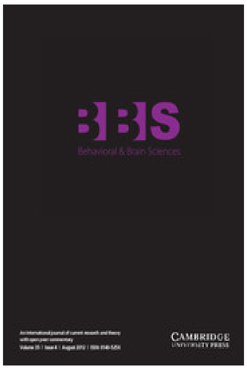


Characteristics of dissociable human learning systems

David R. Shanks and Mark F. St. John

Behavioral and Brain Sciences / Volume 17 / Issue 03 / September 1994, pp 367 - 395
DOI: 10.1017/S0140525X00035032, Published online: 04 February 2010

BEHAVIORAL AND BRAIN SCIENCES (1994) 17, 367–447
Printed in the United States of America

Characteristics of dissociable human learning systems

David R. Shanks
Department of Psychology, University College London, London WC1E 6BT, England
Electronic mail: *david.shanks@psychol.ucl.ac.uk*

Mark F. St. John
Department of Cognitive Science, University of California at San Diego, La Jolla, CA 92093
Electronic mail: *mstjohn@cogsci.ucsd.edu*

Abstract: A number of ways of taxonomizing human learning have been proposed. We examine the evidence for one such proposal, namely, that there exist independent explicit and implicit learning systems. This combines two further distinctions, (1) between learning that takes place with versus without concurrent awareness, and (2) between learning that involves the encoding of instances (or fragments) versus the induction of abstract rules or hypotheses. Implicit learning is assumed to involve unconscious rule learning. We examine the evidence for implicit learning derived from subliminal learning, conditioning, artificial grammar learning, instrumental learning, and reaction times in sequence learning. We conclude that unconscious learning has not been satisfactorily established in any of these areas. The assumption that learning in some of these tasks (e.g., artificial grammar learning) is predominantly based on rule abstraction is questionable. When subjects cannot report the "implicitly learned" rules that govern stimulus selection, this is often because their knowledge consists of instances or fragments of the training stimuli rather than rules. In contrast to the distinction between conscious and unconscious learning, the distinction between instance and rule learning is a sound and meaningful way of taxonomizing human learning. We discuss various computational models of these two forms of learning.

Keywords: artificial grammar; categorization; connectionism; consciousness; explicit/implicit processes; instances; learning; memory; rules

1. Introduction

A classic issue faced by researchers attempting to understand the basic laws of learning is whether there is more than one basic learning mechanism. Can all the phenomena of learning be accommodated by a unitary mechanism, or do we need to posit the existence of independent and dissociable human learning systems? In this target article we consider some of the experimental evidence – much of it very recent – that has addressed this issue.

We will consider two dimensions on which it has been suggested that functionally distinct learning systems differ. The first dimension concerns the role of awareness during learning. Many authors (e.g., Hayes & Broadbent 1988; Lewicki et al. 1987; Reber 1989a) have argued that in addition to having a learning system whose functioning is accompanied by concurrent awareness of what is being learned, humans have a quite separate system that operates independently of awareness. The second dimension, which turns out to be closely related to the first, concerns the content of learning. Distinct learning systems encode very different sorts of information; one system induces rules (e.g., Lea & Simon 1979; Nosofsky et al. 1989), whereas a second system memorizes instances (e.g., Brooks 1978; Medin & Schaffer 1978).

We believe it is important to evaluate the current evidence for and against the multiple-systems view for at least two reasons. First, each of the separate systems that has been hypothesized has tended to encourage researchers to develop a set of explanatory constructs that are unique to that system and that allow its characteristic phenomena to be explained. A drawback, however, is that experimental results are often interpreted exclusively in terms of these restricted concepts, with no consideration of whether they might also be understood (and possibly better understood) in terms of more general principles.

The second and perhaps more pressing reason for evaluating the evidence for dissociable learning systems is that there has been considerable interest, over the last few years, in whether there exist dissociable *memory* systems (for reviews, see Richardson-Klavehn & Bjork 1988; Schacter 1987; 1989; Squire 1992). The mounting positive evidence comes from a variety of sources. For example, amnesic patients have been shown to be dramatically impaired on certain direct tests of memory, such as free recall, but less impaired or even unimpaired on indirect tests of memory, such as motor skills (see Squire 1992). Although dissociations between performance on direct and indirect tests do not force us to conclude that there are dissociable memory systems (e.g., Jacoby & Kelley 1991; Roediger 1990), some researchers have argued at length that the experimental results, together with current understanding of brain functioning, strongly imply the existence of separable underlying systems (e.g., Schacter 1989; Squire 1992).

Few would argue that learning and memory can be studied independently. On the contrary, the possible characteristics of dissociable learning systems should be considered in research on the issue of dissociable memory systems and vice versa. Indeed, if there really are dissociable memory systems, it seems very likely that there are also dissociable learning systems that supply them with information. Yet, as several authors have noted (e.g., Berry & Dienes 1991; Reber 1989a), research on learning and on memory has tended to proceed independently. We hope to help memory researchers in their attempts to understand information storage and retrieval by examining carefully the question of whether distinct learning systems exist and by analyzing the properties of the learning mechanisms that acquire information.

1.1. Proposed distinctions between types of learning

Distinctions between different types of learning have been common in psychology for many years. One such distinction is between declarative and procedural learning, that is, between the acquisition of factual knowledge and the acquisition of skills, respectively (e.g., Cohen & Squire 1980; Morris 1984; Winograd 1975). Other distinctions include the acquisition of "habits" versus "memories" (Mishkin et al. 1984) and "taxon" versus "locale" learning (O'Keefe & Nadel 1978). Of course, if independent memory systems require independent learning mechanisms, then many more distinctions might be needed. For instance, we might require separate learning systems to feed semantic and episodic memory stores (Neely 1989; Tulving 1983; see also multiple book review: *BBS* 7(2) 1984).

Of these distinctions, the one between declarative and procedural learning has probably attracted the most attention, with a variety of empirical phenomena being interpreted in that framework. For example, Cohen and Squire (1980) suggested that amnesics have normal or near-normal procedural learning but impaired declarative learning, a theoretical notion that has been widely taken up by other researchers in the amnesia field. This distinction has in recent years been largely eclipsed, however, by the alternative distinction between "explicit" and "implicit" learning. (Note that some authors have replaced the original declarative/procedural distinction with the terms "declarative" and "nondeclarative" [e.g., Shimamura & Squire 1989; Squire 1992].) The main reason for the shift in terminology and emphasis toward the terms "explicit" and "implicit" is dissatisfaction with the original terminology, the term "procedural" apparently being too narrow to encompass the relevant learning effects. For example, the learning that is preserved in amnesia is not always of a procedural nature: it includes a variety of priming effects involving, for instance, the ability to complete word stems (Graf et al. 1984) and an increase in the likelihood of judging a nonfamous name famous as a result of prior exposure (e.g., Squire & McKee 1992).

The term "implicit learning" was first coined by Reber (1967), who is responsible for much of the recent interest in the issue of distinct learning systems (see Reber 1989a for a review). Different authors have used a variety of definitions to capture the fine detail of the explicit/implicit learning distinction (see Mathews et al. 1989, for examples), but the key factor is the idea that implicit learning occurs without concurrent awareness of what is being learned and represents a separate system from the one that operates in more typical learning situations, where learning does proceed with concurrent awareness (i.e., explicitly). At the same time, it is clear that many authors have been concerned with the possibility that different learning tasks might give rise to different kinds of knowledge (e.g., Mathews et al. 1989; Reber 1989a; Vokey & Brooks 1992), one kind abstract or rule based and the other based on separate fragments or instances. For Reber, implicit learning is not only unconscious but also involves the acquisition of abstract information.

The paradigm case is language learning, where people are assumed to be able implicitly to learn abstract grammatical rules. Few nonlinguists are aware of or are able to articulate the grammatical rules supposed to underlie their linguistic performance, so it makes sense to imagine that those rules are acquired, if at all, without ever being directly represented in consciousness. The rules are abstract in the sense that they apply equally to any linguistic tokens, including novel ones, that come from the appropriate syntactic categories.

Because the aware/unaware and rules/instances dimensions are logically distinct, we believe that they must be treated independently, in this target article we accordingly review evidence for these two dimensions separately. In what follows, we reserve the term "unconscious learning" for learning without awareness, regardless of what sort of knowledge is being acquired. At the same time, we use the terms "rule learning" and "instance learning" to refer to the acquisition of abstract and fragmentary knowledge, respectively, regardless of whether such learning is conscious.

Most of the article is devoted to whether unconscious learning is indeed supported by empirical evidence. In section 2 we survey a wide range of learning paradigms, from subliminal learning phenomena to Pavlovian conditioning to artificial grammar learning and serial reaction-time tasks. The stimuli and specific processes involved in performing and learning each of these tasks differ widely and may share some basic characteristics or may exhibit some basic differences. Across these diverse paradigms we find little actual support for unconscious rule induction (i.e., for implicit learning), or for the unconscious learning of any other type of information. However, in section 3 we do find evidence for a dissociation between a rule-induction system and an instance-memorization system; we review evidence for this dissociation obtained in explicit, or conscious, learning tasks. Within each system, the range of different processes and information is still large, but they nevertheless seem to form two distinct types: slow, effortful hypothesis testing on the one hand, and fast, efficient memorization of instances and fragments of instances on the other.

We concentrate throughout on data from normal subjects. It is clear, however, that amnesic patients have learning difficulties, and these difficulties have been widely interpreted within the explicit/implicit framework (e.g., Squire 1992). For our present purposes, the data from such subjects are tangential, because the question of awareness during learning has not been directly considered in amnesics (but see Knopman 1991). In section 4 we comment briefly on the interpretation of learning data from this population of subjects.

2. Can learning occur without awareness?

Proponents of the explicit/implicit distinction have argued that there are clear demonstrations of subjects' ability to encode new information without being aware of that information, and hence that awareness is the key dimension on which separable learning systems differ. The question of whether learning can occur without awareness goes back many decades (e.g., Adams 1957; Dulany 1961; Eriksen 1960; Krasner 1958; Thorndike & Rock 1934). In addition to the recent work of Reber, which we consider below, in the last five or six years there have been a large number of *sequence learning* reaction time studies that have adopted an interesting and novel technique for assessing the relationship between awareness and learning. A substantial part of our review concerns results obtained using this task. We also consider evidence from a variety of conditioning procedures. We begin with some comments on experimental methodology.

2.1. The logic of dissociations

Almost all studies of unconscious learning have adopted a very constrained version of the logic of dissociation. Separate indices of learning and awareness are used in the attempt to find circumstances in which exposure to a set of stimuli leads to detectable learning unaccompanied by any reliable degree of awareness. On the face of it, such an approach could lead to unequivocal evidence of unconscious learning, but researchers using similar logic to try to establish the existence of unconscious perception have noted several problems (e.g., Reingold & Merikle 1988). What counts as a suitable test of awareness? Can we discount the possibility that our index of awareness is contaminated by unconscious information? Can we be sure it is sufficiently sensitive to detect exhaustively all conscious information? As we shall see, these are deep problems, and researchers have adopted a variety of strategies to try to circumvent them.

Firmer evidence for unconscious learning may emerge from experiments based on alternatives to this particular dissociation paradigm. To test unconscious perception, for example, Reingold and Merikle (1988) have proposed a new and interesting procedure, whereby one looks for greater sensitivity to some variable in an indirect test in which instructions make no reference to the variable as compared with an otherwise identical direct test in which the instructions do refer to the variable. Alternatively, one could try to demonstrate the independence of two learning systems by trying to establish *qualitative* differences between them (e.g., Merikle & Reingold 1992), such that, for example, one system is affected in one way by a variable, the other in the opposite way.

We know of only one study that has even come close to establishing such qualitative differences; this case will accordingly be considered in some detail. Hayes and Broadbent (1988) began by postulating two independent systems: an unconscious system that would slowly accumulate information about predictive events in the environment and a conscious system that would test hypotheses. They further assumed that the conscious system would be highly dependent on a limited-capacity working memory system, and the unconscious system would be independent of it. [See also Broadbent: "The Maltese Cross" *BBS* 7(1) 1984.]

A rather straightforward prediction emerges from this plausible model of the cognitive system. Because the conscious learning mechanism relies on working memory, there should be situations where learning is profoundly affected by loading the working memory system with a secondary task, such as generating random numbers. At the same time, because the unconscious system does not depend on working memory, other (implicit) learning tasks should be unaffected by such a secondary task. Indeed, Hayes and Broadbent went so far as to say that unconscious learning might be facilitated by a secondary task if it prevented the conscious system from exerting an interfering influence on the unconscious system. The importance of the Hayes and Broadbent study is that, in accordance with their model, they appeared to have found two learning tasks that differed in only a minor way, one of which was inhibited and the other facilitated by a secondary task.

In their experiments Hayes and Broadbent contrasted performance in two versions of the computer "person" task. On each trial, the subject entered an attitude (e.g., polite) into the computer, which then responded with its attitude (e.g., unfriendly). The subject's task was to try to get the computer to be friendly. If we designate the 12 possible attitudes – going from very unfriendly to loving – with the numbers 1 . . . 12, then the computer's attitude on each trial was a simple numerical function of the subject's input. In one (No-Lag) condition, the computer's attitude (O_t) on each trial was a function of the subject's attitude (I_t) on the same trial:

$$O_t = I_t - 2 + r \qquad (1)$$

where r is a random number (−1, 0, or 1) and the attitudes have the 12 numerical values mentioned above. In the other (Lag) condition, I_t was replaced by I_{t-1}, so that the computer's attitude was determined by the subject's attitude on the preceding trial:

$$O_t = I_{t-l} - 2 + r \qquad (2)$$

Performance was measured in terms of the number of trials in which the subject's input was one that could (given the random element) have produced a friendly response from the computer person. Although learning occurred in both groups, Hayes and Broadbent found that subjects could give highly accurate verbal reports about the No-Lag task, indicating that their learning had been accompanied by awareness, whereas the verbal reports of subjects in the Lag version were very poor. This result encourages the view that learning in the No-Lag task can be readily achieved by the explicit system, but that the Lag task requires the implicit system. Thus we might predict that a concurrent secondary task would have an effect on learning in the No-Lag condition but not in the Lag condition.

To test this, Hayes and Broadbent (1988) gave subjects a block of learning trials using either Equation 1 (No-Lag group) or Equation 2 (Lag group). After 30 trials in the No-Lag condition and 50 trials in the Lag condition, performance was approximately equated, and at this point Hayes and Broadbent changed the rules by replacing the −2 in the equations with +2. They then presented a further 30 (No-Lag group) or 50 (Lag group) relearning trials. Under single-task conditions (Experiment 1), performance in the Lag condition was affected more detrimentally than in the No-Lag condition by this rule change. In contrast, when subjects were required to perform a concurrent secondary task (generating random letters or digits; Experiments 2 and 3), a change in the rule interfered more with performance in the No-Lag than in the Lag task, exactly the opposite of the result obtained when there was no secondary task. The results conform to Hayes and Broadbent's theory – and hence to their conception of separate implicit and explicit learning systems – if we simply assume that the secondary task occupied the conscious working memory system and therefore interfered with the explicit system, whereas removal of the working memory system allowed the implicit system to operate without any interfering influence from the explicit system.

Unfortunately, Green and Shanks (1993) were unable to replicate Hayes and Broadbent's results. In the single-task groups, Green and Shanks found that the introduction of the equation change had similar effects on performance in the No-Lag and Lag groups, thus failing to replicate Hayes and Broadbent's (Experiment 1) finding that performance was more detrimentally affected in the Lag condition. Under dual-task conditions the situation was the same: performance was approximately equally affected in the two groups. There was not the slightest hint that performance in the Lag group was less affected by the equation change, and hence Hayes and Broadbent's (Experiment 2 and 3) dual-task results were likewise not replicated. Green and Shanks suggest that Hayes and Broadbent may have obtained the results they did owing to the inappropriate inclusion of subjects who had learned very little prior to the equation change.

Hayes and Broadbent's dissociation posed a genuine problem for theories of learning relying on a single learning mechanism. Because the secondary task appeared to have opposing effects on the two primary tasks, Hayes and Broadbent's data seemed to support the claim that there exist dissociable learning systems. Obviously, the fact that their results could not be replicated undermines those conclusions.

With the exception of Hayes and Broadbent's study, implicit learning experiments have universally adopted the dissociation logic of attempting to demonstrate learning in the absence of any detectable degree of awareness. As we shall see, various methodological problems with the dissociation procedure make it doubtful whether unconscious learning has yet been established. It is worth bearing in mind, however, that future experiments using alternative methods may license stronger inferences concerning the dissociability of learning systems. We now begin our discussion of the empirical evidence.

2.2. Unconscious learning with subliminal stimuli

Most studies of unconscious learning have asked whether people can learn about *relationships* between stimuli without being aware of those relationships, but before discussing the results of such studies we will briefly consider evidence from experiments asking a more direct question, Can people learn about stimuli when they are unaware of the existence of these stimuli, that is, when the stimuli are subliminal? A situation in which unconscious learning would, on the face of it, be fairly straightforward to establish is one in which a subject is entirely unaware that the critical stimulus in the learning phase is present at all, yet still shows evidence of leaning something about that stimulus.

There have, of course, been a large number of experiments in which subjects are presented with brief or low-intensity stimuli intended to be below the threshold of awareness and in which an attempt is made to measure effects of such stimulation on subsequent behavior. We ignore much of this literature, for two reasons: first, in some cases such effects may be only tenuously related to learning. For example, many subliminal activation experiments ask whether the way a stimulus is interpreted may be biased by a supposedly subliminal stimulus presented a few hundred milliseconds previously (e.g., Marcel 1983). It is doubtful, however, that such biasing effects would occur over longer intervals: instead, they are typically interpreted as examples of some sort of short-lived facilitation. Needless to say, it is difficult to draw a sharp line between perception and learning, but if unconscious learning is to have any real significance, it must be demonstrable over reasonable intervals of time (at the very least seconds or minutes rather than milliseconds). Second, many subliminal activation experiments that do appear to show longer-lasting effects (e.g., Eich 1984) have already been the subject of extensive criticism in this journal (see Holender 1986, and accompanying commentaries). We have no wish to repeat arguments made previously except to point out that in such experiments it is extremely difficult to be confident that the stimuli are indeed below the threshold of conscious perception.

We accordingly focus in this section on studies that avoid these problems. Andrade (in press), Bornstein (1992), Ghoneim and Block (1992), Greenwald (1992), and Schacter (1987) review a number of relevant studies examining learning with subliminal stimuli. Although there have been some positive results, a corresponding number of negative findings leads us to suggest that unconscious learning with subliminal stimuli has not yet been conclusively demonstrated.

Subliminal stimuli may be presented to awake subjects as auditory messages at extremely low intensity or in some scrambled form, or as images presented for very brief durations or embedded in other figures; alternatively, they may be presented to subjects during sleep or anesthesia. There is a widespread popular belief in the ability of such subliminal messages to condition attitudes or preferences or otherwise to influence behavior. Indeed, this belief is so powerful that the families of two young men who died from self-inflicted gunshot wounds sought more than $6 million in damages from the rock group Judas Priest on the grounds that subliminal messages on one of the group's records had caused the men to commit suicide (see Loftus & Klinger 1992). Recent investigations, however, suggest that the concern is misplaced. Controlled experiments attempting to see whether subliminal messages can influence behavior or whether people can use self-help audiotapes as learning aids have yielded exclusively negative results (British Psychological Society 1992; Greenwald et al. 1991; Vokey & Read 1985). It seems unlikely that unconscious learning can occur in such situations.

Several investigations of spared cognitive functions under general anesthesia have obtained evidence of small but reliable amounts of learning, but these are matched by a comparable number of negative results (see Andrade, in press; Ghoneim & Block, 1992, for reviews). If the anesthetic has been adequately administered and renders the patient entirely unconscious, then spared learning must in turn be unconscious. A typical positive result was reported by Jelicic et al. (1992). They gave anesthetized patients repeated auditory presentations of two words (e.g., yellow, green) from a semantic category. Later, when the anesthetic had worn off, subjects were asked in a priming test to generate members of those categories. Subjects were significantly more likely to produce the preexposed words than were control subjects who had not been read the words during anesthesia. Thus some information does seem to have been encoded while the subjects were unconscious.

Another positive result was reported by Kihlstrom et al. (1990). They gave anesthetized patients lists of strongly associated cue-target word pairs, with each list being presented about 67 times during the operation. Later, when the anesthetic had worn off, subjects were given a cued recall and a recognition memory test; in a third test, they were read the cue words and had to say the first word that came to mind. Although the recall and recognition tests yielded no evidence of retention, on the generation test subjects were more likely to produce target items to preexposed cue words than to nonpreexposed cue words, whether the test was relatively soon after the exposure phase (median 87 min) or much later (median 14 days). Thus, again, some degree of unconscious registration seems to have occurred.

In contrast to this are the many negative results that have been published. Some of these are particularly revealing because they come from experiments using procedures very similar to those of studies that have found positive results. For example, Cork et al. (1992) failed to replicate the Kihlstrom et al. (1990) results using a different anesthetic but otherwise identical procedures. Furthermore, despite the likelihood that sleep renders a person less unconscious than general anesthesia, in a well-controlled experiment Wood et al. (1992) were unable to obtain evidence of learning during sleep, again with procedures similar to those used in the Kihlstrom et al. (1990) study. Similarly, Ghoneim et al. (1992) found no evidence of Pavlovian conditioning in anesthetized patients; they used experimental procedures that did reveal conditioning in nonanesthetized subjects.

This pattern of results might simply indicate that learning under anesthesia is a genuine phenomenon, but that relatively subtle methodological factors determine whether a given study will or will not obtain evidence of it. However, Andrade (in press) discusses a large number of studies, including over 20 published reports of failures, and is unable to find any clear factors that determine whether learning will or will not occur. For example, it does not seem to be especially related to the type of stimuli used. More significantly, it remains an open possibility that many positive results have been due to inadequately administered anesthetic that left some or all of the patients at least partially conscious. It is worth noting that in the Cork et al. (1992) study three subjects were excluded from the analysis because they had explicit memory of the study items! As Cork et al. say, "the extent to which implicit expressions of memory are affected by general anesthesia remains uncertain" (p. 897).

2.2.1. Conclusions. Experiments in which subjects are presented with stimuli that they are likely to be unaware of at the time of exposure yield some evidence of unconscious learning, but this is offset by a substantial body of negative evidence. At present, it would be premature to conclude from the available studies that unconscious learning is feasible.

2.3. Criteria for establishing unconscious learning with supraliminal stimuli

In the rest of this section we focus on situations where the stimuli are above the threshold for detection and identification. In such situations, subjects may be unaware of the relationships between stimuli even though they are aware of the stimuli themselves. Learning of interstimulus *relationships* may therefore be unconscious.

We argue that just about all unconscious learning experiments with supraliminal stimuli can be conceptually reduced to the arrangement shown in Figure 1. The figure illustrates an associative learning episode in which subjects have the opportunity to learn that two events, A and B, stand in a predictive relationship. Event A might be a tone conditioned stimulus (CS) and event B a shock unconditioned stimulus (US); the measure of learning might be a galvanic skin response (GSR) at time t_2 when the CS is presented again. Or event A might be a feature or set of features, event B might be a category, and the measure of learning might be the probability of making the category response at t_2. We are interested in whether subjects can learn the predictive relationship in the absence of concurrent awareness of that relationship. We assume for the sake of simplicity that there is just one learning trial.

Learning itself presumably takes place during or after presentation of event B; we wish to ascertain the subject's state of awareness during this learning episode. Unfortunately, there are likely to be profound technical difficulties in assessing awareness of a predictive relationship at just the moment learning itself occurs. Apart from anything else, asking subjects at time t_1 whether they are aware of the relationship between stimuli A and B is likely to direct their attention to that relationship. As an illustration, in a study by Baeyens et al. (1990a) that will be discussed in more detail later, the proportion of A-B relationships which the subjects appeared to be aware of on a postconditioning recognition test increased from 18% to 77% when subjects *also* gave concurrent estimates of awareness during the learning stage. Clearly, the concurrent index of awareness directed subjects' attention to the relationship and affected the very entity it was designed to measure.

Hence, we will usually have to settle for assessing awareness after the target learning trial. At this time (t_2 in Fig. 1), suppose we present event A (a tone previously paired with shock) and measure the GSR as well as asking subjects whether they have any particular expectancy of event B. If we obtain a GSR but no evidence of a conscious expectancy of event B, we have obtained the crucial finding that lies at the heart of all attempts to demonstrate

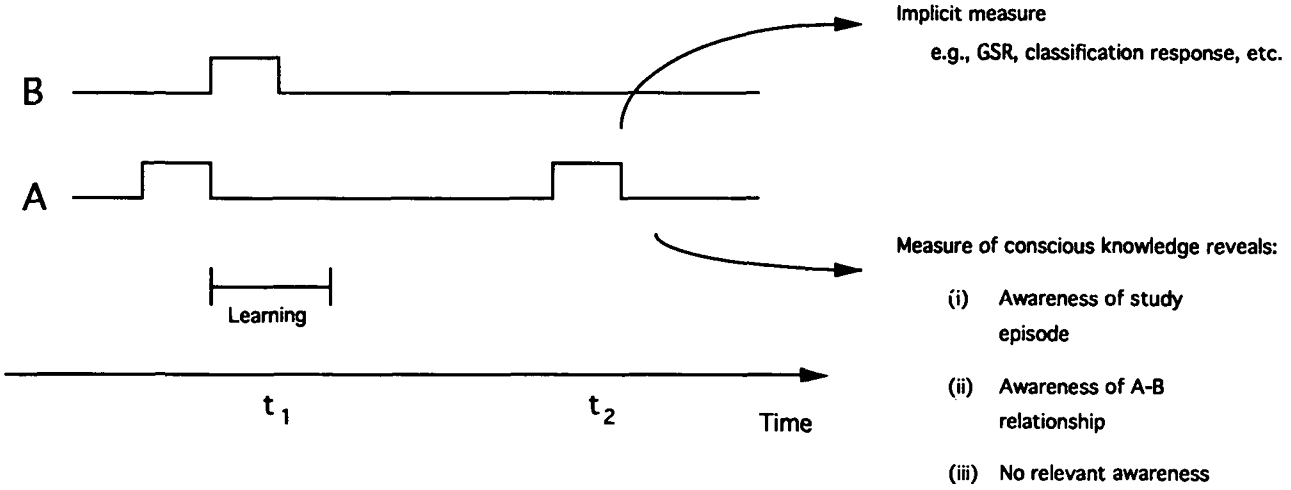

B
A
Learning
t_1
t_2
Time
Implicit measure
e.g., GSR, classification response, etc.
Measure of conscious knowledge reveals:
(i) Awareness of study episode
(ii) Awareness of A-B relationship
(iii) No relevant awareness

Figure 1. Schematic illustration of events in experiments that investigate the role of awareness in the learning of predictive relationships. Subjects witness a predictive relationship between stimuli A and B, with learning presumed to occur during the interval marked t_1. At some time later (t_2) stimulus A is presented again. Performance at t_2 is taken as an index of learning at t_1, whereas a concurrent measure of awareness at t_2 is used to infer the content of the subject's awareness at t_1.

implicit learning with supraliminal stimuli. For if subjects have no expectancy of event B at t_2, we have some basis for inferring that they were not aware of the A-B relationship at t_1.

This might seem to be a very strong inference, but we believe such inferences will have to be accepted if unconscious learning is to be established. It is unavoidably difficult to assess awareness concurrently with learning, so one is forced to rely on some later test. Of course, we also make a backward inference concerning learning itself: if performance at time t_2 is no better than we would expect by chance, we often infer that learning did not occur at t_1. Conversely, if performance is better at t_2 than we would expect by chance, we conclude that learning did occur.

2.3.1. The relationship between unconscious learning and implicit memory. The basic design shown in Figure 1 allows us to see the intimate relationship between unconscious learning and implicit retrieval: demonstrations of unconscious learning are a proper subset of the larger set of demonstrations of implicit retrieval.

Implicit retrieval is defined as occurring when information from some prior episode can be retrieved and can hence influence current processing, but in the absence of conscious recollection of that prior episode (e.g., Schacter 1987; we use the term "implicit retrieval" rather than the more common term "implicit memory" to emphasize that we are specifically considering what happens during the retrieval process). Thus, implicit retrieval requires the absence of a conscious reexperience of the study episode. Now, lack of awareness of a contingency at t_2 presumably means the absence of any consciously recallable episodic memory traces in which that contingency is embedded, and hence any piece of evidence that allows us to infer unconscious learning must also be an example of implicit retrieval: this is case (iii) shown in Figure 1. The converse does not hold, however; an example of implicit retrieval does not necessarily represent evidence of unconscious learning.

Suppose that a subject emits a GSR when presented at test with a tone stimulus. There are three possible scenarios, shown in Figure 1:

1. The subject remembers the study episode, in which case the GSR response does not count as an example of implicit retrieval according to Schacter's (1987) definition. Because remembering the episode entails remembering the content of that episode (i.e., the A-B contingency), the learning could not have been implicit either.

2. The subject does not remember the study episode, but is aware – that is, has semantic knowledge – that this tone predicts shock (cf. source amnesia). Although this qualifies as a case of implicit retrieval, we would not infer that learning itself had been unconscious, since at t_2 the subject is aware that A predicts B. (Note that this ignores the possibility that subjects could have been unaware of the A-B relationship at t_1, but aware of it at t_2, for example, as a result of observing their own behavior. Observation of a GSR in response to the tone might lead the subject to believe that the tone must therefore predict shock. How one might exclude this possibility is a difficult question.)

3. The subject neither remembers the study episode nor has conscious semantic knowledge of the A-B relationship. This final case again qualifies as implicit retrieval. More important, we now have evidence that is relevant to unconscious learning, as lack of awareness of the relationship at retrieval licenses the inference that learning too took place without awareness.

Thus, in order for us to infer unconscious learning from implicit retrieval, the subject must be unaware of the relevant relationship that occurred in the study episode, in addition to being unaware of the episode itself. In summary, an unconscious learning experiment just *is* an implicit retrieval experiment, but with the added component of meeting this further condition. For researchers in the field of implicit retrieval, all that is of interest is whether the subject is unaware of the relevant study episode, as in cases (ii) and (iii). But only case (iii) is relevant to the question of unconscious learning; the subject must also be unaware of the relationship that occurred in that episode. It is for this reason, we argue, that much of the data obtained from amnesics is irrelevant to the question of unconscious learning (see sect. 4).

2.3.2. Dissociation of task performance and verbal reports. Within the dissociation paradigm (Reingold & Merikle 1988), many studies have shown that subjects can acquire information without being able to report it verbally at a later time. Such findings have been taken as support of the claim of unconscious learning. Suppose that subjects are presented with some information at time t_1 and that a subsequent performance test indicates they have encoded this information. We argue that if the aim is to establish what the subjects' state of awareness was at t_1, examining the content of their verbal reports at t_2 is certainly not the only way to do this and may not be the best one.

To illustrate this, note that the condition mentioned above (that the backwards inference must be valid) can be made more specific by dividing it into two further criteria. The first concerns the match between the information responsible for performance changes and the information revealed by the test of awareness. We call this the Information Criterion. The second criterion concerns the sensitivity of the test for awareness. We call this the Sensitivity Criterion.

> *Information Criterion:* Before concluding that subjects are unaware of the information that they have learned and that is influencing their behavior, it must be possible to establish that the information the experimenter is looking for in the awareness test is indeed the information responsible for performance changes.

This criterion is intended to exclude situations such as the following: suppose the experimenter sets up a task in which performance can be improved if the subjects learn information *I*. Performance does indeed improve, and subjects are apparently unaware at time t_2 that they have learned *I*. However, an adequate explanation of the improvement in performance is that subjects are not learning *I*, but *I**. By the experimenter's criteria, awareness of *I** would be disregarded as irrelevant, and so the experimenter would erroneously conclude that the subjects' performance was under the control of some information or knowledge of which they were unaware. The Information Criterion is closely related to the notion of "correlated hypotheses" introduced by Adams (1957) and Dulany (1961) and which will be discussed in section 2.6.1.

Our second criterion is far from new (e.g., Brewer 1974; Brody 1989; Dawson & Schell 1985; Ericsson & Simon 1980; 1984; Eriksen 1960; Reingold & Merikle 1988). It is simply that tests of unconscious learning must achieve an adequate level of sensitivity:

> *Sensitivity Criterion:* To show that two dependent variables (in this case, tests of conscious knowledge and task performance) relate to dissociable underlying systems, we must be able to show that our test of awareness is sensitive to all of the relevant conscious knowledge.

Unless this criterion is met, the fact that subjects are able to transmit more information in their task performance than in a test of awareness may simply be due to the greater sensitivity of the performance test to whatever conscious information the subject has encoded. Let us take as our null hypothesis the claim that there is a single source of conscious knowledge that can manifest itself on both the performance and the awareness test. If performance is above chance, but there is no detectable awareness, an immediate inference is that our test of awareness is simply less sensitive than the performance test to the available resource of conscious information. Or, to put it another way, there is conscious knowledge that is not being detected by the supposed test of awareness but is contributing to task performance.

To rule out this possibility, we must have either (1) some independent reason to believe that the test of awareness is sensitive to all of the potentially relevant conscious information, or (2) some reason to believe that the awareness test is at least as sensitive as the performance test in terms of its ability to detect relevant conscious information. The first of these requires demonstrating that the awareness test is *exhaustive,* something that Reingold and Merikle (1988) have noted is likely to be very difficult to do. In contrast, the second requirement can be met if we try to make the performance and awareness tests as similar as possible in terms of retrieval context, differing only in terms of task instructions. If the instructions in the awareness test encourage the subject to retrieve as much conscious information as possible, and if the retrieval contexts in the two tests are approximately matched, then the Sensitivity Criterion may be met, because it is unlikely that the performance test would elicit the retrieval of more conscious information than the awareness test when the latter has provided subjects with a stronger motivation to do so. If we still obtain a dissociation between performance and awareness under such circumstances, we will have good evidence of unconscious learning.[1]

As an illustration of the application of these criteria, consider a widely cited implicit learning study by Lewicki et al. (1987). In the first phase, each trial consisted of the presentation of a target item in one of the four quadrants of a computer screen (which, for purposes of discussion, we can designate as A, B, C, and D); the subjects' task was simply to press a button corresponding to that quadrant as quickly as possible. The basic idea of these experiments can be simply stated: the choice of target location on each trial was nonrandom, and the question was whether the subjects would be able to detect this nonrandomness.

Subjects were presented with sequences of seven trials, with rules constructed so that target locations on the seventh trial could be predicted from its locations on trials 1, 3, 4, and 6. On each of the first six trials, the digit 6 appeared on its own in one of the quadrants of the screen, but on trial 7 (the "complex" trial), it was embedded in a display containing 36 digits. Reaction time on the seventh trial was the measure of interest. Again, the rules specifying target location were deterministic: thus, if the target appeared in locations C, A, D, and B on trials 1, 3, 4, and 6 respectively, then on trial 7 the target would be in location A.

In common with many other such results (which will be reviewed in sect. 2.7 below), Lewicki et al. (1987, Experiment 1) found that reaction times (RTs) on the target trials decreased significantly across 4,608 complex trials. In addition, RTs increased significantly when, toward the end of the experiment, the rules were changed so that on the complex trials the target now appeared in the quadrant diagonally opposite where it had appeared previously. This latter finding rules out nonspecific factors as the locus of the speedup effect. In a second experiment, Lewicki et al. applied deterministic rules only on two out of three sets of seven trials: on the remaining sets, target location on trial 7 was random. Here, a change in the rules only affected RTs in the sets that were rule determined and not in those that were not.

Lewicki et al. found that none of their subjects came even close to being able to report any of the rules. In fact "none of the subjects were even able to correctly specify which four out of six simple trials were the crucial ones" (1987, p. 529). Thus we appear to have good evidence of a dissociation between performance and reports. It is highly doubtful, however, whether these results meet either the Information or the Sensitivity Criterion. With regard to the former, Lewicki et al. required subjects to try to report "at least one pair of co-occurring elements (i.e., a sequence of four target locations in simple trials and the corresponding location of the target in the subsequent matrix-scanning [complex] trial)" (p. 528). Thus subjects were classified as able to report something about the sequence, and hence as aware, only if they were able to specify a complete sequence of four simple trials and one complex trial. The problem with this classification, however, is that to show a speedup in RT, complete knowledge of the sequences was not necessary.

Analysis of the sequences, for example, shows that even the last simple trial on its own was informative about target location on the seventh trial: if the target was in quadrant A on trial 6, it was twice as likely to be in quadrants A and D on trial 7 as in quadrants B or C. Trial 6 provided a great deal of information on its own about target location on trial 7. Knowledge about trials 4 and 6 provided still more information about target location on trial 7, but if the subjects could report this sort of regularity, it would still not have counted as correct according to Lewicki et al.'s criterion. It is true that knowledge of the sequence across the four relevant simple trials provided absolute certainty about the seventh trial, but our point is that considerable amounts of speedup in RT could be attributable to fragmentary knowledge of "microrules" that Lewicki et al. would not have counted as evidence of awareness, even if the subjects could articulate them.

Turning to the Sensitivity Criterion, we may ask whether the verbal report test is an adequate measure of the subject's awareness in this procedure. We suggest that it is not. First, we cannot be sure that the performance and awareness tests are matched in terms of the conscious information they pick up, because quite different retrieval contexts are provided for the two tests. In the case of RTs, performance is elicited in a context where (1) stimuli are presented on the computer screen, (2) responses are made on the keyboard, (3) a horizontal and a vertical line appear on the screen dividing it into quadrants, (4) a response is made very soon after the preceding response, and so on. All these cues are pertinent, in that they were present during the learning phase (which is just the RT task). In the case of verbal report, none of these cues is present. Instead, the subject is required to retrieve the sequence rules from memory, without the aid of any of the aforementioned cues.

Second, we have little reason to believe that the verbal report test provides an exhaustive index of conscious information, since there are other tests such as recognition that manifestly detect information left undetected by verbal report tests. For example, Nelson (1978) compared the sensitivity of recognition and verbal recall in the following way. Suppose we have two memory tests, A and B. Subjects learn a list of items and are then given test A. Then, test B is applied only to those study items that test A failed to detect. If test B detects any of these items, it is said to be more sensitive than test A. It is important also to apply the tests in the reverse order – test B, then test A – and to fail to observe an increase in sensitivity. Using such a procedure, Nelson showed that recognition tests can detect items not detected by free recall tests, but the converse was not true. Hence, recognition is a more sensitive test than free recall, and the latter is therefore not exhaustive.

Moreover, note that it is possible that subjects misinterpret free report questions to mean they should only report rules. They might believe that fragmentary information is not supposed to be reported. Many researchers have attempted to avoid this problem by asking more and more specific questions about what stimuli may begin or end a sequence, and so on. Such questions are somewhat better from a sensitivity standpoint because they are more specific (and provide more cues), and may be better from an informational standpoint if they ask about the information that subjects actually learn.

In sum, we suggest that the Information and Sensitivity Criteria are not met in Lewicki et al.'s (1987) experiment. The default hypothesis – that there is only a single resource of conscious information – may be correct, with less of that knowledge being detected by the verbal report test than by the RT task. There is no evidence that the knowledge used to perform the RT task is any different or is in any way acquired independently of the knowledge that the subject's reports are based on. Verbal reports are impoverished compared to task performance simply because less of the available information is retrieved in the test of reportable knowledge. If the subject were given enough retrieval cues, there is every reason to believe that this knowledge could be brought to consciousness and reported; it is simply that a normal test of verbal report does not do this. Last, if sufficient cues could make the information conscious, there is every reason to believe that it was conscious at the time of encoding.

It is important to note that we are not denying the empirical fact that performance and verbal reports can be dissociated. On the contrary, we acknowledge that there have been numerous satisfactory demonstrations of this (for example, in Lewicki et al.'s [1987] experiment), and that this has interesting implications for applied psychology. Subjects' performance indicates that they have learned something, yet they are poor at articulating verbally what they have learned. Instead, we are suggesting that this dissociation is only very weak evidence for the claim that the original learning was unconscious, and that it provides no evidence at all for the functional dissociation of conscious and unconscious learning. Its status is exactly the same as the difference that commonly emerges between tests of recall and recognition. For the same reason, amnesic patients' inability to recall information that an earlier test shows they had learned (e.g., Nissen & Bullemer 1987) is not in its own right evidence of unconscious learning. Since we are claiming that a dissociation between performance and verbal report is not compelling evidence for unconscious learning, we place special weight (below) on studies that have tried to use more sensitive tests of awareness.

It is also important to recognize that our criteria do not make unconscious learning undemonstrable. As Bowers (1984) has noted, it is pointless to argue about a possible unconscious process if one's criteria for its existence make it a logical impossibility. But the Information Criterion can readily be met in any study that establishes unequivocally what it is that the subject is learning, and the Sensitivity Criterion can be met by tests that adequately reinstate the learning context or that attempt to be exhaustive with respect to conscious information. Indeed, we will see in section 2.7 below that a replication of Lewicki et al.'s experiment by Stadler (1989) met both of these criteria by using an alternative test of awareness. Furthermore, successful demonstrations of unconscious *perception* have been possible in experiments that use tests of awareness that meet these criteria (e.g., Merikle & Reingold 1990). In sum, Lewicki et al.'s (1987) experiments demonstrate the dangers of asking the wrong questions and of ignoring substantial differences between different types of test.

With these considerations in mind, we now turn to other evidence for learning without awareness. In the following sections, we focus on four areas of experimental evidence: conditioning, artificial grammar learning, instrumental learning, and sequential pattern acquisition.

2.4. Awareness and conditioning

2.4.1. Pavlovian conditioning. We begin with a consideration of whether classical or Pavlovian conditioned responses can be acquired in the absence of awareness of the scheduled contingency of reinforcement. Since many researchers regard conditioning as representing a relatively primitive learning system (see Boakes 1989), it is plausible to imagine that learning without awareness can occur in this context. The conclusion from a huge number of studies, however, is quite the opposite: there is no compelling evidence for conditioning in human subjects without awareness of the reinforcement contingency. This conclusion was first reached in a classic review by Brewer (1974), and more recent studies have not changed the situation (see Boakes 1989; Dawson & Schell 1985, for reviews). Such conclusions have not always been heeded, however, because there are still claims in the literature to the effect that conditioning can occur without awareness (e.g., Musen et al. 1990, p. 1074) and is hence an instance of implicit, unconscious learning.

There have been two general approaches to examining the relationship between conditioning and awareness. First, some studies have sought to ascertain whether instructions to the subject concerning the nature of the relationship between a cue and a reinforcer affect conditioning as measured, for instance, by GSRs. The rationale is that if conditioning is a relatively automatic form of learning that can proceed independently of awareness, then changes in the subjects' conscious beliefs ought to have little effect on their behavior. Using this logic, Grings et al. (1973), for example, presented subjects with two conditioned stimuli (CSs), one of which (CS+) was followed by a shock unconditioned stimulus (US), and one of which (CS−) was not. At the end of the training stage, CS+ elicited a larger conditioned GSR than did CS−. Prior to the second stage, subjects were correctly told that the relationship between stimuli and shocks would now be reversed, with shocks following CS− but not CS+.

As has been observed in many other studies, these instructions had a powerful effect on conditioned responding. Grings et al. found that their subjects responded on the first trial of the second stage to CS− but not to CS+, indicating that their knowledge at least partially controlled their responding. Significantly, the response to CS+, a stimulus that had been paired several times with shock, was no greater than the response to a control stimulus that in the first stage had been presented with uncorrelated USs. Similar results of verbal instruction have been obtained in experiments using phobic stimuli such as pictures of snakes (Davey 1992), where it was once thought that conditioned responding could proceed independently of instructions (e.g., Hugdahl & Ohman 1977).

Although such results are unsupportive of the notion that conditioning can proceed without awareness, they do not address the issue directly because awareness itself is not examined. A recent experiment by Lovibond (1992) exemplifies the approach of eliciting measures of awareness concurrently with conditioned responses. Lovibond presented subjects with two stimuli (slides depicting flowers or mushrooms), one of which (the CS+) was paired with shock while the other (CS−) was nonreinforced. Awareness of the relationship between the stimuli and shock was measured in two ways. First, during the learning phase subjects continually adjusted a pointer to indicate their moment-by-moment expectation of shock (note that asking for a rating of shock expectancy does not specifically direct attention to the A-B relationship); and second, at the end of the experiment they were given a structured interview designed to assess their awareness.

It should be apparent how the design conforms to the basic procedure depicted in Figure 1, except that there are four learning trials. In Lovibond's experiments, each of trials 2–4 in fact represents a new learning trial, an assessment of whether learning occurred on the preceding trial(s), and an assessment of the subject's awareness on the preceding trial(s). The Information Criterion should not raise particular problems here, because there is little doubt that the information the subjects learn (the contingency between the CS and US) corresponds with what the awareness test asks them to report.

In each of the experiments, some subjects gave no indication, on either of the tests of awareness, that they associated A with shock to a greater extent than B. Critically, these subjects also gave no hint of stronger conditioned responding to A than to B. For subjects who were aware of the conditioning contingencies, GSRs were stronger to A than to B. Thus, on the basis of these results we would have to conclude that learning about a CS-shock relationship does not occur in the absence of awareness of that relationship. It is also worth noting that Lovibond's experimental design is well suited to demonstrating that our criteria for implicit learning do not make it a logical impossibility. If his results had been different – something which is simply an empirical matter – the criteria would have been met and implicit learning could have been firmly established.

Other studies have tried to mask the CS-US relationship and again compare awareness and conditioning. The results have been clear: so long as awareness is measured by an immediate test, usually a recognition test, significant conditioning only occurs in situations where the subject is aware of the contingency (see Boakes 1989; Dawson & Schell 1985). One recent experiment serves to illustrate the typical result. Marinkovic et al. (1989) presented their subjects with a recognition memory task for odors. On each trial, one odor was presented for 8 sec as a "target," followed in succession by three further odors. Subjects' primary task was to say which of the three was the same as the target. One of the three recognition odors was in fact either the CS+ or the CS−. If it was CS+, a shock was presented at its offset; skin conductance was measured as the conditioned response. The question of interest was whether acquisition of GSRs could occur without concurrent awareness of the contingency between the CS+ and the shock. Marinkovic et al. measured awareness with a test in which subjects were required to indicate their expectancy of the shock during each odor on a 7-point scale. Because awareness was measured during the CSs, this again represents a concurrent assessment of awareness, rather than a *post hoc* one.

The outcome was that differential conditioning to CS+ was only observed in subjects classified as aware, indicating that awareness is necessary for conditioning. In addition, Marinkovic et al. obtained some evidence that when conditioned responding did occur, it only started after the onset of awareness. In sum, results from conditioning experiments appear to contradict the notion that this type of learning can proceed without concurrent awareness.

For a variety of reasons, some researchers have questioned whether GSRs condition in the same way other responses, such as the eyeblink or salivary reflexes, do. Thus it is worth noting that correspondences between awareness and conditioning seem to occur with other response systems as well (e.g., for eyelid conditioning, Baer & Fuhrer 1982).

The conclusion from these studies is clear, and confirms Brewer's (1974) earlier analysis: Pavlovian conditioning, which is often cited as a fundamental form of learning, does not seem to occur in the absence of awareness of the reinforcement contingency.

2.4.2. Evaluative conditioning. Evaluative conditioning refers to a form of learning that manifests itself in changes in affective response to a stimulus (Martin & Levey 1978). Specifically, it refers to the transfer of affect from a US to a CS. Some authors (e.g., Baeyens et al. 1990a; Martin & Levey 1987) have suggested that – unlike standard Pavlovian conditioning – this form of learning can proceed in the absence of awareness of the CS-US relationship. We briefly review some of the relevant evidence.

Baeyens et al. (1990a) presented subjects with 10 repetitions of a CS-US pair of slides, in which the CS slide had been previously evaluated by the subject as affectively neutral and the US slide as either liked, neutral, or disliked. Evaluative conditioning was observed in that on a postconditioning test of affect, the CS slides became affectively positive (liked) if they had been paired with a liked US, negative (disliked) if they had been paired with a disliked US, and they remained neutral if they had been paired with another neutral stimulus.

As a test of awareness, at the end of the learning phase Baeyens et al. showed the subjects each of the CS pictures and asked them to identify which had been the relevant US. If subjects failed to respond correctly they were then asked whether the US had been liked, neutral, or disliked. They were classified as "unaware" of the CS-US relationship if they failed on both of these questions. Evidence that evaluative conditioning occurred without awareness emerged in the observation that conditioning was the same for CS-US pairs, regardless of whether or not the subject was aware of the relationship.

Of course, the test of awareness may have been an insensitive one. Baeyens et al. accordingly tried to use a more sensitive concurrent measure of awareness. One group of subjects was required to indicate during the 4-sec interval between the onset of the CS and US slides whether they expected a liked, neutral, or disliked US stimulus on that trial. Subjects were classified as "unaware" if they failed to respond correctly on the final three pairings of each stimulus combination. Unfortunately, results from this group undermine the notion of unconscious learning. As discussed in section 2.3, subjects could accurately report most of the pairings, and for those few they could not report, there was no significant evaluative conditioning. Further, in another study, Baeyens et al. (1992) found that groups of subjects given increasing numbers of CS-US pairings showed an increase in both the magnitude of evaluative conditioning and the level of awareness as measured by a postconditioning test. In sum, these studies of evaluative conditioning have failed to show that it can occur unconsciously. (See Shanks & Dickinson 1990, for further criticisms of this research.)

Although they are not usually classified as studies of evaluative conditioning, Lewicki's (1986; Lewicki et al., 1989) experiments on the learning of nonsalient contingencies can be readily conceived as such. Lewicki presented subjects with photographs of people accompanied by personality descriptions such as "kind" or "capable." For some subjects all "kind" people had long hair and all "capable" people had short hair, while for other subjects the opposite was the case. Lewicki reported that on test trials in which subjects had to affirm or disconfirm statements classifying new people as either "kind" or "capable," they responded "yes" more often when the description preserved the study-phase correlation than when it broke the correlation. (They also consistently took longer to answer "yes" when the correlation was preserved.)

Lewicki's (1986) subjects were apparently unaware of the relationship between hair length and personality description, because "not one subject mentioned haircut or anything connected with hair" (p. 138) in a test of verbally reportable knowledge. If we take the personality description as being an evaluative response conditioned to the cue of hair length, the results would again appear to suggest unconscious evaluative learning. However, that conclusion requires us to assume, without any supportive evidence, that the Sensitivity Criterion has been met in these studies. In addition, some of Lewicki's results have proven hard to replicate (see de Houwer et al., in press; Dulany & Poldrack 1991); so we must at this stage reserve judgment on whether this form of learning indeed can occur unconsciously.

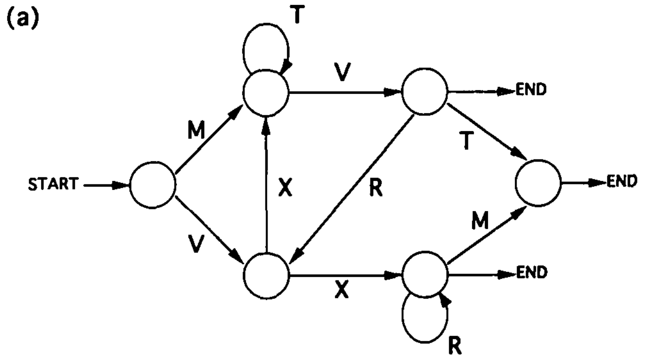
2.4.3. Conclusions. In experiments examining the relationship between learning and awareness in Pavlovian conditioning, researchers have striven to meet the Sensitivity Criterion by using multiple tests of awareness. The Information Criterion does not raise particular problems, because there is little doubt that the information the subjects learn (the contingency between CS and US) corresponds to what the awareness test asks them to report. Thus these studies provide a reasonably good test of the role of awareness in learning. The results we have surveyed give little reason to believe that unconscious learning can occur in these situations. For evaluative conditioning the evidence is less clear-cut, but we have few reservations in suggesting that unconscious evaluative learning has not yet been adequately established.

2.5. *Awareness in artificial grammar learning tasks*

Studies of subjects learning artificial grammars present the classic pattern of unconscious learning: subjects clearly learn something about the input domain, but they appear unable either to report the rules of the grammar or to explain their performance. Such studies provide evidence of unconscious learning if learning involves rule induction. In this section we examine the evidence for unconscious learning of artificial grammars and conclude that memorization rather than rule induction is the principal process involved; we conclude that evidence for unconscious learning is weak. Later, in section 3.5, we review several further studies that have examined conscious hypothesis testing in artificial grammar tasks.

In a prototypical experiment, Reber (1967) required subjects to memorize either a series of letter strings generated from a small finite-state grammar or a series of strings generated at random (see Fig. 2). Subjects who learned the rule-governed strings then performed a grammaticality test in which they were asked to accept novel strings that fit the rules and reject novel strings that did not. They categorized 79% of the 44 test strings correctly, which is significantly above chance. Yet these subjects were unable to report the rules they had apparently learned and then used in the grammaticality task.

Reber's (e.g., 1967; 1989a) account of such grammar-learning results, endorsed by many other investigators since then, proposed that subjects use an unconscious, or implicit, rule-induction mechanism. This mechanism creates a knowledge-base of rules that may be used in a grammaticality task but that is inaccessible to conscious report. As with the other unconscious learning paradigms, we believe that there is another way to interpret the data. We can raise two questions. The first (the Sensitivity Criterion) is whether retrospective verbal report is sufficiently sensitive to test for conscious knowledge of the rules. More sensitive measures of subjects' knowledge, such as concurrent thinking-aloud protocols and recognition tests might reveal marginal or uncertain knowledge. The second question (the Information Criterion) concerns what the subjects are learning from the training strings. If subjects have learned something other than rules, then asking them about rules may lead to erroneous conclusions. On the other hand, if we ask the subjects questions about what they did in fact learn, we may get reasonable answers. It may be that usable knowledge is always both consciously learned and consciously applied. The experimenter's job is to discern what information subjects are aware of during training and whether that information is used to perform the grammaticality task.



(b)

MVT

VXM

MTTV

MTTTVT

VXVRXR

Figure 2. String generator and example strings. (a) Diagram of a finite-state grammar. Strings are generated by selecting one of the possible routes through the network, commencing at "start" and continuing until one of the "end" symbols is reached. (b) Several example strings generated by the grammar.

2.5.1. Types of knowledge. The literature has identified three types of knowledge that might be acquired by subjects: rules, memory for whole strings, and knowledge of the frequency and position of substrings, such as pairs of letters. There are several problems with rules. First, it is not really clear what a "rule" would be like: Is it a rewrite rule or a transition graph? How complex can it be, and how many are there? Second, such rules would be very difficult for any but very sophisticated subjects to articulate even if they did explicitly acquire them. Third, it is not clear what sort of mechanism is capable of acquiring such rules, particularly since it must *ex hypothesi* operate outside consciousness. In the face of these questions, it seems sensible to consider other types of knowledge first, and to determine the extent to which they can account for subjects' performance. We return to the evidence for knowledge of rules in artificial grammar learning tasks in section 2.5.3.

The picture with regard to memory for whole strings and knowledge of substrings seems reasonably clear. Such knowledge is easy to articulate and there is ample evidence that subjects do acquire this information, because they do articulate it. These types of knowledge are also consistent with a variety of contemporary memory models, such as chunking (Servan-Schreiber & Anderson 1990), distributed memory (Cleeremans & McClelland 1991), and memory-array models (e.g., Estes 1986; Hintzman 1986; Nosofsky 1986). In addition, these models have been shown to approximate subjects' grammaticality test performance. For example, Dienes (1992) compared a number of these memory models on a set of grammaticality judgment data and was able to achieve good fits, particularly with distributed memory models. We return to this topic in section 3.3.

With these different knowledge types in mind, we can now ask what sort of information subjects in artificial grammar learning tasks actually acquire, and whether they are conscious of it. A number of studies have asked these questions using several methods and have asked them at various points during training and testing. Mathews et al. (1989) interrupted subjects periodically during training and asked them to instruct an imaginary confederate how to distinguish the grammatical strings. The trained subjects performed better on the grammaticality test than did the yoked subjects, suggesting that not all of the trained subjects' knowledge was explicit and reportable. This verbal report procedure, however, is essentially uncued recall, and so is unlikely to evoke all of the subjects' knowledge of the grammar. More interesting, though, is that the verbal instructions that subjects did report consisted mainly of legal bigrams and other short sequences, sometimes coded by their positions in legal strings.

In a study by Servan-Schreiber and Anderson (1990), subjects were trained on grammatical strings using a recall task. For training, the strings were divided into substrings using gaps (*T PPP TX VS*). Servan-Schreiber and Anderson hypothesized that subjects in all grammar-learning tasks encoded the strings into substring chunks, and the gaps were used to ensure consistent chunkings across subjects. Subjects' written recall preserved these gaps. Servan-Schreiber and Anderson suggested that this phenomenon demonstrates that subjects were in fact encoding the strings as sequences of short strings in accord with the gaps. The subjects' later grammaticality judgments supported this contention as follows. Servan-Schreiber and Anderson constructed ungrammatical strings that consisted of illegal sequences of legal substrings (e.g., *PPPTXTVS*). If subjects were learning just the substrings then these strings would be falsely accepted as legal strings. Indeed, 50% of these strings were mistakenly accepted. On the other hand, test strings that violated specific substrings were correctly rejected; only 26% of these strings were mistakenly accepted. Both subjects' written protocols during training and their test performance, then, support the hypothesis that subjects learn simple substring information in grammar-learning tasks. That only 50% rather than 100% of the strings containing illegal sequences of legal substrings were accepted does not imply that knowledge of substrings is insufficient to account for performance completely. Compared to grammatical strings, these nongrammatical strings (by definition) still contain illegal bigrams (e.g., *XT* in the example above). In addition, subjects' knowledge at test time is clearly incomplete: previously seen grammatical strings were only accepted 70% of the time.

Moreover, Servan-Schreiber and Anderson (1990) went on to build a model that acquired chunks and then used them to evaluate the grammaticality of test strings. The model performed at the level of trained subjects ($r = 0.935$). This result supports their claim that subjects are learning and using chunks by demonstrating that chunks are learnable and sufficient to account for the level of performance of subjects on the grammaticality task.

It is possible that Servan-Schreiber and Anderson's presentation technique, placing gaps in the training strings, biased subjects' learning procedure. A similar experiment by Perruchet and Pacteau (1990), however, used the standard (no gap) format during training and found similar results. Subjects were trained on strings generated from the same grammar that Reber and Allen (1978) used. To test for awareness of simple substrings, trained subjects performed a recognition test on letter pairs present in the training strings. Subjects performed quite well: only 3 out of 25 old pairs were judged less familiar than any new pair. The correlation between recognition scores and the frequency of occurrence of pairs in the training strings was 0.61. According to the results of the recognition test, then, subjects were aware of the relative frequencies of letter pairs. Similarly, Dulany et al. (1984) concluded that a recognition test of awareness could elicit as much knowledge as was projected in the grammaticality test.

Perruchet and Pacteau (1990) also constructed test strings that contained either (1) illegal orders of legal pairs, or (2) illegal pairs. If subjects only had information about legal pairs on which to judge the grammaticality of test strings, then the illegal pairs should have been rejected, but the illegal orders of legal pairs should have been mistakenly accepted as grammatical. This is the pattern of results Perruchet and Pacteau obtained. Discriminability, measured in D scores (zero indicates random responding), was 22 for illegal pairs but only 7 for illegal orders. These results therefore further support the hypothesis that subjects are aware of and make use of only simple substring information.

Perruchet and Pacteau then considered a model that used pair frequency information to make grammaticality judgments. The model produced the same level of performance as subjects, except in one instance. Subjects were sensitive to the beginnings and endings of strings, but the model was not. Perruchet and Pacteau concluded that subjects primarily knew letter pairs, but also which pairs could legally start and end strings. Together with the behavior of Servan-Schreiber and Anderson's (1990) chunking model, these results show that simple fragment-memorization systems can be sufficient to account for subjects' imperfect performance on grammaticality tests.

Dienes et al. (1991) also found evidence that subjects were sensitive to more than just pairs. Following training and a grammaticality task, subjects were given incomplete letter sequences varying in length from zero letters upwards (e.g., *VXT* . . .) and asked to judge which single-letter continuations (*M*? *V*? *X*? *R*? *T*?) were acceptable at the next location in the string. In this sequential letter dependencies (SLD) task, which was hypothesized to be sensitive to conscious knowledge of the grammar, subjects were sensitive to illegal orders of legal pairs even in the middle of strings. Dienes et al. showed that the knowledge that subjects demonstrated in the completion task correlated with their grammaticality judgments and could be used to model the grammaticality judgment data. They found in addition that knowledge gleaned from subjects' free reports also correlated with their grammaticality judgments, but that less knowledge was reported in the free report task than in the continuation task. These correlations suggest that a single knowledge source is tapped by both tasks, but that the free report task, uncued recall, is less sensitive.

Reber and Allen (1978) asked subjects to describe retrospectively their learning experience and, concurrently, to justify their grammaticality judgments. Overall, subjects justified their classifications on 821 out of 2,000 test strings. Subjects reported using a variety of information in making their grammaticality judgments. The violation or nonviolation of bigrams was the most common justification, especially concerning the first bigram of a string. String-initial bigrams accounted for fully 30% of the justifications. Violations of single letters, particularly the first or last letter of a string, and violations of trigram or longer sequences were also reported, as well as recognition of and similarity to whole training strings. The grammaticality responses to the remaining unjustified cases presumably consisted of guessing or of knowledge that could not be elicited by verbal report.

So much for substring knowledge. Vokey and Brooks (1992; Brooks & Vokey 1991) have argued that subjects can encode whole-item information in addition to substring information. They found that the similarity of test strings to specific whole-study strings is an important factor in subjects' grammaticality judgments. When the grammaticality and the similarity of the test strings were varied independently, they were shown to be additive factors on grammaticality judgments. Vokey and Brooks argued that such a result indicates that subjects have encoded the whole strings and can determine similarity relationships between strings.

Brooks and Vokey's evidence for whole-string information raises no particular problems for our interpretation of the artificial grammar learning data, since subjects are clearly aware of their whole-string knowledge just as they are aware of the substring knowledge; the study task, after all, requires the subjects specifically to memorize whole strings. However, as Brooks and Vokey (1991, p. 321) themselves concede, their results can at least in principle be explained without reference to whole-item knowledge. Just as grammatical test strings tend to contain more studied bigrams than nongrammatical strings (Perruchet & Pacteau 1990), so also a test string that is highly similar to a study string will contain more studied bigrams than one that is less similar. In fact, Vokey and Brooks' results have been challenged by Perruchet (1994), who has shown that both the effect of similarity and the apparently independent effect of grammaticality that Vokey and Brooks obtained can in turn be reduced to substring knowledge. Grammatical test strings tend to contain more substring components that were part of the training strings than do nongrammatical items. The same is true for similar and dissimilar test items, with similar items tending to contain more substring components from the study strings.

A final piece of evidence supports the view that grammaticality judgments are controlled by comparison to memorized substring or whole-item information. On such a view, but not on an abstraction account, it is likely that judgments would be relatively susceptible to changes in the superficial characteristics of the studied strings. To test this, Whittlesea and Dorken (1993) required subjects to pronounce the training strings from one grammar and to spell the training strings from another grammar. At test, subjects were asked either to pronounce or to spell test strings and to judge their grammatical status. Subjects were more likely to assign test strings to grammars when the encoding task matched the task for the test string than when they differed. Test strings that were equally similar to strings in both grammars were assigned to the grammar where the encoding and test tasks matched. Such results, although consistent with the idea that judgments are based on a comparison with a set of items in memory that represent the study items in a relatively unanalyzed form, would clearly not be anticipated if what was encoded were the underlying abstract rules of the grammar.

Our conclusion from this section, then, is that subjects use their memory system to acquire knowledge of (possibly) whole strings and (certainly) their parts, and that this simple information is conscious both during acquisition and testing. The results reported by Dulany et al. (1984), Perruchet and Pacteau (1990), and Dienes et al. (1991) show that the knowledge that subjects can consciously retrieve in a recognition test is sufficient to explain their grammaticality judgments. From the evidence we have considered, we do not need to assume the existence of an additional implicit knowledge base, and conclusions to the contrary have arisen because of failures to meet the Information Criterion.

Our interpretation rests on the results of a variety of tests of conscious knowledge that have attempted to address the Sensitivity Criterion. Dienes et al.'s (1991) SLD test, for instance, which required subjects to judge which continuations of a sequence of letters were legal, was actually found in a signal detection analysis to be *more* sensitive than the implicit grammaticality test itself. Thus, if such a test is accepted as a measure of explicit knowledge, no evidence of a dissociation between learning and awareness emerges. Of course, an alternative (see Reber et al. 1985; and the reply, Dulany et al. 1985) is to argue that performance on these explicit tests is contaminated by unconscious influences; subjects may choose a correct continuation on the SLD test as a result of some implicit knowledge to which they do not have conscious access.[2] The problem with this interpretation, however, is that it means we would have to abandon the test as an index of conscious learning and rely instead on verbal reports, in which case it is hard to see how the Sensitivity Criterion can ever be met. And if that criterion cannot be met, then how are defenders of unconscious learning ever going to unconfound test type from sensitivity, and hence establish the existence of unconscious learning?

We believe it is rather unlikely that unconscious influences play a significant role in the SLD test. Presenting subjects with a letter sequence (e.g., *VXT*. . .) and asking them to judge, under no time pressure, whether a given letter (e.g., *M*) could continue the sequence would seem to be a prototypical example of a task requiring conscious reflection, even if it involves mere conscious recollection of studied strings. Nevertheless, to claim that the SLD test is only sensitive to conscious information does require adopting what Reingold and Merikle (1988) call the "exclusiveness" assumption: the assumption that performance on a test of awareness is only affected by conscious influences. This, of course, is a very strong assumption and one that may well be incorrect.

2.5.2. Learning systems. In addition to the question of awareness, a second issue concerns whether whole item, bigram, and possibly rule information are acquired by a single learning system or by separate systems. If they are acquired by separate systems, perhaps those systems interfere with each other's operation? To examine this possibility, Reber and Allen (1978) manipulated the training task. Subjects either observed the strings without any explicit task (observation training), or they performed a paired-associate task, where each string was paired with a different city name. The idea was that the paired-associate task would require better item encoding, thereby facilitating item knowledge but potentially inhibiting other learning processes.

The paired-associate task produced several significant differences from the observation task. Overall, paired-associate subjects were less accurate on their grammaticality judgments: 72.4% versus 81.2% accurate for observation subjects. Paired-associate subjects produced twice as many recognition justifications as did observation subjects (77 vs. 40), and paired-associate subjects' probability of making consistent errors suggested they were more likely to develop unrepresentative knowledge than were observation subjects.

Clearly, the two training tasks affected the quantity of whole-item and substring knowledge that was acquired, but the underlying learning processes do not appear to be in opposition. The verbal reports show that both groups justified their responses with the same knowledge sources, but to differing degrees. It appears, then, that whole-item learning is compatible with substring learning. Vokey and Brooks (1992) examined a range of encoding tasks that produce differences in the extent of item knowledge, but they also found no reliable interference between item knowledge and substring knowledge.

Finally, Dienes et al. (1991) required subjects to generate random digits during training. Their goal was to test whether this task would interfere selectively with subjects given explicit instructions to search for rules that describe the study strings, but not with subjects given implicit instructions simply to observe the study strings. Instead, Dienes et al. found equivalent reductions in learning for both implicitly and explicitly instructed subjects.

2.5.3. Implicit rule induction. Although the considerable evidence presented above supports the conclusion that subjects' knowledge consists of simple substrings (or whole strings), there are two further pieces of evidence that support the conclusion that subjects learn rules. The first piece of evidence supporting rule learning was reported by Reber and Lewis (1977). Subjects were trained on a subset of strings and then solved "anagrams" based on the remaining strings generated from the grammar – that is to say, they took strings of letters and rearranged them to make grammatical strings. The frequencies of bigrams produced by subjects in the anagram task were tabulated and compared with the frequencies of the bigrams in the training set and in the full set of grammatical strings. If subjects were learning bigram frequencies from the training strings, the correlation between the frequencies of bigrams in the training strings and in the solved anagram strings should be high. While this was the case, Reber and Lewis found that the correlation between the frequencies of bigrams in the solved anagrams and in the whole grammar was actually higher. This result suggests that the subjects went beyond the training set to learn the rules of the grammar.

Perruchet et al. (1992), however, argued that Reber and Lewis's (1977) result must hold on statistical grounds alone. The anagrams demand the production of certain bigrams and not others, in fact, exactly those bigrams that are underrepresented in the training set. Suppose, for example, that *VT* is a bigram in the grammar that is underrepresented among the training strings. *VT* must then be overrepresented among the solved anagram strings since the training and correctly solved anagram strings together constitute the complete set of grammatical strings. It is no wonder, then, that the correlation between the frequencies of anagram bigrams and training bigrams is low and that the correlation between anagram bigrams and the full grammar bigrams is higher. Perruchet et al. went on to demonstrate this fact empirically by training subjects only on the individual bigrams from the training strings. Under these circumstances, subjects could not be learning rules because they only saw bigrams, yet as with Reber and Lewis's subjects the frequencies of their anagram bigrams also correlated better with the full grammar bigrams than with the training string bigrams. The original conclusion, therefore, that subjects go beyond the training strings to learn rules appears to have been an artifact of the experimental design.

The second and more compelling piece of evidence for abstraction is the fact that subjects show some degree of transfer to strings governed by the same underlying grammar, but formed from a new set of letters or from a completely new set of stimuli such as tones. Reber (1969) trained subjects to recall grammatical strings, and when he switched to a new set of letters, subjects showed no increase in recall errors. This result suggested that subjects had learned abstract rules that were easily instantiated with different letters. More impressively, Altmann et al. (in press) required subjects to observe a set of letter strings, generated from the grammar shown in Figure 2, prior to making grammaticality judgments concerning sequences of tones. Some of the tone sequences could be generated from the grammar by substituting a tone for a letter (e.g., middle *C* for the letter *M*). Altmann et al. found that exposure to letter sequences allowed grammatical and nongrammatical tone sequences to be discriminated at better-than-chance levels. Although the improvement was generally small (about 5% increase in correct classifications), this result strongly suggests that at least some aspects of the abstract structure of the letter sequences had been isolated and were available to aid classification of the tone sequences.

It is important to note that the change of stimulus set did have a detrimental effect on performance, however. Compared to a situation in which the study and test items were from the same set (both letters or both tones), classification performance was significantly impaired when the study and test sets differed. Thus, abstract knowledge was plainly not the sole source of information that subjects were relying on – specific memorized fragments or strings must also have been playing a role. A study by Mathews et al. (1989) confirms this conclusion. In Mathews et al.'s study, over a series of training sessions subjects were trained either on a single-string set or on different sets generated from the same underlying grammar. Subjects in the same set condition learned better, and a final switch to a new set doubled the error rates in the single-set training condition. Such a result would not be expected if an abstract set of underlying rules were the sole factor guiding classification, because the rules would apply equally to the new and to the original letter set.

What is the significance of these results for unconscious learning? To the extent that subjects might be poor at describing what they have abstracted, such results may imply that unconscious learning is taking place. But given the rather small improvement in classification performance that results from training and testing on different sets of items, it is quite likely that what is abstracted is fairly limited (e.g., only two initial symbols are legal, the first two symbols of a string cannot be the same, etc.), and it is quite possible that subjects, if asked, would be able to report such simple regularities. In sum, although the data from transfer studies do suggest that some aspects of the underlying structure can be abstracted, from the point of view of unconscious learning the significance of these findings has yet to be established.

2.5.4. Conclusions. These studies indicate that relatively simple information is to a large extent sufficient to account for subjects' behavior in artificial grammar learning tasks. In addition, and most important, this knowledge appears to be reportable by subjects. Appreciable knowledge of the grammar does not seem to be acquired by explicit hypothesis testing or other complex analytic processes (although we return in sect. 3.2 to consider some rather different cases where grammars appear to be learned explicitly). Instead, knowledge seems to be mainly accumulated over training by simple memory mechanisms that collect frequency statistics on bigrams, slightly longer sequences, and possibly whole items.

2.6. Awareness in instrumental learning tasks

In contrast to the conditioning and artificial grammar studies described above, which arrange relationships between external cues, instrumental tasks establish some contingency between an action the subject performs and an associated outcome. Learning is measured as a change across trials in the propensity to perform the action. Naturally, the question we may again ask is whether such learning can occur without awareness. As in his review of Pavlovian learning studies, Brewer (1974) concluded that the answer to this question is no. There have recently been some further investigations of the role of awareness in instrumental learning: we consider results separately from tasks in which the instrumental contingency is simple or more complex. By "simple" we mean any task in which there is ostensibly just one action available to the subject.

2.6.1. Simple instrumental learning tasks. Svartdal (1989; 1991) has reported a number of studies in which subjects are led to believe that there is a relationship between a reinforcer and one aspect of responding, when in fact the critical variable is some other aspect of responding. For example, Svartdal (1991) presented subjects with brief trains of between 4 and 17 auditory "clicks." Subjects immediately had to press a response button exactly the same number of times and were instructed that feedback would be presented when the number of presses matched the number of clicks. In reality, however, feedback was contingent on the rate of responding: for some subjects, it was given when the interresponse times (IRTs) were lower than in a baseline phase, while for others it was given when IRTs were higher.

Svartdal (1991) obtained evidence of learning, in that IRTs adjusted appropriately to the reinforcement contingencies, but subjects seemed to be unaware that it was the rate of responding that was important. A structured questionnaire revealed no evidence of awareness of the contingency between response rate and feedback in subjects whose response rate had adjusted appropriately.

Such demonstrations appear at first glance to be quite compelling, especially as the contingency to be learned is such a simple one. It is unclear, however, that the Information Criterion is met in these and similar studies, because it is very difficult to rule out the possibility that subjects acquire "correlated" hypotheses about the reinforcement contingency that are incorrect from the experimenter's point of view but happen to produce response profiles that are difficult to distinguish from those generated by the correct hypothesis. For example, suppose subjects learn that resting their hand in a certain position increases reinforcement rate. This could be a true experienced contingency if that hand position was conducive to a fast or slow response rate. Such an "incorrect" hypothesis would generate behavior that was very similar to what would be produced by the correct hypothesis, yet a subject who reported hand position as the crucial variable would be regarded by the experimenter as "unaware" of the reinforcement contingency.

Although such a criticism is undoubtedly *post hoc*, there is good evidence of subjects' behavior being under the control of such correlated hypotheses. In the 1950s, a number of studies asked subjects to generate words *ad libitum* and established that the probability with which they would produce, say, plural nouns was increased if each such word was followed by the experimenter saying "umhmm" (e.g., Greenspoon 1955); as with Svartdal's (1991) experiment, this result occurred in subjects apparently unable to report the reinforcement contingency. However, in an elegant study, Dulany (1961) proved that subjects were hypothesizing that reinforcement was contingent on generating a word in the same semantic category as the previous one. Although incorrect, this hypothesis was correlated with the true one, in that if the subject said "emeralds" and was reinforced, then staying in the same semantic category meant they were more likely to produce another plural noun ("rubies") than if they shifted categories. Thus the subjects were perfectly aware of the contingency that was controlling their behavior, namely, the contingency between staying in the same semantic category and reinforcement.

In sum, even ignoring possible insensitivity in the test of verbal awareness, results such as Svartdal's (1991) cannot be taken as conclusive evidence of unconscious learning. Subjects may learn a rather different contingency from that explicitly programmed by the experimenter, and the Information Criterion may hence fail to be met. The problem is particularly worrisome in operant studies because, by definition, the experimenter has little control over the subject's behavior and therefore over the contingencies that may be present. In nonoperant tasks, the problem can be avoided because the experimenter can in principle eliminate all reinforcement contingencies except the one of interest. For this reason it seems that clear evidence for unconscious learning is likely to be difficult to establish in instrumental learning tasks.

In contrast to such apparent dissociations between learning and awareness, Shanks and Dickinson (1991) have argued that there are a number of variables that seem to have rather similar effects on performance assessments of learning and on awareness. In two studies, subjects performed a simple operant-learning task in which pressing a key on a computer keyboard was related, via a schedule of reinforcement, to a triangle flashing on the screen. Subjects were exposed to a reinforcement contingency in which they scored points whenever the triangle flashed but lost points for each response, so that they were encouraged to adapt their response rate to the reinforcement schedule. Learning was demonstrated by changes in subjects' rates of responding. As a measure of awareness, subjects were asked to report on a scale from 0 to 100 what they thought was the relationship between the response and the reinforcer.

Shanks and Dickinson (1991) found that response rate was sensitive both to the degree of contiguity between the response and reinforcer and to the degree of contingency between them. At the same time, subjects' judgments were equally sensitive to these factors. Furthermore, certain judgmental illusions likewise manifested themselves in performance measures. For example, subjects frequently judge an action and an outcome to be related when in fact they are not. Shanks and Dickinson found that this effect appears in performance measures such as response rate as well as in verbal judgments. Of course, the appearance of a bias in two behavioral measures strongly suggests that they are mediated by a common underlying process.

The notion that learning and awareness proceed in tandem is corroborated to the extent that they are affected in similar ways by various manipulations. Shanks and Dickinson's results indicate that – at least for the two important factors of contingency and contiguity – this is exactly the case. Shanks (1993) discusses some further apparent concordances.

The human operant-learning literature provides perhaps the most convincing evidence that learning and awareness are *associated* in simple learning tasks. A wealth of data shows concordances between response rate and verbal reports under different schedules of reinforcement (e.g., Catania et al. 1989; Rosenfarb et al. 1992; see also Skinner 1984b, and accompanying commentaries). For instance, Rosenfarb et al. required subjects to press a button either on a differential-reinforcement-of-low-rate schedule, in which reinforcers were delivered for a response provided that 5 sec had elapsed since the preceding response, or on a fixed-ratio schedule in which eight responses were required to earn a reinforcer. Rosenfarb et al. found that subjects' verbal reports concerning the programmed contingency accorded very well with the actual contingencies. Furthermore, there was a strong correlation between the time at which responding became appropriate for a schedule and the time at which verbal reports indicated awareness of the reinforcement contingency operating in that schedule.

2.6.2. Complex instrumental control tasks. Several experiments have investigated the relationship between learning and awareness in more complex instrumental learning tasks where the subject has to learn to control an interactive system. Again, the basic idea is as shown in Figure 1, with some learning episode followed by an assessment of awareness. In most of these tasks awareness at time t_2 is measured by verbally questioning the subject.

Berry and Broadbent (1984) conducted an influential and widely cited experiment in which there was an apparent dissociation between learning and awareness. As in Hayes and Broadbent's (1988) study, one of the tasks they used required subjects to interact with a computer "person." On each trial, the subject entered an attitude (e.g., polite) to the computer, which then responded with its attitude (e.g., unfriendly). The subject's task was to try to get the computer to be friendly. The computer's attitude on each trial was a simple numerical function of the subject's input on that trial and the computer's previous attitude. Inclusion of the computer's attitude on the previous trial makes the task quite a difficult one to learn.

Berry and Broadbent (1984, Experiment 1) found, not surprisingly, that performance improved with practice: significantly more trials on target occurred during a second block of 30 trials than during the first block. However, scores on a structured questionnaire designed to assess the subjects' reportable knowledge of the task were no better after the second block than after the first one. Hence, here we have apparent evidence that learning to perform a task can take place without any change in awareness of the underlying structure of the task. Similar results have been obtained in a number of other studies (e.g., Berry & Broadbent 1987; 1990; Broadbent et al. 1986; Hayes & Broadbent 1988; Stanley et al. 1989).

On the other hand, a detailed examination by Sanderson (1989) found evidence of associations rather than dissociations between performance and reports. Sanderson argued that because subjects often have complex prior beliefs about the interactions within a large system, and because these beliefs may be erroneous in a laboratory version of the system, it is possible for their mental models to undergo considerable revision without yielding an overall improvement in accuracy. It is only with prolonged practice that mental models, and hence the verbal reports based on them, begin to show noticeable improvement. Consistent with this, Sanderson was able to obtain significant performance improvements at the same time as weak improvements in the overall accuracy of verbal reports in a complex transportation task, but she showed that the detailed nature of the verbal reports was changing very considerably.

A further experiment by Berry and Broadbent (1984) found the converse of the previous dissociation, namely, reportable knowledge improving without corresponding improvements in task performance. One group simply completed two sets of trials, while between the two sets another group received detailed verbal instructions about the nature of the input-output relationship. These instructions essentially represented a verbal description of the equation governing the computer's attitude. When questioned at the end of the experiment, subjects who had received instructions outscored those who had not, yet the groups were indistinguishable in terms of number of trials on target. Thus a change in "awareness" (or at least a change in reportable knowledge of the task) was not accompanied by a change in task performance.

What are we to make of such dissociations? One possibility is that it is not only possible for learning to proceed without awareness, but in addition the system responsible for implicit learning is quite independent of another (explicit) system in which learning is accompanied by awareness. Such a "systems" account would then be able to explain why we can obtain double dissociations of the sort reported by Berry and Broadbent (1984): learning to perform the control task involves the implicit system, and proceeds without awareness, while a change in awareness involves the explicit system and can proceed without any benefit in task performance.

While double dissociation results are certainly consistent with the notion that there are two learning systems, one conscious and the other unconscious, we feel that an alternative account is equally feasible: there may be two systems, both of which are conscious, but which encode different types of knowledge. The basic problem is that we do not know that the sort of knowledge the subjects in Berry and Broadbent's experiments acquire when learning to perform the task is at all the same as the knowledge they require to score well on the test of reportable knowledge (i.e., the results may fail to meet the Information Criterion). Suppose, for the sake of argument, that good task performance simply depends on learning an unrelated set of stimulus-response (S-R) pairs or instances (evidence for such a possibility certainly exists; see Cleeremans 1993b). It is then not hard to imagine that although practice provides the subjects with more and more knowledge of this sort, they might be hard pressed to use such knowledge when faced with questions about possible structural rules underlying the task. At the same time, giving the subjects detailed instructions about the task may improve their knowledge of the rules, and hence their questionnaire scores, but might not transfer to better performance on the task itself since S-R knowledge is required for that. But of course the subjects' inability to describe the rules underlying the task does not imply that the S-R learning occurred without awareness: if they were asked to report *that* knowledge, perhaps subjects would be able to do so. In sum, there are ways of interpreting such data that do not appeal to unconscious learning (see Stanley et al. 1989, for an examination of some of the alternative types of knowledge that subjects may encode).

A second problem concerns the sensitivity of the test of awareness. Can we be certain that the questionnaire procedure exhausts the subject's knowledge of the task? Can we be confident that failure to express on the questionnaire any awareness of the nature of the task means that the subjects were unaware at the time they were learning? For example, one alternative strategy would be to ask each subject to instruct a yoked "partner" in how to perform the task. If the partners could then perform the task as well as the original subjects, we would conclude that the original subjects were in fact able to articulate all their task knowledge. Such procedures have been used with other learning procedures (e.g., Mathews et al. 1989, for grammar learning) and have proven highly sensitive.

2.6.3. Conclusions. Instrumental learning experiments arrange some relationship between the subject's actions and certain outcomes. Implicit learning would be demonstrated if learning, as indexed by changes in instrumental behavior, occurred in the absence of awareness of the reinforcement contingencies. Although some studies have found that subjects seem unaware of the relevant contingencies, reliance on verbal report means that the Sensitivity Criterion is unlikely to have been met. Furthermore, because the experimenter necessarily relinquishes a certain degree of experimental control in an instrumental learning task, it is difficult to rule out the possibility that the subject is responding on the basis of a correlated hypothesis, in which case the Information Criterion is violated. Finally, even ignoring these considerations, a surprisingly large number of studies have documented impressive concordances between behavior and awareness.

2.7. Learning and awareness in serial reaction time tasks

Nissen and Bullemer (1987) and Lewicki et al. (1987) introduced a simple and ingenious technique, the serial RT task, in an attempt to demonstrate unconscious learning. In Nissen and Bullemer's version, a stimulus is presented on each trial in one of four locations (A–D) and the subject simply has to press the button corresponding to that location as fast as possible. The subject is given instructions appropriate for a typical choice RT task, but in fact there is a sequence underlying the selection of the stimulus on each trial. The question is, can subjects learn the sequence without being aware of it? With respect to Figure 1, the subject is presented with a series of learning trials in which there are predictive relationships between stimuli. These are accompanied by both a concurrent assessment of learning (RT) and a later assessment of awareness.

Some of the most compelling evidence for unconscious learning using this technique comes from a later study by Willingham et al. (1989); this study is worth considering in some detail because of the heavy reliance placed upon it in recent discussions of conscious and unconscious processing (e.g., Velmans 1991). In their first experiment, Willingham et al.'s subjects performed a 4-choice RT task. The actual sequence of signals was DBCACBDCBA, which repeated many times with no break between cycles. Subjects' RT improved across a total of 400 trials. To see whether this improvement represented knowledge of the sequence or general nonspecific speedup, Willingham et al. compared the speedup of their subjects to that obtained in a group of subjects from the earlier study by Nissen and Bullemer (1987) for whom there had been no structured sequence; for these control subjects, target location was random from trial to trial, except that the same location never occurred on consecutive trials.

The improvement in RT was significantly greater in the sequence group than in the control group, apparently indicating that sequence learning had occurred. Furthermore, this was still true for subjects who subsequently reported no awareness of the existence of a sequence during the RT trials.

2.7.1. Problem of suitable control group. Although such results suggest the possibility of unconscious learning, there are a number of significant problems with such experiments. First, the demonstration of sequence learning has typically involved one of the following two comparisons: (1) a comparison (e.g., Willingham et al. 1989) between a group exposed to the sequence and one for whom the stimulus on each trial is chosen at random (with the constraint that stimuli never repeat on consecutive trials), or (2) a within-subjects comparison (e.g., Hartman et al. 1989) between performance at the end of a long period of exposure to the sequence and performance on a subsequent block of trials where the stimuli are chosen at random, again with the constraint that stimuli never repeat on consecutive trials. The problem with both of these comparisons is that performance can differ between the sequence and random trials without the subject having any knowledge – implicit or otherwise – of the sequence.

As a moment's reflection reveals, faster responses on the DBCACBDCBA sequence compared with a random sequence might simply be due to response biases developing during exposure to the sequence. The stimuli are not equally frequent (B and C occur three times, D and A twice) in the 10-trial sequence. Thus in the sequence, but not in the random conditions, the subject is to some degree able to predict which stimuli are most probable, a fact that – as has been demonstrated extensively (see Broadbent 1971) – will allow fast responses to develop.

Clearly, the appropriate comparison is with a group of subjects who receive a "pseudorandom" series constrained to have the same number of each of stimuli A, B, C, and D per 10 trials as appear in the sequence proper, and in which stimuli never repeat on consecutive trials. Such an experiment was reported by Shanks et al. (1994). One group of subjects was presented with the normal sequence, another with the pseudorandom series, and a third with a "truly random" sequence, in which again there was the constraint that stimuli never repeated on consecutive trials. The stimuli were dots arranged in a horizontal row and the general procedure followed that of Willingham et al. (1989).

After 400 RT trials, subjects in the sequence group were classified on the basis of a structured interview as having no knowledge of the sequence, some knowledge, or full knowledge. The prediction was that if the no-knowledge subjects had indeed learned something about the sequence, they should have speeded up more than the pseudorandom subjects. In all but the truly random group the RT difference between the first and fourth block of 100 trials was significantly greater than zero. The normal-sequence/full-knowledge group speeded up more than any of the others; the difference between the normal-sequence/full-knowledge and pseudorandom groups confirms that the normal-sequence/full knowledge subjects had indeed learned something about the sequence. However, there was no significant difference between the normal sequence/no-knowledge and pseudorandom groups, though both speeded up more than subjects exposed to the truly random series. Thus, we suggest that with Willingham et al.'s stimuli and procedure, most if not all of the supposedly implicit learning in the normal-sequence/no-knowledge group was simply due to the development of response biases reflecting knowledge of the frequencies of the different stimuli.

As a consequence, Willingham et al.'s experiment fails to satisfy the Information Criterion. The subjects may have been unable to articulate information about the sequence verbally simply because they were not learning (in any sense) about the sequence. Instead, they were learning about the frequencies with which the different stimuli occurred. This is information they may, if asked, have been able to report.

2.7.2. Prediction tests as measures of awareness. The second problem is that, even ignoring the above considerations, we cannot rely just on the subjects' informal reports as assessments of their state of awareness some seconds or even minutes previously. Two somewhat different strategies have been advocated with regard to using more sensitive tests of awareness, namely, recognition and prediction tests. We discuss recognition tests in the next section. Prediction tests, introduced by Nissen and Bullemer (1987), require the subject to try to predict the next element of the sequence; such a test was used by Willingham et al. (1989) in addition to their verbal report test. After the RT phase of their experiment, Willingham et al. instructed subjects to try to predict on each trial where the stimulus would next appear, with no requirement for rapid responses. Subjects simply chose response keys on each trial until they picked the correct one, at which point they would then try to predict the next stimulus. Across many blocks of this prediction task, the subject again has the opportunity to learn the sequence. Evidence for explicit knowledge of the rule appears as savings (compared with the control group) in the number of trials required to learn the sequence in the prediction task.[3]

The rationale behind the prediction task is that if subjects are instructed to try to predict events and are able to do so with above-chance accuracy, this is evidence of conscious knowledge, because their predictions must be based on conscious expectancies. As this task requires the subject to act on a conscious expectancy concerning which stimulus will appear next, it is apparently a test of awareness of elements of the sequence. This contrasts with the RT task, in which they have to respond as fast as possible to the current target. The prediction task is a good one in that it is irrelevant whether or not the subjects believe that their performance in the RT phase was being affected by the sequence (indeed, they may not even be able consciously to report having detected a sequence). All that matters is whether any evidence of savings emerges in the prediction task, for, according to the reasoning behind the task, such savings must be due to conscious information about the sequence.

More to the point, failure at the prediction task would demonstrate a subject's inability to draw consciously on information about the sequence, thereby supporting the contention that the information indeed is implicit. It is important to drawing such a conclusion that the prediction task satisfies the Sensitivity Criterion where verbal reports did not. The retrieval cues for the prediction task are virtually identical to those of the reaction time task. Hence we now have two tests that are almost identical, but in one the subject's performance (i.e., RT) is measured and in the other awareness is assessed. This very much follows the rationale of recent experiments on unconscious perception (e.g., Merikle & Reingold 1992), where the test of awareness and the test of perception are designed to differ in little more than the instructions given to subjects. Although the temporal arrangement of stimuli and responses is different in the two tasks, and the response metrics are quite different, the prediction task nonetheless represents an interesting new procedure for assessing awareness.

What are the results obtained from studies using the prediction task? Willingham et al. (1989, Experiment 1) discovered that subjects they had classified as unaware on the basis of their verbal reports not only speeded up in the RT phase but also, according to Willingham et al., showed no evidence of awareness as assessed by the prediction task. Such a result appears to provide quite compelling evidence of implicit learning, even if Shanks et al.'s (1994) data suggest that learning probably involved frequency rather than sequence information. It is important to note that this dissociation of RT speedup and prediction performance only applies to subjects who have been selected as unaware on the basis of their verbal reports. Across all subjects (regardless of their verbally reported awareness), RT speedup and prediction performance tend to be closely associated, as experiments by Cleeremans and McClelland (1991) and Perruchet and Amorim (1992) have shown.

Willingham et al. compared the performance of their normal sequence/no-knowledge subjects in the prediction task with that of a "no-training" group who had not received the RT phase at all. This comparison was the critical evidence that the sequence learning in the normal-sequence/no-knowledge subjects was implicit. However, there are three problems with these results. First, although Willingham et al. claimed that there was no evidence of savings on the prediction task in their "no-explicit-knowledge" subjects, close inspection of their data reveals that these subjects did perform at a better level than naive subjects, though not significantly so. Over each of the first six sets of 10 trials of the prediction task, performance was better in the normal-sequence/no-explicit-knowledge group than in the control group by about 5% in each set (Willingham et al. 1989, Fig. 3). On the first block of trials, the normal-sequence/no-explicit-knowledge group scored 42.6% correct and the control group 38.7%. Although small, this trend is as much evidence for savings as it is for a dissociation between awareness and learning. A similar conclusion may be drawn for the data reported by Hartman et al. (1989), where small but consistent savings are also apparent.

Second, Perruchet and Amorim (1992) pointed out that Willingham et al. did not instruct their subjects that the stimulus sequence in the prediction phase would be the same as that in the RT phase. Hence subjects may not have been maximally motivated to show transfer savings in the prediction task. The third and final problem is that Shanks et al. (1994), in their replication and extension of Willingham et al.'s study, obtained savings that were of a statistically significant magnitude. Shanks et al.'s normal-sequence/no-knowledge subjects performed much better (mean 5.7 correct predictions) than the no-training control subjects (mean 2.7) across the first 10 trials of the prediction phase, indicating that at least some of the knowledge they had acquired in the RT phase but were unable to report verbally was available for transfer to the prediction task. In sum, we conclude that the Willingham et al. study has failed to establish unconscious sequence learning.

A number of other studies have also used the sequence-learning task. Several of these have adopted Willingham et al.'s procedure of classifying some subjects as unaware on the basis of their verbal reports and then examining their prediction task performance. Others have sought to obtain different dissociations between RT speedup and prediction performance. Whatever the strategy used, we suggest that claims for implicit learning in these studies (e.g., Cohen et al. 1990; Hartman et al. 1989; Howard & Howard 1989; Knopman 1991; Lewicki et al. 1987; Lewicki et al. 1988; Nissen & Bullemer 1987; Nissen et al. 1987; Stadler 1989) are difficult to interpret for many of the reasons we have raised concerning Willingham et al.'s experiment. These other studies either (1) fail to show that subjects have acquired any *sequence* knowledge in the RT phase, (2) show small but consistent trends toward savings in the prediction task in supposedly unaware subjects, (3) present control subjects in the prediction task with random rather than pseudorandom sequences, or (4) do not provide feedback in the prediction test and hence run the risk of inducing forgetting of the sequence, which will lead to an underestimation of conscious knowledge. Caution suggests that these studies do not warrant the conclusion of reliable sequence learning in the absence of awareness.

Rather than reviewing all of these studies, we consider two widely cited ones (Lewicki et al. 1988; Stadler 1989) that illustrate some of the problems. Lewicki et al. (1988) presented subjects with blocks of trials that were arranged into sequences of five trials. On each trial a target appeared in one of the four quadrants of the computer screen and the subject had to respond by pressing the key appropriate to that quadrant. RTs were collected from a total of 4,080 experimental trials experienced by each subject. On the first two trials target location was random, except that the target was never displayed twice in the same place. Target location on trial 3 was determined by what had happened on trials 1 and 2. If the movement on the first two trials had been horizontal, then the movement from trial 2 to trial 3 was vertical; if it was vertical, then the next was diagonal; and if it was diagonal, the next was horizontal.

Similarly, target location on trial 4 depended on target locations on trials 2 and 3, and target location on trial 5 depended on its locations on trials 3 and 4. The net effect was that target location on trials 3, 4, and 5 was entirely predictable from the underlying rules, but locations on trials 1 and 2 were random. Hence if the subjects were indeed learning something about the rules, this should have manifested itself in a significantly greater reduction in RTs across blocks on trials 3, 4, and 5 than on trials 1 and 2, and this is exactly what Lewicki et al. found (in fact, they took as their dependent measure the number of correct responses with latencies less than 400 msec). Also, when the rules were changed toward the end of training, reaction times increased on trials 3, 4, and 5 but not on trials 1 and 2.

Lewicki et al.'s (1988) subjects could apparently report next to nothing about the rules determining target location. "None of the subjects mentioned anything even close to the manipulated pattern of exposures" (p. 33), although eight of the nine subjects did seem to be aware that their performance had dropped when the rules were changed. Lewicki et al. concluded that the subjects had learned the rules determining target locations on trials 3–5 implicitly or unconsciously.

Perruchet et al. (1990), however, disputed Lewicki et al.'s (1988) conclusions. The criticism is that the set of possible events that could occur on trials 3–5 was more constrained than the set of events that could occur on trials 1 and 2. By analyzing the rules that determined the permissible transitions from one trial to another, Perruchet et al. were able to show in their replication of Lewicki et al.'s experiment that speedup in RT on trials 3–5 relative to trials 1 and 2 was mainly due to relative speedup only on trials 4 and 5; furthermore, it was almost entirely due to two factors. First, on trials 1 and 2, but not trials 3–5, there were some occasions when the stimulus moved back to a location from which it had just come; these backwards movements led to a slowing of RTs simply because they increased the unpredictability of the movement. Second, on trials 1 and 2 there were infrequent horizontal movements, which again led to a slowing of RTs. On trials 4 and 5 horizontal movements were not permissible. Rather than learn rules such as, "If the movement from trial 1 to trial 2 was horizontal, then the next movement will be vertical," subjects need only have learned that the possible transitions had widely different overall probabilities. Low-probability transitions tended to occur on trials 1 and 2 and hence led to slower RTs.

Clearly, Lewicki et al.'s data fail to show that the knowledge that subjects could articulate in their verbal reports was in any way poorer than the knowledge that underlay their RT speedup. This fails to meet the Information Criterion and hence Lewicki et al.'s belief that knowledge of the composition rules was necessary for RT improvement is almost certainly not correct. Instead, relative RT speedup is simply due to subjects' learning that certain movements of the target occurred with low probability. As Perruchet et al. say, "The fact that subjects do not articulate any of the composition rules no longer applies if improvement in performance turns out to be unrelated to this kind of knowledge" (1990, p. 497). Furthermore, "Subjects' reports on the frequency of occurrence of particular target transitions would have been rejected as irrelevant to the actual manipulation" (p. 512).

Perruchet et al. did not assess their subjects verbally, but on the basis of Lewicki et al.'s results we may assume that on such a test they would also be classified as unaware. However, a verbal report test assessing experimenter-defined sequence knowledge may fail the Sensitivity Criterion. Perruchet et al. therefore used a prediction test in their study to see whether the subjects were indeed unaware. Toward the end of the experiment, they required one group of subjects to predict where they thought the target was due to appear. These predictions only had to be made intermittently, because there were between 4 and 11 normal RT trials between successive prediction trials. On such a trial, a question mark appeared in the center of the screen and subjects pressed the button matching the location of the target that seemed most probable.

Perruchet et al. found that subjects performed at significantly better than chance levels on the prediction trials. On trials 3, 4, and 5 they averaged 55.6% correct predictions, against a chance value of 33.3% (which assumes that subjects have learned – explicitly – that the target never appears in the same location on consecutive trials). On trials 1 and 2, only 29.7% of predictions were correct. If we take prediction responses to be conscious "reports" by the subjects of their expectancies about target location, then these results contradict Lewicki et al.'s (1988) claim that their study supported a dissociation between implicit and explicit knowledge. On the contrary, subjects seemed just as able to "report" target location as they were to show selective RT improvements.

We have already described Lewicki et al.'s (1987) experiments in section 2.3.2 and the ways in which they failed to meet either the Information or the Sensitivity Criterion. Briefly, the verbal report is likely to be an insensitive measure, and subjects were likely to be using a simpler form of knowledge to perform the task than Lewicki et al. considered. Stadler's (1989) replication is worth considering, though, because it made a concerted attempt to meet these criteria. Each trial consisted of the presentation of a target item in one of the four quadrants of a computer screen, designated A–D. The choice of target location on each trial was nonrandom, and the question was whether the subjects would be able to detect this nonrandomness. Subjects were presented with sequences of seven trials, with rules constructed so that target location on the seventh trial could be predicted from its location on trials 1, 3, 4, and 6. On each of the first six trials, the digit 6 appeared on its own in one of the quadrants of the screen, but on trial 7 (the "complex" trial), it was embedded in a display containing 36 digits. Reaction time on the seventh trial was the measure of interest. Again, the rules specifying target location were deterministic: thus, if the target appeared in locations C, A, D, and B on trials 1, 3, 4, and 6 respectively, then on trial 7 the target would be in location A.

Stadler (1989) repeated Lewicki et al.'s (1987) finding that RTs on the target trials decreased significantly across a large number of complex trials. To assess awareness, Stadler used a prediction task in which subjects were free to use whatever knowledge they had acquired, be it of the rules or of fragments of sequences, to predict target location. The task thus meets the Information Criterion, and because it reinstates much of the learning context, should also go a good deal of the way toward establishing equal sensitivity to conscious information in the awareness and performance tasks. In the prediction task, Stadler presented each of the four subjects with 48 sequences of seven trials, with target location on the seventh (complex) trial being determined by its location on the first, third, fourth, and fifth of the simple trials. On the complex trial, instead of presenting the target and distractors and requiring the subjects to locate the distractor as quickly as possible, Stadler presented a question mark in each possible location, and the subjects had to guess in which quadrant the target would appear. No feedback was given.

Stadler found that the four subjects made correct predictions on, respectively, 13, 11, 11, and 11 of the 48 sequences, where chance performance was 12/48. Clearly, there is no hint here of transfer to the prediction task. But in the absence of feedback on the prediction sequences, it seems to us that there is a very substantial likelihood that over the 48 test sequences, subjects would have forgotten a large part of whatever knowledge underlay their RT performance. Without feedback, 48 sequences (including 288 simple trials) represents a vast amount of interfering information. To be fair, Stadler could not, with his design, have given feedback, because better-than-chance performance would have been ascribable equally to savings from the RT stage and to new learning of the sequences. For this reason, designs in which relative savings in the unaware group are compared to those in a novel control group are much to be preferred; they allow a sensitive savings test to be administered without the problem of the forgetting of the sequences. Alternatively, Stadler could have interspersed his prediction trials throughout further RT trials in order to offset any forgetting on the former. At any rate, acceptance of the null hypothesis (of no transfer to the prediction trials) would be better warranted if we knew that there was no hint of transfer on the first few trials, but Stadler did not present these data.

In the meantime, we suggest that cautious readers should not interpret these results, or any of the sequence learning results we have reviewed, as convincing evidence for learning without awareness.

2.7.3. Recognition tests. As an alternative to prediction tests, several researchers have argued that an appropriate and sensitive test of the subject's explicit knowledge – particularly if that knowledge is fragmentary – is to use a recognition memory test. Using the Nissen and Bullemer (1987) task, Perruchet and Amorim (1992), for instance, presented subjects in the test phase with 4-trial sequences such as DBCA, which had either been part of the training sequence or not, and required them to respond exactly as they had in the study phase prior to making recognition decisions about the sequences. The results indicated that the old and new sequences could be discriminated, and furthermore, the recognition scores correlated extremely highly ($r = 0.821$) with RTs. On this basis, there is little evidence that sequence learning was unconscious. However, Perruchet and Amorim did not divide their subjects into aware and unaware groups, and it remains possible that their subjects included an unaware subgroup in whom recognition performance was at chance.

More interesting are data reported by Willingham et al. (1993). Using the standard 4-stimulus task, one-half of the subjects saw a random sequence and one-half saw a 16-trial repeating sequence, equated for overall stimulus frequencies. After the learning phase all subjects were given a detailed verbal questionnaire consisting of five questions and were also given a recognition test, in which five 16-trial sequences were shown, and the subjects had to rate how likely it was that each sequence had been the one used in the RT phase. The mean rating for the distractor sequences was subtracted from that given to the target sequence to yield a recognition measure.

Consistent with the idea that sequence learning was unconscious, Willingham et al. (1993) found only small and nonreliable correlations between the awareness measures (the verbal and recognition measures) and RT speedup in the study phase. This result also confirms, of course, that the recognition measure indexes something different from the performance measure. However, when Willingham et al. examined their individual subjects' data, they found that only two out of 45 sequence subjects could be classified as genuinely unaware of the sequence: the remainder all scored better than the median of the random subjects on one or more of the awareness measures. Thus it is only for these two subjects that we have any evidence of unawareness. When the RTs of these subjects were examined, they did indeed improve significantly more across trials than the random subjects, but interpretation is complicated by the fact that they started with abnormally slow RTs and ended the training phase with RTs similar to the random subjects. Finally, as Perruchet and Gallego (1993) point out, in a sample of 45 subjects, two of them might be erroneously misclassified as unaware simply because of unreliability in the awareness test. Obviously, a much larger sample is needed before strong conclusions can be drawn.

2.7.4. What is learned in sequence-learning experiments? Perhaps it should come as little surprise that, in general, RT performance and awareness tend to correlate (see Perruchet & Amorim 1992). Indeed, it is well established that in choice RT tasks, when subjects make a correct prediction about which stimulus will appear on the next trial, their RT on that trial will be much faster than if they had made an incorrect prediction (e.g., Simon & Craft 1989). But what exactly is the nature of the knowledge that subjects acquire in a sequence-learning experiment? In attempting to answer this question, perhaps we can better understand why there is such a temptation to regard the learning as unconscious.

In the studies we have described, the experimenter arranges that the location of the target is governed by a complex rule or set of rules. For example, in Lewicki et al.'s (1987) and Stadler's (1989) experiments, one of the rules says that if the target appeared in locations C, A, D, and B on trials 1, 3, 4, and 6, respectively, then on trial 7 the target will be in location A. In some studies (e.g., Lewicki et al. 1988) it appears that the experimenter is assuming that significant learning in the RT stage must occur because subjects learn those rules in their entirety. But the Information Criterion cautions us to examine closely whether RT speedup can be due solely to learning of the entire rule: Might performance not simply be due to learning of more fragmentary information? In the Lewicki et al. (1987) study, might RT improvement not arise just from the subjects' learning contingencies between, say, target locations on trials 4 and 6 and target location on trial 7?

The best evidence to date suggests that learning fragments of the training sequences is probably sufficient to explain the available sequence-learning data (see Perruchet, in press). Cleeremans and McClelland (1991, Experiment 1) were able to compare RTs to targets that could only be predicted by knowing the previous three elements of the sequence. They found a reliable difference between RTs to targets that conformed to the rules compared to those that did not, indicating that subjects could indeed maintain three items of temporal context. However, no evidence emerged that they could maintain four items of context. If such a result is generally valid – though the size of the temporal context that can be maintained is likely to be influenced by the exact experimental procedure – it would be very unlikely that Lewicki et al.'s (1987) subjects could learn rules requiring knowledge of six items of context.

Cleeremans and McClelland (1991, Experiment 2) obtained more direct evidence of the constraints on the amount of context subjects can maintain. They set up a task in which the location of the stimulus on a target trial could *only* be predicted by knowing where the stimulus had appeared four (and sometimes more) trials previously. Confirming the results of their first experiment, no evidence emerged that subjects could learn this long-range contingency, even when presented with a massive 60,000 trials. Instead, subjects appeared to be able to predict target location by reference only to the last one, two, or three locations. Thus, asking subjects to report entire rules risks falling foul of the Information Criterion.

2.7.5. Objections to prediction and recognition tests. Given the tentative conclusion we have reached, namely, that prediction tests do in general reveal savings and that recognition tests so far have not yielded clear dissociations from RT speedup, defenders of unconscious learning might say that these tests are not truly tests of awareness. In all probability, they might argue, subjects do not "know" why they press certain keys in the prediction task; perhaps their fingers just get pulled toward certain keys. According to some of Lewicki et al.'s (1988, p. 33) subjects, describing their RT performance, "their fingers were doing the job by themselves." Perhaps this happens as much for prediction as it does for the RT task. Similarly, perhaps subjects say "old" in a recognition test not because they are aware that the test sequence was part of the study sequence but because of perceptual fluency (see Perruchet & Amorim 1992). As we have mentioned already, there is no inherent reason why any behavioral measure should be influenced exclusively by conscious processes. In the language of the subliminal perception literature, maybe these tasks are not *exclusive* (Reingold & Merikle 1988).

If the prediction and recognition tests are not pure measures of awareness, then any conclusion based on them – e.g., that supposedly unaware subjects are in fact aware of the sequence – is called into question. But there are at least three reasons to doubt that unconscious processes do play a significant role. First, remember that in the prediction task, a response is required that is *different* from the one that was performed in the RT task. In the prediction task subjects respond to the *next* stimulus, whereas in the RT task they respond to the *current* stimulus. So, if the claim that prediction responses are under unconscious control is correct, subjects would make erroneous responses on the prediction task. This does not seem to be the case to any significant degree. Second, if some unconscious process is contributing to recognition performance, it seems that RTs should be faster in the recognition test for sequences that were part of the original sequence than for those that were not. Perruchet and Amorim (1992) were unable to find any evidence for this. Finally, Willingham et al. (1993) used three different versions of the recognition task. In one, subjects responded to the test sequences just as they had in the study phase, and then gave recognition ratings. Other subjects merely observed the stimulus sequence prior to making a recognition judgment, and a third subgroup saw the sequence presented in the form of the digits 1–4 rather than in terms of screen locations. The latter two procedures should rule out perceptual fluency as the basis of judgments, yet Willingham et al. observed no difference between the three tests. We believe this strongly suggests that the recognition test is a genuine measure of awareness of the sequence.

Of course, if the prediction task and recognition tests cannot be treated as tests of awareness, we have no recourse other than to examine the subject's verbal reports as the only available index of explicit knowledge, and this, we have argued, is unsatisfactory because it precludes meeting the Sensitivity Criterion. Instead, some authors have suggested that we should abandon the narrow version of the dissociation paradigm that underlies these implicit learning studies (see sect. 2.1) and try to demonstrate *qualitative* differences between conscious and unconscious learning. It could be argued that such a difference would exist if the information that can be expressed in performance (RTs) increases dramatically across learning trials while the information available to awareness only changes marginally. Awareness may not have to be entirely absent. Thus, presumably, one might say that subjects in the Willingham et al. (1989) experiment were able to project more information in their RTs than in their predictions (where the savings tended to be only about 5%), and thus, even though prediction performance was better than chance, this is still evidence for a distinction between implicit and explicit learning. But without measures of (1) the amount of information that is transmitted when a subject shows an RT speedup of x msec, and (2) the amount of information the subject is transmitting when their prediction performance improves by y%, it is very hard to assess such objections. We have no model for how much information is being conveyed by these different measures (for discussion of the same point, see Nelson 1978; Reingold & Merikle 1988).

It is exactly for this reason that one normally looks for cases where awareness of some variable is absent but performance is significantly affected by that variable, because even if we have no formal description of information, we know that a variable (which represents information) is being conveyed in performance but not in reports. For example, we would know that a variable such as the predictability of a sequence is affecting RT but not reports, and therefore RTs are conveying more information. Thus it is impossible to invalidate the null hypothesis (that performance and prediction convey the same amount of information) unless something can be done to show that more information is being conveyed in the implicit than in the explicit measure.

2.7.6. Conclusions. Sequence-learning studies have used prediction and recognition tasks as indices of awareness. These tasks reproduce the stimulus context of the learning stage (hence addressing the Sensitivity Criterion), and can be performed at above-chance levels, whether the subjects' knowledge is of fragments or of the complete sequence (hence meeting the Information Criterion). However, contrary to claims that sequence learning is unconscious, the results to date suggest that in most cases subjects *are* aware of the relevant knowledge, and that their knowledge consists of fragments of the training sequence. We believe that no convincing evidence of implicit learning has yet emerged in sequential RT tasks; nevertheless, this is a very promising field of research that may in the future allow more positive conclusions.

3. Encoding instances versus inducing rules

In the introduction we raised two fundamental issues concerning implicit learning: that of consciousness during learning and the type of knowledge acquired. By crossing these factors we obtain four hypothetical learning systems: unconscious instance/fragment learning, unconscious rule learning, conscious instance/fragment learning, and conscious rule learning. This review has shown that the evidence for unconscious learning of any sort is highly questionable; we accordingly conclude that unconscious learning is unsupported in general. It is time, then, to turn to our second dimension for characterizing dissociable learning systems, namely, the content of the acquired knowledge, and to assess the rules versus instances distinction within the domain of conscious, explicit learning.

In the following discussion we use the term "implicit learning tasks" to refer to the sorts of tasks reviewed in section 2.

3.1. Evidence from studies of concept learning

Perhaps the most important conclusion from our discussion so far is that performance in a variety of different tasks can be well accounted for by reference to fragmentary knowledge or knowledge of instances rather than to abstract rules. In the present section we pursue this idea by showing that to a large extent human performance in more traditional concept-learning studies can also be well understood in such terms. However, the concept learning literature leads us also toward more compelling evidence that people can genuinely learn rules. Thus we begin to see a characteristic that does distinguish different learning systems: whether the knowledge acquired is of instances or rules.

A view of concept learning that had been popular for many years prior to the 1960s was that learning a concept involves the acquisition of a rule specifying the features necessary and sufficient for membership of that category. However, when Rosch (1975) argued that for many natural categories it was impossible to specify the necessary and sufficient features, and when Posner and Keele (1970) showed that subjects could learn to classify random dot patterns that had not been generated by deterministic rules, research began to be dominated by the alternative view that concepts are represented by prototypes. A prototype is an abstraction from a set of training stimuli that corresponds to their central tendency.

On the prototype view, category membership is determined simply by computing which of a series of stored prototypes the test stimulus is closest to. As in the rule-based account, in prototype theories the subject is assumed to abstract something from the training stimuli and not to retain information about those specific instances.

3.1.1. The role of instances. The view that the learning of a concept could be based on little more than the encoding of the separate instances that fall under the concept began to emerge in the late 1970s in the work of researchers such as Lee Brooks and Douglas Medin. They observed that, contrary to what would be expected on a prototype account, subjects do appear to retain information about training instances, in that studied instances can bias subsequent classification decisions (e.g., Brooks 1978; Homa et al. 1981; Malt 1989; Medin & Schaffer 1978). On the basis of such findings, Medin and Schaffer (1978) proposed that one component of a concept is simply a set of memorized exemplars or instances.

Although Medin and Schaffer assumed that both instance storage and prototype abstraction could occur during the learning of a concept, subsequent studies have shown that performance in a great many category-learning studies can be understood in terms of instance storage alone (for a historical review, see Medin & Florian 1992). We can illustrate the power of instance-storage theories by considering the results of a study by Shin and Nosofsky (1992), who examined category learning with dot patterns. Shin and Nosofsky first performed a multidimensional-scaling analysis on subjects' judgments of pairwise similarity for the patterns, which yielded coordinates in psychological space for each of the patterns. Other subjects were then trained to classify some of the patterns into three categories and, following that, they were tested on their classification decisions for the remaining patterns.

The instance view proposes that subjects memorize the actual exemplars seen during training and base their classifications on the similarity between a test item and stored instances. Since Shin and Nosofsky knew the psychological coordinates of all of the patterns, they were able to compare subjects' decisions to test patterns with the predictions of a model that assumed that classification was determined solely by similarity to memorized instances. They found a remarkable degree of concordance between predicted and observed classifications, with over 95% of the variance in the observed classifications being accounted for in a 1-parameter model. A prototype theory, assuming that the training instances formed the basis for an abstracted prototype, performed much more poorly in predicting responses.

The implications of such results for implicit learning cannot be overemphasized. The stimuli used in typical categorization experiments are every bit as complex and difficult to label verbally as are the stimuli used in implicit learning experiments, so the powerful evidence for instance storage that emerges from categorization experiments should encourage us to take very seriously the possibility that the encoding of instances is a major factor in implicit learning experiments as well.

3.1.2. Evidence for rule induction. Despite the wealth of evidence in favor of instance theories of concept learning, we argue that it is also possible for people to classify objects on the basis of a rule or hypothesis. As a particularly dramatic example, consider the evidence for a difference between instance learning and rule learning in the sexing of day-old chicks (Biederman & Shiffrar 1987). It has been estimated that professional sexers, trained with feedback on instances, require 2.4 months of solid practice to reach 95% accuracy. However, naive subjects trained on one simple rule immediately achieved 90%

accuracy. Of course the simple rule misses the rare and subtle exceptions that instance learning can provide, so further accuracy gains will be difficult. On the other hand, the initial difference in training time is immense.

Laboratory demonstrations of contrasts between rule and instance learning have been provided in a number of studies (e.g., Allen & Brooks 1991; Kemler Nelson 1984; Nosofsky et al. 1989; Regehr & Brooks 1993; Smith & Shapiro 1989; Ward & Scott 1987). Consider, for instance, Nosofsky et al.'s (1989) Experiment 1. In this experiment, the stimuli were semicircles with an interior radial line: 16 stimuli were constructed from the combination of four sizes of semicircle (1 . . . 4) with four angles of inclination of the radial line (1 . . . 4). In one condition, subjects learned to classify three of the stimuli into category 1 and four into category 2, and then were tested for their transfer performance across the remaining nine stimuli. The dependent measures were the overall probabilities with which subjects placed each of the 16 stimuli into categories 1 and 2.

Nosofsky et al. (1989) found that 97.9% of the variance in these combined classification probabilities across subjects could be accounted for by a quantitative model that assumed that stored instances were the only basis on which the decisions were made, an impressive fit that confirms the conclusion of the Shin and Nosofsky (1992) study. In addition, however, Nosofsky et al. compared their model to several rule-based descriptions. Specifically, it would have been possible for subjects to learn the categorization problem by inducing a rule for partitioning the stimulus space into response regions. For instance, one rule that would have correctly classified the training stimuli was the rule: *a stimulus is in category 2 if the value of angle is 1, or the value of size is 1, or the value of angle is 4; otherwise the stimulus is in category 1.* Nosofsky et al. constructed a variety of such set-theoretic rules, but found that none of them fitted the overall classification performance of the subjects nearly as well as their instance model. From such data it would appear that subjects rely on nothing more than stored representations of the training items in classifying stimuli.

When they looked at the behavior of individual subjects, however, Nosofsky et al. (1989, Experiment 1) found that some subjects' classification responses conformed to patterns that were quite unlikely according to an instance theory, and yet they matched fairly simple rules. Thus, at the level of individual subjects, some evidence for rule following rather than generalization to stored instances did emerge. In their second experiment, Nosofsky et al. chose two rules that could be used to classify the stimuli accurately, and gave two different groups of subjects explicit instructions to follow one of the two rules in classifying the stimuli. Here, the instance theory failed dramatically, accounting for only 82.4% and 40.9% of the variance in responses for the two groups. In contrast, 99.5% and 93.6% of the variance in the subjects' classifications in the two groups were accounted for by the rules themselves. Similar results were obtained in their Experiment 3, using a different set of stimuli. Thus, here we have clear evidence that subjects are able to learn an abstract rule and hence need not rely just on stored instances.

An even more compelling demonstration of the inadequacy of pure instance storage comes from recognition memory data that Nosofsky et al. collected during the test phase of their experiments. Although in Experiment 1 subjects had no difficulty recognizing the old training stimuli and discriminating them from new test stimuli, when subjects were explicitly instructed to use a rule to classify the stimuli in Experiment 2, no evidence emerged that the subjects could remember which test stimuli had been training stimuli. The implication of this result is that these subjects had encoded nothing in the training stage except the rule: they had not encoded any of the instances. This remarkable finding shows that when appropriate conditions are established, subjects can indeed learn an abstract rule from exposure to instances. Of course, just because subjects had to be given rule-following instructions in Nosofsky et al.'s studies before they would actually engage in rule following does not mean that this will always be necessary. In fact Nosofsky et al.'s stimuli make rule following difficult in that the stimuli do not readily lend themselves to verbal descriptions.

Other compelling evidence for rule learning, from tasks using more complex stimuli, has been reported by Allen and Brooks (1991) and Regehr and Brooks (1993). The rationale of the experiments was as follows: suppose that subjects learn to classify stimuli in a situation where a simple, perfectly predictive classification rule exists and they are then tested on transfer items that vary in similarity to the training stimuli. Observed behavior with the transfer items can be of two contrasting types: (1) transfer items similar to old items from the opposite category (bad transfer items) may be classified as quickly and as accurately as items similar to old items from the same category (good transfer items), or (2) the bad transfer items may be classified much less rapidly and accurately than the good transfer items.

The first case above would be consistent with classification being determined by the speeded application of a rule. In this case all that matters is whether the rule assigns the transfer item to one category or the other. Whether the item is similar or not to a training instance, and whether that instance was in the same or a different category, should be immaterial. On the other hand, the second outcome described above would be consistent with categorization on the basis of similarity to training instances, and there would be no need to cite a rule as being part of the classification process.

Allen and Brooks (1991) and Regehr and Brooks (1993) obtained evidence that both sorts of outcome can occur, depending on the type of stimuli used and the precise nature of the task. They trained subjects to classify animals into two categories. The animals varied in terms of five binary-valued dimensions: body shape, spots, leg length, neck length, and number of legs, but only three of the dimensions were relevant. In some experiments subjects were explicitly told the classification rule (e.g., category 1 is defined by the conjunction of long legs, angular body, and spots). Evidence for rule learning – no difference in latency or accuracy in classifying a new item similar to a training item and in the same category versus a new item similar to a training items but in the opposite category – was related to a number of factors. For example, rule learning was more likely to be the controlling process when subjects were actually told the rule prior to the task, although this was not a necessary condition. It also depended on the nature of the stimuli used: highly individual stimuli tended to elicit instance-based rather than rule-based classification, and stimuli composed of interchangeable features tended to elicit more rule-based classification behavior.

Additional factors that appear to determine the balance between rule learning and instance learning have been investigated. Smith and Shapiro (1989) found that rule learning was less likely when subjects had to perform a secondary task during the learning stage and when the training stage was conducted as an incidental learning task. Smith and Kemler Nelson (1984) found that rule learning was less likely in a speeded than an unspeeded learning task, and also that there seems to be a developmental trend in rule learning: in situations where adults classify according to a rule, children often do so on the basis of similarity to training instances. In sum, laboratory studies have established the reality of rule- or hypothesis-based concept learning and have begun to identify the circumstances that determine when it predominates.

3.2. Rule induction from artificial grammars

In our earlier discussion of artificial grammar learning, we suggested that performance could be well understood on the assumption that subjects principally learn about whole strings (instances) and about legal substrings. Although some evidence for abstraction exists (e.g., Altmann et al., in press), such a view denies that knowledge of abstract grammatical rules plays a major part in performance. However, when subjects are told prior to seeing them that the strings are rule governed, and particularly if they are taught something about finite state grammars (Mathews et al. 1989; Reber et al. 1980), a pattern of results rather different from that seen under normal conditions occurs. Under these conditions, subjects appear to engage in a very explicit form of rule induction or hypothesis testing that produces reportable rules. The dissociation between performance and verbal reports does not occur under these conditions.

Reber et al. (1980) extensively instructed subjects on the nature of finite state grammars before, during, or after presenting the training strings: When instruction occurred before training, subjects performed well on the grammaticality task but showed evidence of having learned unrepresentative rules: rather than random errors from guessing, their errors were consistent. Nonrandom mistakes fit with the idea that subjects were trying out various hypotheses in turn. At test time subjects will have some hypothesis for evaluating test strings, but it may be incorrect and therefore produce consistent errors. Further supporting this claim, Reber et al. found that a well-ordered presentation of training strings, making the structure of the underlying grammar more salient, assisted learning only in the case where the explicit instructions occurred early in training. The conclusion is that the explicit instructions promote an explicit hypothesis testing strategy, and that well-ordered string presentation was particularly helpful for this learning strategy.

Mathews et al. (1989) found that the nature of the grammar also affected subjects' ability to learn it via an explicit strategy. Finite state grammars are difficult to learn explicitly: instructions that promote hypothesis testing but provide no information about the nature of grammar rules, such as instructions to "find the rules," fail to produce good learning (Perruchet & Pacteau 1990; Reber et al. 1980). Indeed, Mathews et al. found that subjects given such instructions learned no better than subjects given incidental learning instructions. Other types of grammars, however, are easier to learn explicitly. For example, subjects in Mathews et al.'s study readily acquired a biconditional grammar in which strings consisted of two sets of four letters, with three rules mapping the letters between the sets. Mathews et al. observed far superior learning in subjects instructed to search for rules than in subjects given incidental learning instructions. Moreover, in contrast to subjects given rule-searching instructions, subjects given incidental instructions did not acquire this grammar at all, a result that strongly argues for the existence of independent learning systems.

How does explicit hypothesis testing differ from the memory-based processing that is the norm in implicit learning paradigms? Smith et al. (1992) compiled a set of eight characteristics of rule following behavior. For example, a defining feature is that rules ought to apply equally well to familiar, unfamiliar, and abstract problems or stimuli (as in Allen & Brooks 1991; Regehr & Brooks 1993). Smith et al. review numerous experiments that demonstrate subjects following rules such as *modus ponens* and the law of large numbers.

Turning to rule *learning*, Smith et al. point to subjects' learning protocols and their ability to report intermediate hypotheses as evidence for rule learning. Indeed, explicit hypothesis testing presumes that the hypotheses are conscious and reportable. Protocols indicate that generating and testing these hypotheses is a slow, labored, and conscious process. Lea and Simon (1979) describe a general framework for understanding the cognitive processes involved. Hypothesis testing, they claim, is a form of problem solving involving search through two problem spaces: the space of hypotheses and the space of experiments or instances to test hypotheses.

Studies of series completion, a task that is similar to artificial grammar learning, provide a clear example of the characteristics of the hypothesis-testing strategy (Kotovsky & Simon 1979). In this task, subjects are shown a single string of letters that contains a short repeating pattern (e.g., *MABMCD*) and are asked to continue the series. Their protocols clearly indicate that they cycle through gathering evidence from the string and generating hypotheses to fit the data. In addition, subjects' occasional mistakes are consistent with their last-considered, but incorrect, hypothesis. They will then produce consistent errors as did subjects in the grammar learning studies (Reber et al. 1980). Finally, unusual hypotheses are difficult to generate and therefore unlikely to be discovered; this effect has been shown in several hypothesis-testing paradigms (Bruner et al. 1956; Klahr et al. 1990; Klayman & Ha 1989; Wason 1968). This unsurprising effect may explain the difficulty observed in getting subjects who are instructed to find the correct grammar rules actually to do so. These rules may be unusual enough that they are difficult for subjects to generate without detailed instruction.

How is rule generation and evaluation performed in this dual problem space? Lea and Simon (1979) offer several alternatives. The simplest alternative is to refrain from any feedback from the evaluation stage to the generation stage except to signal rejection when a hypothesis is disconfirmed by data. Rule generation is then essentially blind to the detailed data of experiments or instances. More sophisticated alternatives allow more information from experiments or instances to inform the generation process. In concept-formation tasks, Lea and Simon found that different subjects' behaviors conformed to different alternatives.

Protocols from Kotovsky and Simon's (1979) series-completion task provide some indication of what alternatives might underlie subjects' behaviors in implicit learning tasks. Kotovsky and Simon's protocols showed that subjects examined the series and discovered periodicities in the letters. Subjects then developed compact rules to describe those periodicities. The explicit hypothesis-testing strategy in grammar-learning experiments might operate similarly. As subjects observe strings, they may discover repetitive sequences. Subjects could then attempt to learn those sequences and develop rewrite rules to describe sequences of those sequences. Unfortunately, we do not know of any protocol data that examine this question.

To conclude, subjects performing explicit hypothesis testing demonstrate in their protocols clear reports of intermediate hypotheses and a relatively slow time course of processing. These phenomena stand in contrast to the obscure and superficial report and relatively fast time course of processing found in protocols of subjects engaged in typical implicit learning tasks. Though we have seen no data, we predict that subjects' behavior in grammar-learning experiments under explicit hypothesis-testing instructions will more closely resemble Bruner et al.'s (1956) and Kotovsky and Simon's (1979) subjects than subjects working under implicit instructions.

It seems clear, then, that there are two separate learning strategies available to subjects, and that these strategies can be invoked by differences in the instructions given to subjects. The rule-induction strategy is characterized by conscious effort to develop and evaluate hypotheses that are often unrepresentative of the actual grammar. This strategy can be invoked by detailed instructions and it can be facilitated by sensibly ordered examples. The "instance" strategy is invoked by instructions simply to memorize or observe the training strings, and it does not seem to be affected by the presentation order of training strings (Reber et al. 1980). Under this strategy subjects encode the whole strings and their basic features – such as pairs and triplets. Both strategies appear to be conscious, but the contents of consciousness vary. The rule-induction strategy trades in hypotheses and is well characterized as problem solving, but the instance strategy trades only in simpler data and seems well characterized as memorization.

3.3. Models for implicit learning tasks

How do computational models fit with the results we have discussed? We first consider the computations involved and then turn to the issue of consciousness.

Artificial grammar research has provided the most detailed analysis of what types of knowledge are acquired during implicit learning tasks and what processes use them at test time. There is memory for whole strings for recognition and similarity judgments, and substring fragments for piecemeal familiarity judgments. Distributed-memory models in general and Parallel Distributed Processing (PDP) models in particular are well suited to these knowledge types and tasks. Networks are capable of learning both types of knowledge simultaneously and in the same set of weights. Dienes (1992) compared a variety of distributed- and memory array models on a set of grammar-learning data and found that an autoassociative network using the delta rule for learning fit the data best. The less competitive alternatives were connectionist models using the Hebb rule, and three memory array models.

The autoassociative network model consisted of a single layer of units that were completely interconnected. Each unit coded for one letter at a particular position in a string. Accordingly, there were five units (one for each letter) for each of six string positions. A variety of other local and distributed encoding schemes were also tested, but no great qualitative differences were found. The model was trained on the set of 20 training strings shown to subjects, and the delta rule (Widrow & Hoff 1960) was used to change the weights until each training pattern could reproduce itself over the input units. In one set of simulations, the model was trained on each stimulus the same number of times as were subjects. Each training trial, therefore, was equivalent for subjects and for the model. The model was tested on the grammaticality task by requiring it to compute a response to each test string. If the model could correctly reproduce the input with minimal error, the response was called "grammatical." If the model reproduced the input with a large error, the response was called "ungrammatical."

The delta rule model produced a high level of performance on the grammaticality test. It also produced the same ratio of random errors to consistent errors as did subjects. Critically, its ranking of the grammaticality of the test strings correlated highly with subjects' rankings. No other model could produce this correlation.

A somewhat similar account has been proposed for sequence learning. Cleeremans and McClelland (1991) developed a simple recurrent network model to simulate their human sequence-learning data. Since much of the benefit in reaction times during the sequence task is presumed to come from predicting the next stimulus, the model was designed to predict the next stimulus at each point in the sequence. The input layer of the model encoded one stimulus of a sequence at a time and used one unit for each of the possible stimuli in the "grammar" of stimulus sequences. As the model stepped through a sequence, the input layer successively encoded each stimulus in turn. At each step, activation from the input layer was fed forward through the model to produce activations in the output layer to represent the model's prediction about the upcoming stimulus. During training, the weights between units were adjusted so as to produce accurate predictions.

The model's behavior corresponded to the human data in three ways. First, after training, the model performed at the same level as subjects on the sequence task. Second, the model learned at the same rate as subjects. Cleeremans and McClelland drew a correspondence between single training trials for subjects and single training trials for the model. The model matched subjects' performance throughout the time course of training. Cleeremans and McClelland divided knowledge of the grammar into sets, depending on how many previous stimuli were required to make accurate predictions at each point in a sequence. Like subjects, the model first acquired knowledge allowing it to predict correctly in cases where only one previous stimulus was important. Both subjects and the model then slowly acquired the longer, more complex dependencies.

The third match between subjects and the model was the length of delay between predicted stimuli and their predicting context. Cleeremans and McClelland created a grammar with a set of stimuli that intervened between a predictive beginning stimulus and a predicted ending stimulus (e.g., 15553 vs. 25554). For both subjects and the model, only three intervening stimuli could be tolerated before the initial stimulus was forgotten.

As discussed above, concept learning also often appears to be mediated by the conscious memorization of instances. Recently, PDP models have been developed to simulate human results on these tasks as well (Gluck & Bower 1988; Kruschke 1992; McClelland & Rumelhart 1985; Shanks & Gluck 1994). These models produce close quantitative fits to a variety of concept identification, classification, and recognition data. Like the other PDP models, they learn in a memorization-like fashion by encoding individual stimuli and modifying their weights in response to each stimulus to process that stimulus better.

What about more fragmentary information? Perruchet and Pacteau (1990) found that subjects recognized frequent bigrams better than infrequent bigrams. This result requires some representation of the frequency of bigrams. An important feature of distributed memories (Cleeremans & McClelland 1991; Dienes 1992) is that they produce frequency statistics, in the form of strengths of encodings, as a by-product of the memorization process. Such a process has two good points: it does not require any covert computations to produce useful knowledge, and the computations it does require – memorization – fit with subjects' reportable experience of the training tasks.

These PDP models, therefore, capture a variety of data from sequence-learning, grammar-learning, and concept-learning tasks. Their details differ, but their basic representational abilities and modes of processing are the same. In this sense, they provide a unifying mechanism for learning both whole item and fragment knowledge and for simulating a wide range of cognitive phenomena from conditioning to sequence and grammar learning. Cleeremans (1993b) discusses at length some of the correspondences between human performance in implicit learning tasks and the behavior of connectionist models.

These models do not, however, say anything about consciousness. A standard idea in the psychological literature is to equate consciousness with the states of a processor, rather than with the processes themselves. In a PDP model, the states are the transitory activations of the units. "We assume that responses and perhaps the contents of perceptual experience depend on the temporal integration of the pattern of activation over all of the nodes" (McClelland & Rumelhart 1981, p. 381). In the models of instance memorization, these activations encode the features of individual stimuli. There is no representation within the models of rules or of the testing of hypotheses. These representations, therefore, fit well with the idea that what subjects are doing in implicit learning tasks is memorizing the stimuli.

The ability to generalize and perform above chance on a grammaticality task arises from the fact that memorization takes place in a distributed system, where every memory contributes to every response. In a connectionist model such as Dienes's (1992), memorization occurs when the weights of the network are changed to encode a stimulus. Similar stimuli produce similar weight changes that establish strong connections over time, whereas dissimilar stimuli, or parts of stimuli, produce dissimilar weight changes that wash out over time. In this way, the central tendency and underlying structure of a set of stimuli is slowly captured by the weights. In both the Dienes (1992) and Cleeremans and McClelland (1991) models, the system is fully "aware" of the stimulus on each trial. In the latter model, the input layer encodes the stimulus one letter at a time, and that letter is fully activated. The input layer in Dienes's model encodes an entire string at once. While the stimulus on each learning trial is fully registered, the weights of the network start out very weak and slowly grow stronger with training, as knowledge is acquired. The models say, therefore, that subjects are fully aware of each stimulus, and that they slowly grow more aware of the underlying structure of the stimulus set as the weights grow stronger. But just as awareness of the structure may be very limited early in training, so will the extent to which that structure actually controls performance: "awareness" and performance will be correlated.

In contrast, connectionist models do not fit well with our understanding of the explicit hypothesis-testing strategies also found in the grammar-learning literature (sect. 3.2). Lea and Simon (1979) describe a possible mechanism, along the lines of the "general problem solver," to perform hypothesis testing. This learning strategy is different enough from the memorization strategy in process and results so that it is not surprising that it calls for a different sort of mechanism to execute the hypothesis-testing computations.

4. Learning and amnesia

In our evaluation of learning systems we have not considered at all the evidence from patients suffering from the classic anterograde amnesic syndrome (Squire 1992). Because these patients are generally considered to suffer from a learning or memory deficit, it is worth considering briefly how the data from this population relate to the data from normal subjects.

We know of no convincing data that would suggest that amnesics are capable of unconscious learning. It is very important to note that most studies have used episodic memory tests and hence fail to meet our condition for inferring unconscious learning from unconscious memory. For example, amnesic patients show normal or near-normal responding to a previously conditioned stimulus although they are apparently unable to recall the conditioning episode (Weiskrantz & Warrington 1979). But this is an example of case (2) rather than case (3) from section 2.3.1 (Fig. 1) and is therefore insufficient to establish that learning had been unconscious: that is, subjects may have been conscious of the reinforcement contingency but unable to remember the episode in which it was learned. The amnesics, if asked, may well have been able to report a conscious expectancy of the US (unconditioned stimulus) given the CS.

It is true that a number of studies have shown that amnesics, like normals, can sometimes acquire information but be unable to report it verbally. Thus Nissen and Bullemer (1987; see also Knopman 1991) found that Korsakoff amnesic patients could speed up on a sequential RT task but report neither the sequence nor awareness of the existence of a sequence. However, as our discussion in section 2 revealed, that subjects are unable to report verbally some information they can otherwise be shown to have learned does not prove that the learning was unconscious, and this is as true for amnesics as it is for normal subjects. Nissen and Bullemer did not use a prediction test for conscious knowledge, and we know of no other studies that have attempted to use more sensitive tests of awareness.

This is not to imply that amnesics' poor performance on explicit memory tests is due to the insensitivity of such tests. On the contrary, this is unlikely to be the case because compared to normals, amnesics appear to be *selectively* impaired on explicit tests such as recognition (but see Ostergaard, 1994, for a contrary view). For example, Knowlton et al. (1992) reported that amnesics can perform as well as normals in judging the grammaticality of new strings generated from an artificial grammar while being very poor at recognizing the training strings. It is hard to see how such dissociations can be explained by differences in sensitivity between the grammaticality and recognition tests, because in that case the effect should be comparable for normals and amnesics. Instead, the evidence suggests that amnesics have a genuine problem with a certain class of memorial experience (i.e., episodic memory; see Humphreys et al. 1989, for a relevant computational model). However, there is no reason to believe that the problem involves awareness at the time of learning, or more specifically that results from amnesics provide any evidence for unconscious learning.

With respect to the learning of rules versus instances, there seems little doubt that amnesics can perform well (possibly at the same level as normals) in tasks that can be mastered by learning instances or fragments. Thus, in addition to Nissen and Bullemer's (1987) demonstration of learning in a sequential RT task, Knowlton et al.'s (1992) report of excellent artificial grammar learning in amnesics suggests that instance or fragment learning is intact. In contrast, evidence for genuine rule learning is sparse, which is not surprising given the problems involved in distinguishing instance from rule learning. Nevertheless, there is some evidence of conscious rule learning in amnesia. For example, Wood et al. (1982) report that although amnesics have some difficulty learning the Fibonacci rule, they are able to do so.

5. Summary

A variety of strategies have been used to assess, more or less directly, the content of a subject's awareness during a learning episode. Evidence for implicit learning would come from (1) demonstrations of learning with subliminal stimuli, and (2) dissociations between task performance and measures of awareness such as verbal reports. The latter dissociations appear at first sight to provide evidence for implicit learning, because the inability to report the relevant stimulus relationship licenses the inference that learning may have occurred without awareness.

We have argued that there is little convincing evidence of learning with subliminal stimuli. On the other hand, with respect to the unconscious learning of stimulus relationships, we have documented a number of dissociations between performance and reports. There are nonetheless two reasons to question whether they establish implicit learning. First, there may be a relatively uninteresting explanation of such dissociations, stemming from the experimenter's failure to address the Information Criterion. If learning involves the acquisition of information I, but the experimenter is focusing on information I^*, then subjects may appear to be unaware of the relevant knowledge when in fact they are aware of it. Second, if the test of reportable knowledge fails to meet the Sensitivity Criterion, it is impossible to know whether a dissociation is genuine or merely reflects inadequate sensitivity to conscious information in the awareness test. Our review suggests that when attempts are made to use sensitive tests, dissociations do not emerge. Finally, if one is unprepared to accept recognition or prediction tests as measures of conscious knowledge, then it is difficult to see how the Sensitivity Criterion can ever be met. It is simply a fact of life that tests of verbal recall tend to be less sensitive to small amounts of knowledge than other behavioral measures. Perhaps alternatives to the simple dissociation logic adopted in almost all experimental tests of unconscious learning need to be explored, as they have in studies of unconscious perception (Reingold & Merikle 1988).

Our evaluation of the results that have emerged is similar to Holender's (1986) conclusion in this journal concerning unconscious semantic activation. Although there are some interesting pieces of evidence, a cautious approach would suggest that unconscious learning has not yet been satisfactorily established. Instead, there is substantial evidence for more than one conscious learning strategy and knowledge type. People certainly can learn and use rules, and they can also memorize instances and fragments. Researchers have begun to identify the factors (e.g., study time, stimulus properties) that are conducive to rule learning and instance learning.

Proponents of implicit learning, which is hypothesized to involve the unconscious learning of rules, have failed to demonstrate that it accurately describes a class of human learning abilities. On the contrary, human learning is almost invariably accompanied by conscious awareness, and in tasks such as artificial grammar learning, where learning is frequently thought to involve rule abstraction, performance is most often based on the acquisition of instances or fragments from the training stage.

ACKNOWLEDGMENTS

The first author was supported by the UK Medical Research Council, and by a SERC/NATO fellowship held at the Department of Cognitive Science, University of California, San Diego. We would like to express our considerable gratitude to colleagues who have commented on this work, especially Axel

Cleeremans, Anthony Dickinson, Zoltan Dienes, Don Dulany, Celia Heyes, Jonathan Kolodny, Phil Merikle, Pierre Perruchet, and Arthur Reber.

NOTES

[1] The mere exposure effect may provide a paradoxical example of how instructions that seem to encourage the subject to rely on available conscious knowledge can result in their doing so to a lesser degree than instructions for a performance task. Several well-known studies (e.g., Kunst-Wilson & Zajonc 1980; see Bornstein 1992, for a review) have given subjects very brief presentations of geometrical figures prior to testing them with pairs of stimuli consisting of one old and one new figure. For each pair, the subjects had to indicate which one was old and which one was preferred. Results indicate that subjects will choose the old stimulus for their preference judgment although their ability to discriminate old from new stimuli in the recognition test is at or close to chance. On the face of it, this provides a powerful dissociation between results obtained from closely matched performance (preference) and awareness (recognition) tests. Ignoring the possibility that the procedure may fail to meet the requirement for inferring unconscious learning from unconscious memory (sect. 2.3.1), we suggest that the result is still not evidence for unconscious learning because, relative to the preference test, the recognition test may encourage subjects to *discount* deliberately a conscious source of information (viz., stimulus familiarity), because they know that familiarity can be a poor index of whether a stimulus has recently been seen. Evidence for this interpretation comes from a recent experiment by Merikle and Reingold (1991), who found that as testing continued, the recognition test gradually became more sensitive to the old/new distinction than the preference test (i.e., hypermnesia occurred). This result is consistent with the idea that subjects began to rely on familiarity in making their recognition judgments because they realized that when they discounted it, they had no other cues on which to base their recognition decision.

[2] A possible empirical way to determine whether unconscious influences do play a role would be to adopt Jacoby's (1991) process dissociation technique of asking subjects to provide letter continuations but to avoid any continuation that they had seen in the study phase. If some studied continuations were given, that would indicate the presence of unconscious influences.

[3] The prediction task bears a striking but presumably unintentional resemblance to the task used by the parapsychologist Schmidt (1969) in his tests of precognition. In Schmidt's experiments, target selection on each trial was determined by random particle emission from a strontium-90 source, yet subjects were apparèntly able to predict at better-than-chance levels where each target would appear! In our discussion of data from the prediction task, we ignore the rather distressing possibility that subjects' performance may be influenced by such precognitive abilities.

Open Peer Commentary

Commentary submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

Is learning during anaesthesia implicit?

Jackie Andrade
Medical Research Council Applied Psychology Unit, Cambridge CB2 2EF, England; jackie.andrade@mrc-apu.cam.ac.uk