Indirect tests of implicit linguistic knowledge

John N. Williams and Albertyna Paciorek
University of Cambridge, UK.

A slightly revised version of this manuscript appeared in A. Mackey & E. Marsden (Eds.), *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 25-42): Taylor & Francis. 2016.

Implicit knowledge is knowledge that we do not know that we know (Dienes, 2008). Much of our first language linguistic knowledge is represented in implicit form, at least in areas of morphology and syntax, and possibly aspects of word meaning as well. Implicit knowledge supports automatic behaviour, in the sense that it acts quickly and without conscious control, and it is these characteristics that make the acquisition of implicit knowledge desirable from the second language learner's perspective. And for the researcher it is clearly important to be able to assess whether a learner is employing implicit knowledge to perform a task since this is regarded as the product of true 'acquisition' as opposed to mere conscious learning (Krashen, 1981). Without reliable measures of implicit knowledge, therefore, it is not possible to develop a theory of second language acquisition (R. Ellis, 2005).

Here we discuss 'indirect' tests as measures of implicit knowledge. What they all have in common is that in the moment of use people have no awareness of at least some critical components of the knowledge that evidently underlies their performance. Obviously whilst producing and comprehending language people have conscious awareness of forms, meanings, and communicative intentions, but by and large they do not have awareness of the particular grammatical regularities, or statistical biases, that they are using in order to speak and understand. It is this kind of knowledge that is usually the target of second language acquisition research, and which we will refer to as the 'knowledge at issue' in particular experimental contexts. The second language learner would not consider themselves fluent or native-like if they had to consciously entertain linguistic rules while speaking or comprehending. And a researcher would not have good evidence for the operation of implicit learning processes if the learner's behaviour in a test were the result of conscious knowledge since this could (but not necessarily) be the result of explicit learning. In this chapter

we argue that so-called 'indirect' tests are a good means of reducing (but not necessarily eliminating) the possibility that participants' performance will be affected by consciousness of the knowledge at issue and that they therefore provide a useful means of assessing implicit knowledge.

Note, however, that whereas for the first language learner the knowledge at issue usually is consciously inaccessible in principle, this may not be the case for the second language learner, particularly if they have received explicit instruction. In this case the point is not that the knowledge cannot be brought into consciousness at all, but rather that the person does not use that knowledge with conscious intent *in the moment* of producing some behaviour[1]. Conscious awareness of knowledge in some situations does not rule out unconscious use of that knowledge in others.

Before introducing indirect tests we will consider other means by which researchers have tried to ensure that task performance is a reflection of unconscious, as opposed to conscious knowledge. One approach is to use speeded tasks, on the assumption that conscious knowledge is slower to mobilise than unconscious knowledge. Empirical support from this idea comes from a classic semantic priming experiment by Neely (1977). He showed that at an interval between a prime and target of 250 msecs the category name BIRD facilitated lexical decision responses to related targets such as 'robin', as one would expect. But in another condition the participants were instructed to expect that the prime 'BODY' would be followed by a part of a building, e.g., BODY – roof. Facilitation of expected targets, and inhibition of unexpected targets (e.g. BODY – table), in this condition was only obtained at a prime-target interval of 750 msecs. This puts a lower limit on the speed with which non-automatic, controlled, retrieval of knowledge can occur, but of course, only in this specific situation where arbitrary associations between categories have just been learned. The problem is that the speed with which conscious knowledge may be mobilised cannot be determined a priori. Besides, even if the knowledge at issue does not influence performance in a speeded task it does not mean that it does not exist in implicit form. In the early stages of acquisition knowledge may be relatively weakly encoded in memory, may be unavailable to consciousness, and may take time to activate (Cleeremans, 2011), so it may only have an effect in unspeeded tasks.

Can we use conscious awareness as a way of isolating the contribution of implicit knowledge to task performance? The problem here, of course, is how to assess awareness at the point of use in a test task. Ronald Leow has been a long-time

advocate of the use of concurrent think aloud (TA) precisely for this purpose (Ronald P. Leow, et al., 2014). Of course, as in scientific inquiry in general, one may worry that the very act of measurement alters the nature of the object being measured, which in the TA literature is known as reactivity. Whilst there is evidence that TA has no effect on second language performance (e.g., Morgan-Short, et al., 2012) Rebuschat et al. (in press) found evidence for the reactivity of concurrent TA in a semi-artificial language learning experiment (we will return to this below). The alternative is to use verbal report measures after the behaviour of interest, say at the end of the test phase of an experiment. One problem here is that if the conscious impressions that affected performance were merely fleeting then they may be rapidly forgotten and difficult to report after a delay. The force of this argument depends on the simplicity of the regularities that are being tested. Whereas for highly complex regularities, such as investigated in artificial grammar research, participants' conscious impressions may be fragmentary, only partially veridical, and likely to go unreported after the event. We would argue that this is less of a problem with regularities that are easier to conceptualise, such as simple agreement patterns based on familiar conceptual categories.

Speed manipulations and verbal report are usually used in the context of grammaticality judgement tasks as measures of linguistic knowledge. In this case the test directly probes the knowledge at issue. Possessing the knowledge will allow the participant to discriminate between grammatical and ungrammatical sentences. This is a 'direct' test of knowledge. As such the task invites the participant to reflect upon the knowledge at issue, and hence performance typically reflects a complex mix of implicit and explicit knowledge (Rebuschat, 2011; this volume). Another approach is to construct test situations in which the knowledge at issue is not directly relevant to the primary task as the person perceives it. These are known as 'indirect' tests. This tactic will reduce the probability that consciousness of the knowledge at issue will influence performance.

The distinction between direct and indirect tests originates in research on implicit and explicit memory. For example, Keane et al. (1995) first asked their participants to rate how much they liked each word on a list. Their memory for the words was then tested in two ways. In the perceptual identification test, words from the list and new words were flashed very briefly and the participants' task was merely to read the words if they could. It was found that the probability of being able to

identify a word was about 6% higher for words that had occurred on the list than words that had not – a reflection of repetition priming. In the second test the participants were presented with the same words as in the perceptual identification test and asked to indicate whether they remembered seeing them in the first liking judgment task. The participants were above chance on this test showing that they had conscious memory for the repeated words. Both tests reveal memory effects. But the first, perceptual identification task is an indirect test of what one might call the 'knowledge in question' (in this case, whether a word occurred in the liking task). The participants' task was simply to identify the words, not indicate whether they consciously remembered them. The second, recognition memory task, was a direct test of memory because the task was aimed directly at the knowledge in question.

It is tempting to argue that in the Keane et al. (1995) experiment when the knowledge in question influenced the perceptual identification test it did so involuntarily, outside of the participants' conscious control. And indeed, priming effects of this kind are generally regarded as reflections of implicit memory (N. C. Ellis, 1994), whereas recognition memory tests are more sensitive to explicit memory. However, the issue is not quite that simple because it is conceivable that performance on the indirect test is contaminated by conscious knowledge. Suppose participants come to realise that words from the liking task are being repeated in the perceptual identification task. Assuming that briefly presented words lead to a number of hypotheses about what the word is, then if a hypothesised word is consciously recognised as having occurred in the liking task there could be a bias towards reporting it as the identified word. In this case, facilitated recognition for repeated words is due to a strategic decision bias based on information in explicit memory. One could try to circumvent this problem by obscuring the relationship between the two tasks, or by increasing the delay between them. Keane et al provide evidence of another kind – when the experiment is carried out on amnesiacs, performance on the recognition memory task is significantly worse than the normal participants, as one would expect, but in the perceptual identification task performance is unimpaired (and in fact there is a numerically larger priming effect for the amnesiacs). This dissociation between performance on direct and indirect tests of memory has been regarded as strong evidence that the two kinds of task tap into different forms of knowledge – explicit and implicit respectively[2]. But when using indirect tests one must always be wary of contamination from decision biases based on explicit

knowledge. Later we will argue that in some cases such biases are detectable from task performance.

In this chapter we will argue that the combination of indirect tests and either or both speed manipulations and verbal report provide a relatively robust method for isolating the influence of unconscious knowledge on behaviour. We provide two examples from the literature on the assessment of implicit knowledge in second language learners followed by two examples from our own work on semi-artificial language learning.

Indirect tests of linguistic knowledge in second language learners

Shadowing

Shadowing tasks are trivially simple – all that is required is that one repeats back immediately what one hears. Yet Marslen-Wilson & Welsh's (1978) early experiments with this paradigm provided the insights that led to the development of an influential theory of spoken word recognition and language understanding (Marslen-Wilson, 1987). These experiments revealed the earliness of spoken word recognition, the importance of top-down lexical and contextual constraints, and how errors in the speech signal may be spontaneously corrected with no disruption in performance, a phenomenon that underlies the logic of elicited imitation tasks in SLA research (Erlam, 2006; this volume).

An experiment by Guillelmon & Grosjean (2001) illustrates how shadowing latencies can be used as an indirect test of linguistic ability, in this case processing of gender agreement in French native speakers and English learners of French. They presented participants with auditory phrases such as "le joli plat" ('the nice dish') and all they had to do was to repeat the noun aloud as quickly as possible. The time between the onset of the noun and onset of vocalisation was recorded. Of interest was what would happen if the incorrect article was used, as in "la joli plat". It was found that for native speakers of French, and early English-French bilinguals, the time to repeat the noun was slower when the incorrect article was used, relative to both the correct article condition, and a neutral article condition ("leur joli plat", 'their nice plate'). But no effects were obtained in very highly proficient late learners of French,

despite the fact that these learners made only 10% errors when subsequently asked to produce the test nouns with the correct article.[3]

Eye movements and the visual world paradigm

The second example addresses a similar issue but using a different methodology, based on the visual world, or "look while listening", paradigm. In Lew-Williams & Fernauld (2007; this volume) 3-year-old native Spanish speaking children would see pairs of objects on the screen, say a ball on the left and a shoe on the right. They would then hear an instruction in Spanish such as "Encuentra la pelota", 'Find the ball', and their eye movements were recorded. In this example, the labels for *ball* and *shoe* have different grammatical gender, *la pelota* and *el zapato*. The time to look at the target object was compared with trials in which the labels for both objects had the same gender, such as *la pelota* and *la galleta* ('cookie'). It was found that the children were faster to orient to the target object when the nouns were of different gender than when they were of the same gender, suggesting that they were using the information in the article to anticipate which object was going to be referred to. Lew-Williams & Fernald (2010; this volume) showed the same effect in adult native speakers of Spanish, but not in adult English late learners of Spanish.

Evaluation – were these good tests of implicit knowledge?

The shadowing task in Guillelmon & Grosjean (2001) (G&G) and the search task in Lew-Williams & Fernauld (2007) (LW&F) are indirect tests because the knowledge at issue is not required in order to perform the task. In G&G the nouns could have been repeated even without any knowledge of French, and in LW&F all that was required was to understand the noun in order to locate the correct picture. Yet knowledge of gender agreement influenced behaviour in both tasks. This knowledge also affected behaviour very rapidly. In G&G shadowing latencies in native speakers were around 500 ms, and in LW&F looking data for the different versus the same gender conditions diverged at around 500 ms after the onset of the article (which had a duration of 300 ms). In both cases it would appear that knowledge of gender was affecting performance quickly and involuntarily, suggesting automaticity of access, and implicitness of representation (although the authors of neither paper explicitly make the latter claim). Yet was the knowledge in question implicit in the sense of being unavailable to conscious report? No, because in G&G the non-native

participants could supply the correct articles nearly all of the time when asked (and we can assume that the participants in LW&F also knew the correct articles given the simplicity of the lexis and the level of the learners). The dissociation between performance on indirect and direct tests might suggest a fundamental difference in underlying representation, such as between procedural and declarative memory (see Ullman, 2001, for evidence relating to this distinction).

But we might still ask whether the effects obtained for native speakers in these tasks were truly automatic, as opposed to being the result of rapid and strategic application of conscious knowledge. For example, in LW&F the pictures were displayed for 2 seconds before the audio was played. In this time the participants could have generated the labels for the objects (with which they had been previously familiarised) and covertly generated the articles. On different gender trials they would merely have had to perceive the difference between *la* and *el* in order to move their eyes to the correct object. A number of considerations count against this explanation however: there was no sign of divergence between the conditions prior to the noun (although this does not rule out the possibility that some participants showed such an effect), there was no mention in the instructions to look at the objects as quickly as possible (and hence no reason for the participants to develop such a strategy), both native and non-native speakers would be equally able to generate expectations in the 2 second preview period (and yet only natives showed the effect), and the same effects were obtained with native 3-year-olds who would have been unlikely to develop such a sophisticated strategy.

In the case of G&G strategic use of conscious knowledge seems even less likely because there was no way that the gender of the article could have been used to strategically anticipate the noun. One possibility is that conscious awareness of the mismatch between article and noun on a particular trial caused a disturbance to behaviour, perhaps by directing attention to the article to check its identity, and hence slowing repetition of the noun. Two considerations count against this explanation, however. First, judging by their post-test performance the late learners would have been able to consciously detect gender mismatches, and yet they showed no effects in shadowing. Second, for native speakers, response times were faster for congruent ("le joli plat") than neutral ("leur joli plat") trials even in blocks that did not contain incongruent trials. Even if in the context of the experiment as a whole participants began to notice gender agreement, as well as disagreement, this could not produce a

difference between congruent and neutral trials, since both are grammatical. In sum, for both experiments, there seem to be good reasons to believe that the effects are due to the rapid, involuntary access and application of the knowledge in question, and not the result of the intentional, strategic, use of conscious knowledge. They make a good case for the relevant knowledge being implicitly represented, at least in native speakers.[4]

Given their simplicity, timed shadowing tasks would seem to provide a promising means for assessing implicit knowledge in SLA research. For example, Hodzik & Williams (under review) show how shadowing latencies can be used to explore the influence of local and global constraints in sentence processing (the results were compared with simultaneous interpreting, which, of course is shadowing plus translation). Hanna et al. (under review) found that highly proficient Chinese late learners of English show slow-downs in shadowing latency when repeating verbs that violate number agreement (e.g., **the chefs bakes*). The size of the effect correlated with the strength of a highly immediate neurophysiological measure, the syntactic mismatch negativity, which is evident around 200 msecs after verb offset, and shows the same localisation as for native speakers performing the same task. No effects in either shadowing or the mismatch negativity were obtained in lower proficiency learners (despite performing well above chance on a grammaticality judgement task on the same structure). The rapidity of the brain response appears to confirm the sensitivity of shadowing to implicit linguistic knowledge.

Indirect tests of implicit knowledge in semi-artificial language learning research

We turn now to the use of indirect tests to evaluate learning of semi-artificial languages. Here there is more of an emphasis on awareness than in the above examples, which arises from an interest in the nature of the underlying learning processes, rather than the product of learning.

The question motivating this research is whether it is possible to spontaneously induce knowledge of linguistic regularities without intention, as opposed to intentionally inducing linguistic regularities through conscious hypothesis formation and testing. What distinguishes these two forms of learning is the involvement of conscious processes. When learning through the process of intentional induction a learning goal is consciously set, strategies are applied for achieving it, and

information is held in working memory - commonly thought of as being closely aligned with focal, and conscious, attention (Cowan, 1999). When learning through incidental induction the knowledge at issue is learned through unconscious processes – be they characterised as associative and statistical learning, or, from a generative perspective, parameter setting within the framework of UG. If we want to establish whether spontaneous or intentional induction is occurring ideally we would have access to the learners' conscious states during the learning process itself, for instance by asking the learners to think aloud. But there is a danger that the very process of thinking aloud will interfere with spontaneous induction. Given that incidental learning is highly sensitive to the distribution of attention (Williams, 2013), and the fact that secondary tasks, if not integrated with the primary task, interfere with implicit learning (Keele, et al., 2003) it seems likely that concurrent think aloud would disturb the incidental learning process. Indeed, Rebuschat et al. (in press) show that asking participants to think aloud during the exposure phase of a learning experiment (based on Williams, 2005) eliminates learning of underlying regularities (see also Rebuschat, this volume).

The alternative to concurrent think aloud during learning is to assess awareness either during or after the test phase; that is, to focus on awareness of the product, rather than the process, of learning. Note, however, that whether or not people are aware of the product of learning is not a primary research question. Rather, the logic is that if people are unaware of the product of learning then it seems likely that they acquired that knowledge through spontaneous induction. In this case we say that 'implicit' learning occurred, a term coined by Reber (1967) to refer to learning that takes place without intention and without awareness of what is learned.

As mentioned above, assessment of awareness is subject to all kinds of problems such as the difficulty of verbalising complex regularities, and forgetting. The former may be alleviated by targeting simple regularities, awareness of which would be easy to verbalise, and the latter by using subjective measures such as confidence and source attributions (such as whether the response was based on guess, intuition, memory, or a rule) at the time of making each decision (Rebuschat, 2011, this volume). Note, though, that Rebuschat et al. (in press) also show that, at least when simple regularities are used, asking people to indicate whether their responses were based on a rule can encourage intentional induction during the test phase, obscuring any underlying effects of unconscious knowledge. It appears, then, that

post-experiment verbal report may be preferable as the least invasive option for assessing awareness of the product of learning.

Post-experiment verbal report may actually be a conservative means of isolating cases of spontaneous induction. This is because it is possible that spontaneous induction could lead to conscious knowledge through spontaneous insight (an "aha" or "eureka" moment). And it must also be born in mind (as we will illustrate below) that just because a participant is able to report relevant conscious knowledge after the experiment does not mean that they necessarily used that knowledge, intentionally, to guide decisions in the test task. This possibility becomes more likely when an indirect test is used, and more likely still when speed is emphasised. Thus, the combination of verbal report and indirect testing provide a conservative methodology for isolating the effects of knowledge acquired through spontaneous induction. We now describe two lines of research that employed this logic, the first using reaction time tasks, and the second a memory task.

Speeded decision tasks as indirect tests

Experiments reported in Leung & Williams (2012) and later variants reported in Leung & Williams (Leung & Williams, in press; available on IRIS) illustrate how speeded indirect tests and verbal reports can be combined as tests of implicit learning. These experiments were a development of previous work by Williams (2005) that examined implicit learning of a semi-artificial animacy-based agreement system. Participants were first introduced to 4 novel words, *gi*, *ro*, *ul*, and *ne*, and told that they were like the English article 'the', but they also encoded the distance between the speaker and the object, such that, for example, *gi* and *ro* are used with nearby objects and *ul* and *ne* with distant objects. What they were not told was that two of the articles, say *gi* and *ul*, were used with living things, and the other two, *ro* and *ne*, with non-living things. In other words, the language had a 2 x 2 factorial structure in which the participants were informed about the distance dimension, but not the hidden animacy dimension. The task in Leung & Williams (2012) involved looking at pictures whilst hearing noun phrases and deciding whether the noun in the phrase referred to a living thing or not. For example, they might see a picture of a crocodile on the left of the screen and a hot air balloon on the right. They would then hear the phrase "gi crocodile" and have to indicate as quickly as they can that the noun refers

to a living thing by pressing the right-hand response key. They then have to repeat and translate the noun phrase aloud by saying "gi crocodile, the near crocodile". Distance was conveyed in the pictures by presenting 'far' referents as relatively small images in the top left corner of the left or right image frame. After 280 such trials the experiment seamlessly continued into a test phase, with no perceptible break from the participants' point of view, in which a block of 32 grammatically correct items, 'control' trials, was seamlessly followed by a block of 32 'violation' items in which the articles had the wrong animacy value, as in *ul kettle*. On all of the Control and Violation trials the pictures on the screen had the same distance, but opposite animacy. We predicted that if the participants had learned the association between *gi/ul* and 'living' and *ro/ne* and 'non-living' then their animacy decision times should be slowed down when this pattern was violated.[5] The items in the Control and Violation conditions were counterbalanced across two groups of participants so that any difference in response time between the two conditions was not confounded with item-specific factors. The nouns in the test phase had not occurred during training so that any difference between conditions would reflect generalised knowledge of the underlying animacy rule.

Of course, the question was whether reaction time effects would be obtained in participants who remained unaware of the relevance of animacy to article use. This was assessed in a post-test in which the participants were asked the following questions: What do you think the experiment was about? Did you notice anything strange towards the end of the experiment? Were you aware during the experiment of factors that determined the use of the articles besides distance? Out of 33 participants 20 did not mention animacy in response to the questions and were classed as "unaware". Ten reported the rule correctly, noticed the change in the violation block, and were classed as "aware". A further 3 either reported the wrong rule (perhaps deriving from the violation block) or worked it out in the test phase. The results showed that for the 20 unaware participants there was indeed a significant difference between the animacy decision times to the nouns in the Control and Violation conditions, with mean response times of 1963 msecs and 2031 msecs respectively, $p < 0.05$. There was also a significant difference for the 10 participants who reported the correct rule, with mean response times of 994 msecs and 1340 msecs respectively, $p < 0.05$. We concluded that the unaware group had implicitly learned the correlation between articles and noun animacy.

Subsequent work reported in Leung & Williams (Leung & Williams, 2014, Experiment 1) replicated this result using a 'double decision' procedure originally developed by Paciorek (2012) and reported in Paciorek & Williams (2015a). This is a much simpler procedure than that used by Leung & Williams (2012) in which noun phrases were presented visually (e.g. 'gi dog' would appear on the screen) and the participant made an immediate animacy decision on the noun by pressing one of two response buttons whereupon the phrase disappeared. The cue 'N/F' then appeared and the participant had to indicate, by button press, whether the article meant near or far. After the training blocks the test block contained 16 control and 16 violation items randomly intermixed. Once again, participants who remained unaware of the relevance of animacy, as assessed by post-experiment interview, and who were even unable to guess the relevance of animacy when inspecting further examples of grammatical items, showed a significant reaction time difference between the Control and Violation conditions. This result was also obtained for Chinese native speakers who received the nouns in Cantonese, but using the same articles as the English version. The effect has also been replicated in Chinese learners of English, although only after a lengthier training phase (Gai, 2013), and in native speakers of English in an ERP experiment (Batterink, et al., 2014).

The advantage of these procedures are that, (i) there is no division between 'training' and 'test' phases, eliminating the possibility that participants will alter their response strategy due to a sense of being tested, (ii) there is no requirement to transfer knowledge from one task to another, respecting the principle of transfer-appropriate processing (Morris, Bransford, & Franks, 1977), and making the test maximally sensitive to the knowledge acquired during training, (iii) the tasks themselves are indirect measures of the knowledge at issue, and (iv) the speeded nature of the task makes it less likely that conscious deliberative processing will influence decisions (but not impossible, as we will discuss below). In one sense it may seem strange to refer to the animacy decision as an 'indirect' measure; after all, the hidden regularity is based on animacy. The point is that the animacy decision can be made perfectly well even in complete ignorance of the regularity, and in this sense the knowledge at issue is not directly relevant to the task.

However, indirect measures do not necessarily preclude the strategic intervention of conscious knowledge, as was discussed in relation to the use of perceptual identification (Keane, et al., 1995) and eye movements (Lew-Williams &

Fernald, 2010). In the present case, it is quite possible that participants who were aware of the hidden regularity could base their animacy decisions on the article alone, and not even pay attention to the noun – at least, that is, until violation trials were encountered, whereupon they would begin to make obvious errors. An aspect of the Leung & Williams (2012) design makes it possible to assess this possibility because the auditory stimuli were created in such a way that the onset of the noun always occurred 1 second after the onset of the article. It is therefore significant that the participants who were classified as "aware" by verbal report showed a mean response time of 994 msecs in the Control block, indicating that many of their responses were indeed made purely on the basis of the article (note also that their error rate rose from 4% to 8% in the Violation block, although this difference was not significant). What is most striking, however, is that amongst the participants classified as "unaware" by verbal report there was not one instance of a response time of less than 1 second in the Control block (and only one in the Violation block). Only "aware" participants showed any evidence of strategic use of the animacy rule in their performance, as one would expect if this knowledge was conscious. This finding validates post-experiment verbal report as a means of identifying participants whose conscious knowledge affected their performance in the test task. Note that it is only possible to identify strategic use of conscious knowledge in this way because an indirect test is used. If participants become aware of the hidden regularity it can alter the way in which they perform the task. This would not be possible with a direct test, such as grammaticality judgement.

Leow & Hama (2013) (L&H) criticised the internal validity of the Leung & Williams (2012) methodology on a number of grounds. One issue concerns the provision of feedback. In our procedure, in the first half of the training, if participants made an error on the animacy decision they received an error signal and the trial was repeated. L&H argue that this may have stimulated explicit learning. This is indeed logically possible, but it would have merely had the effect of increasing the number of participants who became aware of the hidden regularity. The real question is whether post-experiment verbal report was sufficiently sensitive to identify cases where conscious knowledge had an effect on task performance. L&H are sceptical about this on the grounds that delayed report is not sensitive to fleeting impressions that may arise during the task, and that it may be difficult to verbally report the relevant knowledge (a problem that afflicts concurrent think aloud as well of course). In order

to increase sensitivity of the post-experiment report Leung & Williams (2014) invited participants to make guesses about the conditions under which the alternating articles were used. They found significant learning effects even for participants who failed to guess the hidden regularity (see Leung & Williams, 2012, 2014, for further discussion of the L&H critique).

The most compelling evidence for implicit learning using the double decision reaction time paradigm comes from a recent study reported in Paciorek & Williams (2015a) in which we examined a different domain from the above experiments – namely, semantic preferences of novel verbs. Participants were first introduced to four novel verbs – *powter, mouten, gouble*, and *conell* – and told that two of them (*powter* and *mouten*) meant to increase something, and the other two (*gouble* and *conell*) meant to decrease something. What they were not told was that *powter* and *gouble* went with abstract nouns, and *mouten* and *conell* with concrete nouns. They then performed a reaction time task in which they saw phrases such as *powter the clarity, gouble the awkwardness, mouten the junk, conell the money*. Their task was first to indicate by button press whether the noun had a positive or negative connotation for them (e.g., *clarity* would have a positive connotation, whilst *awkwardness* a negative one). Since there was no prescribed response no feedback was provided. They then indicated by button press whether the verb meant 'increase' or 'decrease'. The important innovation in this procedure is that the connotation decision is independent of, and indeed orthogonal to, the hidden regularity. The items were assigned in such a way that half of the abstract or concrete nouns were expected to provoke positive connotation responses and the other half negative ones (subsequent analyses confirmed that this was the case). In the testing block violations of the abstract/concrete rule were introduced with nouns not encountered in training, e.g. *powter the fruit, mouten the laughter*. Reaction time slow-downs were obtained for these items even though none of the 22 native English speaking participants reported any awareness of the relevance of concreteness.[6] Interestingly this time the reaction time effect was obtained on the second, increase/decrease, decision, rather than the connotation decision (the same effect occurred in another experiment where an abstract/concrete decision was required). The delayed effect may reflect slow-acting implicit knowledge that is relatively weakly encoded in memory (Cleeremans, 2011). Whatever the reason for this, the important point is that both the connotation and increase/decrease decisions are highly indirect tests. Even conscious knowledge of the

concreteness rule will not aid these decisions in any way. This task therefore reaches a level of indirectness similar to that in the Guillelmon & Grosjen (2001) shadowing task. In both cases, reaction time slow-downs would appear to reflect general perturbations in processing arising from ungrammaticality, but not strategic use of conscious knowledge.

Indirect memory tests: The false memory paradigm

All of the examples of indirect tests that we have discussed so far involve laboratory-based procedures involving recording of reaction times and eye tracking. In practical SLA research such tests are not always convenient to administer. Here though we will describe an indirect test of implicit knowledge that is simple to carry out even as a pencil and paper task or over the internet. These experiments exploit the phenomenon of false memory – the tendency to believe we have encountered things, that we in fact have not, because they conform to a pattern. For example, if a person studies a list of words such as *cup, beer, bowl, coffee*, they will subsequently falsely believe that they saw the word *mug* as part of the list (Deese, 1959; Roediger & Mcdermott, 1995).

A study by Cleary & Langley (2007) provides an example of the use of a false memory task to study the acquisition of linguistic knowledge. The participants read 40 nonsensical sentences, each conforming to one of 5 underlying syntactic structures (e.g. *Mechanical consumers submit colder songs* and *Efficient dreams write better umbrellas* have the same underlying structure). Subsequently they were given more sentences and asked to rate how confident they were that they had seen them before, in other words they judged their familiarity. The test contained sentences that had actually occurred previously, new sentences which conformed to the same syntactic structures as those seen previously (e.g. *territorial dancers operate fuzzier curtains*) and new sentences with a different syntactic structure (e.g. *the transparent lecture became a private reverence*). Higher confidence ratings were obtained for new sentences that conformed to the syntactic structures received in training than sentences that did not conform, indicating that the participants had retained abstract representations of the syntactic structures to which they had been exposed. This is an example of syntactic priming, a phenomenon that has been utilised, albeit using different methods, to study syntactic development in both the L1 and L2 (Kidd, 2012; Shin & Christianson, 2012). In one sense, the familiarity judgement task looks like a

direct test of memory since the participants are being asked if they consciously remember seeing a particular item in a study phase. The point is that the test is an indirect measure of the actual knowledge in question, which in this case is their memory for the underlying syntactic structure. Since this knowledge ostensibly has nothing to do with whether a sentence was seen before or not, any influence that it has on judgements is generally assumed to be involuntary and unconscious. Whether this is necessarily the case, though, we shall consider below.[7]

Paciorek & Williams (2015b; forthcoming on IRIS) used a false memory task to examine implicit learning of semantic preferences of novel verbs, which as in Paciorek & Williams (2015a) were *powter, gouble, conell*, and *mouten*. This time the participants read sentences containing the novel verbs, such as 'Nightingale worked tirelessly to improve public health and POWTER the status of nurses', 'The fact that TV is still the most popular medium does not GOUBLE the significance of the radio', 'Your compost pile does not require you to MOUTEN minerals. It only needs moisture and heat', 'The body must first CONELL carbohydrates before it can begin to burn fat'. Their task was to indicate, for each sentence, whether they thought the novel verb had a generally 'increasing' or 'decreasing' meaning. They were not told how many different verbs there would be, or that they would be used with consistent meanings, in order to simulate as closely as possible the conditions of naturalistic vocabulary acquisition. However, feedback was given if they gave the incorrect answer so as to help them learn the correct increase/decrease meanings of the verbs. Two of the verbs were always used with an increasing meaning (*powter, gouble*) and two with a decreasing meaning (*gouble, conell*). What they were not told was that two of the verbs always went with abstract nouns (*powter, gouble*) and two of them with concrete nouns (*mouten, conell*). After making the increase/decrease decision they also indicated how important the ideas expressed in the sentence were to them, simply responding 'important' or 'not important'. There were also sentences which did not contain a novel verb, such as 'In clinical trials, peptides derived from food proteins have shown an effect on blood pressure', 'Emergency treatment is indicated if potassium is very high, or if severe symptoms are present.' For these items the participants just indicated importance. After the exposure phase there was a surprise test phase in which verb-noun pairs were presented and the participants were asked to indicate, first, whether they remember seeing the words near to each other in the same sentence, and second, the confidence in their decision (guess, somewhat confident,

very confident). For example, they might see ‘mouten minerals’, which did occur, and conforms to the concreteness rule (‘old grammatical’, OG), ‘conell proteins’, which did not occur together (the words appeared in different sentences) but conforms to the concreteness rule (‘new grammatical’, NG), ‘powter potassium’, which again appeared in different sentences but violates the concreteness rule (‘new ungrammatical’, NU). The participants were probed afterwards about their awareness of the concreteness rule by thinking aloud whilst deciding which verb would be appropriate in further test sentences. In Experiment 1 of Paciorek & Williams (2015b) 26 out of 36 participants did not appear to have awareness of the concreteness rule. Yet for these “unaware” participants the probability of falsely judging NG pairs as familiar was significantly higher than for NU pairs (0.54 vs 0.43, $p < 0.01$), although of course lower than the probability of judging OG pairs as familiar (0.70). Since the effect was obtained for verb-noun combinations that had not actually been encountered together we argue that it reflects implicit learning of the concreteness rule; that is, an involuntary bias deriving from unconscious knowledge of a semantic generalisation. The 10 “aware” participants produced a similar pattern, with familiarity judgement rates in the NG, NU, and OG conditions of 0.51, 0.39, and 0.70 respectively. The false memory effect in unaware participants was replicated over three further experiments – on a slightly different set of items (to eliminate syntactic cues), in Polish (to eliminate possible translation to verbs with similar semantic preferences), and using test nouns with reduced semantic similarity to training nouns (in order to test the robustness of the abstract semantic generalisation). These three experiments were conducted over the internet, with the procedure programmed in Python and Java Script, and participant recruitment via MechanicalTurk.[8]

We must address the question, as in previous applications of indirect tests, of whether conscious knowledge could have influenced responses strategically. Consciously knowing the concreteness rule will not help identify a new grammatical verb-noun pair as familiar; in fact it would be a hindrance more than a help because it would obscure actual memory for the pairing. But what about rejecting a new ungrammatical pair? If this is recognised as violating the rule, and if it is assumed that all pairs in the exposure phase conformed to the rule, then it can be confidently ruled out as not having been seen before. This strategy turns the indirect test into a direct one, at least for ungrammatical items. How could we detect application of this strategy in the data? This is where the confidence ratings become useful. If a

participant is applying the above strategy then if a test pair is consciously identified as being ungrammatical they should be highly confident in rejecting it; that is, in responding 'unfamiliar'. They do not have to consult their episodic memory for training items. But they should be less confident when rejecting a new grammatical item as 'unfamiliar' (as their episodic memory might suggest they should) because it is perceived as grammatical (suggesting that it might have occurred). Thus we would expect to find greater confidence when rejecting NU than NG items. The greater the person's conscious knowledge, that is, the more complete and secure their grasp of the hidden regularity, then the greater this confidence difference will be. But the probability of accepting NU items as familiar will also go down, increasing the size of the false memory effect, as measured by the difference in accepting NG and NU items as familiar. Thus, we would expect a positive correlation between the NU-NG confidence difference and the false memory effect as measured by the NG-NU endorsement difference. And indeed for the participants who were classified as "aware" by verbal report there was such a correlation, $r(36) = 0.59$, $p < 0.001$ (data combined across Experiments 1 to 4). But there was no such correlation for the "unaware" group, $r(147) = -0.09$. The same contrast held for each individual experiment. The fact that even the "aware" group showed a range of confidence differences (from negative to positive) and learning effects (which in some cases were also negative) suggested that the participants varied in the precision of their conscious knowledge, or the extent to which it was actually used strategically in the task. Because the task is only indirectly related to the knowledge at issue, it is quite possible that participants will have conscious knowledge, but not actually apply it strategically when performing the task. In fact, though, amongst the "unaware" group there were also cases where high confidence differences were associated with large learning effects, although this was in the context of a statistically non-significant correlation between confidence differences and learning effects. Could these participants have been responsible for the overall learning effects? No, because for the participants who were classed as "unaware" by verbal report and whose confidence difference was either zero or negative there was still a significant learning effect, the difference in probability of endorsing NG and NU items being 0.09, $t(69) = 4.67$, $p < 0.001$. We regard this as a highly conservative test of implicit knowledge because it triangulates non-concurrent verbal report with a concurrent subjective measure taken in the moment of responding in an indirect test.

Conclusion

Indirect tests provide the possibility of creating situations in which consciousness of the knowledge at issue is either totally irrelevant to the task, or only relevant if deployed strategically. In the latter cases we have shown how it is in fact possible to detect strategic use of conscious knowledge in task performance, either from changes in response times, or patterns of confidence judgments. The issue here is not so much whether participants have conscious knowledge at all, as say evident in verbal report, but whether they actually used it in such a way that it influenced their behaviour in the test task. When indirect tests are used it is quite possible to have conscious knowledge, but not apply it. This perspective moves us on from considerations of whether any conscious knowledge can be detected in the learner towards the question of whether their task performance (and ultimately their natural language comprehension and production) is dependent upon the application of conscious knowledge in the moment of processing. We argue that constructing situations in which the knowledge at issue is not obviously relevant to the task at hand is a useful way of reducing the probability of intervention from conscious knowledge, and can create situations in which implicit knowledge has an opportunity to manifest itself.

---

[1] Whether the reported knowledge is underlyingly the same knowledge that was used unconsciously to influence behaviour is an interesting question (that may only be resolvable through neuro-scientific methods).

[2] The idea that dissociations in performance necessarily reflect distinct underling neural systems has been contested by Kinder & Shanks (2003) – a perceptual identification task might just be more sensitive than a recognition task, and both could tap into the same underlying knowledge.

[3] For late bilinguals response times in the congruent condition were about 125 msecs slower than the monolinguals whereas for the early bilinguals there was no such difference. There is no reason to suppose that this difference in response speed could mask the grammaticality effect, and indeed one could argue that slower responding would increase sensitivity to grammaticality.

[4] Arnauld & Ramscar (2012) provide a plausible explanation for the difference between native and non-native speakers in the domain of gender agreement. Whereas children may learn article-noun combinations initially as whole, unsegmented chunks, adults are more likely to learn them as separate units.

[5] Our original hypothesis was that this would probably be because the article orients attention to the wrong picture on the screen. However, given the results of Lew-

Williams & Fernauld (2010) this is unlikely to be the case, at least as might be reflected in eye movements.

[6] No effect was obtained for high proficiency non-native speakers of English. We attributed this to reduced automaticity of semantic access in the second language.

[7] Hamrick (2014) reports a similar experiment in which new test sentences only differed from (old) trained sentences with respect to syntactic structure. Hence this was a direct test of memory for syntax. Discrimination was above chance indicating conscious memory for structure.

[8] The potential drawbacks of internet-based testing and crowd-sourcing have to be acknowledged (e.g., participants may take breaks, take notes, not concentrate adequately, or not follow instructions correctly). These are generally compensated for by larger than usual sample sizes. Paciorek & Williams (2015) obtained similar results for lab-based and internet-based experiments (see also Schnoebelen & Kuperman, 2010).

References

Arnon, I., & Ramscar, M. (2012). Granularity and the acquisition of grammatical gender: How order-of-acquisition affects what gets learned. *Cognition, 122*, 292-305.

Batterink, L. J., Oudiette, D., Reber, P. J., & Paller, K. A. (2014). Sleep facilitates learning a new linguistic rule. *Neuropsychologia, 65*, 169-179.

Cleary, A. M., & Langley, M. M. (2007). Retention of the structure underlying sentences. *Language and Cognitive Processes, 22*, 614-628.

Cleeremans, A. (2011). The radical plasticity thesis: how the brain learns to be conscious. *Frontiers in Psychology, 2*, 1-12.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 62-101). Cambridge: Cambridge University Press.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17-22.

Ellis, N. C. (1994). Vocabulary acquisition: The explicit ins and outs of explicit cognitive mediation. In N. C. Ellis (Ed.), *Implicit and Explicit Learning of Languages* (pp. 211-282). London: Academic Press.

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

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: an empirical validation study. *Applied Linguistics, 27*, 464-491.

Gai, S. (2013). Implicit Learning by Chinese Adult EFL Learners: Extending Leung & Williams (2012). *Foreign Language Teaching and Research, 2*.

Guillelmon, D., & Grosjean, F. (2001). The gender marking effect in spoken word recognition: The case of bilinguals. *Memory and Cognition, 29*, 503-511.

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

Hanna, J., Shtyrovb, Y., Williams, J. N., & Pulvermüller, F. (under review). Early neurophysiological indices of morphosyntactic processing in second language.

Hodzik, E., & Williams, J. N. (under review). Predictive processes during simultaneous interpreting from German into English.

Keane, M. M., Gabrieli, J. D. E., Mapstone, H. C., Johnson, K. A., & Corkin, S. (1995). Double Dissociation of Memory Capacities after Bilateral Occipital-Lobe or Medial Temporal-Lobe Lesions. *Brain, 118*, 1129-1148.

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

Kidd, E. (2012). Implicit statistical learning is directly associated with the acquisition of syntax. *Developmental Psychology, 48*, 171-184.

Kinder, A., & Shanks, D. R. (2003). Neuropsychological dissociations between priming and recognition: A single-system connectionist account. *Psychological Review, 110*, 728-744.

Krashen, S. (1981). *Second language acquisition and second language learning*. London: Pergamon.
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., & Hama, M. (2013). Implicit learning in SLA and the issue of internal validity. *Studies in Second Language Acquisition, 35*, 545-557.
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.
Leung, J. H. C., & Williams, J. N. (in press). Crosslinguistic differences in implicit language learning. *Studies in Second Language Acquisition*.
Lew-Williams, C., & Fernald, A. (2007). Young Children Learning Spanish Make Rapid Use of Grammatical Gender in Spoken Word Recognition. *Psychological Science, 18*, 193-198.
Lew-Williams, C., & Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *Journal of Memory and Language, 63*, 447-464.
Marslen-Wilson, W. (1987). Functional Parallelism in spoken word recognition. *Cognition, 25*, 71-102.
Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology, 10*, 29-63.
Morgan-Short, K., Heil, J., Botero-Moriarty, A., & Ebert, S. (2012). Allocation of Attention to Second Language Form and Meaning: Issues of Think-Alouds and Depth of Processing. *Studies in Second Language Acquisition, 34*, 659-685.
Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of Processing Versus Transfer Appropriate Processing. *Journal of Verbal Learning and Verbal Behavior, 16*, 519-533.
Paciorek, A. W. (2012). *Implicit Learning of Semantic Preferences.* University of Cambridge, Cambridge.
Paciorek, A. W., & Williams, J. N. (2015a). Implicit learning of semantic preferences of verbs. *Studies in Second Language Acquisition*.
Paciorek, A. W., & Williams, J. N. (2015b). Semantic generalisation in implicit language learning. *Journal of Experimental Psychology: Learning Memory and Cognition*.
Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 855-863.
Rebuschat, P. (2011). Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching. *Studies in Second Language Acquisition, 33*, 131-132.
Rebuschat, P., Hamrick, P., Sachs, R., Riestenberg, K., & Ziegler, N. (in press). Triangulating measures of awareness: A contribution to the debate on learning without awareness. *Studies in Second Language Acquisition*.
Roediger, H. L., & Mcdermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning Memory and Cognition, 21*, 803-814.
Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija, 43*, 441-464.

Shin, J.-A., & Christianson, K. (2012). Structural Priming and Second Language Learning. *Language Learning, 62*, 931-964.

Ullman, M. T. (2001). The neural basis of lexicon and grammar in first and second language: the declarative/procedural model. *Bilingualism: Language and Cognition, 4*, 105-122.

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

Williams, J. N. (2013). Attention, awareness, and noticing in language processing and learning. In J. M. Bergsleithner, S. N. Frota & J. K. Yoshioka (Eds.), *Noticing and Second Language Acquisition: Studies in Honor of Richard Schmidt* (pp. 39-57). Honolulu, Hawai'i: National Foreign Language Resource Centre.