Studies in Second Language Acquisition, 2015, **37**, 299–334.
doi:10.1017/S0272263115000145

TRIANGULATING MEASURES OF AWARENESS

A Contribution to the Debate on Learning without Awareness

Patrick Rebuschat
Lancaster University

Phillip Hamrick
Kent State University

Kate Riestenberg
Georgetown University

Rebecca Sachs
Virginia International University

Nicole Ziegler
University of Hawai'i at Mānoa

The authors would like to thank John Williams, Ron Leow, Kara Morgan-Short, Mandy Faretta-Stutenberg, and Daniel Jackson for thoughtful discussions and feedback. We are grateful to Aidan Dugan and Michael Mungiello for their assistance in data collection and analysis as well as to Ana-Maria Nuevo, Patrick Callier, and Jennifer Sclafani for permitting us to recruit participants in their classes. Finally, we would like to thank our two anonymous reviewers for their extensive comments and suggestions, which significantly improved the quality of the manuscript.

Correspondence should be addressed to Dr. Patrick Rebuschat, Department of Linguistics and English Language, Lancaster University, Lancaster, LA1 4YL. E-mail: p.rebuschat@lancaster.ac.uk

Williams's (2005) study on "learning without awareness" and three subsequent extensions (Faretta-Stutenberg & Morgan-Short, 2011; Hama & Leow, 2010; Rebuschat, Hamrick, Sachs, Riestenberg, & Ziegler, 2013) have reported conflicting results, perhaps in part due to differences in how awareness has been measured. The present extension of Williams (2005) addresses this possibility directly by triangulating data from three awareness measures: concurrent verbal reports (think-aloud protocols), retrospective verbal reports (postexposure interviews), and subjective measures (confidence ratings and source attributions). Participants were exposed to an artificial determiner system under incidental learning conditions. One experimental group thought aloud during training, another thought aloud during training and testing, and a third remained silent, as did a trained control group. All participants were then tested by means of a forced-choice task to establish whether learning took place. In addition, all participants provided confidence ratings and source attributions on test items and were interviewed following the test. Our results indicate that, although all experimental groups displayed learning effects, only the silent group was able to generalize the acquired knowledge to novel instances. Comparisons of concurrent and retrospective verbal report data shed light on the conflicting findings previously reported in the literature and highlight important methodological issues in implicit and explicit learning research.

Implicit learning, essentially the process of acquiring unconscious (implicit) knowledge, is a fundamental aspect of human cognition (Perruchet, 2008). The term *implicit learning* was coined by Arthur Reber (1967) to describe a process during which participants derive knowledge from a complex, rule-governed stimulus environment without intending to and without becoming aware of the knowledge they have acquired. In contrast, the term *explicit learning* usually refers to a process during which participants acquire conscious (explicit) knowledge; this is generally associated with intentional learning conditions (e.g., when participants are instructed to look for rules or patterns).[1]

The field of SLA has a long-standing interest in the topic of implicit and explicit learning. Within this area, two closely related questions have attracted a considerable amount of discussion. The first question concerns the possibility of learning without awareness (see, e.g., Godfroid, Boers, & Housen, 2013; Hama & Leow, 2010; Leow, 1997, 1998, 2000; Leow & Hama, 2013; Leung & Williams, 2011, 2012, 2014; Schmidt, 1990, 1995, 2001; Williams, 2005, 2009). The second question is methodological in nature and refers to the measurement of awareness. The focus here has been on how to measure awareness at the time of encoding (i.e., while participants are engaged in a learning task; e.g., Godfroid & Schmidtke, 2013; Leow, 1997) and awareness of what has been learned (i.e., of the product of learning; e.g., Ellis, 2005; Hamrick & Rebuschat, 2012; Rebuschat & Williams, 2012). The current study aims to contribute to the debate surrounding implicit learning by triangulating three measures of awareness to investigate more valid ways of examining the role of awareness in second language (L2) acquisition. The awareness measures in question—concurrent verbal reports (think-aloud protocols), retrospective verbal reports, and subjective measures—have been widely used in both cognitive psychology and SLA research (for reviews, see Bowles, 2010; Leow, Grey, Marijuan, & Moorman, 2014; Rebuschat, 2013), but this is the first study to directly compare all three.

BACKGROUND: WILLIAMS (2005) AND THE DEBATE ON LEARNING WITHOUT AWARENESS

Although it is accepted that attention and awareness play important roles in learning (see Leow & Bowles, 2005; Schmidt, 2001, for reviews), the early notion that low levels of awareness of linguistic phenomena are necessary for their acquisition (e.g., Schmidt, 1990) has recently been challenged. In a widely cited study, Williams (2005) investigated the acquisition of an artificial determiner system in a meaning-oriented task. Participants were exposed to four new determiners (*gi*, *ro*, *ul*, and *ne*) that encoded both distance (near vs. far) and animacy (animate vs. inanimate). Participants were told that the determiners functioned like English determiners except that they also encoded distance; for example, *ul* and *ne* were used for far objects, whereas *gi* and *ro* were used for near objects.[2] Participants were not informed that the artificial determiners also encoded animacy: *Gi* and *ul* were used with animate objects, whereas *ro* and *ne* were used with inanimate ones. In other words, the role of animacy in determiner usage served as a hidden regularity.

Participants were thus trained on the artificial system under incidental learning conditions, and although they were informed that there would be a "memory test for some of the sentences at the end" (Williams, 2005, p. 282), they were not told that they would be tested on the hidden regularity. In the exposure phase, participants were asked to listen to each training sentence (e.g., "I spent an hour polishing ro table before the dinner party"), to judge whether the novel determiner used in the sentence meant near or far, to repeat the sentence verbatim, and to form a mental image of the general situation described by the picture. The testing phase had two sections. In the first section, participants read part of a novel sentence such as "The lady spent many hours sewing . . ." and then had to select the appropriate segment to complete it from two options that matched in their distance values and differed only according to animacy (e.g., ". . . *gi* cushions / *ro* cushions"). Participants were then interviewed to determine if they were aware of the animacy regularity. In the second section, those participants who were still unaware of the relevance of animacy were given the same test sentences but this time with the instruction to find out the rules that governed determiner use. They were then interviewed again to assess the conscious or unconscious status of any acquired knowledge.

Williams (2005) found that 80% of participants reported themselves to be unaware of the relevance of animacy after the first part of the test, despite performing at 61% accuracy, which was significantly above chance. After the rule-search task, 50% of participants were still unaware of the rule, yet their accuracy was still significantly above chance (58%). The results were interpreted as showing that participants can establish new form-meaning connections without becoming aware of what those connections are. In other words, learning without awareness was taken to be possible.

Williams (2005) inspired several extension studies (Faretta-Stutenberg & Morgan-Short, 2011; Hama & Leow, 2010; Rebuschat, Hamrick, Sachs, Riestenberg, & Ziegler, 2013). In an important first extension, Hama and Leow (2010) adapted the methodology of Williams (2005) to assess whether learning without awareness is indeed possible. According to Hama and Leow, the discrepancy between Williams (2005) and earlier research that had found no evidence for learning without awareness (Leow, 2000) can be explained by methodological differences. Whereas Leow (2000) employed an online measure of awareness (think-aloud protocols) to assess awareness during the exposure and testing phases, Williams (2005) relied on an offline measure (retrospective verbal reports). As a consequence, Leow (2000) investigated the role of awareness at the time of encoding, whereas Williams (2005) examined whether exposure had resulted in conscious or unconscious knowledge. That is, Leow (2000) focused on the process of learning, whereas Williams (2005) focused on the product. Hama and Leow's (2010) extension of Williams (2005) modified the original design by adding think-aloud protocols to the experimental tasks (i.e., participants were prompted to verbalize their thoughts while performing the tasks). They also added a production task and changed the forced-choice test to include four options instead of two. In addition, they kept all of the tasks in the auditory modality, in contrast to Williams (2005), who had used the auditory modality for training and the written modality for testing.

In Hama and Leow's (2010) study, 43 native speakers of English were trained on the artificial determiner system employed by Williams (2005) by means of the same type of exposure task. Afterward, the participants were required to perform two tests (multiple-choice recognition and production). The recorded verbal reports were transcribed and coded as *understanding*, *noticing*, or *no report*. A verbal report was coded as *noticing* when some aspect of animacy was mentioned or commented on, as *understanding* when correct rules related to animacy were mentioned, and as *no report* when the report did not fall under the coding categories of noticing or understanding. Hama and Leow (2010) found no evidence for awareness of animacy during the exposure phase. However, the think-aloud protocols for the test phase clearly provided evidence for awareness at the level of noticing and at the level of understanding (see Schmidt, 1995, 2001). On the basis of the data, nine participants were classified as aware of the hidden regularity and 34 as unaware. Further analyses indicated a significant learning effect in the aware group on both tests but no learning effect in the unaware group. In other words, learning appeared restricted to those participants who became aware of the hidden regularity. These results are supported by another extension study, Faretta-Stutenberg and Morgan-Short (2011).

Despite assessing awareness at different stages of the learning process, the measures employed by Williams (2005) and Hama and Leow (2010) share a fundamental limitation: They both rely on verbalization (or a lack thereof) to distinguish implicit and explicit processes (Leow and colleagues) and implicit and explicit knowledge (Williams). In the case of think-aloud protocols, the assumption is that learning proceeds without awareness when participants are unable to verbalize relevant features of the target system while engaged in the training or test tasks. In the case of retrospective reports, it is assumed that knowledge is unconscious when participants show an effect of training (e.g., above-chance accuracy on a forced-choice task), despite being unable to describe the knowledge that underlies their performance. In either case, these assumptions are probably not warranted. For example, awareness may happen more quickly than concurrent verbalization allows expression of, given that "subjective awareness is fleeting and cannot be completely recorded" (Schmidt, 1995, p. 28). In retrospective verbalization, awareness may have decayed in memory by the time participants are asked to report on it. In addition, participants may fail to report knowledge simply because they lack confidence or do not realize that the knowledge is relevant. When participants have the option of not responding during concurrent or retrospective verbal reports, then conscious knowledge, though present, may simply not be detected. Thus both types of verbal reports may not be sensitive enough to capture all of the relevant conscious knowledge.[3]

Although a lack of verbalization does not provide strong evidence for learning without awareness (in the case of think-aloud protocols) or implicit knowledge (in the case of retrospective verbal reports), it is important to note that the presence of verbalization does not imply that all learning in the study involved awareness or that only explicit knowledge was acquired. Both types of verbal report lack exclusivity in the sense that they may be contaminated by unconscious knowledge (Reingold & Merikle, 1990). When think-aloud data suggest that participants were aware of a given complex L2 phenomenon, this does not mean that other aspects of the same phenomenon have not been acquired without awareness. In addition, one needs to ask what processes contributed to participants suddenly becoming aware of a feature in the first place, with implicit processing (e.g., in the form of statistical learning) a possible candidate in this case (e.g., Cleeremans, 2008). Likewise, when participants verbalize knowledge at the end of an experiment, this does not mean that they have acquired only conscious knowledge. In fact, recent research (Grey, Williams, & Rebuschat, 2014; Hamrick & Rebuschat, 2012, 2014; Rebuschat, 2008; Rebuschat & Williams, 2012; Tagarelli, Borges Mota, & Rebuschat, 2011, 2015) indicates that, under incidental learning conditions, participants are quite likely to acquire both implicit and explicit knowledge.

We recently reported the results of another extension of Williams (2005), Rebuschat et al. (2013), which sought to illustrate the usefulness of a measure of awareness that does not rely on verbalization—namely, subjective measures of awareness (Dienes & Scott, 2005; Rebuschat & Williams, 2006; see Rebuschat, 2013, for review). Rebuschat et al. (2013) exposed participants to the artificial system developed by Williams (2005) and then tested them by means of a forced-choice task that also collected confidence ratings and source attributions. That is, in addition to deciding on the appropriate artificial determiner, participants were also asked how confident they were in their decision and what the basis of their decision was (*guess*, *intuition*, *memory*, or *rule knowledge*; see the Methods section for details). At the end of the experiment, participants also provided retrospective verbal reports. Rebuschat et al. (2013) found that 70% of participants were able to verbalize some knowledge regarding the hidden animacy regularity and that learning in the experiment was restricted to those participants who had acquired explicit knowledge, a result that is in line with Hama and Leow (2010) as well as Faretta-Stutenberg and Morgan-Short (2011). Importantly, however, the analysis of the confidence ratings and source attributions further showed that although participants were aware of having acquired knowledge, at least some of that knowledge remained unconscious, as evidenced by above-chance performance in responses based on guess and intuition. That is, subjects had also developed implicit knowledge as a result of exposure, which supports Williams's (2005) findings.[4]

The present study extends our work by adding two think-aloud groups to Rebuschat et al. (2013) to directly compare the three measures of awareness that have been used in Williams (2005) and subsequent extensions (Faretta-Stutenberg & Morgan-Short, 2011; Hama & Leow, 2010; Rebuschat et al., 2013)—namely, concurrent verbal reports (think-aloud protocols), retrospective verbal reports, and subjective measures of awareness. Our objective is to determine the advantages and disadvantages of each measure to arrive at more sensitive, and more valid, ways of measuring awareness. Specifically, we wanted to establish (a) the relative sensitivity of each measure, (b) whether the measures are potentially reactive,[5] and (c) how the measures can contribute to our understanding of what participants become aware of and when.

METHODS

Participants

Fifty-two undergraduate students (30 women, $M_{\text{age}} = 20$) were randomly assigned to a trained control group[6] (n = 15) or one of three experimental groups: one group that did not think aloud during the exposure phase or the test phase (silent group, n = 14), one group that thought aloud only during the exposure phase (think-aloud exposure group, n = 12), and one group that thought aloud during both the exposure phase and the test phase (think-aloud throughout group, n = 11). The key manipulations that differentiated these groups were thus (a) the presence or absence of a hidden regularity in the exposure phase (controls vs. experimental participants), (b) the use of think-aloud protocols (silent group vs. think-aloud groups), and (c) the timing of the think-aloud protocols (exposure phase only vs. exposure and testing phases).

All participants were native speakers of English. Fourteen participants reported having an additional native language; these included German (n = 3), Mandarin (n = 3), Spanish (n = 2), Cantonese (n = 1), Farsi (n = 1), French (n = 1), Japanese (n = 1), Korean (n = 1), and one student who indicated both Tibetan and Hindi as additional native languages. Forty-seven participants reported studying the following foreign languages: Spanish (n = 34), French (n = 14), German (n = 13), Latin (n = 6), Mandarin (n = 6), Italian (n = 4), Arabic (n = 3), Korean (n = 3), Russian (n = 3), Portuguese (n = 2), Catalan (n = 1), Hindi (n = 1), and Japanese (n = 1). The experimental and control groups did not significantly differ with respect to age, sex, number of linguistics courses taken, number of L2s, or undergraduate level (all *p*s > .31) except in the case of the number of linguistics courses taken by the trained controls (M = 1.87) and the think-aloud throughout group (M = 0.91), $t(16.27) = 2.26$, $p = .04$ (corrected for violation of Levene's test).

Materials

We employed the artificial determiner system from Williams (2005). This system consists of four artificial determiners (*gi, ro, ul,* and *ne*) that encode both distance (near vs. far) and animacy (animate vs. inanimate). The determiners *gi* and *ro* precede nouns whose referents are near, whereas *ul* and *ne* precede nouns whose referents are distant. Moreover, *gi* and *ul* precede nouns that refer to animate (natural, living, or moving) entities, whereas *ro* and *ne* correspond with inanimate (manmade, nonliving, or stationary) ones. As in previous studies, all participants were pretrained explicitly on the near-far distinction before the incidental exposure phase, but none were informed of the animacy regularity. For a more detailed exposition of the stimulus design, including lists of the training and test items, see Rebuschat et al. (2013).

Training Items. The artificial determiners were placed in noun phrases (NPs) that were presented in the context of sentences (e.g., "The park warden reassured us that gi bears were tame enough to pet"). Two training sets were constructed to include 12 animate and 12 inanimate nouns, with half of each category presented in near contexts and the other half in far contexts, amounting to six NPs of each type (near-animate, far-animate, near-inanimate, and far-inanimate). Over the course of the exposure, the 24 determiner-noun combinations were repeated six times, for a total of 144 items, with the order of the sentences randomized within each set and plurality counterbalanced across alternating sets. That is, if *gi bears* appeared in Set 1, then *gi bear* appeared in a different sentence in Set 2; then Set 1 was presented again, and so on. Following Williams (2005), nouns appeared with only one determiner (e.g., near-animate *gi bear/s* but not far-animate *ul bear/s*) to ensure that learning would be of form-meaning connections (e.g., that *gi* is used with animate nouns) as opposed to form-form associations between determiners (e.g., that any noun that takes *gi* [near] also takes *ul* [far]).

Participants in the trained control group were exposed to the same sentences as the experimental groups except that the animacy regularity was removed by changing the determiners. Because all participants received the same explicit pretraining on distance, the near-far meanings of the determiners were maintained for the trained controls, and, as described previously, each NP was presented in both its singular and plural forms. However, each determiner was used half the time with animate nouns and half the time with inanimate nouns so that the trained control group would not be exposed to any reliable determiner-animacy information.

Test Items. All participants completed the same test phase. The test phase consisted of new context sentences, none of which had appeared during training. The sentences were designed to test three types of NPs: trained, partially trained, and new. For the experimental groups, trained items had already occurred in exactly the same form in the exposure phase (e.g., *gi bears*). Partially trained items had occurred during training but in a different determiner-noun pairing. For example, if *ro picture* (the near picture) had occurred in training, then either the singular or plural version appeared in the far context, requiring *ne* on the test. Finally, new items used novel nouns that had not occurred in the exposure phase (e.g., *gi rabbit*). Note that Williams (2005), Hama and Leow (2010), and Faretta-Stutenberg and Morgan-Short (2011) featured only trained and partially trained items (the latter of which they called generalization items), but none of these previous studies used novel nouns. In contrast to the present study, these studies did not contain true generalization items.

Thirty-six test sentences were produced, most of them on the basis of Williams (2005) and Hama and Leow (2010), with six of each type (trained, partially trained, and new) for each animacy class. Plurality and distance values were balanced within each test-item type (e.g., the six trained animate items included three singular and three plural NPs as well as three near and three far NPs), taking care not to confound plurality and distance. The trained and partially trained items were identical to Williams (2005), except for the noun *rat*, which we included instead of *mouse* to avoid irregular plurals. Most of the new NPs were derived from Hama and Leow (2010), plus four other nouns (*hamster*, *camel*, *towel*, and *desk*), which we added for counterbalancing purposes and to test generalization ability.

Procedure

All participants completed (a) vocabulary pretraining on the distance regularity, (b) an exposure phase, (c) a testing phase, and (d) a post-exposure verbal report interview. Participants in the think-aloud exposure group provided think-aloud protocols during the exposure phase, whereas participants in the think-aloud throughout group did so during both the exposure and testing phases. All participants met individually with one of the researchers in a quiet laboratory and were audio recorded while performing the tasks to ensure they had followed the instructions. At the end of the experiment, participants also completed a brief computer-based questionnaire asking for their age, field of study, previous experience in linguistics courses, native language(s), and any foreign languages studied. Where applicable, participants provided additional information regarding their foreign language background, including contexts of instruction, levels of formal schooling, length of study, and self-reported proficiency. All tasks were run on Apple iMacs; the exposure and testing phases were delivered via Cedrus SuperLab Pro (Version 4.0.7b). The entire session took approximately one hour.

Vocabulary Pretraining. The purpose of this activity was to introduce the four novel determiners and their distance meanings in English. The participants were informed that they would be tested on some new words following a vocabulary pretraining activity, which they could complete at their own pace. The activity was administered via Microsoft PowerPoint. Participants were presented with a list of the words (*gi*, *ro*, *ul*, *ne*) and their corresponding English meanings (near or far) and then completed two practice tasks that exposed them to written repetitions of each novel word. They were permitted to repeat the training as many times as they wished, but only three participants (one from the trained control group and two from the silent group) chose to do so, with each repeating it once. After this pretraining, participants were quizzed on whether they had learned the distance meanings of the determiners. The quiz was administered using the online testing website ClassMarker (www.classmarker.com). Participants were required to score 90% or higher on the quiz to move on to the exposure phase, and all were able to do this on their first attempt.

Exposure Phase. The participants were given written instructions explaining the general purpose of the experiment, adapted from Williams (2005, pp. 281–282). They were informed that the four artificial determiners functioned like the English word *the* except that they also encoded the distance meanings from the vocabulary pretraining. Participants were not informed of the animacy regularity in the stimuli. A sample sentence ("The little boy patted gi tiger in the zoo"), which did not recur during the training task, illustrated how the determiners could be used in a sentence context.

Participants were then told that they would be presented with written sentences that included the new determiners from the pretraining phase and that their task was to read each sentence aloud and then indicate, as quickly and accurately as possible, whether the novel word meant near or far by pressing the corresponding key (marked with a sticker) on the computer's keyboard. After each decision, the sentence disappeared, and participants were to repeat aloud the novel determiner together with its noun (e.g., *gi tiger* in the previous example), while simultaneously forming a mental image of the situation.[7] Using the sample sentence as an illustration, participants were instructed to imagine a boy patting a tiger that was close to him. The aim of asking participants to form mental images was to encourage them to process the meanings of the words and the overall situation described in the sentence. This was emphasized as an important aspect of the study by the researcher administering the experiment. All participants completed a short practice session with four sentences that were not repeated in the exposure phase.

Testing Phase. After the exposure phase, all participants completed a two-alternative forced-choice (2AFC) task to assess whether they had learned the hidden regularity, as in Williams (2005) but unlike Hama and Leow (2010), who used four response options. The 2AFC task consisted of 36 new sentences, all presented in the visual modality.[8] For each test item, the computer displayed a sentence context (e.g., "The babysitter poured juice into ___ cups for the children") with two choices of artificial determiners (e.g., *gi* and *ro*) located in the bottom left and right corners of the screen. The determiner choices were always identical in their distance values and differed only in their animacy values. As such, participants could not respond on the basis of distance, as they had done during the exposure phase. Participants were instructed to read through each test sentence in its entirety and to "choose the word that seems more familiar, better, or more appropriate" based on what they had done so far. They entered their choices by pushing the corresponding key (marked with a sticker).

Subjective measures of awareness, specifically confidence ratings and source attributions, were recorded after each response (see Rebuschat, 2013, for a summary of this technique). After reading each test sentence and indicating their decision with regard to the choice of determiner, participants were asked to indicate how confident they were in their decision and what the basis of their decision was. Participants could indicate their level of confidence by selecting one of four response options for each item: *not confident at all* (*guessing*), *somewhat confident*, *very confident*, or *100% confident* (corresponding to the numbers 0, 3, 6, and 9, respectively). Participants were instructed to select the guess category only if they had no confidence whatsoever in their classification decision and truly believed they had been guessing. If they had even slightly more confidence than this, they were asked to select one of the other categories. Confidence ratings were used to determine whether participants had developed conscious or unconscious judgment knowledge (Dienes & Scott, 2005; Rebuschat, 2013). For the source attributions, participants were asked to select one of four options as the basis of their determiner decision: *guess, intuition, memory,* or *rule knowledge*. Participants were instructed to use the guess category only when a decision was based on a true guess (e.g., in the sense that they could just as easily have flipped a coin). The intuition category was to be selected if participants had a gut feeling that they were right but did not know why. They were asked to choose memory when a judgment was based on a recollection of an item from the earlier part of the experiment and rule knowledge for any decision that was based on a rule that they would be able to report at the end of the experiment. Following Dienes and Scott (2005), source attributions were used to determine whether participants had acquired conscious or unconscious structural knowledge.[9]

The 36 test sentences were presented in the same order for all participants. As in Williams (2005), the test sentences were arranged so that animacy comparisons could not be made across adjacent items with the same distance values; for example, test items targeting far-animate NPs (e.g., *ul bees*) were never followed by far-inanimate NPs (e.g., *ne clocks*). Because our study featured three types of test items (as opposed to the two previous studies, which featured two types), we could not use exactly the same item ordering as that employed by Williams (2005). However, we did follow his ordering on the more abstract level of plurality, distance, and animacy features. Before beginning the test itself, participants completed a short practice session with four sentences that were not repeated in the test phase.

Instructions for Concurrent Verbal Reports (Think-Aloud Protocols). As outlined previously, two experimental groups were instructed to think aloud while completing the experiment, with the think-aloud exposure group thinking aloud during the exposure phase only and the think-aloud throughout group doing so during both the exposure and the testing phases. Before the exposure phase, participants in these groups were given instructions regarding how to think aloud. Specifically, they were told that one of our research goals was to obtain a realistic representation of what people are thinking when they are understanding language. They would therefore be asked to "externalize [their] 'inner speech' and speak [their] thoughts aloud" during the experiment. The instructions emphasized to the participants that it was important that they "feel free to say whatever comes to mind . . . in the same way it would normally go through [their] mind[s]" without worrying about giving explanations or using complete sentences. They practiced this by thinking aloud while solving a multiplication problem on scrap paper. Then, the think-aloud participants were given the same instructions and practice items for the exposure phase as those given to the silent group and the trained controls. Later, during the instructions for the testing phase, participants in the think-aloud throughout group were reminded again to think aloud, whereas participants in the think-aloud exposure group were informed that they no longer needed to do so. In transcribing the think-aloud protocols for analysis, all recordings from the think-aloud groups were checked to ensure that these participants had indeed thought aloud as instructed. Two participants assigned to the think-aloud exposure group continued to think aloud during the test, whereas one participant in the think-aloud throughout group did not verbalize his thoughts during the test. For the purposes of analysis, the group designations of these participants were switched to reflect their actual behavior; that is, each was retroactively assigned to the other think-aloud condition.

Retrospective Verbal Reports. Following the testing phase, all participants completed a short interview with one of the researchers. The interview was structured so that the questions gradually became more explicit and direct in probing the participants' awareness of the hidden regularity (animacy). This was to minimize the effect of the interview questions on participants' inferences about the stimuli. The participants were first asked what criteria they had used to make their choices. If they made any references to living-nonliving, moves-does not move, or a similar distinction, they were asked at what point they had become aware of this difference. The participants were then asked whether they had ever indicated rule knowledge as a basis for their decisions. If so, they were asked to describe what they had been thinking and why they had selected the rule knowledge category. If they had not indicated rule knowledge as a source, they were prompted to share any other ways in which they had made their choices, whether on the basis of intuition or on the basis of other sources. If, up to this point in the interview, participants had not mentioned anything related to animacy or had not reported indicating rule knowledge as a basis for their decisions, they were informed that there was a rule and were asked to speculate about what the rule might have been. If animacy was still not mentioned, the researcher explained the system and then asked participants if they had considered the possible relevance of animacy at any point during the exposure or assessment task.

The Coding of Concurrent and Retrospective Verbal Reports

The think-aloud protocols and the retrospective verbal reports were transcribed in spreadsheets that were organized to reflect the stages of the interview process and particular themes that emerged from the data (e.g., explicit mentions of animacy, reported memory of exemplars, metalinguistic hypothesis testing, or indications of positive or negative affect). A coding scheme was then developed that would allow us to categorize participants as having demonstrated different types of evidence of awareness. This was an iterative process that resulted in the simple rubric shown in Table 1, which, in addition to identifying whether animacy was mentioned at all, also includes dimensions of accuracy/completeness and confidence/willingness to speculate. Because an important goal of Williams (2005) and its subsequent replications has been to investigate implicit learning, and because of debates over whether learning has truly been implicit in cases in which participants may have had fleeting or partial awareness (Shanks & St. John, 1994), we considered it important to acknowledge any mention of animals as a category as potentially representing a low level of awareness of a

Table 1. The awareness coding rubric used to analyze concurrent and retrospective verbal reports; participants were classified as fully aware, partially aware, minimally aware, or unaware

Code	Description	Sample quotations from the interviews
Full-confident	Full and accurate characterization of the animacy regularity, with confidence in reporting it	P20: "So . . . like I think I figured out the difference . . . that *ul* and *gi* were for living things? And that *ro* and um *ne* were for like inanimate objects."
Full-hesitant	Full and accurate characterization of the animacy regularity but with hesitation in reporting it (e.g., mentioned only after being prompted to speculate)	R: "Seriously, any hypotheses, anything you were considering." P11: "I can, like, I remember for some of them I considered if it was like an animal or an object, or . . . like if it was like a stool or like a tiger or something I'd be like . . . I don't know, yeah . . . because like I considered those differences, but then I like didn't know how they would correspond with the different, um, words. . . . I still put that it was intuition just because I wasn't entirely sure and it wasn't really based on like some like steadfast rule, it was just like OK, I think like that this, like, sounds good to me, yeah."
Partial	Fragmentary or partial characterization of the animacy regularity	P54: "Um, I almost in some of them like combined certain animals into groups that like I remember *gi cow* from the initial one so I made it like *gi elephant* and a few other ones like that just—I don't know why I drew a similarity between the animals I just happened to. So I almost kinda grouped—since a lot of them seemed to be animals I grouped those into kinda categories like if I came across a new one I kinda put it with the same one as—I put the same uh new word or describer of location with animals that I thought were similar to it. . . . Same thing with tiger and lion. Um the birds and monkeys I think because a lot of their examples they were both in trees so I put them in a similar spot. Um I believe I put the cats with lions and tigers. . . . Eventually I kinda realized I was putting all that together and made it a rule."

Continued

Table 1. Continued

Code	Description	Sample quotations from the interviews
Minimal	Very brief or fleeting mention of a relevant category, such as animals or objects, and/or no mention of animacy until after having been told the rule by the researcher	R: “So any thought along these lines, like ‘oh, that’s alive’ or ‘oh, that’s an animal’ or ‘oh, that’s a dead piece of furniture’—did that consciously come up?” P64: “Definitely ‘that’s an animal’ but I think, I don’t know, I like animals, so the phrases just were more interesting. It’s a lot more interesting to hear about, you know, having a lion behind you than having a box near you, so I think those are just—I just noted them as more exciting.”
Unaware	No evidence at all of awareness of animacy	R: “So . . . did it ever occur to you, the living vs. nonliving thing or the animacy type distinction, did that pop into your mind at all?” P36: “Now it does! [laughs] I mean no, no not at all during the experiment, yeah, like, very, like absolutely nothing.”

feature relevant to the full rule. By being conservative in this way, we could be more (albeit not completely) confident that participants labeled as unaware had not oriented consciously toward animacy during the exposure or testing phase.

RESULTS

Performance on the 2AFC task served as the measure of learning. Subjective measures of awareness (confidence ratings and source attributions) and verbal reports (concurrent and retrospective) were used to determine to what extent participants were aware of having acquired knowledge and whether or not the acquired knowledge was conscious. The reliance on multiple measures of awareness allowed us to determine (a) what participants became aware of, (b) when participants became aware, and (c) differences between the measures in their assessments of awareness.

Performance on the 2AFC Task

The Overall Performance of the Experimental and Control Groups. The analysis of the 2AFC task data showed that the silent group performed best ($M = 73.21$ out of 100%, $SD = 28.69$), followed by the think-aloud exposure group ($M = 65.04$, $SD = 19.77$), the think-aloud throughout group ($M = 61.36$, $SD = 29.54$), and then the trained controls ($M = 49.25$, $SD = 9.59$). One-sample t tests on group mean accuracy revealed that only two groups performed above chance (50%)—namely, the silent group, $t(13) = 3.02$, $p = .01$, $d = 1.67$, and the think-aloud exposure group, $t(11) = 2.63$, $p = .02$, $d = 1.58$. Neither the think-aloud throughout group, $t(10) = 1.28$, $p = .23$, nor the trained control group, $t(14) = 0.30$, $p = .76$, differed significantly from chance.[10] Levene's test indicated that variance was not homogeneous, $F(2, 48) = 3.41$, $p = .02$, so Welch's F was used to determine whether there were significant differences in accuracy between groups. Welch's F revealed a significant effect of group, $F_W(3, 21.97) = 4.51$, $p = .01$. Games-Howell post hoc tests revealed significant differences in performance between the silent group and the trained controls, $p = .04$. However, no other group comparisons were significant. Thus, there were overall learning effects in the silent group and in the think-aloud exposure group but no clear evidence of learning in the think-aloud throughout group or the trained controls.

The Performance of the Experimental and Control Groups across Different Test-Item Types. Table 2 summarizes the performance of participants

Table 2. Mean accuracy (*SD*) for each group across the three test-item types

Group	Trained items	Partially trained items	New items
Silent	76.79* (29.09)	72.02* (29.71)	70.83* (30.27)
Think-aloud exposure	75.00** (18.80)	61.11 (24.48)	59.03 (26.70)
Think-aloud throughout	71.21* (28.96)	56.82 (33.30)	56.06 (31.64)
Trained controls	41.11* (12.38)	52.22 (13.53)	54.44 (18.05)

Note. Significance from chance: * $p < .05$; ** $p < .001$.

across the three types of test items (trained, partially trained, and new). Our analysis indicated that the silent group performed above chance on all test-item types, including new (generalization) items. In contrast, both think-aloud groups performed above chance only on trained items (i.e., on NPs that had already occurred in the exposure phase though in different sentence contexts). The trained controls performed significantly below chance on trained items.[11]

To investigate differences in each group's performance across the test-item types, we first conducted a 4 × 3 mixed ANOVA with group (four levels: silent, think-aloud exposure, think-aloud throughout, and trained control) as the between-subjects variable and test-item type (three levels: trained, partially trained, and new) as the within-subjects variable. The mixed ANOVA revealed a marginally nonsignificant effect of group, $F(3, 48) = 2.76$, $p = .05$, $\eta_p^2 = .15$; a significant effect of test-item type, $F(2, 96) = 3.27$, $p = .04$, $\eta_p^2 = .06$; and a significant Group × Test-Item Type interaction, $F(6, 96) = 4.41$, $p < .001$, $\eta_p^2 = .22$.

To explain the significant interaction effect, we further investigated the differences among groups on each test-item type. An ANOVA revealed significant between-groups differences only on trained items, $F(3, 48) = 7.64$, $p < .001$, $\eta_p^2 = .32$, but not on partially trained items (corrected for Levene's violation), $F_W(3, 48) = 1.81$, $p = .17$, or on new items, $F(3, 48) = 1.07$, $p = .37$. For the trained items, Tukey's HSD identified significant differences between the silent group and the trained controls, $p < .001$; the think-aloud exposure group and the trained controls, $p = .002$; and the think-aloud throughout group and the trained controls, $p = .01$. No other post hoc comparisons were significant.

To further investigate the significant effect of test-item type, repeated-measures ANOVAs were conducted on accuracy within each group. There was no significant effect of test-item type in the silent group, $F(2, 26) = 1.59$, $p = .22$, which suggests that participants performed similarly across the types of test items. The effect of test-item type was significant in the think-aloud exposure group, $F(2, 22) = 3.68$, $p = .04$, $\eta_p^2 = .25$; in the think-aloud throughout group, $F(2, 20) = 4.85$, $p = .02$, $\eta_p^2 = .32$; and in the trained control group, $F(2, 28) = 3.95$, $p = .03$, $\eta_p^2 = .22$.

In the think-aloud exposure group, Bonferroni pairwise comparisons revealed a significant difference in accuracy only between trained and new items, $p = .04$. In the trained control group, Bonferroni pairwise comparisons revealed a significant difference between trained and partially trained items, $p = .05$, and trained and new items, $p = .03$. In the think-aloud throughout group, Bonferroni pairwise comparisons revealed no significant differences for any comparison, all *p*s > .07. Taken together, the results indicate that the silent group was the only group to display significant (and somewhat consistent) learning effects across all item types, whereas the think-aloud exposure and think-aloud throughout groups showed significant learning only on trained items.

Subjective Measures of Awareness

The following analyses focus on the three experimental groups, given that we did not find an overall learning effect in the trained control group.

Confidence Ratings. All groups scored significantly above chance when indicating they were very confident and 100% confident in the accuracy of their 2AFC decisions. In the case of the think-aloud exposure group and the silent group, performance was also above chance when participants reported being somewhat confident. No group scored above chance when reporting to be truly guessing (i.e., the guessing criterion for implicit judgment knowledge was not met; Dienes, Altmann, Kwan, & Goode, 1995). Taken together with the fact that participants tended to be more accurate when indicating higher levels of confidence, this suggests that participants developed conscious judgment knowledge. Table 3 summarizes the findings.

Table 3. Accuracy and proportion of responses (%) across confidence ratings

Group		Guess	Somewhat confident	Very confident	100% confident
Silent	Accuracy	50.00	72.03*	80.12*	73.54*
	Proportion	5.65	23.79	32.46	38.10
Think-aloud exposure	Accuracy	41.82	64.52*	69.44*	76.54*
	Proportion	12.80	43.26	25.11	18.80
Think-aloud throughout	Accuracy	41.18	55.11	84.06*	88.89*
	Proportion	9.12	60.32	18.50	12.06

Note. Significance from chance: * $p < .001$.

Source Attributions. All experimental participants performed above chance when basing classification decisions on the explicit categories (memory and rule knowledge), which indicates that participants acquired conscious structural knowledge during the course of the experiment. With regard to the implicit categories (guess and intuition), performance on intuition-based judgments was above chance in both the think-aloud exposure group and the silent group. This suggests that, in these groups, at least some of the acquired structural knowledge was unconscious. Table 4 summarizes the experimental groups' mean accuracy and proportion of responses across the source-attribution categories.

Verbal Reports

With the wealth of information produced by the participants' verbalizations, it is necessary to be selective in presenting the data. In these sections, to facilitate an evaluation of the relative sensitivity of our awareness measures and to shed light on the potential reactivity of various aspects of the experimental design, we focus on (a) classifying participants as aware versus unaware and (b) obtaining information regarding what think-aloud participants became aware of, when, and how. As previously explained, classifications of awareness (fully aware, partially aware, minimally aware, or unaware) were made for each participant using the rubric in Table 1.[12]

Concurrent Verbal Reports (Think-Aloud Protocols). Our analysis of the concurrent think-aloud data indicated that only one participant (out of 12) in the think-aloud exposure group could be classified as (minimally) aware by that measure. Toward the end of the exposure phase, following a sentence about a near-animate pig, this participant (P14) misread the

Table 4. Accuracy and proportion of responses (%) across source attributions

Group		Guess	Intuition	Memory	Rule
Silent	Accuracy	58.18	72.73**	82.65**	73.11**
	Proportion	10.98	21.96	19.56	47.50
Think-aloud exposure	Accuracy	44.44	69.82**	63.11*	73.81**
	Proportion	16.82	39.49	24.07	19.62
Think-aloud throughout	Accuracy	44.74	52.58	71.08**	72.31**
	Proportion	19.69	25.13	21.50	33.68

Note. Significance from chance: * $p < .05$; ** $p < .001$.

next item (#106), inserting the word *animal* (phonologically similar to the next word, *examine*): *Looking closely the detective* animal *examined ro pictures for clues about the crime. Ro* is *near.* This was not, by any means, clear-cut evidence of P14's awareness but is acknowledged here to be as conservative as possible in relation to the question of whether learning was truly implicit.

In the think-aloud throughout group, the data revealed that six (out of 11) participants could be classified as being fully aware of the animacy distinction, as represented in the examples that follow. One of them showed awareness of animacy during the exposure phase (see Example [1]), and another began to make an incipient generalization involving animals during training (Example [2]), which was elaborated on during the test (Example [3]); however, for the rest of the participants in this group, there was evidence of awareness of animacy only during the testing phase (see Example [4]).

(1) P59, exposure item #49
"These all seem to involve animals for some reason."

(2) P53, exposure item #104
"I feel like *ul* is for monkeys and cats, basically, and flies."

(3) P53, test items #3–4
"*The boy played with . . . gi monkeys in the rainforest.* Monkeys, I felt, went with *ul*. . . . You know what, I think *gi* and *ul* are for animals. I really think so because . . . *gi*—there was *ul*? There were *ul flies, ul cats, ul monkeys,* and *ul birds*, and then there was, and then there was *gi* . . . um . . . there was *gi rat, gi cow, gi bear, gi lion.* OK, so I'm gonna go with, I'm definitely gonna go with *gi*, and I'm gonna say I'm very confident, and I'm gonna go with rule knowledge. OK. *I pushed aside*—OK, so then if *gi* and *ul* are for animals, then *ro* and *ne* must be for um objects, so *I pushed aside ro television.* Again, very confident based on rule memory."

(4) P48, test item #3:
"*The boy played with blank monkeys in the rainforest* . . . uh . . . *gi monkeys.* I feel like *gi* is for . . . like animals and stuff and *ro* is for . . . inanimate objects."

Regarding the types of information available in the concurrent think-aloud data, analyses of the exposure phase data revealed that the most common tendency in both think-aloud groups was for participants simply to read the training item aloud and then repeat the noun or NP, indicating also whether it was near or far (e.g., "*ro box*, *ro* is near" [P5]; "*ul* means far, *ul cat*, the far cat" [P17]; "the far vase . . . the far bird . . . the near cows" [P42]). Several participants used repetitive mnemonics, particularly toward the beginning of the training (e.g., "*gi-ro* is near, so near, *ro picture* . . . *ne-ul* is far, so far, *ne plate*" [P53]; "*ne-ul-gi-ro*, *ro* is near, *ro televisions* . . . *gi bear* is *ne-ul-gi, gi* near, *gi bears* . . . *ne-ul*, *ul* is far, *ul snakes* . . ." [P14]). Occasionally, a participant would mention having noticed a repetition (e.g., exposure items #55–56: "Wait, I swear I've seen this before. *I tried to play dead while gi bear sniffed me.* OK, well, that's . . . near, *gi bear*. . . . Yup, definitely I've seen this before. *Gi rat* is near, *gi rat*" [P53]) or would make metalinguistic comments suggesting an analytic approach or an orientation toward grammar (e.g., exposure item #48, in which capitalization is used to represent prosodic emphasis: "*In the night we heard UL cats fighting on the road—fighting IN—IN the road?* Um? OK [laughing], um, *fighting IN the road outside.* Um, *UL* is far, and *UL cats*"; exposure item #132 [a repetition of #48]: "*In the night we heard ul cats fighting in the road out—in the road outside*, um, this grammar is—[laughs] is killing me, um, *ul cats* is far. . . .Um, *ul cats*" [P20]). However, verbalizations of these types were fairly infrequent.

In relation to the question of how participants became aware of the regularity involving animals and objects, interestingly, the think-aloud data suggest that items involving animals may have been more engaging than those involving objects, and certain animals may have been particularly salient. Several participants had affective reactions to sentences about animals, expressing either (amused) disgust or (feigned) concern. For instance, P20 from the think-aloud exposure group laughingly vocalized sounds of disgust ("ew, that's disgusting!" and "euh!") on items about rats and flies, and P55 from the think-aloud throughout group engaged with the content of several sentences involving animals (e.g., saying, "that's not safe!" "that's not nice!" and "that's horrible!" on items about lions, a monkey, and a rat) but did so only once for a sentence involving an object ("*the girl stayed up late watching ro television*—good!"). These concurrent data illuminate and substantiate many of the comments participants made during the interviews regarding how they developed awareness of the animacy regularity, as we discuss this in more detail later.

In the think-aloud data from the testing phase, some participants' concurrent verbalizations pointed toward sources of reactivity in the experimental design. These sources of reactivity were not explicitly related to the process of thinking aloud but, rather, concerned aspects of the 2AFC task and the subjective measures, which they interpreted as clues to search for other rules besides the distance regularity on which they had knowingly been trained. Specifically, on realizing that both options for each test item had the same distance value, some inferred early in the test phase that they therefore had to choose a determiner based on something other than distance (e.g., test item #1: "Oh, what? *Gi* and *ro* both mean near! [whimper] *The babysitter poured . . . juice into* . . . Oh no! I feel like I was supposed to notice a pattern . . . between *gi* and *ro*. [whimper] Oops" [P53]). This, combined with the fact that each test item was followed by a request for a source attribution, one of whose options was rule knowledge, made some participants wonder what the rule might be (e.g., "Why would it be rule knowledge?" [P55]) and in some cases may have prompted them to search for a rule, as suggested by the sequence of verbalizations from P59 presented in Example (5):

(5) Test item #2: “I don’t think I picked up the rule quite yet unfortunately, sorry.”
Test item #3: “Clearly I wasn’t paying close enough attention to the distinctions between *gi* and *ro.*”
Test item #23: “Again, intuition, not very confident, I wish I had paid more attention now. I had an inkling that there were rules but I didn’t even think to figure them out.”
Test item #28: “So now I’m starting to suspect that there might be a distinction between animate and inanimate objects.”

Here also, the think-aloud protocols provided real-time data corroborating comments that participants made retrospectively during the interviews.

Retrospective Verbal Reports. The analysis of the retrospective interview transcripts indicated that 10 (out of 14) participants in the silent group, eight (out of 12) participants in the think-aloud exposure group, and 10 (out of 11) participants in the think-aloud throughout group displayed at least some awareness of animacy, if only minimally, following the training and testing phases. In each group, only the aware participants (as a subgroup) performed above chance in the 2AFC classification task, in contrast to Williams (2005), who found evidence for above-chance performance in unaware participants.[13] In what follows, we discuss only those participants who completed both concurrent and retrospective verbal reports; that is, we exclude participants from the silent group. This is because our objective is to directly compare the use of the two types of verbal reports as measures of awareness. For a discussion of the silent group, please refer to Rebuschat et al. (2013).[14]

In the think-aloud exposure group, five of the aware participants were classified as fully aware (three of them mentioning the animacy regularity with confidence), two as partially aware, and one as minimally aware. In the think-aloud throughout group, eight participants were classified as fully aware (four of them mentioning the animacy regularity with confidence). Two participants were judged to have experienced low levels of awareness of a feature relevant to the animate-inanimate distinction, with one participant being classified as partially aware and another as minimally aware on the basis of their retrospective verbal reports.

Complementing the information obtained through the think-aloud protocols, the postexposure interviews provided many additional details regarding what the participants became aware of, when, and how. For example, they revealed considerable variation in participants’ awareness of animacy. Whereas a few stated the full rule at the outset of the interview (see P20 in Table 1), and others mentioned having used a distinction between animals and objects to guide at least some of their responses, some appeared to realize only in retrospect that many of the stimuli had involved animals or used the word *animals* to describe their responses to test items (e.g., “I would mostly go by if it sounded right or if, um, it was something that was seen in a previous sentence, like a certain animal or something like that” [P8]) but, after the rule was explained, denied having considered the potential relevance of animacy previously during the training or test (“I didn’t realize that at all” [P8]). Many of the participants who showed awareness of the animacy distinction during the interview hesitated when reporting this to the researchers, suggesting a lack of confidence (e.g., “I don’t know, for some reason I felt like—like *gi* and *ul* were for, like, animals” [P48]).

The interviews also uncovered other hypotheses about the stimuli that participants had generated during the experiment, suggesting a tendency toward metalinguistic analysis that might be expected of students of languages and linguistics.[15] Participants also acknowledged a preference for sentences about animals (see P64 in Table 1) and reported having recognized the relevance of animacy by first noticing that animals appeared frequently in the sentences and then realizing that the determiners *gi* and *ul* appeared only with animals. In an interesting implicational relationship, participants who showed awareness that *ro* and *ne* were used with inanimate objects always (and often first) also mentioned that *gi* and *ul* were used with animals but not vice versa. In addition to being more salient and memorable, animals may have seemed a more natural or obvious category.

From a methodological perspective, and similar to the think-aloud protocols, some of the interviews suggested possible reactivity of the test phase, which was mentioned by four participants, and of the source attributions, which was mentioned by two participants, as the following excerpt in Example (6) illustrates.

(6) P14

> “Well, I thought that it was gonna ask the difference between near and far and it just gave two random ones, like whichever one, and then I was thinking, like, it’s not—it doesn’t matter if it’s near or far because they’re giving me two of the same ones, like, two of the near words and two of the far ones. So then I was like well there’s something about it, but I don’t know, and then it took like maybe six or seven and I saw that there could be a rule, and then I thought maybe it is a rule.”

Not all participants remained unsuspecting until the test phase, however. At least one participant (P20) reported having wondered during the exposure, “Why would there be two words for the same thing?” Moreover, for some participants, there seemed to be a strong tendency during the testing phase to expect and seek out a rule and even to use a hypothesized rule to override their intuitions. For instance, regarding a hypothesized singular-plural rule, one participant stated the following, as reported in Example (7):

(7) P14

> "Um, there was a couple when the sentences didn't sound right with the rule that I wanted to change it, but then I just thought that like it doesn't matter if it sounds right . . . when I thought of the rule, I tried to stick to the rule more, and use like if something was more ambiguous just like go with the rule?"

The interviews were also useful in helping to identify and exclude participants who had not responded in good faith, so to speak, as with the participant who is cited in Example (8).

(8) P60

> "Well, both had the exact same meaning for all of them, so I felt like it didn't matter which one I chose. I just went with whatever was on this pinky. I just went with this side always. . . . I was kinda confused, so I was thinking they were gonna throw a trick in there, but I think all of them were the same, so I just kept on clicking this one. Sometimes I would switch it up, but . . . I mean it was the same meaning, so I just chose whichever."

In short, the retrospective interviews were an enlightening source of information regarding participants' intentional response strategies.

DISCUSSION

This experiment sought to contribute to the current debate surrounding learning without awareness by triangulating measures of awareness. Specifically, our objective was to investigate more valid ways of examining whether adult learners are able to establish novel form-meaning connections under incidental learning conditions and without awareness of the product of learning (implicit knowledge). In contrast to previous research (Faretta-Stutenberg & Morgan-Short, 2011; Hama & Leow, 2010; Williams, 2005), the present study employed three types of test items (trained, partially trained, and new NPs) and utilized three independent measures of awareness (concurrent verbal reports, retrospective verbal reports, and subjective measures of awareness). In what follows, we address how evidence for both implicit and explicit knowledge was contingent on evidence from these multiple measures, as the various measures used in the current research sometimes provided conflicting assessments of awareness. In addition, we discuss specific methodological issues in administering and interpreting awareness measures, including unexpected differences in sensitivity and frequently expressed concerns that concurrent measures may be reactive (Bowles, 2010).

Learning Effects

Overall, both the silent group and think-aloud exposure group performed significantly above chance on our measure of learning, whereas the think-aloud throughout group and the trained controls did not. However, when compared against the trained control group as a baseline (as opposed to chance), only the silent group demonstrated a significant learning effect. Our evidence for learning replicates previous findings in this paradigm and confirms that adult learners can learn novel form-meaning connections under incidental learning conditions, without feedback, and after a relatively short exposure period. At the same time, these findings also point to important methodological considerations. One is the possibility of reactivity influencing differences between the silent and the think-aloud groups, which is discussed later in this section. The other issue concerns whether evidence for learning should be judged relative to chance (e.g., Williams, 2005), relative to untrained controls (e.g., Hamrick & Rebuschat, 2012, 2014; Rebuschat & Williams, 2012), or relative to trained control groups (e.g., Hamrick, 2014a, 2014b; Reber & Perruchet, 2003). Each approach has its strengths and weaknesses. For example, because participants in a group naturally perform differently from one another, adding a trained control group introduces variance around a baseline mean, which, when combined with the variance in the experimental groups, may obscure learning effects in smaller sample sizes. Using chance as a baseline minimizes such variance, arguably making learning effects more likely to be found in smaller sample sizes. On the other hand, using a trained control group can help identify experimental confounds, such as unintended regularities in the stimuli or unexpected response tendencies across participants. On the basis of the current findings and the methodological considerations they raise, we recommend employing both chance and trained controls to triangulate the data and provide more reliable results.

Learning across Test-Item Types

As we reported previously, only the silent group performed above chance on all test-item types (trained, partially trained, and new), indicating that they were the only group to show true generalization. Meanwhile, the think-aloud groups performed above chance only on trained items (i.e., on NPs that had already occurred in the exposure phase), which indicates that they were unable to generalize their knowledge. It also suggests that the think-aloud participants relied mostly on stored exemplars and/or inaccurate rules while performing on the 2AFC task. We can only speculate as to why this occurred, but a few explanations seem plausible. One explanation would be that participants developed at least some (partially) inaccurate rules, as the think-aloud protocols and the retrospective verbal reports indicate, and that thinking aloud may have reinforced these inaccurate rules. This may have led participants to apply their rules in the 2AFC task when they did not have memory for exemplars. Another alternative would be that thinking aloud may have affected generalization by inhibiting (or otherwise influencing) the development of abstract knowledge during the exposure phase, for example, by consuming attentional resources that silent participants were able to allocate toward deeper processing of the stimuli.

Finally, it is worth noting that the trained control group performed significantly below chance on trained items in the test phase. This suggests that they also stored exemplars in the exposure phase. However, many of these exemplars were not consistent with the animacy regularity that was tested and, as such, led to low accuracy on the test. Taken together, the results suggest that all groups recollected exemplars from the exposure phase, but only the silent group was able to acquire generalizable knowledge. In addition, it seems clear that the generalizable knowledge acquired by the silent group was probably a result of the exposure phase. If this knowledge had been developed only during the test phase, then one would expect generalizable knowledge in other experimental groups as well (or at least in the think-aloud exposure group, given that thinking aloud may have interfered with test performance in the case of the think-aloud throughout group).

Learning and Awareness during the Test Phase

The evidence from the think-aloud protocols and retrospective verbal reports indicated that, during the test phase, several participants began to actively search for rules and to report awareness. These findings led us to question whether accuracy and/or awareness changed over the course of the test phase. To further investigate this issue, which is rarely reported, we first examined accuracy by (arbitrarily) dividing the test phase into three blocks of 12 sentences each. The results are illustrated in Table 5. A mixed ANOVA on accuracy with

Table 5. Overall mean accuracy (*SD*) across the three blocks of the test phase

Condition	Group	Block 1	Block 2	Block 3
Experimental	Silent	72.58 (10.71)	71.42 (12.00)	75.58 (8.05)
	Think-aloud exposure	59.67 (17.36)	62.92 (13.57)	71.83 (11.46)
	Think-aloud throughout	61.58 (14.17)	59.25 (18.14)	63.92 (10.32)
Control	Trained controls	43.89 (12.53)	50.56 (15.68)	53.33 (12.71)

group (four levels: silent, think-aloud exposure, think-aloud throughout, and trained controls) as the between-subjects variable and block (three levels: one, two, and three) as the within-subjects variable revealed a marginally nonsignificant effect of block, $F(2, 88) = 2.99$, $p = .06$, and no Group × Block interaction, $F(6, 88) = 0.47$, $p = .83$, but a significant main effect of group, $F(3, 44) = 26.19$, $p < .001$, $\eta_p^2 = .64$, reflecting the overall group differences in general accuracy. Although the effect of block was not significant, it was marginal, and the descriptive statistics in Table 5 suggest that there may have been improvements in accuracy in the think-aloud exposure group and in the trained controls as the test phase progressed. These results do not provide definitive evidence of learning during the test phase, but it is possible that some learning may have occurred in at least some groups across test blocks. Learning during the test phase is not uncommon (see, e.g., Rohrmeier, Rebuschat, & Cross, 2011), but it is more common when the artificial system is relatively simple, as in this experiment. In contrast, Grey et al. (2014), Hamrick (2013), Rebuschat (2008), Rebuschat and Williams (2006, 2012), and Tagarelli et al. (2011, 2015) did not find evidence of learning during the test phase in experiments that employed more complex artificial languages.

To investigate whether subjective awareness changed during the test phase, we also examined confidence ratings and source attributions across the three blocks of the test phase. However, because we had relatively few responses per category, we were unable to run inferential tests, so what follows is speculative. In the case of the confidence ratings, the analysis indicated that the proportion of low-confidence responses decreased as the test phase progressed, suggesting, in combination with the possible increase in response accuracy, that conscious judgment knowledge may have developed during the test phase. In other words, participants' awareness of having acquired knowledge could well be a result of performing the 2AFC task. In the case of the source attributions, the analysis indicated that there were consistently fewer guess responses in Block 3 than in Blocks 1 and 2. The data also show a steady decrease in memory attributions and a steady increase in rule knowledge attributions as the test phase progressed. This suggests that, as experimental participants were completing the 2AFC task, they felt themselves to be relying progressively more on explicit structural knowledge when making their decisions. The increased reports of using rule knowledge stand, again, in contrast to previous research that used subjective measures with more complex artificial languages (e.g., Rebuschat & Williams, 2012, in which the proportion of responses for each category did not increase over time). For a more comprehensive discussion of the issue and related analyses, see Rebuschat (2008).

Evidence for Implicit Knowledge

Evidence from our measures of awareness demonstrated that many participants acquired explicit (conscious) knowledge. For example, examining the confidence ratings, participants' accuracy was above chance only when they reported some degree of confidence, and higher confidence decisions tended to be more accurate, particularly in the think-aloud groups. These findings are consistent with previous research (Hama & Leow, 2010) and suggest not only that participants were very likely to develop explicit knowledge but also, generally speaking, that this explicit knowledge was associated with above-chance performance (see also Rebuschat & Williams, 2012). However, examining the source attributions, we found that participants in the silent and think-aloud exposure groups also performed significantly above chance when basing decisions on intuition. This suggests that, in addition to developing explicit knowledge, participants had also developed at least some implicit structural knowledge, a finding consistent with Williams (2005), albeit via a different measure.

Methodologically speaking, this observation illustrates one of the advantages of subjective measures of awareness. Because subjective measures are taken on a trial-by-trial basis, they can reveal implicit knowledge even in the presence of explicit knowledge (and vice versa), without a need to exclude one in the presence of the other. This was not possible in previous studies that used retrospective verbal reports (Hama & Leow, 2010; Williams, 2005) and concurrent verbal reports (Hama & Leow, 2010), because verbal report protocols employ an all-or-nothing strategy of classification. In verbal reports, participants who verbalize something relevant to the system in question are classified as aware, whereas those who do not are classified as unaware.[16] This leads to well-documented problems in sensitivity, because verbal reports may not exhaust all of participants' relevant explicit knowledge and because implicit knowledge may still be used even when participants can verbalize the rule (Reingold & Merikle, 1990; Shanks & St. John, 1994).

The subjective measures used in the present study indicated that participants were aware of having acquired some knowledge (explicit judgment knowledge) and of the content of some of that knowledge (explicit structural knowledge), yet they were at least partially unaware of the content of some of the knowledge that they had acquired (implicit structural knowledge). Thus, our findings that incidental exposure can result in both implicit and explicit knowledge are consistent with previous research on the acquisition of L2 vocabulary (e.g., Hamrick & Rebuschat, 2012, 2014), L2 morphology (e.g., Grey et al., 2014), and L2 syntax (e.g., Rebuschat, 2008; Rebuschat & Williams, 2012) and may shed light on the seemingly divergent findings of Williams (2005) and subsequent extensions (Faretta-Stutenberg & Morgan-Short, 2011; Hama & Leow, 2010).

The Sensitivity of the Awareness Measures

Overall, the results seem to indicate that interviews (retrospective verbal reports) were a more exhaustive measure of awareness in the context of this experiment.[17] The think-aloud protocols collected during training and testing identified many (six out of 10)—but not all—of the participants in the think-aloud throughout group who were also classified as aware via their interviews. The think-aloud protocols that were collected only during training identified just one participant as being (minimally) aware, and this low level of awareness was not captured again in the interview. Several participants (i.e., eight) from the think-aloud exposure group who were shown to be aware of animacy through the postexposure interview were classified as unaware on the basis of their concurrent verbalizations.[18] The interviews were also an excellent source of information about participants' response strategies and ways of figuring out the rule, clarifying the significance of sometimes more abbreviated comments made during the think-aloud protocols.

On the other hand, the think-aloud protocols were able to reveal online thought processes not captured in the interviews. For example, several participants had affective reactions to some of the less appealing animals, such as rats and flies, which may have been one factor contributing to the salience of animals in this experiment. The think-alouds collected during training also occasionally suggested that at least some awareness of a regularity had emerged earlier than participants reported remembering in the interviews. Compare, for instance, quotations from P53's think-aloud (Example [9]) and interview (Example [10]):

(9) Think-aloud
Training item #104:

"I feel like *ul* is for monkeys and cats, basically, and flies."
Test item #1:
"Oh, what? *Gi* and *ro* both mean near! [whimper] *The babysitter poured* . . . *juice into* . . . Oh no! I feel like I was supposed to notice a pattern . . . between *gi* and *ro.* [whimper] Oops."

(10) Interview
P53: "I didn't realize there was a pattern honestly until the second part . . . because the options in the first were either near or far, and so I was just, you know, going based on that, and it never occurred to me that there was a pattern until I had to do the fill in the blank. Literally, when I was looking at the sentences. . . . I was like near-far-near-far-near-far. . . . I didn't like register it because I would just see, like, *ne*, so I'd be like 'far' and like, *ul*, far, and I would just like see that and just press it. . . . I just did it from memory, but I never . . . kind of put together that there was a pattern then."
R: "OK, so it sounds like the first time that it sort of consciously came into your mind that being an animal or not could be relevant was in the second [i.e., the testing phase]."
P53 [laughing]: "Yeah, you can probably hear how . . . I was like, 'Oh no, there was supposed to be a pattern and I missed it!' . . . and then I was like, 'OK, I remember there was like *gi bear*, *gi tig—gi lion*, *gi rat* . . . and like *ul cat, ul* whatever, *ul c*—um *bird* and *ul snake*, I think, and *fly.*'"

P53 accurately reports her realization from the testing phase but appears not to recall having made any partial generalization associating *ul* with monkeys, cats, and flies during the training. The think-aloud data from the testing phase substantiate that this was when most participants apprehended the relevance of animacy. Even though some incipient awareness of animacy may have been present during the exposure phase in some participants who were later able to describe the hidden regularity during debriefing, this was rarely captured via their think-aloud data. During the testing phase, when confronted with a clear problem, participants began to explore potential solutions aloud, revealing their awareness of various aspects of the task and suggesting that certain features of the experimental design may have been reactive.

Reactivity

The results of this study have interesting implications regarding various potential sources of methodological reactivity in research on awareness and implicit learning. Whereas the silent group performed above chance on all test-item types, the think-aloud groups were above chance only on trained items. This suggests that thinking aloud may have interfered with participants' ability to generalize to new items and may help to explain some of the conflicting results in previous studies. Hama and Leow (2010) used think-aloud protocols and retrospective verbal reports to measure awareness throughout their exposure and testing phases and reported no evidence of implicit learning, whereas Williams (2005) used only retrospective verbal reports (with a less conservative coding scheme than the one employed in this study) and reported evidence of implicit learning.[19]

In our study, the subjective measures were useful in allowing us to identify implicit knowledge even in the presence of explicit knowledge. Notwithstanding this methodological advantage, the think-aloud and interview data revealed that the source attributions and even the 2AFC task itself may also have been reactive. The fact that participants had the option to choose rule knowledge as the source of their answers clued many of them into the fact that there was a rule, and several participants in the think-aloud throughout group who had shown no evidence of awareness of animacy during the exposure phase began to search for a rule once they realized that they did not know how to answer the new type of item presented in the test phase. However, there are ways of addressing the reactivity issue. Although perhaps not completely solving the problem, three basic steps may help reduce potential reactivity effects. First, as mentioned previously, using a more complex linguistic system could be helpful because it may discourage participants from looking for rules or patterns during the test phase. As Reber (1993) pointed out, if participants feel they can "crack the code" (p. 26), they will attempt to do so. Second, the rule knowledge category can be avoided. In studies such as Hamrick and Rebuschat (2012, 2014), which looked at implicit and explicit knowledge of L2 vocabulary (see Dabrowska, 2014, for an extension to first language vocabulary), there is of course no need to include a rule knowledge category. Finally, we recommend carefully tracking the performance of the experimental and control groups during the test phase to check for changes in response patterns (see, e.g., the analyses in Rebuschat, 2008). Do the mean accuracy and the proportions across the source categories change as participants complete the test (e.g., from more responses based on guessing and intuition to more responses based on rule knowledge)? If they do not, then the inclusion of a rule category might not have encouraged them to look for rules. If they do, it can be reported. We suggest that future studies involving the three measures employed in this study analyze (and report) in more detail the performance during the test phase to ensure that learning occurs during the exposure phase and not as a result of the testing phase.

CONCLUSION

The main objective of this study was to contribute to the current debate on implicit learning by triangulating three measures of awareness—concurrent verbal reports (think-aloud protocols), retrospective verbal reports, and subjective measures—to determine their validity and usefulness for the investigation of implicit and explicit learning. Our study confirmed that learners are able to rapidly acquire novel form-meaning connections under incidental learning conditions and without the benefit of feedback (see also Faretta-Stutenberg & Morgan-Short, 2011; Hama & Leow, 2010; Williams, 2005). Given our inclusion of true generalization items in the test phase, our study further showed that at least some participants were able to generalize the acquired knowledge to novel instances but that this ability was restricted to those who did not think aloud during exposure. This points to a potential drawback in the use of think-alouds in the present paradigm, and future studies need to consider more carefully whether a given task is well suited to the think-aloud procedure (see Leow et al., 2014, for a comprehensive review).

The study also indicated that incidental exposure can result in both implicit and explicit knowledge of language (see also Grey et al., 2014; Hamrick & Rebuschat, 2012, 2014; Rebuschat & Williams, 2012), which may partially explain the conflicting results obtained by previous studies. Williams (2005) relied on a comparatively insensitive measure of awareness (retrospective verbal reports with a higher threshold for counting a participant as aware), which could have led him to overestimate the number of unaware participants. Hama and Leow (2010), on the other hand, may have overestimated the role of explicit knowledge because they were unable to assess whether implicit knowledge was also present in those participants classified as aware on the basis of their think-aloud data.

In terms of measuring awareness, our comparison of the three measures suggests the following. The inclusion of the think-aloud procedure revealed, occasionally, that awareness had emerged earlier than participants reported in the interviews, confirming that retrospective recall can be unreliable. The think-aloud data collected during the test phase also revealed that the test phase was when most participants became aware of the hidden regularity, and these results shed light on the potential reactivity of other parts of our design. The retrospective reports were useful in that they revealed partial rules or microrules that participants may have formed and that could explain their performance (for discussion, see Hamrick, 2013), though of course it is unclear when participants developed conscious knowledge, and it is important to consider that this could be a direct result of prompting participants to verbally describe rules or patterns at the end of the experiment. Finally, the subjective measures of awareness have the advantage of allowing the detection of both implicit and explicit knowledge. Without the use of subjective measures in the current study, we would have failed to detect that participants in the silent group and in the think-aloud exposure group had developed implicit knowledge, as evidenced by their above-chance performance on responses based on intuition.

The debate on learning without awareness that has followed the publication of Williams (2005) has led to important insights, and it clearly demonstrates the importance of replication and extension studies in SLA research. Although it is still subject to debate whether learning without awareness is possible, it is clear that methodological refinements are necessary to move the topic forward. As Hama and Leow (2010) conclude, future studies in this area would benefit from collecting as much data as possible from "multiple sources and stages, including both online and offline measures and different tasks" (p. 487). What is also required are studies like the present one, whose primary purpose is to systematically compare different measures to develop more valid ways of assessing the role of awareness.

Received 11 December 2014

NOTES

1. As one of our reviewers pointed out, this does not mean, of course, that all participants will follow the instructions, nor does it mean that they will acquire rules or patterns. It is also worth noting that it is possible to develop explicit knowledge unintentionally and to develop implicit knowledge under explicit, intentional learning conditions, as will be discussed later.

2. In Williams (2005), half the participants were told that *gi* and *ro* were used for near objects and *ro* and *ul* for distant ones, whereas the other half were told the opposite.

3. As pointed out in Leow et al. (2014), this limitation may be dependent on the type of task.

4. To clarify, as recommended by one of the reviewers, the evidence for implicit knowledge in the case of Williams (2005) rests on the fact that participants who were unable to verbalize the appropriate rule system still performed above chance on the classification task. In our case, the evidence is based on above-chance performance on those test items for which subjects were guessing or relying on their intuition in the same classification task.

5. The issue of reactivity can be described as follows: Does the addition of a secondary task impact (positively or negatively) on the cognitive processes involved in the primary task? If so, then the secondary task is reactive. For example, the think-aloud procedure can be considered reactive if, "by thinking aloud, participants' internal processes . . . differ from what they would have been had they not performed the verbalization" (Leow & Morgan-Short, 2004, p. 38).

6. As described in Rebuschat et al. (2013):

> trained control groups receive training conditions that are identical to experimental groups but with the relevant independent variables randomized and balanced, rather than removed altogether. The logic behind this procedure stems from the notion that all participants have unforeseen response biases in test phases based on their prior knowledge (Reber & Perruchet, 2003). These biases are "noise" that influences test performance beyond what is learned during training. The use of trained controls ensures that such noise can be identified and accounted for, allowing the effects of the independent variable(s) to be isolated. (pp. 254–255)

7. As pointed out by one reviewer, the exposure task used here (as well as in Williams, 2005, and Hama & Leow, 2010) may not be particularly conducive to gathering think-aloud data (see Leow et al., 2014).

8. Rebuschat et al. (2013) and the present study used the written modality for both training and testing. In contrast, Williams (2005) used the auditory modality for training and the written for testing, and Hama and Leow (2010) used the auditory modality for both training and testing.

9. Note that we had only one test phase, in contrast to Williams (2005), which featured two.

10. One participant in the think-aloud throughout group had zero accuracy. This, combined with the low sample size for the think-aloud throughout group, may be the cause of the nonsignificant result.

11. It is noteworthy that the trained controls performed significantly below chance on trained items, presumably due to having been trained on stimuli with unreliable animacy information, which, in some cases, would be counted as incorrect on the test. In effect, for certain NPs, if the trained control participants remembered exemplars from the training, they were penalized essentially for learning what they had been exposed to.

12. Only one participant was classified as unaware by both types of verbal reports. Because inferential statistics are not possible, we will not report this further.

13. The values for the different groups are as follows: silent group: unaware participants, *M* = 53.46%, *SD* = 10.72, and aware participants, *M* = 90.83%, *SD* = 10.06; think-aloud exposure group: unaware participants, *M* = 49.31%, *SD* = 11.19, and aware participants, *M* = 75.31%, *SD* = 18.86; think-aloud throughout group: unware participant, 0%, and aware participants, *M* = 67.49%, *SD* = 22.56.

14. In the silent group, nine participants were fully aware, and one participant was minimally aware.

15. For instance, at least eight participants mentioned having considered a singular-plural distinction, and nine mentioned phonological issues such as euphony or ease of articulation. Others speculated about subject-object marking, possessive pronouns, different types of demonstrative determiners, verbal aspect, and even classifiers, which one participant was familiar with due to her study of Chinese.

16. This applies to both concurrent and retrospective reports. Subjects can be coded for different levels of awareness, but they are still assigned to one of the categories. For example, a subject who only provides evidence of low levels of awareness is placed in the aware category, disregarding the existence of unconscious knowledge. The advantage of subjective measures is that they acknowledge that a subject can have developed both conscious and unconscious knowledge during the experiment.

17. Of course, it could be argued here that participants perhaps developed conscious knowledge only after the experiment (i.e., while being prompted to verbalize any rules or patterns they might have noticed).

18. This could be because the subjective measures led participants to become aware of the regularity.

19. The coding scheme in Williams (2005) was less conservative in the sense that participants were classified as aware if they stated the rule and not just mentioned *animal* or *animacy*. Our minimally aware participants, and possibly some of our partially aware participants, would be considered unaware had we followed the coding scheme employed by Williams (2005).

REFERENCES

Bowles, M. A. (2010). Concurrent verbal reports in second language research. *Annual Review of Applied Linguistics*, *30*, 111–127.

Cedrus SuperLab Pro (Version 4.0.7b) [Computer software]. San Pedro, CA: Cedrus Corporation.

Cleeremans, A. (2008). Consciousness: The radical plasticity thesis. *Progress in Brain Research*, *168*, 19–33.

Dabrowska, E. (2014). Implicit lexical knowledge. *Linguistics*, *52*, 205–223.

Dienes, Z., Altmann, G., Kwan, L., & Goode, A. (1995). Unconscious knowledge of artificial grammars is applied strategically. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1322–1338.

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

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition*, *27*, 141–172.

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* (pp. 18–28). Somerville, MA: Cascadilla Proceedings Project.

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye tracking. *Studies in Second Language Acquisition*, *35*, 483–517.

Godfroid, A., & Schmidtke, J. (2013). What do eye movements tell us about awareness? A triangulation of eye-movement data, verbal reports and vocabulary learning scores. In J. M. Bergsleithner, S. N. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 183–205). Honolulu, HI: University of Hawai'i, National Foreign Language Resource Center.

Grey, S. E., Williams, J. N., & Rebuschat, P. (2014). Incidental exposure and L3 learning of morphosyntax. *Studies in Second Language Acquisition*, *36*, 1–34.

Hama, M., & Leow, R. P. (2010). Learning without awareness revisited: Extending Williams (2005). *Studies in Second Language Acquisition*, *32*, 465–491.

Hamrick, P. (2013). *Development of conscious knowledge during early incidental learning of L2 syntax* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

Hamrick, P. (2014a). Recognition memory for novel syntactic structures. *Canadian Journal of Experimental Psychology*, *68*, 2–7.

Hamrick, P. (2014b). A role for chunk formation in statistical learning of second language syntax. *Language Learning*, *64*, 247–278.

Hamrick, P., & Rebuschat, P. (2012). How implicit is statistical learning? In P. Rebuschat & J. N. Williams (Eds.), *Statistical learning and language acquisition* (pp. 365–382). Berlin, Germany: Mouton de Gruyter.

Hamrick, P., & Rebuschat, P. (2014). Frequency effects, learning conditions, and the development of implicit and explicit lexical knowledge. In J. Connor-Linton & L. Amoroso (Eds.), *Measured language: Quantitative approaches to acquisition, assessment, processing and variation* (pp. 125–139). Washington, DC: Georgetown University Press.

Leow, R. P. (1997). Attention, awareness, and foreign language behavior. *Language Learning*, *47*, 467–505.

Leow, R. P. (1998). Toward operationalizing the process of attention in SLA: Evidence for Tomlin and Villa's (1994) fine-grained analysis of attention. *Applied Psycholinguistics*, *19*, 133–159.

Leow, R. P. (2000). A study of the role of awareness in foreign language behavior: Aware versus unaware learners. *Studies in Second Language Acquisition*, *22*, 557–584.

Leow, R. P., & Bowles, M. (2005). Attention and awareness in SLA. In C. Sanz (Ed.), *Mind and context in adult second language acquisition* (pp. 179–203). Washington, DC: Georgetown University Press.

Leow, R. P., & Hama, M. (2013). Implicit learning in SLA and the issue of internal validity: A response to Leung and Williams (2011). *Studies in Second Language Acquisition*, *35*, 545–557.

Leow, R. P., Grey, S., Marijuan, S., & Moorman, C. (2014). Concurrent data elicitation procedures, processes, and the early stages of L2 learning: A critical overview. *Second Language Research*, *30*, 111–127.

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition*, *26*, 35–57.

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*, 33–55.

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.

Perruchet, P. (2008). Implicit learning. In J. Byrne (Ed.), *Learning and memory: A comprehensive reference: Vol. 2. Cognitive psychology of memory* (pp. 597–621). Oxford, UK: Elsevier.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *6*, 317–327.

Reber, A. S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. Oxford, UK: Oxford University Press.

Reber, R., & Perruchet, P. (2003). The use of control groups in artificial grammar learning. *The Quarterly Journal of Experimental Psychology*, *56A*, 97–115.

Rebuschat, P. (2008). *Implicit learning of natural language syntax* (Unpublished doctoral dissertation). University of Cambridge, UK.

Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*, 595–626.

Rebuschat, P., Hamrick, P., Sachs, R., Riestenberg, K., & Ziegler, N. (2013). Implicit and explicit knowledge of form-meaning connections: Evidence from subjective measures of awareness. In J. Bergsleithner, S. Frota, & J. K. Yoshioka (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 249–269). Honolulu, HI: University of Hawai‘i Press, National Foreign Language Resource Center.

Rebuschat, P., & Williams, J. (2006). Dissociating implicit and explicit learning of syntactic rules. In R. Sun (Ed.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (p. 2594). Mahwah, NJ: Erlbaum.

Rebuschat, P., & Williams, J. N. (2012). Implicit and explicit knowledge in second language acquisition. *Applied Psycholinguistics*, *33*, 829–856.

Reingold, E. M., & Merikle, P. M. (1990). On the inter-relatedness of theory and measurement in the study of unconscious processes. *Mind & Language*, *5*, 10–28.

Rohrmeier, M., Rebuschat, P., & Cross, I. (2011). Incidental and online learning of melodic structure. *Consciousness & Cognition*, *24*, 214–222.

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

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (pp. 1–63). Honolulu, HI: University of Hawai‘i, National Foreign Language Resource Center.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge, UK: Cambridge University Press.

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

Tagarelli, K., Borges Mota, M., & Rebuschat, P. (2011). The role of working memory in implicit and explicit language learning. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2061–2066). Austin, TX: Cognitive Science Society.

Tagarelli, K. M., Borges Mota, M., & Rebuschat, P. (2015). Working memory, learning context, and the acquisition of L2 syntax. In W. Zhisheng, M. Borges Mota, & A. McNeill (Eds.), *Working memory in second language acquisition and processing: Theory, research and commentary* (pp. 224–247). Bristol, UK: Multilingual Matters.

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

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). Bingley, UK: Emerald Press.