositions onto the English *in* versus *on* distinction. Given the partial overlap between this distinction and the Czech *v* versus *na* distinction, we ran further analyses to establish to what extent transferring the English rule could have aided participants in the testing phase. We conducted an online norming study with participants recruited from the same population as our experimental sample ($N = 49$), which showed that 80% of the items in the testing phase could be answered correctly by mapping the *in* versus *on* distinction onto the prepositions governed by the hidden rule. Therefore, participants who explicitly transferred the *in* versus *on* rule were included in the aware group.

Recall Test: Trained Items

Descriptive statistics for the training phase data can be found in Appendix S4 in the Supporting Information online. Both groups had high overt rule accuracy, with no significant differences between groups (Table 3 and Figure 3), which is not surprising because all participants included in the analysis had become aware of the overt rule by the end of the experiment. There was more variation in hidden rule accuracy (Table 3 and Figure 4). Even though the hidden rule technically applied to system items only, we refer to "hidden rule accuracy" for both system and random items, by which we mean how accurate participants

Table 3 Means (standard deviations) for overt and hidden rule accuracy for trained items in the recall test

	Overt rule accuracy		Hidden rule accuracy	
Group	System	Random	System	Random
Aware ($n = 16$)	.80 (.24)	.74 (.35)	.75 (.27)	.53 (.15)
Unaware ($n = 17$)	.79 (.23)	.76 (.26)	.58 (.15)	.49 (.13)

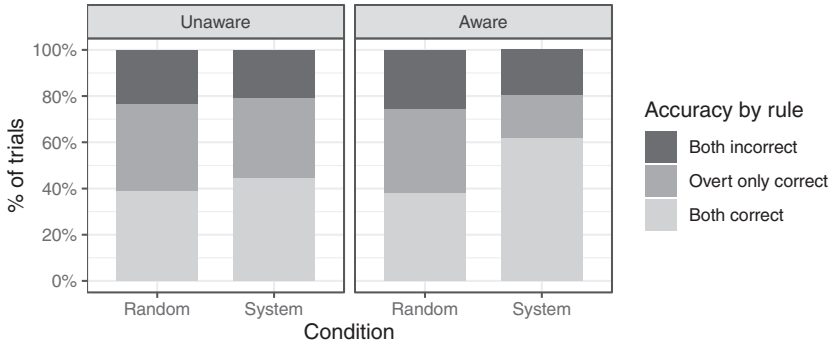

Figure 3 Combined overt and hidden rule accuracy in the recall test (trained items).

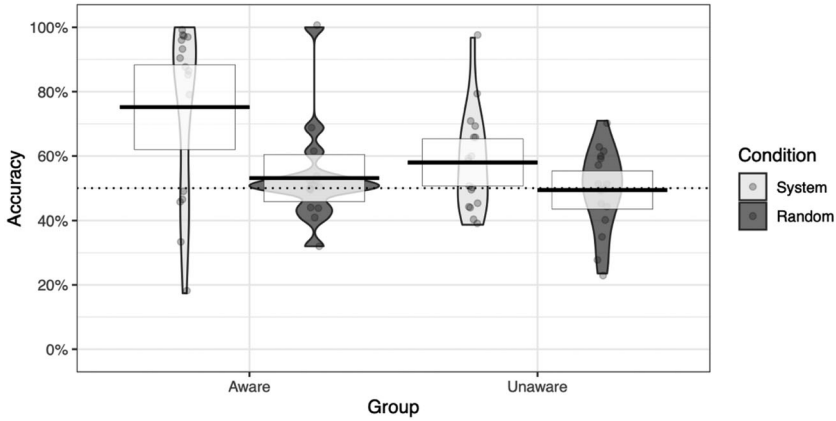

Figure 4 Hidden rule accuracy in the recall test (trained items). Horizontal bars represent group means, shaded rectangles 95% confidence intervals, and dotted line 50% chance level.

Table 4 Final model for accuracy in the recall test (trained items) by the aware group and equivalent model for the unaware group

	Recall accuracy (aware)			Recall accuracy (unaware)		
Predictors	*OR*	95% CI	*p*	*OR*	95% CI	*p*
(Intercept)	1.00	[0.65, 1.55]	.991	1.03	[0.80, 1.33]	.807
Condition (System)	3.69	[2.65, 5.14]	< .001	1.27	[0.97, 1.67]	.087
Observations		792			149	
Marginal/conditional R^2		.100/.231			.000/.097	
Random effects		(1 | subject)			(1 | subject)	

were in recalling the specific preposition for a given item, in cases where they had got the overt rule correct (i.e., the preposition they produced was one of the two allowed for that condition). In the random condition, we would expect their hidden rule accuracy to be at chance level (50%), due to the lack of a systematic relationship between prepositions and places that may help participants in recalling the correct preposition.

Both groups showed higher hidden rule accuracy for system compared to random items (Figure 4 and Table 4). The size of the effect was greater for the aware group ($OR = 3.69$, 95% CI [2.65, 5.14]) than for the unaware group ($OR = 1.27$, 95% CI [0.97, 1.67]), and it only reached statistical significance for the aware group. Expressed as Cohen's *d*, the effect sizes were 0.97 for the aware and 0.61 for the unaware group, which in the context of L2 research constitute a medium and small effect, respectively (Plonsky & Oswald, 2014; see Appendix S5 in the Supporting Information online for details). Since trained items in the system and random condition were closely matched (they all included the same place photographs), we interpret the difference in recall accuracy between conditions as evidence of hidden rule knowledge. Knowing the hidden rule would have helped participants to "reconstruct" the correct preposition in the system condition because it was cued by the picture, but not in the random condition. Performance on random items was at chance level for both aware, $t(15) = 0.80$, $p = .400$, and unaware, $t(16) = -0.20$, $p = .800$, groups. This indicates that participants had on average no reliable memory of individual items. By contrast, performance on system items was above chance for both aware, $t(15) = 4.00$, $p = .002$, and unaware, $t(16) = 2.00$, $p = .050$, groups. This further suggests that above-chance performance in system items was driven by rule knowledge: Given that accuracy in the random condition was at chance level, it seems reasonable to take this as the null hypothesis in

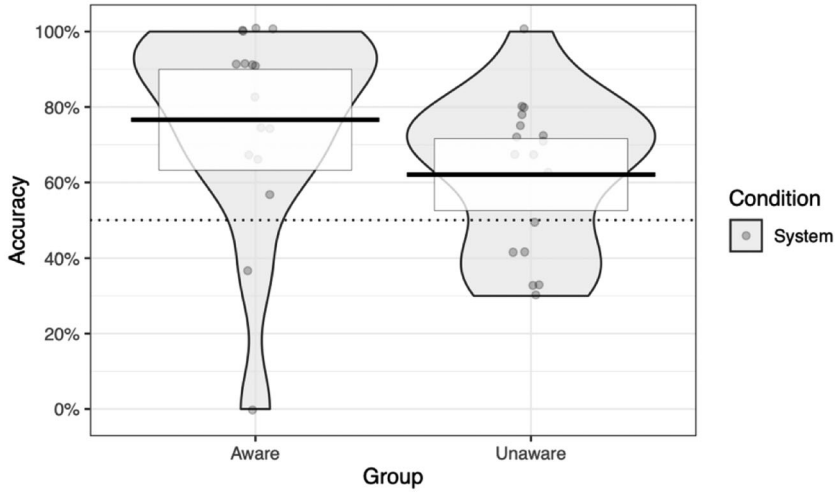

Figure 5 Hidden rule accuracy in the recall test (generalization items). Horizontal bars represent group means, shaded rectangles 95% confidence intervals, and dotted line 50% chance level.

a one-sample test. Split-half reliability for trained items was .91 in the system condition, compared to .45 for the random condition, suggesting that system items were reliably tapping into hidden rule knowledge. Breaking down scores by group, reliability was higher for the aware group (.94 for system items) than for the unaware group (.78 for system items).

Recall Test: Generalization Items

Generalization items were novel system items that were created from place photographs previously seen in training, but in the opposite condition relative to the one used in training (i.e., with the character in a different position), and they were presented intermixed with trained items during the recall test. Hidden rule accuracy on generalization items was .77 ($SD = .27$) for aware participants and .62 ($SD = .20$) for the unaware (Figure 5). Both groups were significantly above chance in their accuracy (Table 5), but again the size of the effect was greater in the aware group ($OR = 4.59$, 95% CI [1.78, 11.84]) than in the unaware group ($OR = 1.66$, 95% CI [1.05, 2.63]). The effect sizes for each group were similar to those observed with trained items ($d = 0.96$ for aware and $d = 0.60$ for unaware), again constituting a medium and small effect, respectively, based on the guidelines proposed by Plonsky and Oswald (2014). Since the picture combinations to which participants were responding were

Table 5 Intercept-only models for hidden rule accuracy in the recall test (generalization items)

	Accuracy (aware)			Accuracy (unaware)		
Predictors	*OR*	95% CI	*p*	*OR*	95% CI	*p*
(Intercept)	4.59	[1.78, 11.84]	.002	1.66	[1.05, 2.63]	.030
Observations		150			149	
Marginal/conditional R^2		.000/.411			.000/.097	
Random effects		(1 | subject)			(1 | subject)	

novel, the sentences produced by participants were novel, too, even though they believed that they were doing a recall task. Therefore, above-chance accuracy in producing novel system items indicates that participants had knowledge of the hidden rule, and that they were relying on this knowledge to select the appropriate prepositions when producing new sentences.

Split-half reliability for generalization items was lower than that for trained items, possibly due to the smaller number of items. Overall reliability for these items was .67, but breaking down the score by group showed a large difference between groups, with .47 reliability for the unaware and .79 for the aware participants. This difference could be symptomatic of the fact that the two groups were relying on different kinds of knowledge to provide their answers: Whereas the aware group had access to explicit knowledge, the unaware group could only rely on implicit knowledge acquired through brief exposure, which was likely to be quite weak and fragmentary.

Post Hoc Analyses

Following our planned analysis, we conducted two post hoc analyses to gain further insight into the nature of the knowledge acquired by participants, and the nature of the learning process itself. First, we looked at correlations between recall of system items and generalization accuracy, to establish whether recall provided a measure of rule knowledge. Second, we analyzed recall of training items based on whether they contained a pseudoword place noun, to establish whether this manipulation may have affected participants' ability to recall the items.

Correlation Between Trained and Generalization Items

To explore the extent to which recall of trained system items in the recall task provided a measure of rule knowledge, we examined the correlation between hidden rule accuracy in recall of trained system items and hidden rule accuracy

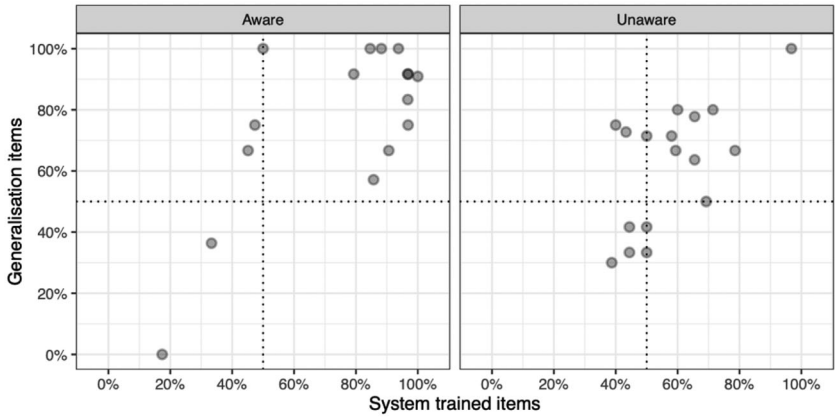

Figure 6 Correlations between system trained item recall accuracy and generalization accuracy. Dotted lines mark 50% chance level.

in the production of generalization items. A correlation between accuracy on these items would indicate that performance was driven by the same underlying factor, suggesting that recall of system trained items was tapping into hidden rule knowledge in the same way as generalization items were. On the other hand, we would not expect to see a correlation between generalization item accuracy and recall for random trained items, because the latter could only be driven by item memory, and not by knowledge of the hidden rule (which did not apply to random items).

That is what we found, in both groups. Accuracy on generalization items positively correlated with average recall accuracy for system items in the recall test for both aware participants (Pearson's $r = .69$, $p = .003$) and unaware participants ($r = .61$, $p = .009$; Figure 6). By contrast, there was no significant correlation between generalization accuracy and recall of random items for either the aware group ($r = .44$, $p = .090$) or the unaware group ($r = .15$, $p = .600$; Figure 7). Taken together with at-chance recall performance on random items, these results suggest that accuracy in recall for system items was at least partially driven by hidden rule knowledge.

Effect of Training Item Manipulation (English Versus Pseudoword Place Nouns)

During the training phase, a subset of training items (32 out of 112) had place nouns replaced by pseudowords (e.g., "Harry is gi *dreet*" for "Harry is gi kitchen") in order to increase memory load. To determine whether this feature

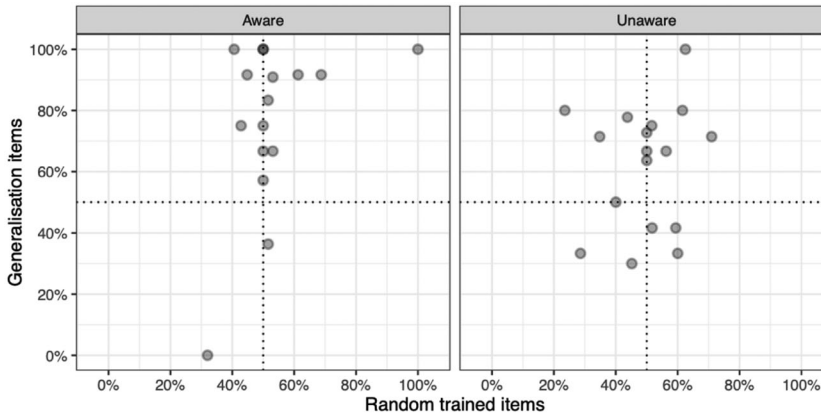

Figure 7 Correlations between random trained item recall and generalization accuracy. Dotted lines mark 50% chance level.

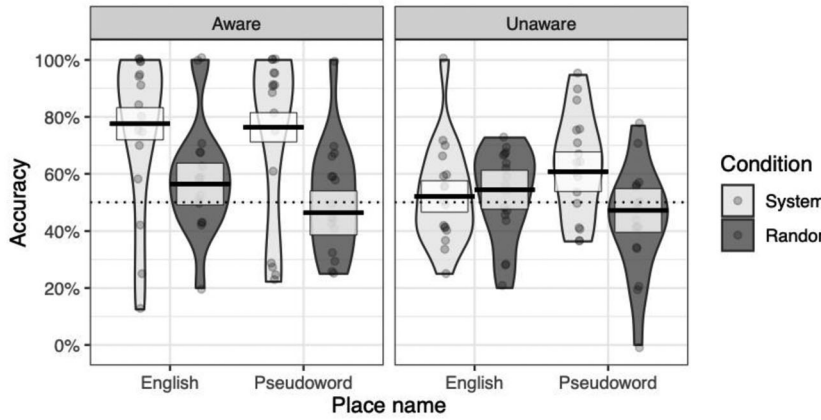

Figure 8 Hidden rule (recall) accuracy for trained items by type of place noun used during training (English vs. pseudoword). Horizontal bars represent group means, shaded rectangles 95% confidence intervals, and dotted line 50% chance level.

had any impact on the recall of prepositions used in those specific items during the recall test, we ran an additional post hoc analysis on trained items. We compared recall accuracy for items that were encountered during the training phase with the relevant English place noun to those where the English noun had been replaced by a pseudoword (Figure 8). Details of the analysis and results can be found in Appendix S4 in the Supporting Information online.

The key finding was an interaction between condition and place noun (English vs. pseudoword): In both groups, the effect of condition on recall was greater for items in which the English place noun was replaced by a pseudoword. At the group level, unaware participants were significantly more accurate in their recall of trained system items relative to random ones if they had encountered them with pseudoword place nouns during training, $\chi^2(1) = 7.98$, $p = .019$, but not if they had seen them in the English version, $\chi^2(1) = 0.00$, $p = 1.00$. We discuss this unexpected finding below.

Discussion

The research question we investigated in this study was whether participants could develop implicit knowledge of a linguistic rule and use it in production without being aware of it. In a test administered after the training phase, participants had to recall items from training (trained items), and produce new ones (generalization items) under the guise of a cued recall test. We found that participants who were unaware of the correlation between prepositions and the nature of the space (as open or enclosed) in the system items nevertheless performed significantly above chance in producing the correct preposition for both trained and generalization items, suggesting that they had acquired implicit knowledge of the distinction and could use it in production. By contrast, their recall of prepositions in the random condition (where there was no regularity in preposition usage) was at chance, showing a lack of verbatim memory for preposition usage. The use of generalization items allowed us to find evidence of productive knowledge in the absence of awareness, even after limited exposure. This suggests that the knowledge acquired through implicit learning is immediately available to production, provided that an adequately sensitive task is used. It also supports our interpretation of findings from previous research that did not see learning effects in production (Denhovska & Serratrice, 2017), which we hypothesized were due to low task sensitivity, rather than a need for explicit knowledge in production.

Considering this together with the finding that there was a correlation between recall accuracy for trained and generalization items, we conclude that performance in both cases was at least in part a reflection of generalized, and yet inexpressible, knowledge of the rule. Together, these findings support our hypothesis that recall of trained system items would tap into rule knowledge by a process of sentence regeneration. Even though we did not use the elicited imitation task, our findings also support the regeneration hypothesis (Lombardi & Potter, 1992; Potter & Lombardi, 1990) of sentence recall. The regeneration hypothesis maintains that recalling a sentence in a familiar language involves

decoding and then reencoding the sentence upon recall, whereas a sentence in an unknown language can only be stored and recalled as a string of phonemes. The fact that our participants performed above chance in recalling items that could be reencoded from rule knowledge (system prepositions) but not those that could only be stored as phonological strings (random prepositions) shows that recall does involve regeneration, at least at the level of difficulty we employed in our recall test.

Our results are compatible with recent findings on the role of memory and prior linguistic knowledge in elicited recall (Culbertson et al., 2020; Isbilen et al., 2018, 2020). In a study with L2 Spanish learners, Culbertson et al. (2020) tested participants' written recall of naturalistic sentences from Spanish videos. Results showed that recall performance was a better predictor of the learners' translation abilities than a standardized Spanish test using multiple-choice comprehension questions. Furthermore, auditory memory ability (measured by an English recall task) played a small part in Spanish recall and only weakly correlated with translation abilities, adding to the evidence that elicited imitation is sensitive to grammatical knowledge and does not simply rely on phonological short-term memory.

In addition to recall, we used a false memory paradigm to test for productive generalization of the hidden rule to new instances. Isbilen et al. (2018) used elicited recall of novel strings to demonstrate generalization of a learned grammar to a new syllable set. In this study, we went beyond that and tested productive generalization by inducing participants to produce entirely new utterances following the rule, even though they were not aware that they were producing new sentences and believed that they were still doing a recall task. Our findings show that the knowledge acquired by participants was sufficiently strong to not only aid reconstruction during recall (where it could potentially still benefit from the scaffolding provided by short-term memory for the string just heard), but also generate entirely new utterances based on the rule acquired, even in the absence of rule awareness. Generalization performance positively correlated with recall of strings from training (as also observed by Isbilen et al., 2018), providing further evidence of the sensitivity of recall tasks to grammatical knowledge.

The effect size we observed for the unaware group is typical of studies on implicit learning where comprehension is measured. The observed effect for generalization in the unaware group ($d = 0.60$) was based on a mean accuracy score of 62%. This is very close to the findings of similar studies (Chen et al., 2011; Williams, 2005), which recorded mean generalization accuracy scores ranging from 58% (Chen et al., 2011) to 60.8% and 64% (Williams, 2005)

among unaware participants. These scores are typically much lower than those achieved by participants who have developed explicit knowledge of the rule. However, they are still remarkable if we consider that they emerge after only limited exposure to a novel language, and without the help of conscious knowledge, insofar as we could detect it. The effect of condition (system vs. random) on recall of trained items for the unaware group, although statistically not significant, was of similar magnitude ($d = 0.61$). The effect on trained items, although of similar magnitude to the generalization ones, was not statistically significant due to greater variability in the data: The generalization effect was assessed against chance, whereas the trained item effect was evaluated against the random condition, so sensitivity was reduced by variability in the baseline. A sensitivity power analysis carried out after the study suggests that the study may have been underpowered with respect to recall, specifically for the effect size observed in the unaware group (see Appendix S5 in the Supporting Information online for details). Given that the sample size was not large enough to reliably detect an effect of this size, if present in the population, we can speculate that it might have emerged as a significant effect had we used a larger sample size. Further research on this topic should employ larger sample sizes in order to have adequate power to detect the size of effect we found.

Nature of the Knowledge Acquired

The results of the recall test showed that participants had acquired knowledge of the hidden rule and could use it in production. To establish whether that knowledge was the result of implicit learning, we tested participants' awareness of the rule at the end of the experiment by means of a debriefing questionnaire. Even participants who did not show awareness of the rule (the unaware group) showed an effect of rule knowledge in the test, suggesting that they had learned the hidden rule implicitly.

The use of debriefing questionnaires (retrospective verbal report) as a measure of awareness, however, is not without problems. We know, for instance, that participants vary in their ability to verbalize their knowledge and may not be able to fully verbalize it, leading to potentially less sensitivity (Reber, 1967; Rebuschat, 2013). Additionally, it is possible that the knowledge participants are asked about in the questionnaire (and are unable to report) may not be the same knowledge they draw upon when they are tested on the rule (Shanks & St. John, 1994). For instance, in a grammaticality judgment task following exposure to an artificial grammar, participants may rely on their memory of grammatical substrings from training items ("chunks") to help them determine which strings are grammatical (Knowlton & Squire, 1996). In such

cases, above-chance performance on the test coupled with a lack of awareness of the grammatical rules may not be enough to conclude that participants acquired implicit knowledge of the grammatical rules.

Such problems are arguably less of a concern in our case. With regard to sensitivity, two features of our study helped to mitigate any potential issues deriving from our use of retrospective verbal report as a measure of awareness. First, the rule we employed, unlike finite-state grammars such as those employed in artificial grammar learning studies, was very simple. Once a participant has developed awareness of it, a distinction between indoor and outdoor spaces should not be difficult to articulate, unlike a set of grammar rules used to concatenate letters into strings. Second, we scored the debriefing questionnaire in a way that avoided confounds from other factors explaining accuracy. Participants who reported even partial awareness, and those who transferred a related first language category ("in" vs. "on") that could have conferred on them an advantage in the test, were all included in the aware group. This resulted in a relatively high number of participants being classed as aware (roughly 50%) and increased the likelihood that those in the unaware group did, in fact, have no explicit knowledge of the rule.

With regard to the second potential drawback, that the knowledge tapped by the test may not be the same as that targeted in the debriefing questionnaire, our study design again minimized the risks associated with retrospective verbal report. Although "chunking" was not a possibility in this study due to the nature of the rule, it is theoretically possible that production may have been driven by factors that participants found difficult to report, such as "microrules" based, for example, on perceptual similarities between training and test items. Of course, this may just be characteristic of the kind of low-quality implicit knowledge that is acquired from brief exposure, but the issue remains whether lack of reportability is indicative of lack of awareness. Trial-by-trial subjective measures of confidence or knowledge "source" (e.g., Rebuschat et al., 2015) may shed light on this issue, although, as noted in the Background Literature section, some studies have found that explicit knowledge is not detectable by these measures when it is present by the criterion of verbal report (Rebuschat et al., 2015; Zhao et al., 2021), suggesting that the latter may be, if anything, more conservative (perhaps because it identifies people who become aware late in the test phase or only at the point of being debriefed). Nevertheless, unreportable conscious knowledge will always remain a logical possibility that has to be acknowledged, although it could be argued that the higher the proportion of participants classified as aware of the rule (implying that it is readily reportable), the less likely this becomes.

Effect of Training Paradigm

In a post hoc analysis, we found that the advantage observed in recall accuracy for trained system items was largely driven by the items containing a pseudoword instead of an English noun. This was not predicted by our experimental hypotheses, as this feature of the design was simply intended to increase memory load for participants. Indeed, based on this we may have predicted items with pseudowords to be harder to remember, due to the increased amount of novel material. Instead, we saw that participants in both groups showed better recall for these items or, more specifically, better recall for the prepositions used in these items, which was how we assessed recall accuracy. It is possible that the introduction of pseudowords led to better recall for system items, but not for random ones, because it forced learners to pay more attention to the relevant properties of the places pictured. They had to do so because a potential aid to recall (the English label for the place) had been removed: Therefore, to correctly recall the preposition, they had to associate it with the visual depiction of the place, which would lead to greater activation of features relevant to the rule (open vs. enclosed space). For system items, this had a beneficial effect, because the same prepositions were always associated with similar kinds of place images, thus making it easier to encode and recall them. Therefore, replacing place nouns with pseudowords may have constituted something akin to "desirable difficulties" in the practice literature (Bjork, 1994; Bjork & Kroll, 2015): a manipulation that increased difficulty in the immediate term, but ultimately aided acquisition by stimulating the processes necessary for it (in this case, forming an association between prepositions and relevant physical places). However, this is merely a tentative explanation; further research would be needed to replicate this phenomenon and potentially explore it further.

Limitations and Future Directions

One limitation of this study was low power to detect the learning effect in the unaware group. Further replications should employ larger sample sizes to ensure validity. Another limitation of the study is that we only focused on the effect of production practice during exposure on subsequent cued recall. This was to maximize the chances of obtaining an effect on production. Within memory research, the principle of transfer-appropriate processing dictates that the more similar the encoding and retrieval conditions are, the greater the probability of retrieval becomes (Morris et al., 1977). We also have to acknowledge the possibility that, at the implicit level, training effects are skill-specific; in the case of this study, the trained skill would be establishing a mapping from a semantic condition (enclosed/open space) to preposition selection in production.

There is evidence for such skill specificity in the effect of long-term practice (DeKeyser & Sokalski, 1996), which presumably reflects the process of proceduralization, but it is not clear whether the same applies to implicit knowledge gained after brief exposure. Knowledge at that point may not be skill-specific, as shown by syntactic priming from comprehension to production over small time scales (Bock et al., 2007; Branigan et al., 2000). Transfer from production practice to comprehension has been shown in artificial language learning research after just one session (Hopman & MacDonald, 2018). As in the present study, participants in the study by Hopman and MacDonald (2018) learned form–meaning associations with the aid of visual scenes. However, the training procedure used in that study encouraged conscious processing, and whereas the study showed learning effects on comprehension with a range of different tests, the relative contributions of implicit and explicit knowledge to these effects are not known. It therefore remains an open question whether implicit knowledge gained after brief training in a purely comprehension-based task would transfer to production. Here, we have taken a first step by providing evidence that implicit knowledge acquired after brief exposure during a primarily production-based task can produce grammatical biases in production without awareness.

Conclusion

In this study, we expanded the understanding of implicit language learning by examining the nature and availability of knowledge acquired through an implicit learning paradigm. Specifically, we investigated whether newly acquired implicit knowledge can be used in spoken production. Production in implicit learning is an underresearched field, partly for methodological reasons: not only the inherent difficulty of eliciting production, but also the difficulty of eliciting production of a specific rule, while ensuring that participants remain unaware of that rule. We tackled these challenges with a methodology based on recall and false memory, which allowed us to test for implicit knowledge in two ways: first, by measuring its effect on sentence recall (as already done in proficiency tests such as the elicited imitation task), and second, by inducing participants to produce novel instances of the rule, while believing that they were still doing a recall task. Our results show that newly acquired implicit knowledge of a language rule can be recruited in production immediately after learning. These findings have important implications for understanding the role of implicit knowledge in language learning. In L2 acquisition, production normally follows comprehension, suggesting that more established representations may be needed for production. However,

our findings show that under the right circumstances, even new implicit knowledge—which is presumably still weak and fragmentary—can emerge in spoken production, guiding participants' lexical choices in the absence of awareness.

Final revised version accepted 13 September 2022

Open Research Badges

This article has earned an Open Materials badge for making publicly available the components of the research methods needed to reproduce the reported procedure. All materials that the authors have used and have the right to share are available at https://doi.org/10.17605/OSF.IO/86b7c and http://www.iris-database.org. All proprietary materials have been precisely identified in the manuscript.

Note

1 On the other hand, if a matched item did *not* include a pseudoword, participants would encounter the English name for the place pictured in the item twice (once when it occurred in the system condition and once when it occurred in the random condition).

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Batterink, L. J., Cheng, L. Y., & Paller, K. A. (2016). Neural measures reveal implicit learning during language processing. *Journal of Cognitive Neuroscience*, *28*(10), 1636–1649. https://doi.org/10.1162/jocn_a_00985

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). MIT Press.

Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *The American Journal of Psychology*, *128*(2), 241–252. https://doi.org/10.5406/amerjpsyc.128.2.0241

Bock, K., Dell, G. S., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition*, *104*(3), 437–458. https://doi.org/10.1016/j.cognition.2006.07.003

Bovolenta, G., & Williams, J. N. (2022). *Materials. Datasets from "Implicit learning in production: Productive generalization of new form–meaning connections in the*

absence of awareness" [Collection: Stimuli and experiment files]. IRIS Database, University of York, UK. https://doi.org/10.48316/m377-zb85

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition*, *75*(2), B13–25. https://doi.org/10.1016/s0010-0277(99)00081-5

Chambers, K. E., Onishi, K. H., & Fisher, C. (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition*, *87*(2), B69–B77. https://doi.org/10.1016/s0010-0277(02)00233-0

Chen, W., Guo, X., Tang, J., Zhu, L., Yang, Z., & Dienes, Z. (2011). Unconscious structural knowledge of form–meaning connections. *Consciousness and Cognition*, *20*(4), 1751–1760. https://doi.org/10.1016/j.concog.2011.03.003

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*(1), 3–42.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences*, *2*(10), 406–416. https://doi.org/10.1016/S1364-6613(98)01232-7

Cleeremans, A., & Jiménez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. In R. M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness* (pp. 401–418). Psychology Press.

Culbertson, G., Andersen, E., & Christiansen, M. H. (2020). Using utterance recall to assess second language proficiency. *Language Learning*, *70*(S2), 104–132. https://doi.org/10.1111/lang.12399

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22. https://doi.org/10.1037/h0046671

DeKeyser, R. (2003). Implicit and explicit learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 312–348). Blackwell. http://doi.wiley.com/10.1002/9780470756492.ch11

DeKeyser, R., & Sokalski, K. J. (1996). The differential role of comprehension and production practice. *Language Learning*, *46*(4), 613–642. https://doi.org/10.1111/j.1467-1770.1996.tb01354.x

Dell, G. S., Reed, K. D., Adams, D. R., & Meyer, A. S. (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1355–1367. https://doi.org/10.1037/0278-7393.26.6.1355

Denhovska, N., & Serratrice, L. (2017). Incidental learning of gender agreement in L2. *Journal of Psycholinguistic Research*, *46*(5), 1187–1211. https://doi.org/10.1007/s10936-017-9487-x

Dienes, Z., & Perner, J. (1999). A theory of implicit and explicit knowledge. *Behavioral and Brain Sciences*, *22*(5), 735–808. https://doi.org/10.1017/S0140525X99002186

Ellis, N. C. (2011). Implicit and explicit SLA and their interface. In C. Sanz & R. Leow (Eds.), *Implicit and explicit language learning: Conditions, processes, and knowledge in SLA* (pp. 35–47). Georgetown University Press.

Ellis, R. (2012). *Language teaching research and language pedagogy*. Wiley. https://doi.org/10.1002/9781118271643

Erlam, R. (2009). The elicited oral imitation test as a measure of implicit knowledge. In R. Ellis, S. Loewen, C. Elder, H. Reinders, R. Erlam, & J. Philp (Eds), *Implicit and explicit knowledge in second language learning, testing and teaching* (pp. 65–93). Multilingual Matters. https://doi.org/10.21832/9781847691767-005

Faretta-Stutenberg, M., Morgan-Short, K., Granena, G., Koeth, J., Lee-Ellis, S., & Lukyanchenko, A. (2011). Learning without awareness reconsidered: A replication of Williams (2005). In G. Granena, J. Koeth, S. Lee-Ellis, A. Lukyanchenko, G. Prieto Botana, & E. Rhoades (Eds.), *Selected proceedings of the 2010 Second Language Research Forum: Reconsidering SLA research, dimensions, and directions* (pp. 18–28). http://www.lingref.com/cpp/slrf/2010/paper2612.pdf

Francis, A. P., Schmidt, G. L., Carr, T. H., & Clegg, B. A. (2009). Incidental learning of abstract rules for non-dominant word orders. *Psychological Research*, *73*(1), 60–74. https://doi.org/10.1007/s00426-008-0138-6

Fukuta, J., & Yamashita, J. (2021). The complex relationship between conscious/unconscious learning and conscious/unconscious knowledge: The mediating effects of salience in form–meaning connections. *Second Language Research*. Advance online publication. https://doi.org/10.1177/02676583211044950

Goo, J., Granena, G., Yilmaz, Y., & Novella, M. (2015). Implicit and explicit instruction in L2 learning: Norris & Ortega (2000) revisited and updated. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages* (pp. 443–482). John Benjamins. https://www.jbe-platform.com/content/books/9789027268723-sibil.48.18goo

Hama, M., & Leow, R. P. (2010). Learning without awareness revisited. *Studies in Second Language Acquisition*, *32*(3), 465–491. https://doi.org/10.1017/S0272263110000045

Hopman, E. W. M., & MacDonald, M. C. (2018). Production practice during language learning improves comprehension. *Psychological Science*, *29*(6), 961–971. https://doi.org/10.1177/0956797618754486

Hulstijn, J. H. (2005). Theoretical and empirical issues in the study of implicit and explicit second-language learning: Introduction. *Studies in Second Language Acquisition*, *27*(2), 129–140. https://doi.org/10.1017/S0272263105050084

Isbilen, E. S., Frost, R. L. A., Monaghan, P., & Christiansen, M. H. (2018). Bridging artificial and natural language learning: Comparing processing- and reflection-based measures of learning. In C. Kalish, M. Rau, J. Zhu, & T. T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1856–1861). Cognitive Science Society. https://cognitivesciencesociety.org/wp-content/uploads/2019/01/cogsci18_proceedings.pdf

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2020). Statistically induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science*, *44*(7), Article e12848. https://doi.org/10.1111/cogs.12848

Kerz, E., Wiechmann, D., & Riedel, F. B. (2017). Implicit learning in the crowd: Investigating the role of awareness in the acquisition of L2 knowledge. *Studies in Second Language Acquisition*, *39*(4), 711–734. https://doi.org/10.1017/S027226311700002X

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), 169–181. https://doi.org/10.1037/0278-7393.22.1.169

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon.

Lee, M.-W., & Williams, J. N. (1997). Why is short-term sentence recall verbatim? An evaluation of the role of lexical priming. *Memory & Cognition*, *25*(2), 156–172. https://doi.org/10.3758/BF03201109

Leung, J. H. C., & Williams, J. N. (2011). The implicit learning of mappings between forms and contextually derived meanings. *Studies in Second Language Acquisition*, *33*(1), 33–55. https://doi.org/10.1017/s0272263110000525

Leung, J. H. C., & Williams, J. N. (2012). Constraints on implicit learning of grammatical form-meaning connections. *Language Learning*, *62*(2), 634–662. https://doi.org/10.1111/j.1467-9922.2011.00637.x

Leung, J. H. C., & Williams, J. N. (2014). Crosslinguistic differences in implicit language learning. *Studies in Second Language Acquisition*, *36*(4), 733–755. https://doi.org/10.1017/S0272263114000333

Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, *39*(4), 593–620. https://doi.org/10.1017/S0272263116000358

Lombardi, L., & Potter, M. C. (1992). The regeneration of syntax in short term memory. *Journal of Memory and Language*, *31*(6), 713–733. https://doi.org/10.1016/0749-596X(92)90036-W

Marsden, E., Williams, J. N., & Liu, X. (2013). Learning novel morphology: The role of meaning and orientation of attention at initial exposure. *Studies in Second Language Acquisition*, *35*(4), 619–654. https://doi.org/10.1017/S0272263113000296

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

Paciorek, A., & Williams, J. N. (2015a). Implicit learning of semantic preferences of verbs. *Studies in Second Language Acquisition*, *37*(2), 359–382. https://doi.org/10.1017/S0272263115000108

Paciorek, A., & Williams, J. N. (2015b). Semantic generalization in implicit language learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(4), 989–1002. https://doi.org/10.1037/xlm0000100

Pacton, S., Perruchet, P., Fayol, M., & Cleeremans, A. (2001). Implicit learning out of the lab: The case of orthographic regularities. *Journal of Experimental Psychology: General*, *130*(3), 401–426. https://doi.org/10.1037/0096-3445.130.3.401

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*(6), 633–654. https://doi.org/10.1016/0749-596X(90)90042-X

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org

Radulescu, S., Wijnen, F., & Avrutin, S. (2020). Patterns bit by bit: An entropy model for rule induction. *Language Learning and Development*, *16*(2), 109–140. https://doi.org/10.1080/15475441.2019.1695620

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*(6), 855–863. https://doi.org/10.1016/S0022-5371(67)80149-X

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*(3), 595–626. https://doi.org/10.1111/lang.12010

Rebuschat, P., Hamrick, P., Riestenberg, K., Sachs, R., & Ziegler, N. (2015). Triangulating measures of awareness: A contribution to the debate on learning without awareness. *Studies in Second Language Acquisition*, *37*(2), 299–334. https://doi.org/10.1017/S0272263115000145

Rebuschat, P., & Williams, J. N. (2012). Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics*, *33*(4), 829–856. https://doi.org/10.1017/S0142716411000580

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Rogers, J., Révész, A., & Rebuschat, P. (2016). Implicit and explicit knowledge of inflectional morphology. *Applied Psycholinguistics*, *37*(4), 781–812. https://doi.org/10.1017/S0142716415000247

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Shanks, D. R., & St. John, M. F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*(3), 367–395. https://doi.org/10.1017/S0140525X00035032

Slobin, D. I., & Welsh, C. A. (1973). Elicited imitation as a research tool in developmental psycholinguistics. In C. A. Ferguson & D. I. Slobin (Eds.), *Readings in child language acquisition* (pp. 485–497). Holt, Rinehart and Winston.

Spada, N., Shiu, J. L.-J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, *65*(3), 723–751. https://doi.org/10.1111/lang.12129

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning*, *60*(2), 263–308. https://doi.org/10.1111/j.1467-9922.2010.00562.x

Wang, F. H. (2020). Explicit and implicit memory representations in cross-situational word learning. *Cognition*, *205*, Article 104444. https://doi.org/10.1016/j.cognition.2020.104444

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition*, *27*, 269–304. https://doi.org/10.1017/S0272263105050138

Williams, J. N. (2009). Implicit learning in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *The new handbook of second language acquisition* (pp. 319–353). Emerald.

Williams, J. N., & Rebuschat, P. (2012). Statistical learning and syntax: What can be learned and what difference does meaning make? In P. Rebuschat & J. N. Williams (Eds.), *Statistical learning and language acquisition* (pp. 237–264). De Gruyter Mouton. https://doi.org/10.1515/9781934078242.237

Wilson, B. A., Baddeley, A., Evans, J., & Shiel, A. (1994). Errorless learning in the rehabilitation of memory impaired people. *Neuropsychological Rehabilitation*, *4*(3), 307–326. https://doi.org/10.1080/09602019408401463

Zhao, C., Kormos, J., Rebuschat, P., & Suzuki, S. (2021). The role of modality and awareness in language learning. *Applied Psycholinguistics*, *42*(3), 703–737. https://doi.org/10.1017/S0142716421000023

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary
Appendix S1. Materials and Procedure.
Appendix S2. Comprehension Test.

Appendix S3. Scoring and Analysis.
Appendix S4. Additional Data.
Appendix S5. Effect Sizes and Sensitivity Power Analyses.