Studies in Second Language Acquisition (2024), **46**, 1231–1257
doi:10.1017/S0272263124000500

REPLICATION STUDY

Learning without awareness revisited and reconsidered

A conceptual replication and extension

John N. Williams and Yuyan Xue

University of Cambridge, UK
Corresponding author: John N. Williams; Email: jnw12@cam.ac.uk

(Received 06 September 2023; Revised 11 June 2024; Accepted 24 June 2024)

Abstract

Is it possible to acquire a sensitivity to a regularity in language without intending to and without awareness of what it is? In this conceptual replication and extension of an earlier study (Williams, 2005) participants were trained on a semiartificial language in which determiner choice was dependent on noun animacy. Participants who did not report awareness or recognition of this rule were nevertheless above chance at selecting the correct determiner in novel contexts. However, further analyses based on trial-by-trial subjective judgments and item similarity statistics were consistent with the possibility that responses were based on conscious feelings of familiarity or analogy to trained items rather than unconscious knowledge of a semantic generalization. The results are discussed in terms of instance-based approaches to memory and language, and the implications for the concept of "learning without awareness" are considered.

Keywords: animacy; artificial language learning; awareness; implicit learning

Introduction

In 2005, Williams (2005, henceforth W2005) published an article titled "Learning without awareness" claiming to show that it was possible for adults to pick up meaning-based regularities in language incidentally (without intending to) and without awareness of what they were. A specific sense of "awareness" was intended in the title of that article (and in those of subsequent replication studies)—awareness of a regularity or pattern in linguistic input, what Schmidt (1990) referred to as "awareness at the level of understanding," as opposed to "awareness at the level of noticing" actual forms. The original W2005 study examined whether learning without awareness at the level of understanding could be experimentally demonstrated using a miniature artificial determiner system in which the use of article-like elements (translated as "the-near" and "the-far") depended on the animacy of the accompanying noun (yielding

4 determiners, e.g., gi = the-near animate, ro = the-near inanimate, ul = the-far animate, ne = the-far inanimate). The participants were first instructed on the near/far meanings of the novel determiners, no mention being made of the potential role of noun animacy (which remained the "hidden" dimension). During training these novel forms were embedded in sentences in the participants' L1, English, such as "When I was out for a walk I patted gi dog and it bit me," "The researchers studied ul bees from a safe distance," "As I was passing I knocked over ro vase," "I looked up at ne clock on the church and realised that I was late." After exposure to 144 such sentences, there was a surprise test phase in which participants were provided with nouns that they had received in training, but in a new sentence context, and with a choice between two articles, neither of which had occurred with the noun in training, e.g., "While sitting by the wild flowers I heard the sound of gi/ro bees" (correct answer, "gi"). Accuracy was significantly above chance for these "generalization" items, indicating, it was claimed, sensitivity to the animacy agreement rule. Crucially, this was the case even for participants who were unable to report the animacy rule in a postexperiment debriefing. It appeared that there was learning without "awareness at the level of understanding" of the correlation between determiner choice and noun animacy. This artificially induced experimental result is, anecdotally at least, consonant with many language learners' experience of acquiring aspects of an L2 without being aware of how they did so, and with the general view of language acquisition as being in some sense an unconscious process.

Several subsequent studies have claimed to demonstrate similar "semantic implicit learning" (henceforth, SIL) phenomena across a range of different regularities and procedures—see Paciorek & Williams (2015b) for a review (and for more recent examples see Bovolenta & Williams, 2022; Cayado & Chan, 2022; Fukuta & Yamashita, 2021; Li, Zhao & Li, 2020; Pham, Kang, Johnson & Archibald, 2020). Yet other studies have failed to replicate the original W2005 result using an animacy-based regularity and a forced choice test task (see Table 1). These latter studies are more consistent with Schmidt's (1990) original scepticism over learning of abstract generalizations without awareness at the level of understanding in SLA, and as argued by others more recently (Leow, 2015).

But why is this issue important? From an applied linguistic perspective, implicit learning is of interest because it appears to open a window on unconscious, nonintentional learning processes that will hopefully reveal properties of incidental acquisition as opposed to intentional and explicit learning. For example, one might ask, as studies have begun to do so, whether all semantic regularities are equally learnable in this way, and whether these differences are the result of L1 influence, universals, or general principles of conceptual salience (Fukuta & Yamashita, 2021; Leung & Williams, 2012, 2014; Pham et al., 2020). Yet without proper means of assessing awareness the validity of the "window" itself remains open to question.

One criticism launched against many SIL studies is that since awareness is evaluated during, or after, a postexposure test phase, the degree of awareness that occurred during the exposure phase is unknown, hence making the claim that these studies establish "implicit learning" as a process subject to doubt (Leow, 2015). This focus on awareness of the product learning following exposure, as opposed to awareness of products or processes during the learning process itself, is merely characteristic of the tradition of implicit learning research in psychology dating back to Reber's (1967) seminal artificial grammar learning studies. The assumption is that, at least in the context of laboratory studies with brief exposure, if explicit learning processes were to deliver veridical conscious knowledge during training then this would be detectable as conscious

Table 1. Summary of animacy-based implicit learning studies using the W2005 procedure or similar.

W2005 procedure						
Study	Semantic feature	Design / procedure	Test task	Unaware/ total	Unaware by verbal report gen accuracy %	Accuracy by source judgment (generalization items and whole sample unless stated)
1. Hama and Leow (2010)	animacy	W2005 TAP during training & test	4AFC (animacy and distance)	34/54[1]	48.5	na
2. Faretta-Stutenberg and Morgan-Short (2011)	animacy	W2005	2AFC	21/30	53.87	na
3. Chen, Guo, Tang, Zhu, Yang, and Dienes (2011), Expt 1	animacy	Mandarin materials based on W2005 (Expt 1) W2005 design and procedure	2AFC with source judgments	(40/40)	**56.0***	G ≅ 0.53 **I ≅ 0.55*** **G + I = 56.0***
4. Chen et al. (2011), Expt 2	animacy	Mandarin materials based on W2005 (Expt 2) W2005 design and procedure	2AFC with source judgments	(30/30)	**58.0***	G ≅ 0.56 **I ≅ 0.59*** **G + I = 58.0***
5. Rebuschat, Hamrick, Riestenberg, Sachs, and Ziegler (2015)	animacy	Materials based on W2005 (Expt 2) W2005 procedure + TAP during training/ test	2AFC with source judgments	Silent: 4/14	Silent[2]: 53.5	Silent[2]: G = 58.2, **I = 72.7**** **M = 82.6**** **R = 73.11 ****
				TAP in training: 4/12	TAP in training[2]: 49.3	TAP in training[2]: G = 44.4 **I = 69.8 **** **M = 63.1*** **R = 73.8****
				TAP throughout: 1/11	TAP throughout[2]: 0.0	TAP throughout[2]: G = 44.7 I = 52.6 **M = 71.1*** **R = 72.3***

(*Continued*)

Table 1. (*Continued*)

W2005 procedure						
Study	Semantic feature	Design / procedure	Test task	Unaware/ total	Unaware by verbal report gen accuracy %	Accuracy by source judgment (generalization items and whole sample unless stated)
6. Sachs, Hamrick, McCormick, and Leow (2020)	animacy	Rebuschat et al. (2015) materials W2005 procedure	2AFC with source judgments	4/10	na	G = 45[2,3], I = 58 M = 62 R = 82
7. Zhao, Kormos, Rebuschat, and Suzuki (2021)	animacy	Rebuschat et al. (2015) materials W2005 procedure with modality manipulations	2AFC with source judgments	36/88	53.0	**G+I[2] = 55.5*** **
8. Kim, Maie, Suga, Miller, and Hui (2023)		W2005 Academic sample		33/107	52.0	
Other procedures						
9. Kerz, Wiechmann, and Riedel (2017)	animacy	W2005-like system 8 items presented in a story	2AFC with confidence ratings	Expt 1: 61/112	Expt 1: **53.0***	Guess[4] = 0.5 Somewhat = 0.55 Undecided = 0.52 **Very = 0.59*** Certain = 0.52
				Expt 2: 34/80	Expt 2: **57.0*** **	Guess[4] = 0.51 Somewhat = 0.51 **Undecided = 0.60*** **Very = 0.60*** **Certain = 0.60***

(*Continued*)

Table 1. (*Continued*)

W2005 procedure						
Study	Semantic feature	Design / procedure	Test task	Unaware/ total	Unaware by verbal report gen accuracy %	Accuracy by source judgment (generalization items and whole sample unless stated)
10. Fukuta and Yamashita (2021)	animacy	Markers combined in a complex morpheme along with thematic role and explicitly taught number marking.	2AFC with source judgments. Immediate and delayed tests.	31/40	Immediate: 53.1 (LCI = 44.2)	G[4] = 0.52 (LCI = 0.43) I = 0.51 (LCI = 0.38) M = 0.56 (LCI = 0.28) R = 0.60 (LCI = na)
		TAP *during training* as measure of awareness.			Delayed: **58.1** (LCI = 0.502)	G = 0.41 (LCI = 0.30) I = 0.50 (LCI = na) **M = 0.67** (LCI = 0.54) R = 0.75 (LCI = 0.43)
11. Brown, Smith, Samara, and Wonnacott (2021)	Animacy (animals vs. vehicles)	One marker for animals, one marker for vehicles. 6-year-old children. Consistent marking or partially consistent marking.	2AFC	17/30	Consistent marking, not significant (β = 0.10) Partially consistent marking, (β, na) not significant.	

Notes: TAP: Think-aloud protocol.
Source judgments: G = guess, I = intuition, M = memory, R = rule.
LCI = lower credible interval (above chance level indicates significant).
*$p < 0.05$

[1] 9 aware, 11 reported some other rule that would lead to chance performance.

[2] Combined for trained, partially trained (as in W2005 generalization items) and new test items.

[3] No inferential statistics reported.

[4] Source/confidence data for unaware participants.

knowledge during immediately subsequent test performance or debriefing.[1] Of course, ideally one would probe awareness during the exposure phase itself. Requiring thinking aloud is one possibility (bearing in mind the potential disturbance to the learning process), but interestingly, of the studies in Table 1 that applied think aloud to the W2005 paradigm, instances of awareness at the level of understanding during training have either been completely absent [1] or very rare [5]. Hence, we will assume that the lack of awareness of the product in the W2005 paradigm is indicative of implicit learning as a process.

Given the practice of assessing awareness of the products of learning, the debate over the reality of SIL, and implicit learning more generally, centres on the existence of awareness of the product of learning; that is, on the measurement of implicit knowledge. Can learners show sensitivity to knowledge that they do not know that they know? Simple postexperiment verbal reports (as used in W2005) have been criticized for being subject to underreporting of low-confidence knowledge or forgetting of fleeting impressions (Rebuschat et al., 2015; Rebuschat et al., 2013; Shanks & St. John, 1994). In this view, demonstrations of above-chance test performance in "unaware" learners might merely reflect undetected conscious knowledge (though how that knowledge was acquired, which, as argued above, is the essential issue, remains an open question).

To obviate problems with postexperiment verbal reports, trial-by-trial subjective measures (Dienes & Scott, 2005) have been adopted in many studies. For example, for each decision, participants indicate whether they guessed, had an intuitive feeling of being correct without knowing why, relied on memory for items seen before, or applied rules. This technique has the advantage of being sensitive to the participant's state of awareness in the moment of making their decision. Therefore, if previous results are due to underreporting of conscious knowledge, one would expect that studies that use subjective measures should be less likely to show SIL effects than those that rely simply on postexperiment verbal reports. But if anything, the reverse is the case. The summary of studies using the W2005 paradigm or similar (Table 1) shows that only 2/11 studies found a learning effect for participants who were unaware by postexperiment verbal report (studies 9, and 10, delayed test). In contrast, all 6 studies that used subjective measures (and report statistics) found an effect at least for the category of intuition (studies 3, 4, 5, 7, 9, 10). Assuming that intuitive responses reflect unconscious structural knowledge, that is, lack of awareness of the relevant regularities (Dienes & Scott, 2005), these results could be regarded as indicative of implicit knowledge. However, in four of these studies, subjective measures were analysed for the whole sample, regardless of awareness by postexperiment verbal report (studies 3, 4, 5, 7). In study 5's silent condition, 71% of the participants were aware by verbal report, and in study 7, 59% were. In Studies 3 and 4, it is reported that none of the participants were aware by verbal report but given the high awareness rates in other studies using similar procedures and materials, one may suspect that the debriefing method was relatively insensitive here.[2] Also, in the case of study 10, awareness was determined with respect to the training phase, leaving open the possibility that awareness developed during testing. Hence, if some of the participants included in the source judgment analyses

[1] Given that there may be cases where unconscious and nonintentional learning processes result in conscious knowledge through spontaneous insight, this practice may actually underestimate the extent to which implicit learning has occurred.

[2] As the authors point out "a problem with free oral report is that participants can avoid reporting any rules unless they're quite confident. In our case, they seemed not willing to report at all and this is why use of trial by trial structural knowledge attributions was important," Chen et al., 2011, p. 1755).

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

actually acquired conscious knowledge of the rules, can we be sure that responses they classified as "intuitive" were really a reflection of unconscious structural knowledge at the time? Could some of these responses have been based on low confidence conscious knowledge, and would this be sufficient to account for above chance accuracy over the whole sample? It is also striking that in no study in Table 1 was the guess category significantly above chance, violating the "guessing criterion" for establishing unconscious knowledge (Dienes & Scott, 2005). Taken with ambiguity over the status of intuitive responses, the evidence for learning without awareness from studies that have applied subjective measures is not totally convincing.

The purpose of Experiment 1 was to carry out a conceptual replication of the W2005 study to further examine the issue of awareness measurement. Here, "conceptual replication" refers to studies "where there is intentional adaptation of the initial study to investigate generalizability to new conditions, contexts, or study characteristics" (Marsden, Morgan-Short, Thompson & Abugaber, 2018: 325–326). We chose to conduct a conceptual replication of W2005 following the principle that if the primary research question concerns the methodology of awareness measurement then it is advantageous to maintain comparability with related studies at least in terms of the learning target (given that, as just noted, semantic regularities may vary in learnability). The learning target of W2005—the semantic regularity concerning animacy—has been employed by a large number of studies (as summarized in Table 1). Therefore, W2005 seems a proper candidate for conceptual replication.

In the present experiment, unlike most previous studies, accuracy in the different source categories will be analysed only for participants who were classified as unaware according to the postexperiment debriefing. Even though postexperiment debriefing and subjective measures may each have their shortcomings (e.g., postexperiment debriefing may suffer from memory decay; subjective measures may be subject to each individual participant's criteria about how this task should be performed, see Chan and Leung, 2018 for review), combining them should provide sensitivity to in-the-moment decision-making processes in the test phase while minimising contamination from awareness at the level of understanding in the implicit source categories (guess and intuition). To anticipate, the results showed that even after the exclusion of aware participants, responses made by intuition and memory were above chance for generalization items, but guesses were at chance, in line with previous results. So, a further question arose—if participants were basing their judgments on conscious feelings of correctness that were not rule-based (taking the postexperiment verbal report at face value), then what kind of conscious knowledge were they using?

Experiment 2 therefore constitutes an extension of the original W2005 study—it investigated the possibility that responses were based on feelings of familiarity arising from similarity or specific analogies between test and training items. Perhaps the most robust finding from decades of artificial grammar learning research is that measures of similarity between test and training items influence grammaticality judgment accuracy (Ziori & Pothos, 2015). Some studies have even found that there is no residual effect of grammaticality after exhaustive similarity metrics have been applied (Johnstone & Shanks, 1999; Scott & Dienes, 2008). Of course, this is in the context of complex finite state grammars where it is difficult to conceptualise what an abstract rule would be, unlike the case of the simple animacy rule in W2005. But if, nevertheless, responses reflect computations of item similarity then this would have implications for our conception of what constitutes "learning," and also for the very notion of "learning without awareness," in the present context (see General Discussion). These considerations prompted a post hoc investigation (i.e., the extension study reported as

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

Experiment 2) of the role of similarity between test and training items in determining response accuracy at the item level as elicited in Experiment 1 (i.e., the conceptual replication of W2005).

Experiment 1

Experiment 1 used the same materials as the original W2005 (Experiment 2) study with some procedural changes other than the addition of source judgments in the test phase. Below we outline these procedural changes: First, since the focus here is on the participants who are unaware in the postexperiment debriefing it was important to maintain a low awareness rate. As Table 1 shows, studies based on the W2005 design vary widely in the proportion of participants who remain unaware, with studies using sizeable samples evidencing quite low proportions of unawareness (41% in [7], 31% in [8]). In W2005 the set of 48 training items was in fact repeated across 3 blocks of training, yielding 144 trials, providing plenty of opportunity for rule awareness to emerge. Here only a single block of the same 48 trials was employed in an attempt to reduce awareness rates. Second, a variation that may increase the strength of semantic implicit learning effects was introduced—accompanying each sentence with a picture of a scene depicting the critical noun concept in a roughly equivalent context (see Figure 1). Since what is at issue here is the existence of semantic implicit learning, it stands to reason that it is important that a sufficiently rich semantic interpretation of the training sentences is computed. Although the W2005 procedure required participants to imagine the situation described by the sentence in anticipation of a test of gist memory, participants may vary in the degree of elaboration of their sentence interpretation. Here it was reasoned that the provision of pictures would ensure a more consistently elaborated semantic representation of the sentences across participants (for evidence of the facilitative effect of elaboration on memory see, for example, Bower & Winzenz, 1970). Third, the post-experiment debriefing included a rule recognition component in which participants were informed of the rule and asked if they had been aware, vaguely aware, or unaware of it during the experiment. The purpose of this was to guard against potential underreporting of rule knowledge in response to open questions (such as in W2005). Finally, the participants for the present study were more heterogenous in terms of academic background and had far less knowledge of languages other than English (as detailed below), providing a test of the generalisability of the W2005 findings. Given these procedural changes, the present study constitutes a conceptual replication of W2005.

Participants

A total of 90[3] participants were recruited through Prolific Academic, mean age 21.6 years (range = 18–34), 61% identifying as female. All had positive student status and were native speakers of English with either British, Irish, Australian or American nationality. They were located either in the UK or Ireland at the time of testing, had an approval rate on Prolific of at least 90, and had not participated in previous experiments

[3] This sample size was determined from the two previous studies that used subjective measures, Chen et al. (Experiments 1 and 2, 2011) and Zhao et al. (2021) (studies 3, 4 and 7 in Table 1). These studies yielded on average 35 unaware participants. Assuming the unawareness rate of 41% obtained in Zhao et al. for native English speakers, 85 participants would need to be tested, plus an allowance of 5 for the potential of having to exclude participants due to noncompliance under online testing.

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

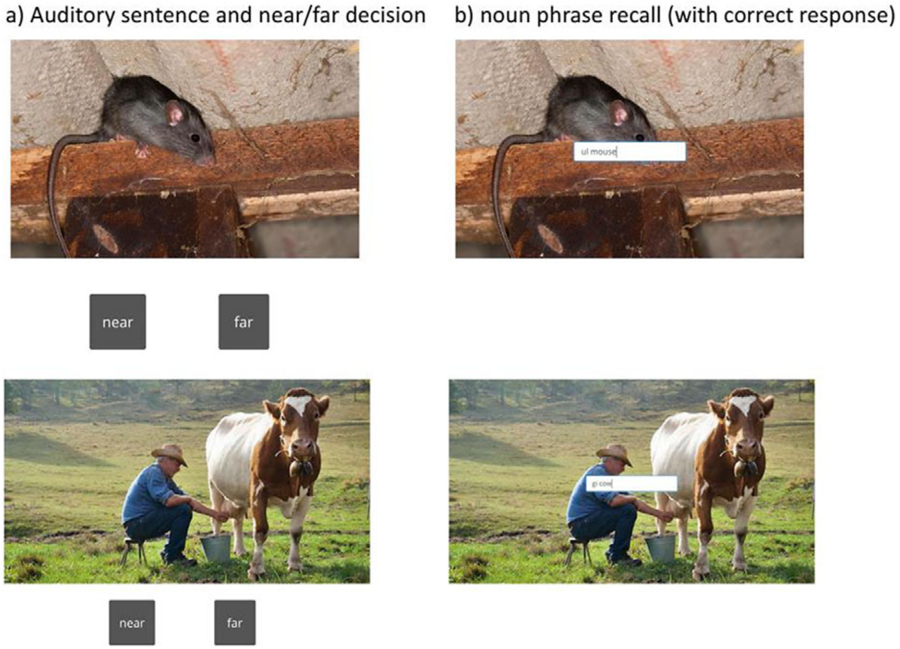

a) Auditory sentence and near/far decision
b) noun phrase recall (with correct response)
ul mouse
near
far
gi cow
near
far

Figure 1. Two example items from the training phase. The accompanying auditory sentences were: “I could hear ul mouse scurrying around in the roof”, and “The farmer was kicked by gi cow when he tried to milk it.”

of this type by the first author. Participants knew a mean of 0.244 (SD = 0.481) languages other than English to a level of intermediate or better, including a mean of 0.211 (SD = 0.462) languages that encode grammatical gender. A total of 27 indicated already holding an undergraduate degree, 7 had a master’s degree, and one had a PhD. The most common areas of study were psychology (16), economics/finance/business (11), law (8), computer science/IT (6), maths (4), and physics (4). Only one person indicated studying languages.

The present sample differs markedly from that in Williams (2005, Experiment 2) where all the participants were nonnative speakers of English who were studying at the University of Cambridge, U.K. They had a wide range of first languages, three of them being brought up bilingually, and there were 21 instances of intermediate or advanced knowledge of a foreign language other than English amongst the 24 participants. A total of 13 out of the 24 participants (54%) were studying language-related disciplines. Hence, the present sample is more heterogeneous than that in Williams 2005 (Experiment 2) in terms of academic background, while at the same time much less diverse in terms of language background.

Methods

Materials

The same training and test items were used as in W2005 (Experiment 2) including the original sound recordings for the training phase. In addition to the original materials, for each sentence, a picture was sourced from the internet which depicted the critical

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

noun in a roughly equivalent context (see Figure 1 for examples). Pictures were included to encourage deeper semantic processing, which is obviously critical in the context of a SIL study, particularly when conducted remotely. The sentences and pictures are provided on the OSF site.[4]

Training materials

For training, there were 24 nouns, 12 living things (animals) and 12 inanimate objects. Half of each type were included in "near" contexts and half in "far" contexts. Each noun appeared in two distinct sentences, being singular in one and plural in the other. The resulting 48 sentences were divided into two equal blocks such that each noun appeared only once in each block and the proportions of animate/inanimate and singular/plural nouns were equal. Pictures were sourced from internet searches to illustrate the noun object in a roughly relevant context.[5]

Test materials

For the test phase, the first 8 items were generalization items (Gen 1) featuring a noun from training used in a different distance context and a unique sentence accompanied by a unique picture. For example, *gi bees* is a generalization item (with a forced choice between the "near" articles *gi* and *ro*) but the training set only contained *ul bee* and *ul bees* (far bee(s)).

The same nouns that were used for Gen 1 were also used to form a second set of generalization items, Gen 2, in which the nouns appeared in a different number (e.g., Gen 1 presented a choice between *gi* and *ro bees*, and in Gen 2 the choice was between *gi and ro bee*). Unique sentence contexts and pictures were used for these items. Part-way through testing (after testing 46 participants) it was noted from the debriefing data that participants had a tendency to erroneously regard number as a factor governing determiner usage. This may have been due to the contrast in the number of the same nouns in the Gen 1 and Gen 2 items. Therefore, for the remaining 44 participants, the Gen 2 items were changed to a different set of nouns which again had occurred during training, but as for Gen 1, with different distance determiners. New sentence contexts and pictures were used with these nouns. To distinguish this change in the Gen 2 items, the test containing the original Gen 2 items (as used in W2005) will be referred to as "test version 1" and the test containing the modified test items will be referred to as "test version 2."

Both test versions also contained a set of 8 Trained items—determiner–noun pairings that had occurred in training but were now presented in new sentence contexts. For each set of test items (Gen 1, Gen 2, Trained) proportions of living/nonliving and singular/plural were balanced. The sentence contexts were such that the noun always occurred as the last word. New pictures were sourced from the internet to depict the critical object in a relevant context (therefore, note that even for "trained" items it was only the specific determiner–noun combination that was familiar, the sentence and picture being entirely novel).

[4] Text materials, data, R code, and an appendix containing additional analyses are available on the Open Science Framework at https://osf.io/aeyxz/?view_only=ea475af6b6eb4decaca7f009731b1abf

[5] For 11 of the 48 items a picture could not be found that conveyed the near/far meaning directly (as in the "ul mouse" example in Figure 1), and in a further 10 cases distance in the picture was ambiguous. But in all cases the sentence context conveyed the near/far meaning.

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

Two lists were created with opposite assignments of determiners to animacy (list 1: gi = animate near, ro = inanimate near, ul = animate far, ne = inanimate far; list 2: ro = animate near, gi = inanimate near, ne = animate far, ul = inanimate far). This was to control for any sound-based biases in determiner selection in the test phase (e.g., ro bees may be preferred to gi bees on the basis of sound alone). There were sentences and pictures for two practice trials using trained nouns that did not appear in the test (gi/ro cushions and ul/ne mouse).

Sentence recognition

At the end of the experiment, participants would be required to judge whether a sentence with the same meaning had occurred during training. The purpose of this was to motivate attention to the overall sentence meanings during training, deemed critical for semantic implicit learning. However, performance on this test was not a critical measure.

Twelve sentences were prepared. Six of them had occurred in training, and the other six contained minor modifications of the original sentence, though the determiner was never changed (e.g., "The workers threw darts at ne picture of their manager" became "The workers admired ne picture of their manager"). No pictures or sounds accompanied these sentences.

Procedure

The entire experiment was built and conducted online using Gorilla Experiment Builder (Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed, 2020). The components of the experiment occurred in the order below.

Introduction. After providing consent and filling in a biographical questionnaire, participants were introduced to the four novel determiners and their function as article-like elements that also encode distance between speaker and object.

Determiner pretraining

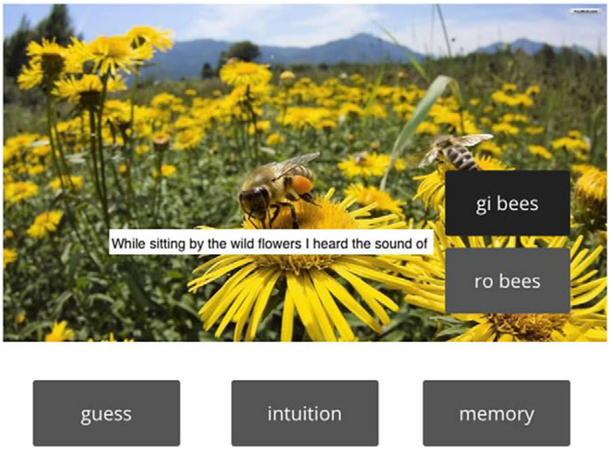
Participants were pretrained on the near/far meanings of the articles. In each of the 36 trials, they heard and saw an article (e.g., ul) and below had a choice between "near" and "far" response options, and were given feedback (tick or cross) on their accuracy.

Training phase. The training phase procedure was as follows: (1) central fixation cross (participant presses space to proceed); (2) simultaneous presentation of the picture and auditory sentence; (3) two seconds after sound offset "near" and "far" response options appear below the picture (the participant clicks on one option, Figure 1a) if response incorrect a red cross appears and the sound is played again and response options appear again; (4) on a correct near/far response a text box appears in the centre of the picture and the participant types in the determiner phrase from the sentence they just heard (Figure 1b), if the response is incorrect a cross appears and the correct response appears under the picture if the response is correct a green tick appears; (5) next trial. The training comprised 48 trials divided into 2 blocks with a 30-sec forced break in between (note this was one third of the training in W2005). The trial order was randomized within each block independently for each participant.

Test phase

The test phase procedure ran as follows (Figure 2): (1) Presentation of a picture with a text box in the middle containing a sentence (e.g. "While sitting by the wild flowers I

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press



While sitting by the wild flowers I heard the sound of

Figure 2. An example trial from the testing phase following selection of "gi bees."

heard the sound of …") and two response options positioned above and below the sentence-final position (e.g. gi bees and ro bees). The participant clicks an option. No feedback is provided. (2) The options "guess," "intuition," and "memory" appear below the picture and the participant makes a selection.[6] There were 24 trials with the order fixed to avoid the same correct option on successive trials. Participants were told to "decide which is the correct phrase based on what you heard previously." In test version 1 the sets of trials followed the order Practice trials (two trained determiner–noun combinations), Gen 1, Trained, Gen 2, and in test version 2 the order was Gen 1, Gen 2, Trained (in case positioning of Gen 2 and Trained items had any effect). For the two practice trials, the procedure was supplemented with onscreen instructions, a reminder that distance is not relevant to the decision and that participants should not worry if they feel they are guessing or using intuition. Guess was explained as "I could have flipped a coin," intuition as "I have a feeling but I don't know what it is based on," and memory as "I remember this phrase from the previous part." None of these instructions were present for the test trials proper.

Sentence recognition test

The 12 sentences were presented visually (with no sound or picture) one at a time in a random order. The participant indicates by clicking the "yes" or "no" buttons whether a sentence with the same meaning had occurred during training. Since this task was not used to measure any critical learning effects, feedback was provided as a tick or a cross so that participants could get a sense of how they were doing.

Debriefing

The debriefing contained the following questions: (1) what the participant had learned about the use of the four novel words (free text response), (2, version 2 only) what they noticed beyond the near/far distinction (free text response) and at what point during

[6] "Rule" was not included because it has been shown to encourage explicit hypothesis formation during the test phase (Rebuschat et al., 2015; Sachs et al., 2020).

the experiment they did so (training phase, test phase, sentence recognition, debriefing, not aware), (3) rule recognition—the version of the animacy mapping that the participant had received is explained and the participant rates their level of awareness of that rule (unaware, vaguely aware, very aware) and when in the experiment that first occurred, (4) participant states how they became aware (free text response), (5, version 2 only) whether the experiment seemed similar to any experiment they have done in the past (yes/no and free text box). All fields were mandatory. The entire experiment took around 25 minutes to complete.

Data coding and analysis

We first analyzed the accuracy rate in the training task. Based on this, we excluded from further analysis one participant who clearly was not performing the training adequately (see "Training task performance" subsection in the Results section).

Next, we classified participants into "aware" and "unaware" based on their responses in the post-experiment debriefing. Participants were classified as "aware" if they (1) in response to the rule recognition question, selected "very" or "vaguely" aware and indicated that awareness occurred prior to the debriefing, or (2) in response to the open questions, reported some relevant knowledge which they attributed to the training or testing task[7]. Otherwise, they were classified as "unaware." Further analysis reported in this manuscript focused on the "unaware" participants, while analysis on the "aware" participants and "unaware" and "aware" pooled together can also be found in Appendix 2 and 3 (OSF files, and summarized at the end of the Results section, below).

To analyse the test performance, logit mixed effect models were run using the lme4 package in R (version 4.3.2) (R Core Team, 2023). The dependent variable of these models was the likelihood of choosing the correct phrase in the test (correct was coded as 1, while incorrect 0). The "anova" function was used to compare models using a likelihood ratio test. A confirmatory analysis strategy was adopted focusing on the fixed factors of theoretical interest that, on the basis of previous research, would be predicted to have an effect: test item type (Gen 1, Trained, Gen 2), source (guess, intuition, memory), and their interaction. For all models in this study, levels of the fixed factors were coded using R's default treatment coding (i.e., at each time, one level is taken as the reference level against which the other two levels are compared. We altered the reference level to obtain all pairwise comparisons). Random factors were participant (because different participants may show greater or smaller overall learning effects) and noun (e.g., "bee" and "bees" are coded as different nouns because they occur in different contexts and hence constitute different items). It was also reasoned that the design factor "list" could modulate the selection between the two alternatives due to sound preferences.[8] To account for this potential interaction between noun and list, the list

[7] Participants who reported the rule or attributed recognition of it to the sentence recognition test (N = 7) were included in the "aware" group because of potential ambiguity in which part of the experiment they were referring to (i.e., confusing what was referred to in the questionnaire as the "third task: sentence memory" with the "second task: choosing between phrases").

[8] Data from an untrained control condition for another study is relevant to this point. Twenty-four participants performed the present determiner pretraining task and then progressed immediately to the Gen 1 test. They were asked to select the response option that they "preferred." Only one participant reported using animacy-based criteria. Of the remainder (12 on list 1 and 11 on list 2), mean "accuracy" (according to the present scoring system) was not significantly different from chance ($M = 0.495$, $SD = 0.158$), but an analysis by items showed that there was a strong negative correlation between accuracy for the items on the

was added as a random slope for a noun. The *emmeans* package was used to estimate the marginal means and 95% confidence intervals for each level of the fixed effects of the models, and the equation P = exp(x)/1+exp(x) was used to convert logit to probability. No corrections were applied to the output of the *emmeans* package. Being significantly higher than chance is indicated by the lower bound of the confidence interval of the estimated marginal mean (in probability) being greater than 0.5. Full results of all models mentioned below can be found in Appendix 1 (see OSF files).

To be comparable to the W2005 study, we first ran models with test item type as the single fixed effect. Successful conceptual replication of W2005 was defined as follows: the lower bound of the 95% confidence interval of the estimated marginal means (in probability) of Trained, Gen 1, and Gen 2 items being higher than 0.5 respectively, and that of Gen 1 and Gen 2 combined being higher than 0.5. Next, we added source into the fixed effect structure to explore the probability of choosing the correct phrase for each source type and the potential interaction between source and test item type.

Besides the analysis of the accuracy in the test phase, we also report the proportions of responses in each source category and test item type (Figure 4).

Results

Training task performance

Near/far decisions and determiner recall were highly accurate (accuracy for each participant was calculated as the number of correct responses divided by the total number of training trials) for all participants with the exception of one participant. This participant's error rate was 0.35 for the near/far decision and 0.89 for determiner recall. This person's entire data were removed from the sample because they were clearly not performing the task properly. Over the remaining participants (n = 89) the mean near/far error rate was 0.02 (SD = 0.03, range = 0–0.12) and the mean determiner recall error rate was 0.01 (SD = 0.02, range = 0 = 0.15).

Awareness scoring

Based on the criteria outlined above, 35 participants (21 who received test version 1, and 14 who received test version 2) were classified as "aware." Among them, twenty of them did not actually report any relevant content in the first and second open questions about regularities they noticed beyond the near/far distinction, and only 8 of them attributed rule recognition to the training task. The following analyses reported in this manuscript will focus on the remaining 54 unaware participants (24 who received test version 1, and 30 who received test version 2).

Test performance

Test version did not significantly affect performance on the Trained or Gen 2 items (which were differentially placed), although performance was numerically lower for

two lists, $r(7) = -0.817$, $p = 0.013$. For example, on List 1, where "ul lion" was "correct," accuracy was 0.25, but on list 2 where "ne lion" was "correct," accuracy was 0.818. Participants predominantly reported going by which option "sounded" better in the context of the sentence. These item-specific biases introduce unwanted variation into the 2FAC data making the task less sensitive to underlying animacy biases.

Table 2. Estimated marginal means for the unaware participants (*N* = 54).

	EMM	LCL	UCL
Gen 1	0.596	0.538	0.650
Trained	0.646	0.590	0.698
Gen 2	0.534	0.480	0.587
Gen 1+2	0.562	0.521	0.603

Note: EMM = estimated marginal mean, LCL = lower confidence limit, UCL = upper confidence limit.

Trained items when placed last (0.612 versus 0.672, $t(52) = 1.316$, $p = 0.194$) and performance on Gen 2 was also numerically slightly lower when placed last (0.516 versus 0.542, $t(52) = 0.556$, $p = 0.580$). Since the test version had no significant effect, it will not be considered further.

The first analysis focused on response accuracy disregarding source judgment data to be comparable with the original W2005 analysis. An initial model had test item type (Gen 1, Trained, Gen 2) as fixed factor and participant and noun as random factors. Adding list as a random slope for noun (Model 1) improved the fit of the model considerably ($X^2 = 22.11$, df = 2, $p < 0.001$) indicating the anticipated interaction between noun and list. The estimated marginal means and confidence limits for each test item type based on Model 1 are shown in the first three rows of Table 2. In addition, we ran another model (Model 1A) which only differs from Model 1 in that Gen 1 and Gen 2 are combined. The estimated marginal means and confidence limits for the Gen 1 and Gen 2 combined conditions based on Model 1A are shown in the last row of Table 2. The lower confidence limits indicate that accuracy in both the Gen 1 and Trained conditions was significantly above chance, whereas this was not the case for Gen 2. Model 1 indicates that Gen 2 accuracy was significantly lower than Trained ($Z = -2.852$, $p < 0.004$, odds ratio = 0.653) but Gen 1 accuracy was not ($Z = -1.262$, $p = 0.207$, odds ratio = 0.808). Model 1A indicates that the accuracy for the Gen 1 and Gen 2 conditions combined (Gen 1+2) was significantly above chance, but the Trained condition was significantly more accurate ($Z = 2.314$, $p = 0.021$, odds ratio = 1.419). Nevertheless, there is evidence of an ability to generalise even for unaware participants, and this was especially evident in the Gen 1 items (which always came first, and hence were least likely to be affected by explicit hypothesis testing during the test phase).[9]

Next, source was added to the model as an additional fixed factor (Model 2), which improved model fit, though not significantly ($X^2 = 5.615$, $df = 2$, $p = 0.060$). According to Model 2, overall accuracy for responses based on guesses was significantly lower than for memory ($Z = -1.984$, $p = 0.047$, odds ratio = 0.711), as was accuracy for responses based on intuition ($Z = -2.134$, $p = 0.033$, odds ratio = 0.731). Adding the interaction between condition and source improved model fit still further ($X^2 = 10.201$, $df = 4$, p = 0.037). To break down this interaction individual models were constructed for each test item type (as above, but with Condition removed. See Model 3–5 in Appendix 1)[10]. The only significant effects of source category were in the Trained condition where guess accuracy was lower than memory ($Z = -3.219$, $p = 0.001$, odds ratio = 0.388), and likewise for intuition ($Z = -2.991$, $p = 0.003$, odds ratio = 0.471). The estimated marginal

[9] Bayesian one sample t-tests (performed in JASP) support this conclusion. Bayes factors (BF_{10}) were 372.6 for Gen 1, 272,973 for Trained, 0.590 for Gen 2, and 39.14 for Gen1+2. Following widely used interpretation guidelines there is strong evidence that accuracy was above chance for Gen 1, Trained, and Gen 1+2, but for Gen 2 there was weak evidence for the null hypothesis.

[10] The Gen 1 model failed to converge but did so after list was removed as the random slope for noun.

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

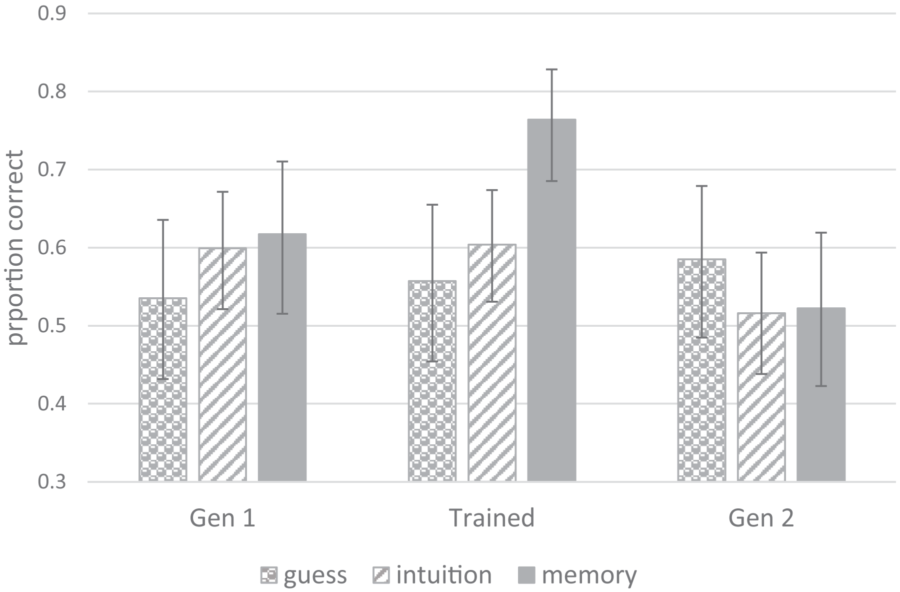


Figure 3. Estimated marginal means (with 95% upper and lower confidence intervals) for the proportion of correct responses for each subjective judgment category and test item type for unaware participants (N = 54).

means derived from Models 3–5 are shown in Figure 3. Guess responses were not significantly above chance for any test item type, not even trained items. Intuition and memory responses were above chance for Gen 1 and Trained items, but not Gen 2.

The mean proportions of total responses for each source category and condition are shown in Figure 4. The patterning of response proportions was similar across the different test item types. Inferring from the 95% confidence intervals, for Gen 1 and Gen 2, intuition was more frequent than guess and memory, which did not differ. For trained items, intuition was still more frequent than guess, while memory was not different from either condition, being somewhat elevated compared with Gen 1 and Gen 2, as one would expect.

Although the focus here is on the participants who were classed as "unaware" according to the debriefing, for completeness the results for the 35 participants who were classified as "aware" will be briefly described (see Appendices 2 and 3, OSF supplementary files). The data were analyzed in a parallel fashion to the unaware participants. Estimated marginal means (and upper and lower 95% confidence limits in parentheses) for test phase performance were as follows: Gen 1, 0.752 (0.647–0.833); Trained, 0.763 (0.661–0.842); Gen 2, 0.726 (0.621–0.812). As one would expect amongst participants who indicated being aware of the animacy rule, performance was uniform across the three test item types. Adding Source to the model improved fit significantly ($X^2 = 7.887$, $df = 2$, $p = 0.019$), whereas further adding test item type did not ($X^2 = 6.135$, $df = 4$, $p = 0.189$). Overall, accuracy based on memory was significantly higher than intuition (estimate = -0.608, $se = 0.218$, $z = -2.788$, $p = 0.005$), a pattern that was only evident in the unaware participants for the Trained items (see Figure 3). Accuracy was significantly above chance for all source categories within all test item types (ranging from 0.70 to 0.80).

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

Discussion

Experiment 1 conceptually and partially replicates the original W2005 result—participants who did not report any awareness of the animacy rule nevertheless displayed sensitivity to it in a forced choice test, where generalization items were responded to with significantly above chance accuracy. The effects were weaker than in the W2005 (Experiment 2) study, however. In W2005 the mean proportion correct over Gen 1, Trained, and Gen 2 was 0.640, 0.847, 0.640, significantly above chance in all cases, and significantly above chance for the Gen items combined. Here the corresponding proportions were 0.596, 0.646, and 0.534, only significantly above chance for Gen 1 and Trained, and not Gen 2, but significantly above chance for the Gen items combined. The fact that the present participants received each training item only once, as opposed to three times, could explain the weaker effects. But there is also the difference in participant pool and experimental setup to be considered (Cambridge University graduate and undergraduate students tested individually in the lab by the author of the research versus crowd-sourced UK University students tested over the internet; greater diversity and number of languages known, and greater academic specialisation in languages in the case of the W2005 participants). Still, the effects obtained here for participants who were unaware by post-experiment verbal reports were stronger than in other partial or complete W2005 replications (Table 1), at least for the Gen 1 test items. The reasons for this cannot be ascertained without controlled comparisons between different procedural variants. The most obvious difference between this and previous studies is the use of rich pictorial contexts which could have made the degree of semantic elaboration more consistent across participants.

Analyses of test accuracy by source judgment category (guess, intuition, memory) indicated above chance accuracy for the first set of generalization items (Gen 1) for responses based on intuition and memory. Unlike most previous studies that have found such effects (studies 3, 4, 5, 7, 9 in Table 1), this analysis was based only on participants who were unaware of the rules according to post-experiment verbal report and rule recognition. Hence intuitive and memory-based responses were less likely to have been contaminated by conscious rule knowledge than in previous studies. Following standard logic (Dienes & Scott, 2005), in the case of generalization test items, above-chance intuitive responses reflect some confidence in the decision (judgment knowledge) generated by unconscious (structural) knowledge of the animacy rule. Memory-based responses might be assumed to reflect false memory—a conscious belief that the item had been presented before, generated by its conformity to unconscious structural knowledge of the generalization (Paciorek & Williams, 2015a).

However, if knowledge of the animacy rule were being applied unconsciously then accuracy should have been above chance when participants claimed to be guessing. But guesses were not above chance for any test item type, even trained items. This null result is in line with the absence of above-chance guess responses in previous W2005-related experiments (studies 3, 4, 5, 6, 9, 10, Table 1) even when including substantial numbers of participants who are aware by verbal report (studies 5 and 6, Table 1). One may suspect that in the present case, the chance accuracy for guess responses was due to low statistical power, but for Gen 1 the mean number of responses that participants made of this type was not much lower than for the memory category, where a significant effect was obtained (Figure 4). Although the lack of an effect for guesses on even trained items is striking, the most that one can say is that, overall, the guess

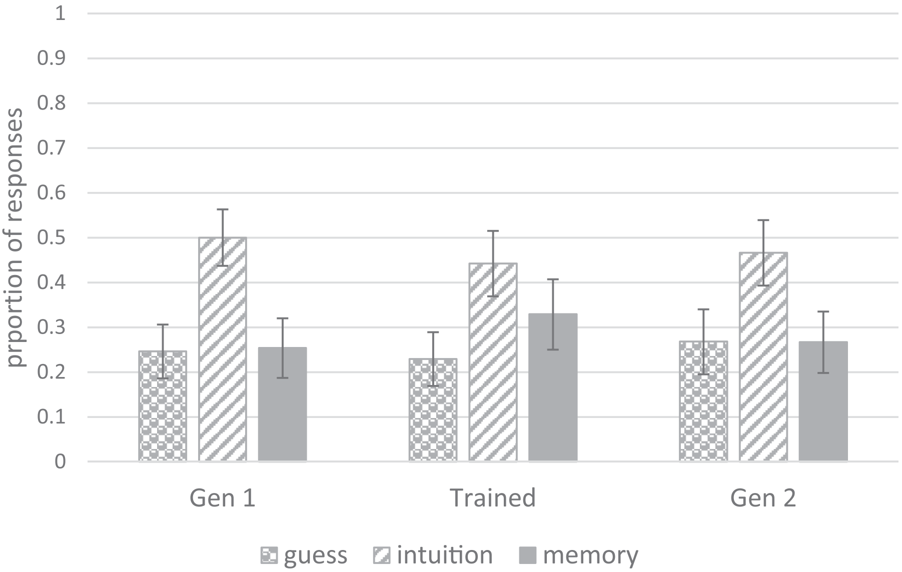


Figure 4. The mean proportion of responses (with 95% upper and lower confidence intervals) regardless of correctness for each subjective judgment category and test item type for unaware participants ($N = 54$).

data are indeterminate,[11] but are consistent with previous findings. If the "guessing criterion" (Dienes & Scott, 2005) is taken as the benchmark for establishing unconscious structural knowledge, then the lack of convincing evidence that it is satisfied here calls into question whether the unaware participants possessed unconscious knowledge of the animacy rule, despite showing above chance accuracy for Gen 1 items when source judgments were not considered.

The question arises as to what kind of structural knowledge underlies correct responding under intuition or memory given that the above findings were based on participants who showed no ability to state the rule, and, more tellingly, no recognition of it when it was presented to them. One possibility is that for both trained and generalization items, correct responding depends on conscious feelings of familiarity. For trained items, this could reflect an exact match to a training item. For generalization items, it could reflect the similarity between individual test and training items, or even consciously recalled analogies. Intuition could be a conscious feeling of familiarity based on aggregated similarity to training items, and memory could be a more specific consciously recalled analogy or analogies. A guess would be when there is no feeling of familiarity with the correct response, hence chance accuracy. Participants could be aware that there is some basis for their decision in familiarity or analogy, but their inability to report a rule may simply reflect their lack of any knowledge, conscious or otherwise, of one. These considerations prompted an exploratory analysis of the relationship between accuracy rates to Gen 1 items (which showed the strongest overall

[11] A Bayesian one sample t-test on the guess accuracies (performed in JASP) gave Bayes Factors (BF_{10}) of 0.147 for Gen 1, 0.430 for Trained, 0.302 for Gen 2, and 0.127 for Gen 1+2. Following a widely used interpretation (van Doorn et al., 2021) there is moderate evidence for the null hypothesis (that guess accuracy was not above chance) for Gen 1 and Gen 1+2 data, and weak evidence in the case of Trained and Gen 2.

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

learning effect when considered alone and were consistent across versions) and their similarity to training items.

Experiment 2

How would item-level similarities affect performance in the test phase? Consider the Gen 1 test item "While sitting by the wild flowers I heard the sound of gi bees / ro bees." The most obvious possibility is that the test item (the sentence meaning and picture combination) activates memories of training items that are similar to it and the prevalence of gi and ro in those items guides the decision. In this case, the test item is clearly more similar to the "gi" trained items (which are about animate things in predominantly natural contexts) than the "ro" trained items (which are about inanimate things in predominantly man-made contexts). Given this contrast between the two types of training items, one would expect a strong preference for the correct choice, and little variation between items since the overall living/nonliving contrast is consistent over the test items. But if this were the case then effects would perhaps be expected to be numerically larger, and more robust, than they appear to have been over replications (conceptual or otherwise) of the original W2005 study. An alternative possibility is that the relationship between similarity and accuracy is nonlinear and that only relatively high levels of similarity are sufficient to influence determiner selection. For example, for the test item "While sitting by the wild flowers I heard the sound of gi bees" similar training items might be "Sitting under the tree I was bothered by gi flies" or perhaps "I was amazed when gi bird ate from my hand." The greater the number of such strong "attractors" the greater the likelihood of selecting a certain determiner. That item-level similarities only have an effect at high values is consistent with instance-based computational models that use the k-nearest neighbour metric (Daelemans & van den Bosch, 2010)—k being a parameter that determines the number of examples (or distances), ranging from nearest to furthest in similarity from a new instance, to be used as the basis for categorisation.

To quantify test-to-training item similarities a separate similarity rating study was run on a new group of participants drawn from the same population as Experiment 1. Given that none of the training items containing the incorrect determiner were likely to be strong attractors, and to reduce the number of judgments required, participants only judged the similarity between a test item and each of the training items that contained the correct determiner (e.g., all of the training items that contained "gi" for the above example, none of which, note, contained the test item noun *bees*). This was to identify the test items with the highest overall similarity to training items, the prediction being that they should have the highest determiner selection accuracies. Items with lower similarities may have accuracies around chance, even if they might actually be more similar to training items containing the correct than the incorrect determiner. The experiment focused on the Gen 1 items since these were constant over the two versions and hence could provide the most reliable estimates of item-level accuracy rates (the accuracy of each item being calculated over a total of 54 observations, as opposed to either 24 or 30 for Gen 2 items).

Participants

Thirty-four participants were recruited via Prolific Academic using the same screening criteria as in Experiment 1 with the addition of not having participated in Experiment 1. Their mean age was 23.3 years (range = 20–31), 68% identifying as female.

Methods

Materials

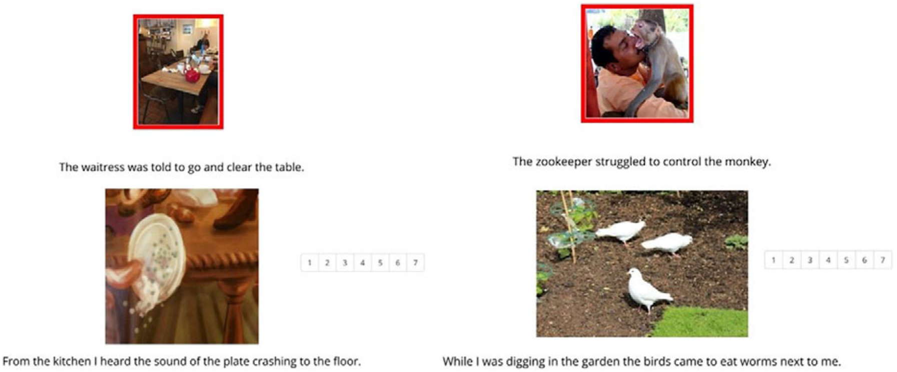
The 48 sentence-picture combinations from the training phase of Experiment 1 and the eight sentence-picture combinations from the Gen 1 test phase were used in this experiment. The novel determiners were replaced with "the" so that decisions would be based on semantic, rather than sound, similarity. The sentences were presented visually and appeared below the corresponding picture (see Figure 5).

Procedure

Participants compared each of the eight Gen 1 test items to each of the 12 training items containing the correct determiner for a total of 96 judgments. In each trial, participants were presented with a "cue" generalization test item picture + (written) sentence combination and a "comparison" item, which was one of the trained picture + sentence combinations that had contained the correct test item determiner (see Figure 5 for two examples). Participants were required to judge how likely they thought it would be that the "cue" item would remind them of the "comparison" item using a seven-point scale ranging from one (very unlikely to remind me) to seven (extremely likely to remind me).[12] There were eight practice trials featuring each of the cue items with a randomly selected comparison item. The experimental items were then presented in blocks containing the same cue item, for a total of eight blocks with 12 comparisons in each block, with the order of blocks and trials within blocks randomized for each participant. The experiment was run on Gorilla SC.

Data coding and analysis

We calculated a mean similarity score for each Gen 1 test item in the following way: first, the mean over participants for each comparison was calculated. Then, the


The waitress was told to go and clear the table.

From the kitchen I heard the sound of the plate crashing to the floor.

The zookeeper struggled to control the monkey.

While I was digging in the garden the birds came to eat worms next to me.

Figure 5. Two example trials from the similarity judgment task (cue item in red box, comparison item below). Mean rated similarities were 5.09 (*SD* = 1.51) for the table-plate comparison and 2.36 (*SD* = 1.54) for the monkey-birds comparison.

[12] Participants were not asked to rate "similarity" because this invites the question as to what the dimensions of similarity should be. "Likelihood of reminding" was used instead because it corresponds more directly to the psychological process that is at issue. However, for convenience "similarity" will be used to refer to the variable derived from this task.

12 comparisons for each cue item were averaged to derive each cue/test item's mean similarity to the training items that contained the correct determiner (e.g., the mean similarity for the "gi bees" test item was 3.242, SD = 0.835). Next, we calculated Pearson's correlation between mean item similarity and mean item accuracy elicited in Experiment 1; we also added mean item similarity to the fixed effect structure of Model 4 (where source was the only fixed effect) to see if this addition could significantly improve model fit and if similarity could significantly predict the likelihood of choosing the correct phrase in Experiment 1. We also report the correlations between similarity and item accuracy for each source category.

Results

There were 33 participants after the exclusion of one participant who had a mean decision time of 666 ms, which was deemed unreasonable for this task (mean over the remainder, 3087 ms, *SD* = 1182). Over the eight Gen 1 test items the correlation between mean item similarity and item accuracy in the main experiment was significant, $r\,(6) = 0.784$, $p = 0.021$. When similarity was entered into the lme model for Gen 1 (Model 6) it produced a statistically significant effect, $Z = 2.733$, $p = 0.006$, and the model showed a significantly improved fit compared with Model 4 where only source was entered as the fixed effect, $X^2 = 6.54$, $df = 1$, $p = 0.010$. Therefore, the overall similarity (or, more exactly, the likelihood of reminding) between a test item and the training items containing the correct determiner was a significant predictor of the proportion of selections of the correct determiner in the test task, accounting for 61.5% (derived as r-squared) of the variance in test item accuracy. The relevant scattergram is shown in Figure 6. The two items with the highest accuracy and similarity are *gi/ro bees* and *gi/ro monkeys.* The item with the lowest similarity is *ro/gi clocks.*

The correlations between similarity and item accuracy for each source category were as follows: for guess, $r\,(6) = 0.784$, $p = 0.065$; for intuition, $r\,(6) = 0.309$, $p = 0.551$; for

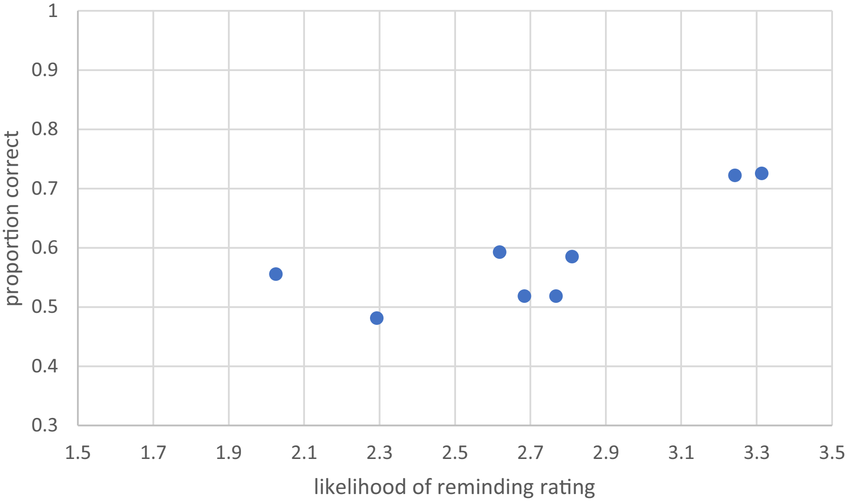


Figure 6. The relationship between similarity (likelihood of reminding rating) and accuracy for the Gen 1 test items (54 unaware participants).

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

memory/knowledge, r (6) = 0.608, p = 0.200. Bearing in mind the relatively small amount of data involved when accuracy is broken down to this level, and the small number of items, there is no evidence that the similarity effect is confined to any particular source categories.

Discussion

When participants/learners perform with above-chance accuracy on items that they have not encountered before it is tempting to conclude that they are utilising knowledge of a rule. But, as the history of AGL research shows us (Ziori & Pothos, 2015), this can only be established if any effects of analogy and similarity on response accuracy are excluded or controlled for. As a follow-up study to Experiment 1, Experiment 2 gathered similarity statistics for the first set of generalization items (Gen 1), requiring a new group of participants drawn from the same pool to judge how likely each test item was to remind them of the training items that contained the correct determiner. It was found that this similarity measure accounted for a significant 61.5% of the variability in the probability of selecting that determiner in Experiment 1 for participants who were classified as being unaware of the animacy rule. Hence it would appear that, in the absence of conscious rule knowledge, these participants' responses were based, to a large extent, on similarity to trained items.

It has to be acknowledged that the present similarity analysis is post hoc and based on just 8 test items. Clearly a larger number of items should be sampled, and ideally intentionally manipulated as part of the experiment design before strong conclusions can be drawn (for example, following the logic of "balanced chunk strength" designs in artificial grammar learning research, Knowlton & Squire, 1996). Note, also, that there is no way of knowing from the present data what the basis for the similarity judgments was. Participants were simply asked to judge how likely it would be that the target item would "remind" them of the comparison (training) item. While in part this would have been determined by the semantic similarity between the critical nouns (e.g., bees and flies) it was also likely a function of the context. For example, the "table-plate" item in Figure 3 had the highest mean similarity rating of any item at 5.30, but this seems to derive more from the context than any similarity between tables and plates as such (but note that the context implies plates—only saucers are evident in the cue picture). Of course, there is no reason why what is designed as a semantic generalization over nouns could not, in fact, be learned, at least in part, as a generalization, or set of partial generalizations, over contexts (see Bovolenta & Williams, 2022, for evidence of implicit learning of an open/enclosed space distinction governing the distribution of prepositions). Nor can it be claimed on the basis of this data that there was no residual effect of grammaticality independent of similarity. It remains to be seen whether a test item set with intentionally constructed low item similarities would produce above-chance responses in the absence of a similarity correlation.

Within SIL the only study to examine the potential impact of item similarity so far is Paciorek & Williams (2015a). The overall similarity between test and relevant training items seemed to pattern with false memory effects across three experiments—Experiments 1, 2, and 4 showed decreasing false memory effects across item sets with decreasing similarity to training items. Unlike the present experiment though, there were no correlations between similarity and the false memory effect at the individual item level in any experiment. However, in Paciorek & Williams (2015a) item similarities were derived computationally from distributional information in a large corpus

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

and were based on the similarity between nouns, disregarding sentence context (equivalent here to considering the similarity between the target noun "bees" and the training noun "flies"). It is possible that the subjective behavioral approach adopted here, using the complete training and test items, provides a more relevant measure of item similarities.

General discussion

The W2005 study has served as a particular focus of debate over the possibility of learning without awareness, in the specific sense of awareness at the level of understanding (Schmidt, 1990), in adult SLA. Can learners show sensitivity to a linguistic generalization—in this case, a correlation between novel determiners and noun animacy—without being aware of it? Experiment 1 set out to address this question using a strategy for minimising the influence of conscious rule knowledge that is arguably more stringent than in previous replication attempts—selection of participants who do not show even vague recognition of the rule in postexperiment debriefing and within that group, selection of test phase responses subjectively judged as being based on the implicit source categories of guess or intuition. Contrary to several previous replication attempts (studies 1, 2, 5, 7, 8, Table 1), but consistent with W2005 (and studies 3 and 4, Table 1), accuracy for the first set of generalization items (Gen 1) was significantly above chance for participants who were unaware of the animacy rule according to the postexperiment debriefing. However, when the responses for this unaware group were analyzed by source category, accuracy for Gen 1 items was above chance for responses based on intuition and memory, but not guesses. Previous W2005 replications have also failed to find a significant effect for guess responses (studies 3, 4, 5, 6, 9, 10, Table 1) even when including substantial numbers of participants who are aware by verbal report (studies 5 and 6, Table 1). It appears, therefore, that in this paradigm, responses tend to be significantly above chance only when participants have what Dienes & Scott (2005) refer to as some degree of confidence, or judgment knowledge—a feeling that their response is correct without knowing why (intuition) or because of a feeling of familiarity (memory, or for a generalization item, false memory).

It was hypothesized that in Experiment 1 intuition and memory-based responses were not a reflection of unconscious rule knowledge (which if it had been present should have elevated guess responses) but conscious feelings of familiarity based on the similarity between test items and specific training items. In Experiment 2 these similarities were found to account for a significant 61.5% of variability in response accuracy. Since the analysis was only based on the 8 items appearing in the first generalization test this result can only be taken as indicative of a potentially more general relationship between response accuracy and item similarity in this paradigm. Nevertheless, in the following, we discuss the theoretical implications of there being such a relationship, as a spur to consider similarity as a factor in future implicit language learning studies.

Instance-based memory models have a relatively long, but continuous, history in cognitive psychology (Hintzman, 1986; Jamieson, Johns, Vokey & Jones, 2022; Johns, Jones & Mewhort, 2012) and provide a simple and intuitively plausible framework for thinking about how similarity effects arise in cognition. The essential insight is that decision behavior in a test phase (e.g., recognition memory or grammaticality judgment) can be explained by nothing more than resonances between test items and memories of training items stored as instances in memory, with the strength of resonance determined by similarity. The challenge that such models present to our conception of learning is that there are no reorganizational or abstraction processes. Episodes of experience are simply laid down as traces in memory; that is, "learning" consists of nothing more than remembering. Test performance only reflects processes occurring in the moment of retrieval, and behavior that may be described as "rule-like" is in fact underlyingly analogical. In a linguistic context, such models have profound implications because they challenge the idea that abstract rules are represented in the mind (Ambridge, 2019). Of course, the tension between analogical and rule-based mechanisms is very familiar to linguists (and is increasingly topical given the apparent success of Large Language Models). The present microanalysis of item behavior is not intended to inform this larger debate about the nature of linguistic knowledge. But at least in the context of small-scale, initial-stage learning studies such as the present, with limited training, immediate testing, and no opportunities for consolidation, an instance-based and analogical explanatory framework seems quite plausible.

In W2005 the notion of "learning without awareness" referred to learning, and applying, a semantic generalization or form-meaning connection without being aware of it (at the level of understanding, Schmidt, 1990). But if all that is learned are instances of input then there is nothing to be unaware of in that sense, and no "hidden" meaning with which a form is associated. And whereas participants may be genuinely unaware of the generalization that, say, "gi" goes with living things, in the moment of making decisions at test they may be aware that the "gi bees" test item reminds them of the training item about "gi flies" (Hintzman's, 1986, "echo content"), selecting "gi" on that basis, with attribution to memory. Or the effect of similarity may aggregate over several test-training item similarities, leading to a response based on a conscious feeling of familiarity ("echo intensity") that might be attributed to intuition.

In fact, conscious feelings of familiarity may drive grammaticality judgments even when participants report they are guessing. In an artificial grammar learning (AGL) experiment, Scott & Dienes (2008) found that when participants claimed they were guessing (and had no confidence) their decisions were still related to their perception of the familiarity of the item (which was correlated with objective similarity statistics), even though the accuracy of guess decisions overall was not above chance. A similar pattern was found in the present Gen 1 data, where guesses were at chance but showed no sign of being more weakly correlated with similarity than responses based on intuition or memory. This pattern can arise because "guess" indicates a lack of judgment knowledge with respect to whether or not the item is grammatical but the participants can still base their decisions on conscious feelings of familiarity (for want of any other criteria) and yet have no confidence that this is an indication of grammaticality. The learning experience may provide a conscious basis for a decision even under guessing, just not with respect to grammaticality as such. Hence, previous failures to find above chance responding when guessing in SIL research cannot be taken as evidence of a complete lack of any effect of the training experience unless it can be shown that there is no underlying correlation with similarity statistics.

In implicit language learning research, it should be possible to reveal the complex interplay of confidence, familiarity, and rule knowledge through think-aloud protocols and debriefing. But interesting as it may be to probe further into the level of awareness of analogical processing at test, awareness of this kind is not relevant to the issue of implicit learning in the sense of learning without awareness at the level of understanding. If the goal is to probe the existence of this form of learning then the essential point of awareness assessment is still to remove contamination from explicit rule knowledge. The influence of analogy or familiarity may or may not be conscious, but this is not relevant to the evaluation of learning without awareness at the level of understanding.

Limitations and future directions

An obvious limitation of the present study is that the similarity analysis was post hoc (in response to the unexpected source judgment results) and based on very few items (just the eight Gen 1 items, where the learning effect was strongest). While the results can only be taken as suggestive, we believe that they call for further investigation, not necessarily within this paradigm, but in any implicit learning scenario where the similarity between test and training items can be estimated along relevant dimensions. What is clear is that it is rather naïve to assume that because participants do not consciously know a linguistic rule that their marginally above-chance generalization performance is a reflection of unconscious knowledge of that rule. If future implicit language learning research were to suggest that responses are in fact based on no more than resonances with memories of training items, then this would be either an indictment of small-scale laboratory learning experiments, or a reflection of the reality of language learning, depending on one's point of view. Or if it can be shown that behavior is truly a reflection of unconscious generalizations, then the rule-based approach will be vindicated. Alternatively, both processes may be operative at the unconscious level and difficult to disentangle (Ziori & Pothos, 2015). Still, analogical and rule-based processes may be differentially affected by parameters such as the "critical mass" of training items, opportunities for consolidation, degree of variation in input (e.g., more or less bland items)[13], or fine details of item presentation regime or task set up. To explore these issues we need to combine appropriate item-level statistics with the means of removing contamination from awareness at the level of understanding, as opposed to awareness at the level of familiarity and analogy. And if it turns out that performance reflects nothing more than item-level analogies then the issue of whether there is learning without awareness of rules disappears, for there is nothing to be unaware of.

Competing interest. The authors declare no conflicts of interest.

References

Ambridge, B. (2019). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*, 509–559.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407.

Bovolenta, G., & Williams, J. N. (2022). Implicit Learning in Production: Productive generalization of new form–meaning connections in the absence of awareness. *Language Learning, n/a.*

Bower, G. H., & Winzenz, D. (1970). Comparison of associative learning strategies. *Psychonomic Science*, *20*, 119–120.

Brown, H., Smith, K., Samara, A., & Wonnacott, E. (2021). Semantic cues in language learning: an artificial language study with adult and child learners. *Language, Cognition and Neuroscience*, 1–23.

Cayado, D. K. T., & Chan, R. K. W. (2022). The influence of prior linguistic knowledge on second language semantic implicit learning: Evidence from Cantonese–English bilinguals. *Language Learning, n/a.*

Chan, R., & Leung, J. (2018). Implicit knowledge of lexical stress rules: Evidence from the combined use of subjective and objective awareness measures. *Applied Psycholinguistics*, *39*, 37–66.

[13] We thank Padraic Monaghan for this suggestion.

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

Chen, W. W., Guo, X. Y., Tang, J. H., Zhu, L., Yang, Z. L., & Dienes, Z. (2011). Unconscious structural knowledge of form-meaning connections. *Consciousness and Cognition*, *20*, 1751–1760.

Daelemans, W., & van den Bosch, A. (2010). Memory-based learning. In A. Clark & C. Fox & S. Lappin (Eds.), *Handbook of computational linguistics and natural language processing* (pp. 154–178). Oxford, UK: Wiley-Blackwell.

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: distinguishing structural knowledge and judgment knowledge. *Psychological Research*, *69*, 338–351.

Faretta-Stutenberg, M., & Morgan-Short, K. (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). Somerville, MA: Cascadilla Proceedings Project.

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*, 02676583211044950.

Hama, M., & Leow, R. P. (2010). Learning without awareness revisited. *Studies in Second Language Acquisition*, *32*, 465–491.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, *93*, 411–428.

Jamieson, R. K., Johns, B. T., Vokey, J. R., & Jones, M. N. (2022). Instance theory as a domain-general framework for cognitive psychology. *Nature Reviews Psychology*, *1*, 174–183.

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology*, *65*, 486–518.

Johnstone, T., & Shanks, D. R. (1999). Two mechanisms in implicit artificial grammar learning? Comment on Meulemans and Van der Linden (1997). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 524–531.

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*, 711–734.

Kim, K. M., Maie, R., Suga, K., Miller, Z. F., & Hui, B. (2023). Learning without awareness by academic and nonacademic samples: An individual differences study. *Language Learning*, *73*, 1087–1126.

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*, 169–181.

Leow, R. P. (2015). Implicit learning in SLA: Of processes and products. In P. Rebuschat (Ed.), *Implicit and Explicit Learning of Languages* (pp. 47–67): John Benjamins.

Leung, J. H. C., & Williams, J. N. (2012). Constraints on implicit Learning of grammatical form-meaning connections. *Language Learning*, *62*, 634–662.

Leung, J. H. C., & Williams, J. N. (2014). Crosslinguistic differences in implicit language learning. *Studies in Second Language Acquisition*, *36*, 733–755.

Li, F., Zhao, C., & Li, W. (2020). Implicit learning of semantic preferences of English words by Chinese learners. *Consciousness and Cognition*, *84*, 102986.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning*, *68*(2), 321–391.

Paciorek, A., & Williams, J. N. (2015a). Semantic generalization in implicit language learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*, 989–1002.

Paciorek, A., & Williams, J. N. (2015b). Semantic implicit learning. In P. Rebuschat (Ed.), *Implicit and Explicit Learning of Languages* (pp. 69–90): John Benjamins.

Pham, T., Kang, J. H., Johnson, A., & Archibald, L. M. D. (2020). Feature-focusing constraints on implicit learning of function word and meaning associations. *Applied Psycholinguistics*, *41*, 401–426.

R Core Team. (2023). R: A language and environment for statistical computing (Version 4.3.2) [Computer software]. http://www.R-project.org

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

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*, 299–334.

https://doi.org/10.1017/S0272263124000500 Published online by Cambridge University Press

Rebuschat, P., Hamrick, P., Sachs, R., Riestenberg, K., Ziegler, N., Bergsleithner, J., Frota, S., & Yoshioka, J. (2013). Implicit and explicit knowledge of form-meaning connections: Evidence from subjective measures of awareness. *Noticing and second language acquisition: Studies in honor of Richard Schmidt, 255*, 23.

Sachs, R., Hamrick, P., McCormick, T. J., & Leow, R. P. (2020). Exploring the veridicality and reactivity of subjective measures of awareness: Is a "guess" really a guess? *Studies in Second Language Acquisition*, *42*, 919–932.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 129–158.

Scott, R. B., & Dienes, Z. (2008). The conscious, the unconscious, and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1264–1288.

Shanks, D. R., & St. John, M. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*, 367–447.

van Doorn, J., van den Bergh, D., Bohm, U., Dablander, F., Derks, K., Draws, T., … & Wagenmakers, E. J. (2021). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin and Review*, *28*, 813–826.

Williams, J. N. (2005). Learning without awareness. *Studies in Second Language Acquisition*, *27*, 269–304.

Zhao, C., Kormos, J., Rebuschat, P., & Suzuki, S. (2021). The role of modality and awareness in language learning. *Applied Psycholinguistics*, *42*, 703–737.

Ziori, E., & Pothos, E. (2015). Artificial grammar learning: An introduction to key issues and debates. In P. Rebuschat (Ed.), *Implicit and explicit learning of languages*. (pp. 249–273). Amsterdam: John Benjamins.

Cite this article: Williams, J. N., & Xue, Y. (2024). Learning without awareness revisited and reconsidered: A conceptual replication and extension. *Studies in Second Language Acquisition,* 46: 1231–1257. https://doi.org/10.1017/S0272263124000500