Journal of Experimental Psychology: General
1996, Vol. 125, No. 2, 123-138

0096-3445/96/$3.00

Transfer in Artificial Grammar Learning: A Reevaluation

Martin Redington and Nick Chater
University of Edinburgh

This article covers methodological and theoretical issues in artificial grammar learning. Arguments that such tasks are mediated by abstract knowledge (e.g., A. S. Reber, 1969, 1990) are based primarily on evidence from transfer experiments, where the surface vocabulary is changed between learning and test items. Because of a number of methodological concerns, the small magnitudes of artificial grammar learning effects generally are difficult to interpret. Possible solutions are offered here. Furthermore, even reliable transfer effects imply neither that subjects have acquired abstract knowledge of the underlying grammar nor that they are performing a process of abstract analogy from memorized whole exemplars. Models that learn only surface fragments of the training stimuli and perform abstraction at test rather than during learning are wholly consistent with transfer phenomena.

One of the most fundamental questions in cognitive psychology is whether the knowledge is stored in terms of abstract rule-like descriptions or as sets of specific instances. According to the first view, novel items or events are dealt with by applying the stored abstract rules to the novel case. According to the second view, there is some process of comparison or analogy between stored examples and the current event. The controversy between these points of view arises in the study of memory (Hintzmann, 1986), categorization (Barsalou, 1990; Reeves & Weisberg, 1994), and analogical reasoning (Gentner, 1989), and aspects of language learning (Pinker & Prince, 1988; Plunkett & Marchman, 1991; Rumelhart & McClelland, 1986). Artificial grammar learning appears to be a domain in which the case for stored abstract rules is especially strong. It is argued that in artificial grammar learning experiments, subjects are able to learn the grammatical rules underlying the training stimuli (Reber, 1967, 1990). Results from so-called "transfer experiments," in which the training and test stimuli have different surface forms but the same underlying abstract structure, have been assumed to unequivocally rule out any instance-based account, because these accounts are inevitably tied to surface forms.

In this article, we show that the argument from transfer experiments to abstract knowledge is flawed in two ways. First, a number of methodological problems render the interpretation of many transfer experiments difficult—the results of many experiments are consistent with no transfer having occurred at all. For instance, almost all studies lack adequate control groups. This is of particular concern because control subjects who have received no training whatsoever have been observed to perform at the same above-chance levels found in typical transfer studies. However, on the basis of the small number of relatively well-controlled studies that have been conducted, it seems likely that transfer is a genuine phenomenon. Our second argument is that in any case, transfer does not imply that subjects acquire abstract rules during learning. We demonstrate this by providing a simple fragment-based account (i.e., the subject's knowledge consists purely of letter pairs and triples found in the training stimuli) that can attain levels of transfer performance well in excess of those obtained by experimental subjects. This shows that successful transfer does not imply that knowledge is stored in the form of abstract rules.

The Artificial Grammar Learning Paradigm

In a typical artificial grammar learning experiment (e.g., Dulany, Carlson, & Dewey, 1984; Perruchet & Pacteau, 1990; Reber, 1967), subjects are instructed to memorize a set of learning strings generated by a finite state grammar such as that shown in Figure 1. Subsequently, when informed that the learning strings were generated by a set of rules and asked to distinguish between test strings that follow those rules and test strings that violate them, subjects perform at above-chance levels. However, subjects are typically unable to articulate much of the knowledge that allows them to perform this task. This is the archetypal implicit learning effect. If the actual letters of the strings are changed between learning and test, subjects are still able to

Martin Redington and Nick Chater, Department of Psychology and Centre for Cognitive Science, University of Edinburgh, Edinburgh, Scotland.

Martin Redington was supported by Economic and Social Research Council Research Studentship R00429234268, and Nick Chater was partially supported by Research Grant SPG9029590 from the Joint Councils Initiative in Cognitive Science/Human–Computer Interaction. We thank Jim Morrison, Zoltan Dienes, Don Dulany, David Shanks, and the Center for Research in Language, University of California, San Diego, where some of this research was conducted. Axel Cleeremans, Zoltan Dienes, Mark St. John, Pierre Perruchet, David Shanks, and Arthur Reber all helpfully commented on earlier drafts of this article.

Correspondence concerning this article should be addressed to Martin Redington or Nick Chater, who are now at Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, United Kingdom. Electronic mail may be sent via Internet to Martin Redington at fmr@sable.ox.ac.uk or to Nick Chater at nick@psy.ox.ac.uk.

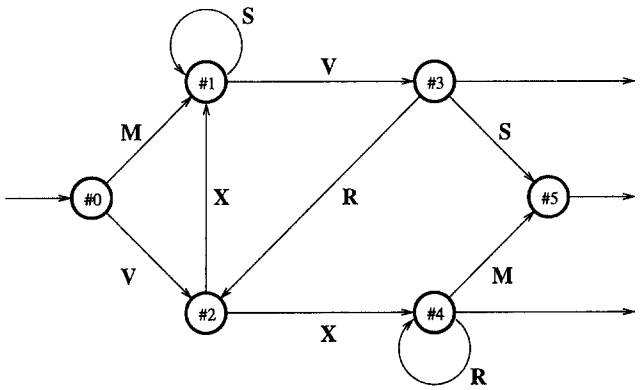

Figure 1. The standard finite state grammar. Grammatical strings are generated by following the paths starting at #0 and continuing until one of the three exiting paths is taken, with each path generating the letter that labels it. This grammar, although not necessarily with these particular letters, was used to generate the stimuli for Reber and Allen (1978), Dulany, Carlson, and Dewey, (1984), Perruchet and Pacteau (1990), Dienes, Broadbent, and Berry, (1991), and Redington and Chater (1994).

perform at above chance on the discrimination task—they are able to transfer their knowledge to a different letter set (e.g. Brooks & Vokey, 1991; Whittlesea & Dorken, 1993), although at a slightly lower level of performance than in the standard nontransfer condition.

A controversial claim in implicit learning in general and artificial grammar learning in particular is that subjects acquire "an unconscious abstract representation of the structure in the information given" (Reber & Lewis, 1977, p. 355). The issues of whether the knowledge used in these tasks is unconscious and how this hypothesis might be tested have been discussed extensively in a recent review (Shanks & St. John, 1994). The concern of this article is whether, and to what extent, the knowledge that subjects acquire is abstract.

There has been an extensive empirical debate concerning abstract knowledge in artificial grammar learning. However, there has been relatively little precise definition of exactly what abstract knowledge could be, except that whatever it is, it is more than knowledge of surface fragments of the training items, and transfer experiments are evidence for it. We therefore begin by delimiting the scope of abstract knowledge that we take to be at the heart of this debate.

Notions of Abstract Knowledge

Many different notions of abstract knowledge have been discussed in the interpretation of artificial grammar learning experiments (e.g., Dulany et al., 1984; Mathews, 1990; Mathews, Buss, Stanley, Blanchard-Fields, Cho, & Druhan, 1989; Reber, 1969, 1989; Shanks & St. John, 1994; Whittlesea & Dorken, 1993, and many others). Here we do not attempt the ambitious project of a taxonomy of notions of abstraction; we simply aim to separate out those notions of abstract knowledge that are generally accepted from those at issue here. We consider three different notions, or degrees, of abstraction:

1. In any learning process, there is clearly abstraction from the stimulus, in a relatively trivial sense. The fact that a stimulus is interpreted as, for example, a string of letters, rather than as a pattern of light and dark, indicates abstraction to the level of letters. The reality of this kind of knowledge is not at issue in the artificial grammar learning literature.

2. A second notion of abstract knowledge refers to abstraction over surface properties of individual training items. For instance, Whittlesea and Dorken (1993) point out that if subjects simply code the training items inefficiently (for instance, retaining only a random pair of adjacent letters from each exemplar) "such degenerate coding of particular items could ... make subjects' subsequent performance sensitive to the underlying structure of the set of exemplars ... because non-grammatical items often contain pairwise violations and grammatical items do not" (p. 228). Similarly, mechanisms that learn common chunks of letters in the training stimuli (Servan-Schreiber & Anderson, 1990) are thereby abstracting some information across the training items. Again, this is not the notion of abstraction that has been controversial in the literature. Even those commentators who lie furthest from the unconscious, abstract position are comfortable with knowledge of this kind (e.g., Perruchet, 1994; Shanks & St. John, 1994).

3. The third notion of abstract knowledge, which appears to be central to artificial grammar learning research, concerns knowledge that abstracts away from the specific vocabulary used in the training set. As distinct from the previous notions, subjects are seen as possessing knowledge of the grammatical structure underlying the training stimuli that is independent of the actual surface letters of the training strings. The principal source of support for this position is taken to be the fact that knowledge acquired from training items with one surface vocabulary can be transferred to test items with a different vocabulary but the same underlying structure. Unless otherwise indicated, this is the sense of *abstract knowledge* that we use below.

Theories of Artificial Grammar Learning

In parallel to these three kinds of knowledge are three accounts of artificial grammar learning and the knowledge that subjects acquire during training:

1. Exemplar-based accounts, which propose that subjects' knowledge primarily consists of relatively unprocessed representations of whole strings (Brooks, 1978; Brooks & Vokey, 1991);

2. Fragment-based accounts, which posit that the primary knowledge acquired is of two- and three-letter "chunks" of the training strings (e.g., Perruchet & Pacteau, 1990);

3. Accounts based on the acquisition of abstract knowledge, in the third of the senses described above (Reber, 1969, 1990; Manza & Reber, 1994).

These accounts differ considerably in their explanation of the transfer phenomenon: In the earliest account of implicit

learning, Reber adopts the third of these notions of abstraction, proposing that subjects are learning "abstract structures to which a particular set of symbols are assigned. They are not simply learning to string together explicit symbols" (1969, p. 119). These abstract structures can be applied relatively easily to strings that share the same underlying structure, allowing these strings to be discriminated from those with a different underlying structure.

Explicit descriptions of abstract knowledge are provided by Whittlesea and Dorken (1993), who characterize it as knowledge of "deep structure"—the pattern of repetitions occurring within each string, encoded in terms that are not tied to the surface features of the string, and by Roussel and Mathews (as cited in Altmann, Dienes, & Goode, 1995), whose THIYOS (THe Ideal YOked Subject) model explicitly encoded whether each letter of the training strings was identical or different to the previous letter.

The abstract knowledge account does not rule out the learning of some surface fragments, in particular the starts and ends of strings and common and highly salient fragments. Indeed, Manza and Reber (1994) claim that some fragmentary, explicit knowledge is acquired, but that only abstract implicit knowledge can be used during transfer, thus explaining the decrement in performance between the standard (same letter set) and transfer (changed letter set case). The very existence of above-chance transfer performance is taken as prima facie evidence of abstract knowledge.

Exemplar-based accounts (Brooks & Vokey, 1991; Vokey & Brooks, 1992) propose that training items are relatively unprocessed during training. Grammaticality judgments, in both the standard and transfer case, are assumed to be made on the basis of similarity of test strings to memorized training strings.

In the standard case, this is relatively straightforward. The string MXVVVM is similar to the test string MTVVVM. Strings that are sufficiently similar to memorized training items (on an unspecified scale) will be accepted as grammatical. In the transfer case, this similarity is based upon a process of abstract analogy; the test string BDCCCB can be seen as analogous to the training string MTVVVM. Note that although the process of abstract analogy posits *abstraction,* this takes place at *test* rather than during training. There is no acquisition of abstract knowledge per se. The problem of deciding whether the abstraction underlying generalization occurs at training or test is methodologically difficult, as has been discussed in other domains such as categorization (Barsalou, 1990). In line with this, Brooks and Vokey do not accept generalization at test as necessarily indicating abstraction during the acquisition phase.

The evidence for whole item memorization is that items that are similar to specific training items are more likely to be accepted by subjects as grammatical (Brooks & Vokey, 1991; Vokey & Brooks, 1992). As well as this similarity effect, there is also some evidence of a separate, additive grammaticality effect (i.e., similar grammatical items are more likely to be accepted than similar nongrammatical items). Brooks and Vokey propose that a process of pooling across multiple training items at *retrieval,* or a broader notion of similarity (than the one-letter-different criterion used by Brooks and Vokey) might account for more of this variance.

One blow to the exemplar account comes from recent findings by Knowlton and Squire (1994). They based their stimuli on Brooks and Vokey's, with test items that were similar to (i.e., only one letter different) or different from (i.e., more than one letter different) particular training items. They also controlled for the frequency of particular bigrams and trigrams, ensuring that these were equally common in both similar and different test items. When chunk frequencies were carefully controlled for in this manner, the effect of similarity between test items and specific training items disappeared. Therefore, there is no evidence that subjects are comparing test items with memorized whole training items, as the exemplar account proposes. However, Knowlton and Squire did not run a transfer condition with these controls, and it is still possible that effects for similarity to specific training items might occur under transfer conditions.

The core claim of fragment-based accounts (e.g., Perruchet & Pacteau, 1990; Servan-Schreiber & Anderson, 1990) is that subjects' knowledge primarily consists of chunks of letters (two, three, or more letters long). At test, they classify as nongrammatical those strings that contain unfamiliar chunks.

As Perruchet (1994, p. 226) observes, "the occurrence of transfer to a new letter set raises some problems for the [fragment knowledge] account." One solution to this has been the proposal (Shanks & St. John, 1994) that the low levels of observed transfer (typically .55–.6 of classifications correct, where chance would be .5) can be accounted for by subjects' explicit knowledge of the deep structure (repetition patterns). For instance, with the standard materials (see Figure 1 and Table 1), subjects might note that the first two letters of a string are always different from each other. Classifying all strings for which this is not the case as nongrammatical and all others as grammatical would result in a score of .58 of classifications correct.

Perruchet's solution (1994) to the problem of transfer is somewhat different. He argues that transfer effects have not been conclusively demonstrated, on the grounds that the small effects observed in transfer studies are not larger than effect sizes observed for control subjects, in the relatively rare cases in which control conditions have been run. Furthermore, Perruchet (1994, p. 226) argues that even if transfer effects are real, "the effect does not appear to be strong enough to prompt questioning" of fragment-based accounts.

We agree with Perruchet's point that transfer experiments should be run with proper controls. Indeed, we argue that this is particularly important, as the observed above-chance levels of control performance are easily explicable, and may be routine, rather than exceptional. We also discuss a number of important related methodological issues, which may have implications for artificial grammar learning in general but are of particular concern for transfer studies, given the magnitude of their observed effects.

Nonetheless, we suggest that transfer is probably a real phenomenon and is observed in relatively well-controlled

Table 1
The Standard Set of Acquisition and Test Strings, First Used by Reber and Allen (1978)

Training	Test	
	Grammatical	Nongrammatical
MSSSSV	VXSSSV	VXRRS
MSSVS	MSSSV	VXX
MSV[a]	MSSVRX	VXRVM
MSVRX	MVRXVS[a]	XVRXRR
MSVRXM	MSVRXV	XSSSSV
MVRX	MSVRXR	MSVV
MVRXRR	MVRXM	MMVRX
MVRXSV	VXVRXR	MVRSR
MVRXV	MSSSVS	MSRVRX
MVRXVS[a]	VXRM	SSVS
VXM	MVS	MSSVSR
VXRR	MSVS	RVS
VXRRM	MSSV	MXVS
VXRRRR	MVRXR	VRRRM
VXSSVS	VXRRR	VVXRM
VXSVRX	VXSV	VXRS
VXSVS	VXR	MSRV
VXVRX[a]	VXVS[a]	VXMRXV
VXVRXV[a]	MSV[a]	MSM
VXVS[a]	VXRRRM	SXRRM
	VXSSV	MXVRXM
	VXV	MSVRSR
	VXVRX[a]	SVSSXV
	VXVRXV[a]	XRVXV
	MVRXRM	RRRXV

Note. For nongrammatical strings, underlining indicates the point of grammatical violation. The actual identity of the letters is not necessarily consistent across studies; for instance, most studies used T instead of S.
[a] Indicates strings that are present in both the training and test sets.

studies (e.g., Altmann et al., 1995). However, we argue, and demonstrate, that the vexed question of transfer effects does not have the theoretical significance that it has been accorded: Transfer implies neither abstract knowledge nor the memorization of whole exemplars.

Methodological Issues

The Importance of Controls

Artificial grammar learning studies have largely failed to use any controls to guard against the possibilities that some (or all) of subjects' advantage over chance performance may be due to (a) learning during the test phase, as opposed to knowledge acquired during training, or (b) grammatical and nongrammatical strings being distinguishable from each other, without benefit of prior training. Instead, it has been assumed that subjects could not learn anything of value during the test phase and that grammatical and nongrammatical strings could not be distinguished from each other without prior exposure to the training strings. The obvious baseline, against which experimental subjects have been compared, is chance performance.

In the first study to use controls, Dulany et al. (1984) observed control performance at .56,[1] performance reliably above chance. Dulany et al. comment on the importance of using control subjects, but do not discuss this finding much further. D. E. Dulany (personal communication, May 11, 1994) has suggested that the effect might be due to the inclusion of certain items (such as MTV—Music TeleVision, and MTM—Mary Tyler Moore) that were both familiar to the subjects and grammatical. It is indeed quite possible that the effects of individual stimulus items are important, as we argue below. However, the majority of subsequent studies have not used controls, presumably because it was assumed that this was an aberrant observation.

To our knowledge, four other relevant studies have used control groups. Perruchet and Pacteau (1990) report control groups who did not significantly differ from chance performance, as do Altmann et al. (1995), and St. John and Shanks (in press). Redington and Chater (1994) found control subjects performing at .57, reliably above chance.[2] Thus, of these five studies, two have found nonchance control performance. Additionally, Z. Dienes (personal communication, June 1, 1994) has observed control subject performance as high as .60.[3]

These nonchance findings suggest that control (and possibly experimental) subjects can learn something about the distinction between grammatical and nongrammatical items during the test phase.

How Control Subjects Could Learn

During the test phase, control (and experimental) subjects are effectively in an unsupervised learning situation. They must make their grammaticality judgments in the absence of

[1] With the exception of Gomez and Schvaneveldt (1994), all of the studies we discuss used equal numbers of grammatical and nongrammatical items and reported performance in terms of the proportion of items correctly classified as grammatical or nongrammatical (where the performance expected from purely random responding is .5). When discussing Gomez and Schvaneveldt's results, performance is reported in terms of *D* (the proportion of correct rejections minus the proportion of misses, see Perruchet & Pacteau, 1990). Here, the value expected from purely random responding is 0.

[2] The task used was a guessing game paradigm, which is similar to Dienes et al.'s (1991) stem letter detection task. Subjects reconstructed each test item by successively guessing the next letter in the sequence and then made a grammaticality judgment for the item, which is the measure of interest here. On conventional performance measures experimental subjects' grammaticality judgments were comparable, both quantitatively and qualitatively, to judgments obtained in the standard paradigm (for instance, Dienes et al., 1991; Perruchet & Pacteau, 1990; Reber & Allen, 1978).

[3] Subjects performed only the grammaticality judgment task, and the effect was observed both for subjects who were informed that they were untrained controls and for subjects to whom the task was presented as a discrimination task, with no mention of training (Z. Dienes, personal communication, August 15, 1995).

any error signal, or feedback, as to the correctness of their responses.

Subjects are aware that 50% of the test items are grammatical, and 50% are not. Thus, as regards the underlying grammar, they are exposed to positive evidence, albeit very errorful; half of the items are known to contain violations of the underlying rules or grammar.

Some types of grammatical violation, however, may be relatively obvious. If subjects are gradually acquiring knowledge about what a typical[4] item looks like, in the sense that it matches their knowledge of previous items, as they proceed through the test set, then subsequent items will fit this notion of typicality to a greater or lesser degree. Since most nongrammatical items (in the test sets usually employed) differ very little from the grammatical ones, it seems reasonable to suppose that control subjects will reject those items that seem less typical and accept those that seem more typical. As long as a subject's notion of typicality is correlated, even weakly, with grammaticality, then above-chance performance may be observed.

In what we shall term the *standard* materials (see Table 1), grammatical items always commence with the letters V or M. Of 25 nongrammatical items, 17 obey this constraint. Thus, it is possible that the 16% of test items that do not follow this pattern are likely to appear atypical to the subject and to be rejected as nongrammatical. In Redington and Chater's (1994) data, of the 8 test items violating this constraint, 6 are amongst the 10 easiest nongrammatical items for control subjects to classify correctly. It is known that experimental subjects are sensitive to initial bigrams (e.g., Perruchet & Pacteau, 1990; Reber & Allen, 1978). These initial bigrams are only an obvious example of the many cues to which control subjects may attend and that might be correlated with grammaticality.

It is possible to argue that, if effects of learning during testing are observed for control subjects, then they should also be observed for experimental subjects. However, a ceiling effect is likely to operate here—experimental subjects have already had good evidence of what grammatical items are like, in the training phase, and hence can gain little from learning from the much less reliable evidence available at the test phase. In addition, experimental subjects are presumably much less motivated to attempt to use this source of information, as compared with control subjects, who have no other information to draw upon.

It is plausible that transfer subjects might rely much more on the information presented in the new letter set than whatever (old-letter-set-based) knowledge they acquired during training. This possibility strengthens the case for the routine use of untrained controls in transfer studies.

One might expect that if control subjects are learning during the test phase, it will be possible to see order effects. However, these may be highly dependent on the strategies that subjects use and the particular items that they have seen. It is also possible that once the subjects have seen many items, their notion of typicality may grow to include previously rejected items. Thus no clear pattern of performance can be predicted, aside from chance performance on the first item. Additionally, it is possible that subjects' notions of typicality will be correlated negatively with grammaticality, and thus that subjects will perform below chance. Where this is the case, a more stringent comparison for experimental subjects is against chance performance, rather than against that of the controls. Otherwise, chance performance by the experimental subjects could be taken as evidence for learning, when they are compared against below-chance controls. Thus, to show learning from exposure to the training materials, experimental subjects should perform reliably above both chance and controls.[5]

We conjecture that the crucial difference between the performance of control subjects in the above-chance instances and Altmann et al. (1995), Perruchet and Pacteau (1990), and St. John and Shanks (in press) is that in the former cases, subjects' tasks forced them to pay attention to the structure of the items during the test phase (in the Dulany et al. study, subjects underlined the part of the string that they perceived as relevant to grammaticality; in Redington & Chater, 1994, they performed the guessing game task, in which test items were reconstructed through a sequence of guesses), whereas in Altmann et al. and Perruchet and Pacteau's studies, subjects performed only the standard grammaticality judgment task. It is therefore possible to argue that control subjects in many previous studies would not in fact have deviated reliably from chance performance. Dienes's above chance results lessen this possibility, as control subjects were observed to perform reliably above chance in the absence of a second task (Z. Dienes, personal communication, June 1, 1994, and August 15, 1995). Additionally, even when control subjects have been observed to perform at chance, this may simply be because control subjects are influenced by different motivational factors from experimental subjects, and this may crucially affect performance, as we now argue.

Motivation and Belief as a Factor in Artificial Grammar Learning Studies

Control subjects who have not seen any previous material have little reason to believe that anything but a random strategy of responding is worthwhile, especially when they have no task other than to make grammaticality judgments. Hence they are unlikely to examine the test items with any care. This might suggest that some control subjects respond completely at random, whereas others seriously engage the task. Thus, it may be that the observed performance of control groups significantly underestimates the level of per-

[4] It is only necessary that this knowledge involve abstraction in the second sense described above, that is, of features that are common across items, or that subjects are memorizing exemplars and matching subsequent items to them; it could be any kind of knowledge concerning the items they have seen. This is by no means necessarily the sense of typicality suggested by Vokey and Brooks (1992).

[5] Altmann, Dienes, and Goode (1995) made some comparisons with controls whose performance is slightly below chance, but their effect sizes were such that this is probably not of importance.

formance it is possible to obtain without exposure to the training items. Unlike control subjects, experimental subjects have better reason to believe that they possess information relevant to the task and to be motivated to attempt good performance. In particular, they may be expected to examine the test items in detail and consider their responses carefully. The potential influence of these factors is unknown, but in view of the small size of many artificial grammar learning effects, it may be of considerable importance to control for both exposure to the relevant training stimuli and perceived self-competence and level of motivation.

How is it possible to control for these factors? An obvious suggestion is to give control subjects a random string of letters (as Altmann et al. did for a second control group), or strings from some other grammar, in place of the training phase received by the training subjects. However, this precaution alone may not be an adequate control, because experimental subjects will see items to which the knowledge they have acquired is applicable, whereas control subjects will see a baffling array of relatively novel items, hence they may be somewhat discouraged relative to the experimental subjects. In our experience, all subjects, and in particular control subjects, find the task confusing, and on occasion, distressing. Attempting to classify a set of stimuli on the basis of irrelevant training is likely to be equally discouraging.

As we see below, it may not be possible to control for all important factors simultaneously. We present one possible approach, using a crossover design, which attempts to provide relatively stringent controls for many factors.

A Crossover Design for Artificial Grammar Learning Studies

We propose that some motivational factors can be controlled for by the use of a crossover design: Two experimental groups are trained on items from distinct grammars (i.e., which generate nonoverlapping sets of items) with similar properties (e.g., length of string, same letter set[6]). In the test phase, each subject group receives a test set half composed of items from the grammar their training items were drawn from (equivalent to the grammatical items in the standard paradigm), and half composed of items from the other group's training grammar (equivalent to nongrammatical items). Similar designs have been used by Brooks and Vokey (1991), Dienes and Altmann (in press), and Vokey and Brooks (1992).

Each subject group acts as the other group's control, because they both see the same test items, and the congruence between the training and the test items is the same for each group. One crucial observation is whether the subjects from the different experimental groups categorize the test set differently (relative to the grammar to which they have been exposed). If so, then this is presumably due to their initial training experience. An additional control group is also required, who would receive no training experience, in order to assess the baseline discriminability of items from the two grammars. For instance, if the strings of one grammar were always symmetrical, and those of the other were not, then this baseline might be well above chance performance. If experimental subjects are observed to perform at above control group performance,[7] then it seems reasonable to conclude that this is again due to their initial training.

Of course, it is still possible that untrained control subjects might perceive themselves as ill-equipped to perform the task and so not make as great an attempt as experimental subjects. However, test instructions that framed the task as, for instance, a problem in distinguishing between items from two different grammars might lessen this difficulty, by making the task meaningful for control as for experimental subjects. A second implication of this design, pointed out by P. Perruchet (personal communication, June 17, 1994), is that in many cases it may be difficult, or impossible, to design two opposing grammars that differ by permitting or forbidding single letters in particular positions—the kind of violations that have heretofore characterized nongrammatical test items.

We do not propose this design as an absolute prescription. Indeed, there may be no ideal control for artificial grammar learning, and experimenters will obviously have to tailor their controls to the particular question under investigation. Rather it is an attempt to stringently control for learning during the test phase and for possible motivational factors. Although the importance of such factors and the conditions under which they play a role in implicit learning are effectively unknown, this is precisely why experimenters should try to control for them. It is conceivable that more stringent controls would not radically alter the conclusions of the existing experimental literature, and we suspect that this is true. However, the amount of care and attention that subjects pay to the test items and subjects' motivation are potentially so variable, and the effect sizes observed in experimental subjects are typically so small (especially in the context of transfer), that these factors cannot be ignored. We believe that many present and past studies of artificial grammar learning are potentially fatally flawed by failing to provide controls of any kind.

The Item-as-Fixed-Effect Fallacy

The discussion of possible learning in control subjects highlights the fact that the strategies that subjects may successfully use to tackle discrimination tasks may be highly dependent on the specifics of the training and test stimuli. This raises a further possible source of methodological concern with artificial grammar learning studies: that the use of the same stimuli for all subjects means that

[6] We discuss a range of criteria it may be worthwhile to control for in artificial grammar learning experiments below.

[7] Obviously individual control subjects performance must be assessed in terms of the distinction between the underlying grammars—they cannot classify the strings with reference to previous training strings.

conclusions cannot be drawn with respect to learning the grammar in general, but only for those particular stimuli.

The specific grammatical strings used in training and grammatical test items are a subset of the strings allowed by that grammar; the nongrammatical test items are drawn from a larger population of nongrammatical items. For us to be able to draw any general conclusions concerning the learning of a particular grammar, it is desirable that these populations are specified and sampled from and that experimental design and statistical analysis is conducted accordingly. In artificial grammar learning experiments, experimenters have tended not to treat variability that might be due to particular choices of item as a random factor, varied across subjects, and taken into account in statistical analysis (Vokey & Brooks, 1994, p. 1509, make a similar point).

Nonetheless, the stimuli used have generally been chosen with some care. Grammatical items are selected to be "representative" of the grammar, in some intuitive sense; and the nongrammatical violations are specified by a few relatively simple distortions of grammatical stimuli. It is also the case that because much experimentation has been concerned with comparing different tests of explicit knowledge, and other procedural differences, the same (nonrandomized) stimuli have been used across studies to facilitate this comparison.

Strictly speaking, this is legitimate as long as the conclusions drawn from the experiment are taken to apply only to the particular materials used (including such factors as the meaningfulness to the subjects of fragments such as MTV and MTM) rather than used to draw conclusions about performance with the grammar in general. That such general conclusions cannot legitimately be inferred has been pointed out by Clark (1973) and has been accepted as standard in many areas of psychology.

It can be argued (Z. Dienes, personal communication, June 3, 1994) that one can generalize from current data on grounds of plausibility—given the observed effects with a variety of grammars, it seems highly implausible that these effects are due solely to the particular choices of items or order of presentation that experimenters have adopted. But given that effect sizes in typical transfer studies are so small, even very small effects of, for example, particular vocabulary items, could lead to spurious transfer results, unless randomization is carried out.

It is not possible, of course, to randomize *all* factors that might conceivably influence the experimental outcome (e.g., time of year, time of day, sex of subjects). As in other areas of experimental psychology, it is therefore important to choose which factors should be randomized and which can simply be ignored. In the case of artificial grammar learning, assessing whether or not the particular choice of grammar or vocabulary is important is difficult, because experiments tend to focus on a small number of standard stimuli. We therefore suggest that rather than choosing which factors to randomize arbitrarily, experimenters should choose these factors on the basis of empirical research concerning the importance of different grammars and letter sets (e.g., letters, numbers, symbols, Chan, 1992; tones, Altmann et al., 1995). Until this research is conducted, it is difficult to assess the extent to which the problem of fixed effects casts doubt on current empirical studies.

We believe that it is generally preferable to err on the side of caution and that even where the constraints of the grammar are such that the number of test items is very limited, it is typically straightforward to randomize the division of grammatical items between training and test sets, the choice of nongrammatical items, and the order of presentation of training and test items. We believe that these steps should be taken where appropriate in order to minimize the possibility of spurious effects.

What Should Be Controlled for in Artificial Grammar Learning Experiments

Let us summarize our points so far. We stress that experimenters should take great care to stringently control for possible sources of contamination. We further suggest that experimenters might routinely use two grammars and have subjects exposed to either one learn to discriminate within a test set of items from both grammars and that control subjects who are exposed to no training items be included to test the discriminability of the two grammars without prior experience. We also argue that potential effects of motivation between conditions must be borne in mind and minimized if possible. Furthermore, we suggest that materials should be randomized across subjects when appropriate, rather than treated as a fixed effect, in order to obtain results of the maximum generality.

We now turn to the rather different, but equally important, issue of controlling properties of the learning and test stimuli. Specifically, where the focus of research is what is learned in artificial grammar learning tasks, then in addition to the precautions outlined above, it is necessary to select grammars and to sample stimuli from them so that the manipulations of interest are not confounded with simple alternative hypotheses. Just as psycholinguists routinely control for word frequency, cloze value, and the like, so researchers should routinely control for factors such as bigram and trigram frequencies, if they intend to eliminate hypotheses based on knowledge of such simple fragments, rather than, for example, memory for whole strings, the extraction of an underlying grammar, and so forth. Perruchet (1994) and Knowlton and Squire (1994) have elegantly demonstrated how Vokey and Brook's (1992) manipulations were confounded with such factors.

Are Transfer Effects Real?

As Perruchet (1994) has observed, the magnitude of practically all transfer studies is so small that the absence of control subjects makes the interpretation of the results extremely difficult. We have noted a variety of additional difficulties that reinforce this concern. We agree with Perruchet that the studies by Brooks and Vokey (1991) and Mathews et al. (1989) do not satisfactorily demonstrate

transfer, and we also follow Perruchet (1994) in limiting ourselves to the standard paradigm. Hence, we do not consider Reber's (1969) study, which measured memorization advantage rather than discrimination performance. However, five more recent studies claim to show significant positive transfer.

Whittlesea and Dorken (1993), Experiment 5c, reported transfer performance of .53, with the nontransfer experimental group performing at .59. They concluded that this "indicates availability of deep-structural knowledge" (p. 243). However, because they had no control group, and given the marginal above-chance level, this claim is impossible to assess.[8]

Gomez and Schvaneveldt (1994) demonstrated reliable transfer performance in subjects taught strings, as compared with subjects taught only isolated bigrams, who showed no such transfer effect. They used *D,* the percentage of nongrammatical items correctly rejected minus the percentage of grammatical items incorrectly rejected (see Perruchet & Pacteau, 1990, p. 267) as their measure of performance and reported that for strings containing illegal pairs, subjects' *D* scores were 13.68, whereas comparable subjects taught strings from a different grammar scored only 2.35 (Experiment 4, p. 406).

Altmann et al. (1995), Experiment 1, in an investigation of cross-modal transfer, found transfer effects of .56 when subjects trained on letter strings were tested on sequences of musical tones and .54 when transfer was from tone sequences to letter strings, with tone sequence controls performing at .49 and string controls performing at .50. A control group trained on random tones performed at .48 when tested on strings of letters. Although these effects are small, they appear to be reliable and relatively well controlled. Altmann et al. demonstrated similar reliable transfer effects in three further experiments.

Similar effects were shown by Dienes and Altmann (in press), with transfer between sequences of color names and sequences of color patches and between letter strings and sequences of color patches. Transfer performance in either direction was in excess of .6. This experiment used a crossover design similar to that described above, although there was no untrained control group. This leaves open the possibility that the stimuli from the two grammars may have been highly discriminable in the absence of any training, but it does strongly suggest the occurrence of positive transfer.

St. John and Shanks (in press) reported transfer subjects performing at .59 of classifications correct, whereas controls performed at .51 (nontransfer subjects performed at .60). This would appear to be another strong demonstration of transfer.

Manza and Reber (1994) reported six experiments, with transfer performance ranging from .53 (Experiment 5) to .61 (Experiment 3). Although the levels of performance for some of their experiments are suggestive of transfer, they did not run control groups of any kind, rendering interpretation of their results difficult.

In addition, in our own research, we have also observed transfer effects, with memorization to criterion as training, and the standard discrimination task, at levels of .60, which is reliably above chance. Controls with no training materials performed at .47 and controls trained on a different grammar performed at .49 (Morrison, 1994).

Although many of the methodological concerns expressed above apply to aspects of all of these studies, the results of Gomez and Schvaneveldt (1994), Altmann et al. (1995), and St. John and Shanks (in press) appear to provide fairly conclusive evidence of a genuine transfer effect.

The Theoretical Implications of Transfer Effects

The main reason for the interest in transfer effects is that they are seen to have implications for the nature of the subjects' representation of the knowledge acquired during learning. In particular, the claim is that transfer effects must be mediated by abstract knowledge. We propose that although transfer effects, by definition, imply *abstraction,* this might be very different from the standard conception of abstract knowledge.

Two distinct types of accounts of transfer performance have been proposed. One uses the third notion of abstract knowledge introduced above: knowledge of the structure of the strings that is independent of their surface vocabulary. Whittlesea and Dorken (1993) call this the "deep structure" of the string, which includes the patterns of repetition within items and the commonality of these patterns across items. For example, the deep structure of the string VXSSV might be represented as □◇△△□. A common deep structural feature of the grammar in Figure 1 is that no grammatical strings begin with the repetition of the same symbol. Similar views of what is learned are vague, but appear to imply that common structural features of the training items are represented in terms of a grammar in some way akin to that which was used to generate the training items (Reber, 1969), or in terms of abstract rules (e.g., Mathews, 1990). An alternative view, advanced by Brooks and Vokey (1991), uses the first notion of abstract knowledge introduced above: that a very shallow representation is formed, which encodes the surface structure of whole training items. Subjects are assumed to base transfer performance on a process of "abstract analogy" between the novel test item and memorized whole training items. In this process, deep structure is not encoded during training but is only computed for purposes of making the analogy.

[8] In a footnote, Whittlesea and Dorken (1993, p. 242) argued that they did not run a control group because their interest in these experiments was primarily in the difference in success between different conditions of test, for example, the interaction between training task and original or novel test letter set. Furthermore they argued that because their "incidental repetition" subjects (whose training consisted of naive repetition of grammatical strings as distractors in a task irrelevant to the experiment) performed at only 51.5% accuracy on the transfer task, this value can be taken as a ceiling for control performance. But if this is the case, they should have compared the performance of the relevant transfer group here against this ceiling rather than against chance.

Both these approaches suggest that the knowledge that mediates transfer is something more than that of surface fragments.

We show below that transfer experiments, even where they show large effects (e.g., .60 of classifications correct, St. John and Shanks, in press) are entirely consistent with various accounts based on the memorization of surface fragments, for which there is good experimental evidence (Dienes, Broadbent, & Berry, 1991; Perruchet & Pacteau, 1990; for a review, see Shanks & St. John, 1994), together with simple processes of abstraction occurring at test, rather than during learning. This alternative account is similar to Brooks and Vokey's, apart from the emphasis on the primacy of fragments rather than whole exemplars. This view is also suggested by Altmann et al.'s (1995) notion of "domain-independent processes." We suggest that this family of relatively simple and empirically supported fragment-based hypotheses should serve as "null hypotheses" against which hypotheses involving more elaborate representations should be compared.

What Knowledge Is Required to Account for Transfer?

First, let us draw a distinction between knowledge that is abstract and knowledge that is "abstractable." Suppose we assume, hypothetically, that during memorization of the training stimuli, subjects perfectly memorize each of the training items. If subjects truly abstract deep structure (let us assume for the moment that they discard the surface structure; the actual identity of the symbols in the learning items), then the representation of the strings could be conceived as something like Figure 2A. Conversely, if subjects learn only the surface structure, then the representation might look something like Figure 2B. However, as Brooks and Vokey's (1991) abstract analogy process makes clear, there is no information in the first representation that is not also available from the second; the deep structure is "implicit within," or "abstractable from" the representation of the surface structure.

Furthermore, this distinction applies whether one takes "abstract" knowledge to mean, for instance, a partial but veridical representation of the grammar underlying the strings (Reber, 1967, 1969), or a set of rules defining what is and isn't grammatical, or any knowledge that allows one to make grammaticality judgments at better than chance. Such knowledge could always be abstracted from knowledge of the surface structure of the training items. This point applies equally in other cognitive processes that involve generalization from experience. For instance, Barsalou (1990) argues that exemplar-based models of categorization and memory may, with certain restrictions, be indistinguishable from accounts involving abstraction during acquisition.

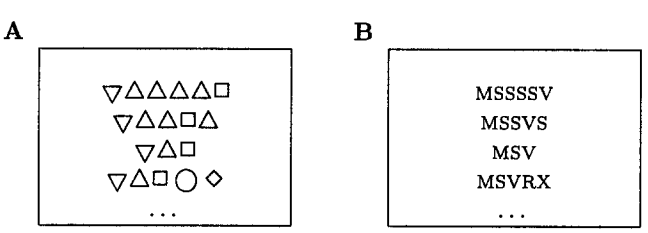

Figure 2. Deep structure (A) and surface (B) representations of the learning strings from Table 1.

Just as whole exemplars, such as MSSSSV, can be considered analogous to the transfer test item, such as JDDDDB, in that they both share a common deep structure, the string initial fragment MS can be considered analogous to the initial fragment JD of a (transfer) test string. This analogy implies that within this test string, J corresponds to M, and D corresponds to S. Thus, one could imagine that subjects might try to fit other surface fragments to the string, respecting these mappings. The "goodness" of the string might then be assessed in terms of how extensively it could be fitted to the surface fragments, given the implied mapping. Thus abstraction processes are necessitated only for the purpose of comparing memorized surface fragments with surface level representations of the test items. Although definitions of abstraction are hard to pin down in the literature, it would appear that no researchers count surface-based processing of this kind as involving abstract knowledge. Indeed, if theorists were to view this as involving abstract knowledge, then this notion would appear to be so bland that it could hardly be a point of theoretical contention.

As we mention above, we believe that the well-supported fragment learning theories should serve as null hypotheses for those wishing to propose more elaborate claims. We do not propose the specific "analogy" processes described above as serious psychological proposals. However, we illustrate in detail below that knowledge of simple surface fragments together with simple mapping/abstraction strategies are capable of matching the observed performance of subjects in the transfer literature and, in many cases, of significantly exceeding it. We therefore conclude that hypotheses invoking abstract knowledge or whole-item representations are not required to account for transfer performance.

Two Kinds of Transfer Experiment

All except one of the transfer studies involves a single change of vocabulary between the training and the test items; the mapping from old to new symbols applies throughout the entire test set. If subjects can find this mapping, then transfer performance can rely purely on translation from surface level memories for aspects of the training items into the vocabulary of the test stimuli. We refer to this single change of letter set as *simple transfer.* In the other kind of study, the mapping between the original training vocabulary and the symbols of the new vocabulary is randomly assigned for each and every test item, rather than just once, between the training and test phase. We refer to this change of letter set for each test item as *randomly changing transfer.* The Whittlesea and Dorken (1993) study is the sole published example using this approach, although

it is difficult to interpret their results because they did not use a control group of any kind. However, recent unpublished studies by Z. Dienes (personal communication, June 1, 1994, and, August 15, 1995) and by Redington and Chater (1996) which do use controls (a cross-over design and untrained controls, respectively), suggesting that subjects can transfer their knowledge under these conditions. Successful performance on the randomly changing transfer task would appear to provide a more stringent test for the presence of abstract knowledge. Transfer performance on both kinds of task can be accounted for by surface fragment knowledge, together with a process of on-line abstract analogy.

Simple Transfer

In the simple (single letter set change) transfer task, finding the correct mapping can be surprisingly easy. We give here a trivial example of one of the numerous strategies that could discover the mapping without recourse to abstract knowledge.

We do not intend to argue that subjects are consciously or unconsciously pursuing such strategies; only that the very possibility of such a multitude of simple strategies that do not require abstract knowledge in the sense at issue in the literature shows that the inference from transfer effects to abstract knowledge is unsound.

Let us assume that the standard training and test items (see Table 1) are in use. We assume that the subject recalls only bigram information from the initial training set. We here use only the string initial bigrams in discovering the mapping. In the training set, MS, MV, and VX are the initial bigrams.

The first test item that the subjects see is JDHBHF, and they respond at random. The second test string is BFHHHH, and again they respond at random. The third string is JBHFJ. The subject now deduces that an initial J can be followed by one of two letters, and assuming that the three test strings do not contain any initial bigram violations, J = M, B = V (because it can both follow M and commence a string), and F = X, because it is the only letter that can follow an initial V. Additionally, D = S (because it is the only letter, apart from V, that can follow an initial M, as in the first test string). Four out of five mappings are now known, so H = R by elimination.

Of course, the subjects may be wrong in their assumption that there is no initial bigram violation, but first, many subjects will be correct, because only 16% of test strings have such violations, and second, the error may rapidly be corrected as more information accumulates. Given more information than initial bigrams, or more examples, there are many possible constraints that would allow the subject to solve the problem.

Randomly Changing Transfer

In the randomly changing transfer task, where the mapping changes for each new test item, the only relation between training items and grammatical test items is their deep structure. It might be assumed that transfer to such stimuli must show that this structure is being stored (or that entire training items are being encoded, and abstract analogy used, as Brooks and Vokey, 1991, suggest). However, we constructed a number of simple "toy" models, based on surface fragments, in the spirit of previous such models (Perruchet, 1994; Perruchet & Pacteau, 1990). Below, we show that these models are capable of demonstrating significant transfer effects, and even better same letter set effects with the Whittlesea and Dorken stimuli.

Additionally, the models apply to the simple transfer case as well as to randomly changing transfer. We shall show that this class of models can perform similarly well on stimuli from the other relevant transfer experiments.

Toy Models

We call the models in this section *toy* models because their purpose is to show that information about fragments is sufficient to account for transfer performance. Thus they constitute a feasibility proof that fragment knowledge, together with abstraction at test, is compatible with current experimental findings. They are not intended as detailed psychological models, although the basic assumptions that they make—concerning fragment knowledge and abstraction at test—are intended to be psychologically plausible.

The models below work on the following principles: they learn some surface features of all of the training strings (e.g., the bigrams that occur in the training strings). When judging a test string, they consider it a *good* string if it contains no novel features (e.g., a bigram that did not occur in the training strings), and otherwise they consider it a *bad* string. Thus, if only the bigram VX and XV occurred in the training strings, the string VXVX is good, but the string VXXV is bad, as the bigram XX is unknown.

In the transfer case, if there is any possible mapping between the original letter set and the new letter set so that the string can be considered good (i.e., it contains no novel features after the mapping has been performed), then the string is considered good. If no such mapping is possible, the string is considered bad. The possible mappings between letter sets are constrained so that each old letter maps to one, and only one, new letter (the mapping is consistent and unique). This constraint, though rarely explicitly stated, is implicit within the description of the task. Thus, again given the bigrams VX and XV, BFBF is a good string, because given the mapping V = B and X = F, all of its bigrams are known, whereas BFFB is bad, because there is no possible mapping that results in no unknown bigrams.

The decision criterion. Additionally, a decision rule is required, specifying how strings are to be classed as grammatical or nongrammatical. For the purpose of these toy models, we assume that all of the strings can be accessed simultaneously. This is not generally the case in these experiments; subjects are at the very least encouraged to proceed through the response set one string at a time and are often prevented from, or encouraged against, consulting

their previous responses. To make these models more realistic, we could have used an on-line decision criterion (for instance considering the goodness of the current string relative to a number of previous strings, e.g., Servan-Schreiber & Anderson, 1990) and obtained similar results to these. However, we wanted to keep these models as simple as possible, and this would be an additional and largely arbitrary and irrelevant technical point.

The criteria used for all models, except for the St. John and Shanks (in press) and the Gomez and Schvaneveldt stimuli, were as follows:

1. Always ensure that half of the responses are grammatical, and half nongrammatical.
2. If half or less of the test items are good, accept these as grammatical and allocate the remaining grammatical responses, and the nongrammatical responses, to the remaining items at random.
3. If more than half of the test items are good, reject all of the bad test items and allocate the remaining nongrammatical responses and all of the grammatical responses to the good items on a random basis.

Thus, on average, a subject who viewed g grammatical items as good, and ng nongrammatical items, where $g + ng$ was half or less of the total number of test items, T, would score as follows:

$$p(\text{correct}) = \left\{ g + \frac{1}{T - (g + ng)} \left[\left(\frac{T}{2} - g \right) \left(\frac{T}{2} - (g + ng) \right) + \left(\frac{T}{2} - ng \right) \left(\frac{T}{2} \right) \right] \right\} / T. \quad (1)$$

In the cases where $g + ng$ was greater than $T/2$, subjects would on average score as follows:

$$p(\text{correct}) = \left\{ \frac{T}{2} - ng + \frac{1}{g + ng} \left[g \left(\frac{T}{2} \right) + ng \left(g + ng - \frac{T}{2} \right) \right] \right\} / T. \quad (2)$$

These formulas of course assume that half of the test items are grammatical, and half nongrammatical.

In the St. John and Shanks (in press) study, subjects' grammaticality judgments were a forced choice between two test items, instead of a grammatical–nongrammatical decision about a single item. Of each item pair, one item was grammatical and one nongrammatical, the pairs being constructed randomly without replacement from the set of test strings. The model was assumed to choose the good string of a mixed pair, or randomly if both strings were good or bad. Where the numbers of grammatical and nongrammatical items are equal, g and ng are the numbers of good grammatical and nongrammatical strings, and T is the total number of strings, the expected proportion of correct decisions is given as follows:

$$p(\text{correct}) = .5(1 - ng/N + g/N). \quad (3)$$

For the Gomez and Schvaneveldt (1994) stimuli, the numbers of grammatical and nongrammatical stimuli were unequal. They reported their results in terms of D. Here, we assumed that subjects accepted all good strings and rejected all bad ones. Thus, where g and ng were the numbers of grammatical and nongrammatical good strings, and T_g and T_{ng} were the total number of grammatical and nongrammatical strings,

$$D = \left(\frac{T_{ng} - ng}{T_{ng}} - \frac{T_g - g}{T_g} \right) \times 100. \quad (4)$$

Classes of models. We used three different classes of models, dividing them in terms of the type of knowledge that they were assumed to acquire in the training phase.

1. Here the model knows either all of the bigrams or all of the trigrams that occur in the training items, in terms of their surface letters. Some models are assumed to represent beginnings or ends of strings (or both) with explicit START and END symbols. For example, bigrams might include START V and trigrams might include SV END. We considered the following models:
   a. bigrams alone;
   b. bigrams, with an explicit START symbol;
   c. bigrams, with an explicit END symbol;
   d. bigrams, with an explicit START and END symbol;
   e. trigrams alone;
   f. trigrams, with an explicit START symbol;
   g. trigrams, with an explicit END symbol;
   h. trigrams, with an explicit START and END symbol.
2. The second class of models also acquire surface bigrams or trigrams, but here they acquire only those fragments that occur at the start or end (or both) of the training items. Thus the model might represent, for instance, that MV is a legal initial bigram or that RRM is a legal final trigram. We considered the following models:
   i. initial bigrams only;
   j. final bigrams only;
   k. both initial and final bigrams;
   l. initial trigrams only;
   m. final trigrams only;
   n. initial and final trigrams.
3. The final class of model has only one instance:
   o. exact match,

and acquires perfect, surface representations of each whole training string. Thus the model knows, for instance, that MVRXVS occurred amongst the training items.

Simulation Results

Table 2 below shows the proportion of correct responses obtained by our models in both standard and transfer conditions for the stimuli used by Brooks and Vokey (1991), Whittlesea and Dorken (1993), and St. John and Shanks (in press). Table 3 shows similar results for the stimuli used by Altmann et al. (1995), and Table 4 shows the results for Gomez and Schvaneveldt's (1994) stimuli. For each set of

Table 2
Proportion of Classifications Correct From Observation and for Each of the Toy Models for the Stimuli From Brooks and Vokey (1991), Whittlesea and Dorken (1993, Experiments 4 and 5), and St. John and Shanks (in press, Experiment 1)

	Stimuli						
	Brooks & Vokey		Whittlesea & Dorken			St. John & Shanks	
				Exp. 5			
Source	Same	Dif.	Exp. 4	Same	Dif.	Same	Dif.
Observation	.60	.56	.57	.59	.53	.60	.59
Model							
1. Bigrams	.61	.50	.50	.54	.50	.78	.50
2. Bigrams w/start	.61	.50	.50	.54	.51	.78	.50
3. Bigrams w/end	.63	.50	.50	.54	.50	.78	.50
4. Bigrams w/both	.63	.50	.50	.54	.51	.78	.50
5. Trigrams	.68	.62	.59	.71	.51	.90	.68
6. Trigrams w/start	.68	.69	.73	.71	.54	.90	.70
7. Trigrams w/end	.65	.63	.62	.71	.58	.90	.70
8. Trigrams w/both	.65	.67	.84	.71	.67	.90	.72
9. Initial bigrams	.55	.50	.50	.57	.51	.60	.50
10. Final bigrams	.60	.50	.50	.56	.51	.60	.50
11. Initial/final bigrams	.66	.50	.52	.63	.53	.70	.50
12. Initial trigrams	.53	.51	.50	.59	.51	.74	.54
13. Final trigrams	.67	.56	.50	.61	.51	.70	.50
14. Initial/final trigrams	.65	.61	.80	.71	.53	.82	.62
15. Exact match	.50	.52	1.00	.50	.50	.50	.48

Note. The observed scores cited are taken from the original experiments. *Same* and *different* (Dif.) indicate whether the letter set was changed between training and test (i.e., they refer to standard and transfer conditions). Whittlesea and Dorken's (1993) Experiment 4 was an unusual randomly changing transfer condition (the assignment of letters to elements of their pseudogrammar varied for every single item, during both training and testing). Their Experiment 5 was a randomly changing transfer condition, with the letter set being changed for every single test item. Transfer in Brooks and Vokey (1991) and St. John and Shanks (in press) was simple transfer (although the models do not differentiate between simple and randomly changing transfer).

materials, we also present figures for observed human performance. It should be noted that the latter are in fact generally low for the artificial grammar learning task. For example, with the standard materials, Dulany et al. (1984) observed experimental subjects with a mean proportion correct of .64 and a range of .63 to .70, and others have found performance of around .8 with these materials on the same letter set (nontransfer) task (e.g., Reber & Allen, 1978).

These results clearly show that many of the simple models can perform exceedingly well on the transfer task. For certain stimuli, some models perform at chance. This is particularly evident in the bigram-based models, confirming the empirical findings of Gomez and Schvaneveldt (1994) and Manza and Reber (1994); with these stimuli, bigram knowledge is not sufficient to support transfer (because it does not place strong enough constraints on the mapping between the old and new letter set). However, this finding does not disconfirm the fragment learning hypothesis in general; when larger fragments (e.g., trigrams) are considered, it can be seen that in all cases, the trigram based models, or some simple variant, are sufficient to support relatively high degrees of transfer. These results suggest that transfer is entirely consistent with a knowledge base consisting essentially of fragments of two and three letters, together with some knowledge of starts and ends of strings. This is not to say that subjects might not learn larger fragments, but in the main, knowledge of legal letter pairs and triples is sufficient to account for performance. This knowledge is not abstract in the sense under consideration here.

An important point to bear in mind is that for all except the Whittlesea and Dorken (1993) observations (Table 2), the empirical transfer results are for the simple transfer versions of the task, whereas these models have been applied to the randomly changing transfer case. To simulate performance with simple transfer, these models might, for instance, simply retain the first mapping found, changing it only if it contradicts successive input more than might be expected (intuitively, if the same mapping does not fit approximately 50% of the test items, it is likely to be wrong). In the simple transfer case, the figures for the nontransfer case are essentially an upper bound on performance. It seems likely, given the ease of finding the correct

Table 3
Proportion of Classifications Correct, From Observation, and for Each of the Toy Models, for the Stimuli From Altmann, Dienes, and Goode (1995)

	Stimuli					
	Exp. 1 & 2		Exp. 3		Exp. 4	
Source	Same	Dif.	Same	Dif.	Same	Dif.
Observation	.58	.55	—	.58	.71	.65
Model						
1. Bigrams	.83	.51	.85	.53	.89	.52
2. Bigrams w/start	.96	.58	.89	.63	.89	.54
3. Bigrams w/end	.83	.51	.85	.53	.89	.52
4. Bigrams w/both	.96	.58	.89	.63	.89	.54
5. Trigrams	.88	.53	.80	.61	.83	.58
6. Trigrams w/start	.89	.63	.81	.70	.83	.59
7. Trigrams w/end	.76	.52	.80	.61	.83	.58
8. Trigrams w/both	.78	.66	.81	.70	.83	.59
9. Initial bigrams	.68	.54	.68	.50	.71	.50
10. Final bigrams	.58	.50	.62	.50	.75	.50
11. Initial/final bigrams	.85	.56	.78	.51	.95	.52
12. Initial trigrams	.75	.54	.80	.59	.81	.50
13. Final trigrams	.61	.50	.73	.50	.74	.50
14. Initial/final trigrams	.75	.58	.71	.68	.71	.59
15. Exact match	.56	.48	.50	.69	.51	.61

Note. The observed scores for Experiments 1 and 2 are the average across conditions for Experiment 1. Experiments 1 and 2 used the standard materials (see Figure 1 and Table 1) with transfer across modalities (letter strings to auditory tones, and vice versa), Experiment 3 used a simple phrase structure grammar, with transfer from spoken sequences (the grammar generates strings of nonsense words instead of letters) to graphical symbols. Experiment 4 used the same grammar, but different training and test sets, with transfer from strings of graphical sequences to written syllables. In each case, subjects performed simple transfer.

mapping, as described above, that the performance of these models would closely approach this bound.

At a relatively coarse level, these models provide a good overall match with the patterns of observed data, with superior performance on the same letter set task as compared with the transfer task (which appears to be the pattern for human subjects), and in terms of the range of performance; the majority of the models are within the acceptable human range and would cause little comment if reported as empirical observations. Where the models exceed human performance, the simple assumption that only some, rather than all, of the relevant features from the training set are retained would appear to be an obvious remedy.

It would be inappropriate to attempt to find a precise match to the empirical data with models of this kind. The variety of possible models is so wide that there will be many different combinations consistent with any observed pattern of data, and for the transfer task the empirical database is relatively poor, at least as far as the fine-grained data needed to assess particular models (e.g., rankings of difficulty of various strings, see Dienes, 1992). Given the variety of even these toy models, it is also hard to imagine future empirical results from the transfer paradigm disconfirming the class of fragment-based models as a whole. Aside from the simple models we have considered here, it is easy to imagine a host of more complex fragment-based accounts, which might make use of frequency information or be able to make partial matches with remembered information, rather than insisting on a complete match, and so on. It seems inappropriate to explore such models at present, given that even the simplest ones appear able to account for transfer performance as well as, if not better than, accounts based on abstract knowledge or abstraction from whole exemplars.

Discussion

We argue that the experimental methodology used in many artificial grammar learning tasks has a number of potential flaws. Furthermore, we have shown that even where the experimental evidence concerning transfer can be taken at face value, it does not necessitate the hypotheses either of abstract knowledge or of whole-exemplar memorization. A simple alternative is that subjects simply acquire knowledge of fragments of the training items and, explicitly or implicitly, attempt to abstract from these to the test strings, accepting as grammatical those for which a good fit can be found. Given that subjects' knowledge of fragments of the training strings is well documented (e.g., Dienes et al., 1991), this simple hypothesis appears to take precedence

Table 4
D Scores From Observation and for the Toy Models for the Stimuli From Gomez and Schvaneveldt (1994)

Source	Stimuli			
	Same		Different	
	NPP	NPL	NPP	NPL
Observation	21	17	14	7
Models				
1. Bigrams	100	0	12	0
2. Bigrams w/start	100	0	47	0
3. Bigrams w/end	100	0	29	0
4. Bigrams w/both	100	0	65	0
5. Trigrams	94	65	47	35
6. Trigrams w/start	94	65	76	53
7. Trigrams w/end	94	65	53	53
8. Trigrams w/both	94	65	82	65
9. Initial bigrams	29	0	0	0
10. Final bigrams	12	0	0	0
11. Initial/final bigrams	41	0	12	0
12. Initial trigrams	41	35	29	41
13. Final trigrams	24	12	6	0
14. Initial/final trigrams	59	47	47	47
15. Exact match	18	18	29	35

Note. NPP (nonpermissible pair) and NPL (nonpermissible location) refer to two different types of grammatical violation and correspond to the presence of illegal bigrams and trigrams, respectively (see Gomez & Schvaneveldt, 1994, p. 400, for details).

over more complex ones. Additionally, it seems likely that the family of simple models presented here may serve as a starting point for more detailed psychological models, capable of empirical test.

Concerning the issue of whether subjects' fragment knowledge and mapping or abstraction processes during testing are conscious, we remain neutral. It is possible to argue that because subjects are generally informed of the rule-governed nature of the stimuli prior to testing, any abstractive processes taking place should be ascribed to conscious analytical reasoning (Perruchet, personal communication, 1995). Although we are sympathetic to this viewpoint, given the current controversy over consciousness in implicit learning (see Shanks & St. John, 1994, and commentaries) and the wide variety of definitions and proposed tests for conscious awareness, relatively few of which have been applied to the transfer paradigm (although see Dienes & Altmann, in press), we believe that it would be premature to make any assertion either way at this time.

The transfer phenomenon in artificial grammar learning is a startling and counterintuitive one, and the hypothesis of abstract knowledge (especially in conjunction with unconscious processes) is particularly seductive, as Perruchet and Pacteau (1990) have pointed out. One reason why simpler hypotheses have been somewhat ignored may be that researchers simply did not realize how powerful very simple mechanisms, relying on only (literally) fragmentary knowledge, could be. We hope that our toy models have shown clearly that a little knowledge can go a long way. The second possible reason for the neglect of simple hypotheses may be the failure to consider the locus of abstraction (this point is also anticipated by Perruchet & Pacteau, 1991). The same considerations are equally applicable in other areas of psychology.

In general, whenever one posits the existence of a representation that is abstract with respect to the learning stimulus, for example, a categorization rule, an abstract schema underlying an analogy, deep structural knowledge, or abstract rules acquired during implicit learning, it is not enough simply to demonstrate that subjects' performance requires the existence of such a representation. Assuming that the latter can be shown, this says little about whether subjects acquired the abstract knowledge in the course of learning or whether it is a manifestation of processes resulting from the requirements of the test, acting on representations that are less abstract with regard to the original learning stimuli.

Although transfer experiments, in themselves, do not provide convincing evidence that subjects have acquired abstract knowledge from exposure to the initial learning stimuli, this does not imply that such evidence could not be found. Three sorts of tests of the representation suggest themselves. First, can the subjects verbalize the abstract knowledge that they are hypothesized to possess? For instance, in implicit learning, we assume that subjects abstract from the patterns of light and dark that they see to the level of letters during learning. If subjects are asked to name the letters constituting the stimuli, they can do so with little difficulty. Similarly, in categorization tasks, subjects can often easily describe the rule on which they base their decision. Unfortunately, in implicit learning, subjects' spontaneous verbal reports typically do not reveal knowledge sufficient to explain their level of performance on the task. A more general difficulty with this approach is that just as subjects may only derive the more abstract representation for the purposes of performing the task, they might similarly derive an abstract representation in the process of providing a coherent explanatory verbalization; even verbalization is no guarantee that the representation was acquired during learning. The second approach avoids this by attempting to utilize indirect, or incidental, tests of the knowledge that subjects acquire. Here, the intent is to avoid placing demands on the subject that might lead to abstraction beyond that which has already taken place. Some steps in this direction have been taken by Whittlesea and Dorken (1993), who used old–new discrimination instead of the grammaticality judgment task. Using this task, subjects need not be told that experimental materials are governed by a rule or that their responses should attempt to conform with that rule. A third approach is to argue from processing constraints, or parsimony. For instance, all theories of past tense learning propose that language learners extract regularities and exceptions from their exposure to many examples of past tense formation (Pinker & Prince, 1988; Plunkett & Marchman, 1991). It is possible that learners simply store all of the past tense formations they have ever experienced and then compute the appropriate regularity or exception

every time they are called on to form the past tense. However, this would seem to be a prohibitively expensive approach, in terms of both storage and computation—it is more parsimonious to assume that they do abstract from their experience to form a set of rules (which may be embodied in a symbolic or connectionist fashion) governing past tense formation.

In the case of transfer in implicit learning, subjects are unable (easily) to verbalize the knowledge that they use to perform the discrimination task, and there do not appear to be any convincing computational or processing reasons why abstraction should take place during learning, as opposed to at test. In the absence of evidence from indirect or incidental tests, there appears to be little convincing support for the abstract knowledge hypothesis.

References

Altmann, G. T. M., Dienes, Z., & Goode, A. (1995). On the modality independence of implicitly learned grammatical knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 899–912.

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Srull & R. S. Wyer, Jr. (Eds.), *Advances in social cognition* (Vol. 3, pp. 61–88). Hillsdale, NJ: Erlbaum.

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 170–211). Hillsdale, NJ: Erlbaum.

Brooks, L. R., & Vokey, J. R. (1991). Abstract analogies and abstracted grammars: Comments on Reber (1989) and Mathews et al. (1989). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 316–323.

Chan, C. (1992). *Implicit cognitive processes: Theoretical issues and applications in computer systems design.* Unpublished doctoral thesis, University of Oxford, England.

Clark, H. C. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359.

Dienes, Z. (1992). Connectionist and memory-array models of artificial grammar learning. *Cognitive Science, 16,* 41–79.

Dienes, Z., & Altmann, G. (in press). Transfer of implicit knowledge across domains: How implicit and how abstract? In D. Berry (Ed.), *How implicit is implicit learning?* Oxford, England: Oxford University Press.

Dienes, Z., Broadbent, D. E., & Berry, D. C. (1991). Implicit and explicit knowledge bases in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 875–887.

Dulany, D. E., Carlson, R. A., & Dewey, G. I. (1984). A case of syntactical learning and judgment: How conscious and how abstract? *Journal of Experimental Psychology: General, 113,* 541–555.

Gentner, D. (1989). The mechanisms of analogical learning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 199–233). Cambridge, UK: Cambridge University Press.

Gomez, R. L., & Schvaneveldt, R. W. (1994). What is learned from artificial grammars? Transfer tests of simple associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 396–410.

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

Knowlton, B. J., & Squire, L. R. (1994). The information acquired during artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 79–91.

Manza, L., & Reber, A. S. (1994). *Representation of tacit knowledge: Transfer across stimulus forms and modalities.* Manuscript submitted for publication.

Mathews, R. C. (1990). Abstractness of implicit grammar knowledge: Comments on Perruchet and Pacteau's analysis of synthetic grammar learning. *Journal of Experimental Psychology: General, 119,* 412–416.

Mathews, R. C., Buss, R. R., Stanley, W. B., Blanchard-Fields, F., Cho, J. R., & Druhan, B. (1989). Role of implicit and explicit processes in learning from examples: A synergistic effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 1083–1100.

Morrison, J. (1994). *Transfer effects in artificial grammar learning.* Unpublished final honors thesis, Department of Psychology, University of Edinburgh, Scotland.

Perruchet, P. (1994). Defining the knowledge units of a synthetic language: Comment on Vokey and Brooks (1992). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 223–228.

Perruchet, P., & Pacteau, C. (1990). Synthetic grammar learning: Implicit rule abstraction or explicit fragmentary knowledge? *Journal of Experimental Psychology: General, 119,* 264–275.

Perruchet, P., & Pacteau, C. (1991). Implicit acquisition of abstract knowledge about artificial grammar: Some methodological and conceptual issues. *Journal of Experimental Psychology: General, 120,* 112–116.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28,* 73–193.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multilayered perceptron: Implications for child language acquisition. *Cognition, 38,* 43–102.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 5,* 855–863.

Reber, A. S. (1969). Transfer of syntactic structure in synthetic languages. *Journal of Experimental Psychology, 81,* 115–119.

Reber, A. S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General, 118,* 219–235.

Reber, A. S. (1990). On the primacy of the implicit: Comment on Perruchet and Pacteau. *Journal of Experimental Psychology: General, 119,* 340–342.

Reber, A. S., & Allen, R. (1978). Analogy and abstraction strategies in synthetic grammar learning: A functional interpretation. *Cognition, 6,* 189–221.

Reber, A. S., & Lewis, S. (1977). Implicit learning: An analysis of the form and structure of a body of tacit knowledge. *Cognition, 5,* 331–361.

Redington, M., & Chater, N. (1994). The guessing game: A paradigm for artificial grammar learning. In A. Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Meeting of the Cognitive Science Society* (pp. 745–749). Hillsdale, NJ: Erlbaum.

Redington, M., & Chater, N. (1996). *Randomly changing transfer in artificial grammar learning.* Manuscript submitted for publication.

Reeves, L. M., & Weisberg, R. W. (1994). The role of content and abstract information in analogical transfer. *Psychological Bulletin, 115,* 381–400.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing:*

Explorations in the microstructure of cognition (pp. 216–271). Cambridge, MA: MIT Press.

Servan-Schreiber, E., & Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 592–608.

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

St. John, M. F., & Shanks, D. R. (in press). Implicit learning from an information processing standpoint. In D. Berry (Ed.), *How implicit is implicit learning?* Oxford, England: Oxford University Press.

Vokey, J. R., & Brooks, L. R. (1992). Salience of item knowledge in learning artificial grammar. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 328–344.

Vokey, J. R., & Brooks, L. R. (1994). Fragmentary knowledge and the processing-specific control of structural sensitivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1504–1510.

Whittlesea, B. W., & Dorken, M. D. (1993). Incidentally, things in general are particularly determined: An episodic-processing account of implicit learning. *Journal of Experimental Psychology: General, 122,* 227–248.

Received December 5, 1994
Revision received August 30, 1995
Accepted August 30, 1995 ■