Salience of Item Knowledge in Learning Artificial Grammars

John R. Vokey
University of Lethbridge, Lethbridge, Alberta, Canada

Lee R. Brooks
McMaster University, Hamilton, Ontario, Canada

Both the specific similarity of test items to study items and the grammaticality of test items were found to be major determinants of performance under task conditions common in the literature. Results bearing on the issue of how item-specific effects are coordinated with knowledge pooled across items are: (a) Better item memory resulted in smaller rather than larger effects of specific similarity on judgments of grammaticality, suggesting that items can be too well differentiated to support transfer to new items. (b) Variation in the effect of specific similarity did not result in compensatory variation in grammaticality, suggesting that any scheme that tightly links the effects of the two variables is insufficient. (c) Differential reliance on the 2 knowledge resources was not under good instructional control, which poses a problem for accounts that use functional task analyses to coordinate functionally different memories.

People are continually learning complex regularities about language and the world around them. Artificial grammars have been taken as a particularly convenient set of materials with which to study the nature of such learning (e.g., Mathews et al., 1989, and Reber, 1989, who reviews his extensive previous laboratory work). In these studies, letter strings generated from an artificial grammar (such as that shown later in Figure 1) are presented to learners. The extent to which the learners become sensitive to the regularities in the generating grammar is subsequently tested in various ways (e.g., Gordon & Holyoak, 1983; Howard & Ballas, 1980, 1982; Miller, 1958; Reber, 1967, 1969, 1976; Reber & Lewis, 1977) but in most recent studies by requiring the learners to sort new items generated by the grammar from similar strings that violate the grammar.

There are two general positions about what is learned from exposure to these grammars and, by extension, to many other types of complex materials: abstractive and distributive. According to the former position, the learner more or less automatically acquires an abstract representation of the structure of the experienced material, which becomes the major basis of grammaticality judgments. Reber (1989) argued strongly for a separate, sophisticated, and tacit system that accomplishes this task for complex domains. Other approaches to the learning of complex domains that centrally rely on abstraction are given by Broadbent (1989; Broadbent, Fitzgerald, & Broadbent, 1986) and by Lewicki (1986; Lewicki, Czyzewska, & Hoffman, 1987; Lewicki, Hill, & Bizot, 1988).

The contrasting view is that a major part of the learner's knowledge of the domain is fundamentally distributive, consisting in routine reference to similar prior episodes. This position assumes that there is not a central core of abstracted knowledge that is so stable and so available that it can be used reliably to characterize the operative knowledge that people have about the domain. Experience with the distributional properties of the domain (e.g., frequency, characteristic features, privileges of occurrence) can remain distributed in the knowledge base and can be exploited by a process of forming immediate local models or analogies as the need arises. Arguments for the adaptive value of "postcomputation" on a distributed knowledge base have been presented by Brooks (1978, 1987, 1990), Jacoby and Brooks (1984), and Kahne-man and Miller (1986). Models that rely on distributed representations have been presented by Estes (1986), Hintzman (1984, 1986, 1988), Logan (1988), McClelland and Rumelhart (1981, 1985), Medin and Schaffer (1978), and Nosofsky (1984, 1986, 1989) among others. Evidence supporting the effect of prior interpreted instances on tasks that have often been taken as relying primarily on abstracted knowledge has been presented by S. W. Allen and Brooks (1991), Jacoby (1983a, 1983b), Jacoby, Baker, and Brooks (1989), Jacoby, Kelley, Brown, and Jaseckho (1989), Ross (1984, 1987, 1989), Vokey, Baker, Hayman, and Jacoby (1986), Whittlesea (1987), Whittlesea and Brooks (1988), and Whittlesea and Cantwell (1987). A comparable shift toward considering less reliance on abstracted knowledge has been evidenced in other areas of cognitive science in work on case-based reasoning (e.g., Kolodner, 1988; Shank, 1982).

This article presents data relevant to attempts to reconcile these different approaches. One of these attempts, which we refer to as the "dual knowledge" approach, grew out of the abstractive position. In response to evidence suggesting that there are effects of prior instances under some circumstances, several proposals have been made that performance in artificial grammar tasks is dependent on both knowledge of individual items and on knowledge abstracted across items (e.g., Mathews et al., 1989; McAndrews & Moscovitch, 1985; Reber, 1989; Reber & Allen, 1978). Such proposals raise the question of how such resources are coordinated. For example, as suggested by Reber and Allen (1978) for the results of the paired-associate training procedure used by Brooks (1978) and by Reber (1989) more generally, it is possible that evidence for reliance on instances in the test phase occurs mainly.
under training conditions that stress close attention to the details of the items. Under less restrictive conditions, such as merely observing the items during training, concentration on details would no longer interfere with apprehending the general structure, and greater reliance on abstracted structure would be expected. More generally, it is possible that there is a compensatory relation between reliance on abstract knowledge and on similarity to instances, so that conditions that favor one tend to discourage the other. Regardless of whether the relationship is compensatory, proposals that performance depends on different types of knowledge require that some principles be specified for how they are coordinated.

However, questions regarding the relative importance of similarity to individual items and of knowledge pooled across items need not be limited to models that assume two fundamentally different knowledge sources. For example, formal instance models, as well as many models of recognition memory, assume that the probe item is compared with all the items in memory (e.g., Estes, 1986; Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1984, 1986, 1989; as well as the less formal models of Kahneman & Miller, 1986; Whittlesea, 1987; although not formally instance models, parallel distributed memory models also have the relevant characteristic of nonlinear generalization around prior instances: e.g., Gluck & Bower, 1988; McClelland & Rumelhart, 1985; Whittlesea, 1988). In all of these models, because new grammatical items are likely to resemble a large number of old items, all of which are also grammatical, a difference between grammatical and nongrammatical items could be explained by "retrieval time averaging" or "chorus of instances" (termed "echo" by Hintzman, 1986, and "computing norms on the fly" by Kahneman and Miller, 1986). In addition, however, all of the instance models cited have nonlinear generalization gradients, which means that an item in memory that is very similar to a probe will have a disproportionately large effect. This is the means deliberately built into these models to obtain item-specific effects. The consequence for judgments of grammaticality is that if a memory item is similar enough to a probe, then it will have a strong effect regardless of whether or not the probe item also matches the average properties of all of the training items, that is, regardless of whether or not the item is grammatical. Thus, any of these instance models could predict effects of close similarity to specific old items that need not be the same as the effects of grammaticality.

Regardless of whether knowledge resources are thought of as two different kinds of knowledge, as in the dual knowledge models, or as two different degrees of similarity, as in the distributed models, there is reason to be interested in the circumstances that affect the relative predictiveness of similarity to individual instances and of grammaticality of the test items. However, in most of the work in the literature, these two variables have been confused. In fact, some of the effects that have been attributed to the grammaticality of the items could be due to very close similarity to individual items, which in both approaches could be independent of an item’s consistency with the average characteristics of the training items. Alternatively, the contribution of close similarity could be small under the conditions common in the literature. To address these problems, the experiments in this article used materials in which the similarity of test items to specific training items was unconfounded from the grammaticality of the test items. This design makes it possible to compare the effectiveness of these two variables for controlling subjects' judgments of grammaticality of new items and recognition of old items.

Empirical Variables of Specific Similarity and Grammaticality

To unconfound the variables of specific similarity and grammaticality, both grammatical and nongrammatical transfer items are needed, with some of each type being similar to particular training items and some of each type being relatively dissimilar to any of the training items. The materials used in Experiments 1 and 4 are shown in Table 1.

As a criterion of similarity, we called any item similar or close if it differed from a training item by only one letter and dissimilar or far if it differed by more than one letter from the most similar training item. Reading across any row of Table 1, it can be seen that for each training item there was one close transfer item that was grammatical according to the grammar in Figure 1 and one close transfer item that was not. Any particular subject was trained on only one of the training lists but was tested on all of the transfer items. In this way, we ensured that half of the transfer items for each subject were similar to items on the training list given, and half were dissimilar to the items in that training list. For example, for a subject trained with List 1, the transfer items in the top half of the table are similar according to our criteria, and the items in the bottom half are dissimilar. For counterbalancing, half of the subjects were trained with the second training list. Consequently, for these subjects, the transfer items in the bottom half of the table were similar to the training list and the top half, dissimilar. In sum, each subject was tested on both grammatical and nongrammatical items; half of each type was similar to specific training items for that subject and half was dissimilar to those training items.

It is important to note that close transfer items differ from far transfer items not by being similar to any general characteristic of the training list but by being similar to a specific item of the training list, a property we refer to as specific similarity, in contrast to the similar and dissimilar items in a prototype study that are defined in terms of proximity to the center of the distribution of training items. In fact, in each of the experiments reported, with the exception of the specific match, close items for a given subject were no more similar to the items of a given training list than were the far items.

The principle of unconfounding specific similarity and grammaticality as a method of diagnosing the locus of effect of other variables was developed in Vokey (1983). Portions of this work were reported in the manuscript "Training the Clever Unconscious: Analogic and Abstraction Strategies in Artificial Grammar Learning," which was accepted for publication in the journal Cognition only to be lost somewhere in the publication process. Because of the resulting long delay in publication, the manuscript was retracted from the journal. It is this paper to which McAndrews and Moscovitch (1985) referred in their valuable extension of the technique.
Thus, not only is the variable of specific similarity unconfounded with the generative grammar, but it also is unlikely to be confounded with any version of a local grammar that reasonably could be constructed from experience with a specific training list (cf. Dulaney, Carlson, & Dewey, 1984, 1985).

This core design was used in four experiments that manipulated variables likely to change the relative effectiveness of grammaticality and of specific similarity in controlling judgments about the items. In Experiment 1, the transfer tasks required of the subjects were manipulated. Subjects were given a standard recognition memory task, and then some were asked to judge grammaticality and others to judge each item’s specific similarity to the training items. Experiments 2 and 3 varied the encoding of items during the training phase as a method of rendering items more or less available to support generalization at test. Experiment 4 directly manipulated retrieval conditions at test to determine whether differences observed in categorical transfer are due to differential acquisition of the sources of knowledge or to the test conditions differentially retrieving individual items. The results of each experiment are evaluated in terms of their effects on the separate factors of specific similarity and grammaticality.

Experiment 1

In this experiment, the relative effects of the variables of grammaticality and specific similarity were observed after training conditions common in the literature. In addition, we address the issue of degree of control by the particular transfer tasks. Proposals that performance is dependent on two different kinds of information usually associate the use of each

Table 1
Training and Transfer Items Generated From the Artificial Grammar Shown in Figure 1 and Used in Experiments 1 and 4

<table>
<thead>
<tr>
<th>Training items</th>
<th>Grammatical</th>
<th>Nongrammatical</th>
</tr>
</thead>
<tbody>
<tr>
<td>List 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MXRVXT</td>
<td>MXRMXT</td>
<td>MXRRXT</td>
</tr>
<tr>
<td>VMTTRRR</td>
<td>VMTTRRX</td>
<td>VMTTRRT</td>
</tr>
<tr>
<td>MXTTRRR</td>
<td>VXTTRRR</td>
<td>TXTTRRR</td>
</tr>
<tr>
<td>VXVRMXT</td>
<td>VXVRRXT</td>
<td>VXVRTXT</td>
</tr>
<tr>
<td>VXVVM</td>
<td>VXVRRV</td>
<td>VXVRT</td>
</tr>
<tr>
<td>VRMVRVV</td>
<td>VMRVVVM</td>
<td>YMRVVVR</td>
</tr>
<tr>
<td>MXTMTVR</td>
<td>MXTMTXR</td>
<td>MXTMTTR</td>
</tr>
<tr>
<td>VXRMRXTR</td>
<td>VMRRVXT</td>
<td>VMRTXT</td>
</tr>
<tr>
<td>MVR</td>
<td>MVRVXV</td>
<td>VMVXVV</td>
</tr>
<tr>
<td>VRVV</td>
<td>VRVVM</td>
<td>VRVMX</td>
</tr>
<tr>
<td>MRVRMVR</td>
<td>VMRMRVX</td>
<td>VMRMVRX</td>
</tr>
<tr>
<td>VRMVX</td>
<td>VRMVXT</td>
<td>VRMXT</td>
</tr>
<tr>
<td>MXTMXT</td>
<td>MXTMXT</td>
<td>MXTMXT</td>
</tr>
<tr>
<td>VXTX</td>
<td>VXVXT</td>
<td>VXVXT</td>
</tr>
<tr>
<td>List 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MVXRM</td>
<td>MVXRM</td>
<td>MVXRM</td>
</tr>
<tr>
<td>MVXRMVR</td>
<td>VXREMXR</td>
<td>VXREMTR</td>
</tr>
<tr>
<td>VXVT</td>
<td>VXVR</td>
<td>VXVM</td>
</tr>
<tr>
<td>MVRMVRV</td>
<td>MVRMVRV</td>
<td>MVRMVRV</td>
</tr>
<tr>
<td>MXRTVM</td>
<td>MXRTVM</td>
<td>MXRTVM</td>
</tr>
<tr>
<td>VMT</td>
<td>VMR</td>
<td>VMM</td>
</tr>
<tr>
<td>VXVRMR</td>
<td>VXVRMT</td>
<td>VXVRMX</td>
</tr>
<tr>
<td>VXVRVV</td>
<td>VXVRVM</td>
<td>VXVRVR</td>
</tr>
<tr>
<td>VRMVMX</td>
<td>VRVMM</td>
<td>VRVMM</td>
</tr>
<tr>
<td>VRVTX</td>
<td>VRVVT</td>
<td>VRVVT</td>
</tr>
<tr>
<td>VXVRMV</td>
<td>VXVRMV</td>
<td>VXVRMV</td>
</tr>
<tr>
<td>MXTTRRX</td>
<td>VXTRRX</td>
<td>VXTRRX</td>
</tr>
<tr>
<td>VXVTTRX</td>
<td>VXVTTRX</td>
<td>VXVTTRX</td>
</tr>
<tr>
<td>MXTTRX</td>
<td>VXTRX</td>
<td>VXTRX</td>
</tr>
<tr>
<td>VMRTMXT</td>
<td>VMRTVXT</td>
<td>VMRTXT</td>
</tr>
</tbody>
</table>

Figure 1. Artificial grammar used in Experiments 1 and 4. (Grammatical items are produced by following the arrows between numbered nodes and picking up the letters associated with the transitions. Note that the items MVR and MXR are legal according to this grammar, but the similar item MTR is not.)
kind with different specific transfer tasks. At a minimum, recognition memory should differentially rely on specific item memory and grammaticality judgments on abstracted structural knowledge. However, it is possible that other variations in the explicit task required will influence differential reliance on the two knowledge sources. This experiment tested explicit task control in two ways.

The first way of testing task control is to vary the instructions given for the grammaticality judgment. All of our subjects were given the usual incidental learning instructions during the training phase (e.g., Dulaney et al., 1984; Reber, 1976). One half of the subjects—the grammar transfer subjects—were then told of the existence of the grammar and were asked to sort a set of new items according to that grammar, the usual procedure in the literature. The other half of the subjects—the similarity transfer subjects—were asked to sort the transfer items with respect to each item's specific similarity to the training items. Both judgment tasks used an appropriately marked rating scale to keep the instructed basis of classification salient to the subjects. The scale ranged from sure obeys the rules to sure does not obey the rules for grammar transfer and from sure similar to at least one training item to sure not similar to any training item for similarity transfer. An abstractive position would be most comfortable with the sorting in the similarity transfer condition, being less under the control of the grammaticality variable and more under the control of specific similarity than when the subjects were asked to sort on the basis of grammaticality. There is no reason for an abstractive position to assume perfect control by these instructions, but a shift would be consistent with the need for some degree of task control given the hypothesized two different sources of information. A simple distributive position, assuming no separate knowledge of grammaticality, suggests that both tasks would be performed in approximately the same way. In particular, it would predict no difference in the effect of grammaticality across the two tasks.

Mathews et al. (1989) also used a design in which a contrast was made between subjects who were informed about the existence of the grammar and those who were asked to sort on the basis of similarity. This contrast was continued through several cycles of training and testing, providing valuable information about the robustness of the results. However, because their experimental design did not unconfound specific similarity and grammaticality, they were unable to comment on the differential influence of the contrast on the effects of these variables.

The second way of testing task control is to have the recognition memory task performed before the grammatical sorting task and before the subjects are informed of the existence of the grammar. Typically in artificial grammar experiments, the subjects are not informed of the existence of the underlying grammatical structure until immediately before the transfer task. Any learning of the grammar during the training phase is, by necessity, incidental to the training task, which would be consistent with the abstractive claim (e.g., Reber, 1989) that such learning is automatic. However, subjects of necessity are told of the existence of the grammar before they are asked to sort the items according to the rules used to generate the training items. There is no such logical necessity in the case of recognition memory. Having the recognition phase before the subjects are informed of the grammar accomplishes two goals for us in testing the abstractive position. First, it is a hard test of the automaticity of the abstraction: The subject is aware of the existence of the grammar neither at training nor at test. Second, it removes any contamination of recognition memory performance from having previously explicitly considered a grammar. Because the abstractive position would most naturally suggest greater control of recognition memory by specific similarity than by grammaticality, this task provides a cleaner set of conditions under which to observe it.

Method

Materials: Two lists were selected without replacement from the pool of all possible letter strings, three to seven letters in length, generated from the artificial grammar shown in Figure 1. Each list consisted of 16 pairs of letter strings, in which the members of each pair were of equal length and differed from each other by a single letter in either the initial, middle, or terminal letters of the string. One member of each pair was used as a training item; the other was used as a grammatical transfer item. Construction of the two lists was constrained so that the members of any pair differed by at least two letters in position from the members of every other pair in both lists and so that factors such as item length, position of change across training and transfer items, and use of the 24 possible node-to-node transitions of the grammar were balanced as completely as possible across training lists and across training and transfer items. These constraints ensured that the two training lists were equally representative of the grammar, that the transfer items did not differ systematically from the training items, and that every transfer item would have at most one close training item.

The nongrammatical transfer items were produced from the grammatical transfer items by the substitution of a single letter. At the same letter position in which each grammatical item differed from its matched training item by the letter M, R, V, X, or T, the corresponding nongrammatical item was produced by substituting the letter R, M, T, T, or X, respectively, to produce an item that cannot be generated by the grammar. For example, the nongrammatical item MTR was produced from the grammatical item MVR by substituting a T for the V, resulting in a string that cannot be generated by the grammar but that differs from the matched training item, MVX, by a single letter in the same position as that of the grammatical item. Thus, for each training item, there was a grammatical transfer item and a nongrammatical transfer item, each differing from each other and its matched training item by a single letter but differing from all other items by at least two letters. The two lists of training items and their matched grammatical and nongrammatical transfer items are shown in Table 1.

Each training list was divided into four sublists of 4 items each, and four random orders of the items in each sublist were generated. The 64 transfer items were randomly divided twice into four 16-item lists, producing four transfer lists for an initial transfer test and another four transfer lists for a retest, with each item occurring once for each test. One half of the transfer items were used as distractors on the recognition test, selected so as to maintain the orthogonal structure of the transfer test. Thus, there were 8 items from each of the combinations of grammaticality and similarity for a total of 32 distractor items. Two recognition tests were constructed from these items, one containing the 16 List 1 training items and the other containing the 16 List 2 training items. The resulting 48 items on each recognition test were randomly divided into three lists of 16 items each.
Subjects. Thirty-two university undergraduates served in each of the two transfer conditions for a total of 64 subjects. Most participated
in exchange for a course credit in introductory psychology; the
remainder were paid $3.00 each to participate.

Procedure. The materials and instructions were presented to the
subjects in booklets. For the training phase, the pages of the booklets
were back printed with a solid mask to prevent the subjects from
viewing the contents of previous and subsequent pages, and for all
phases subjects were not allowed to flip back to previous pages.

Training consisted of a free-recall procedure similar to that used
in many of Reber’s studies. Each ordering of the items for a given
sublist was printed on a separate page. Subjects were instructed to
study the four items on each page and then to recall the items on a
subsequent blank page. One half of the subjects in each transfer
condition received List 1 and the remainder, List 2. Order of the four
sublists within each training list was systematically rotated across
subjects within each transfer condition. Subjects received all four
orderings of a given sublist before proceeding to the next sublist.
Between sublists, the training instructions were repeated, and the
subjects were informed that they would be receiving a new list of
items. Rate of presentation was subject paced. Subjects were not
informed about the number of training sublists they would receive or
that their task consisted of anything other than the training phase.

Subjects received the recognition test appropriate to the training
list presented. Two orders, counterbalanced across subjects, of the
three 16-item recognition lists were used. Each list was presented on
a separate page, and each item was preceded by a blank in which the
subjects were to respond using a 6-category response scale representing
the judgments sure old (1), fairly sure old (2), guess old (3), guess
new (4), fairly sure new (5), and sure new (6).

Each transfer item was presented twice for a total of 128 separate
judgments, although the subjects were not informed of this fact. Each
16-item transfer list was presented on a separate page, and each item
was preceded by a blank in which the subjects were to respond.
Grammar transfer subjects read instructions to the effect that the
training items had been constructed according to a complex set of
rules and that their task was to indicate which of the transfer items
were probably constructed according to the same set of rules. As in
most experiments with these materials, this was the first mention of
the existence of the rules that underlay the letter strings. Subjects
were instructed to use a 6-category response scale similar to the one
used during recognition: sure obeys the rules (1), fairly sure obeys the
rules (2), guess obeys the rules (3), guess does not obey the rules (4),
fairly sure does not obey the rules (5), and sure does not obey the rules
(6).

Subjects in the similarity transfer condition read instructions to
the effect that some of the transfer items were constructed to be very
similar to at least one of the training items, whereas the remainder
were constructed to be dissimilar to any of the training items, and
their task was to indicate for each transfer item whether they thought
the item was similar or dissimilar to the training items. They were
asked to respond on a 6-category response scale representing the
judgments sure similar to at least one training item (1), fairly sure
similar to at least one training item (2), guess similar to at least one
training item (3), guess not similar to any training item (4), fairly
sure not similar to any training item (5), and sure not similar to any
training item (6).

One half of the subjects in each transfer condition received the
transfer test and retest in one order, and the remainder received them
in the opposite order. The sequence of items within each transfer test
and retest was fixed across subjects. During neither the recognition
phase nor the transfer phase were the subjects informed about the
frequency of the various alternatives.

Results

Recognition. The major results of the recognition phase are summarized in the first panel of Table 2. The proportion of old responses (scale responses ≤ 3) to the various types of items served as the dependent variable. For this and all
subsequent analyses, reliability was assessed at the α = .05

<table>
<thead>
<tr>
<th>False positives</th>
<th>Grammatical</th>
<th>Nongrammatical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Condition</td>
<td>Close</td>
<td>Far</td>
</tr>
<tr>
<td>Experiment 1: Grammar transfer</td>
<td>.73</td>
<td>.35</td>
</tr>
<tr>
<td>Similarity transfer</td>
<td>.71</td>
<td>.34</td>
</tr>
<tr>
<td>Combined</td>
<td>.72</td>
<td>.35</td>
</tr>
<tr>
<td>Experiment 2: Observation training</td>
<td>.62</td>
<td>.18</td>
</tr>
<tr>
<td>Free-recall training</td>
<td>.62</td>
<td>.03</td>
</tr>
<tr>
<td>Mnemonic training</td>
<td>.63</td>
<td>.05</td>
</tr>
<tr>
<td>Combined</td>
<td>.62</td>
<td>.08</td>
</tr>
<tr>
<td>Experiment 3: Label</td>
<td>.63</td>
<td>.07</td>
</tr>
<tr>
<td>Unique mnemonic</td>
<td>.73</td>
<td>.03</td>
</tr>
<tr>
<td>Combined</td>
<td>.68</td>
<td>.05</td>
</tr>
<tr>
<td>Experiment 4: Mnemonic test</td>
<td>.96</td>
<td>.07</td>
</tr>
</tbody>
</table>

Note. Effect size estimates are point-biserial correlation coefficients.
level. As expected given that the transfer conditions were treated similarly to this point, the two transfer conditions did not differ significantly in either hits or false positives (Fs < 1).

To analyze these data as a categorical transfer task, the proportions of false-positive responses were subjected to a three-way analysis of variance (ANOVA), with subjects nested within transfer condition as the random variate. The combination of two levels for each of the independent variables of grammaticality and specific similarity resulted in four cells per subject and eight responses per cell. The mean false-positive rate for subjects in both transfer conditions was significantly greater for close items (M = .33) than for far items (M = .25), F(1, 62) = 15.31, MSr = 0.024, and for grammatical items (M = .38) than for nongrammatical items (M = .20), F(1, 62) = 75.44, MSr = .027, with the two effects being additive (F < 1). As is shown in the effect sizes column in Table 2, grammaticality rather than specific similarity was the significantly (z = 2.62) larger of the two effects. None of the remaining effects was significant.

Categorical transfer. The major results of the categorical transfer phase are shown in the first panel of Table 3. The proportion of items labeled grammatical or similar (scale response ≥ 3) were subjected to a three-way ANOVA, with subjects nested within transfer condition as the random variate. The factorial combination of two levels for each of the independent variables of grammaticality and specific similarity resulted in four cells per subject with 32 responses per cell.

The two transfer conditions did not differ significantly from one another in overall positive responses (F < 1). Grammar transfer subjects identified an average proportion of .48 of the transfer items as grammatical, and similarity transfer subjects identified an average proportion of .45 of the items as similar. In fact, the only significant difference between the two transfer conditions resulted from the interaction of transfer condition and specific similarity, F(1, 62) = 8.32, MSr = .010: Similarity transfer subjects were more sensitive to the specific similarity of the items than were grammar transfer subjects.

The only other significant effects were the main effects of both grammaticality and specific similarity. Subjects in both conditions produced significantly more positive responses to grammatical items (M = .53) than to nongrammatical items (M = .40), F(1, 62) = 110.49, MSr = .010, and to close items (M = .52) than to far items (M = .41), F(1, 62) = 78.92, MSr = .010, with the two effects being additive (F < 1). As in the recognition phase grammaticality was the larger of the two effects, although in this case not significantly (z = 0.72). Subjects in both conditions evidenced a large effect of grammaticality and did so to a statistically identical degree. Subjects in the similarity transfer condition discriminated grammatical from nongrammatical items with a proportion correct of .57, which rises to .65 (a value similar to that found in many of Reber’s investigations) if specific similarity is confounded with grammaticality (i.e., close grammatical items and far nongrammatical items), whereas actively attempting to discriminate grammatical items from nongrammatical items, as subjects in the grammar transfer condition were asked to do, did not improve this discrimination beyond that obtained by the similarity transfer subjects who were never informed that such a distinction existed.

**Discussion**

In this experiment, as in that of McAndrews and Moscovitch (1985), both specific similarity and grammaticality were major determinants of performance. Because the training conditions in this experiment are similar to those used in the

<table>
<thead>
<tr>
<th>Table 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean Proportion of Items Given Positive Responses as a Function of the Grammaticality and Specific Similarity of the Items and the Corresponding Effect Sizes for Each of the Experimental Conditions in Experiments 1–4</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Condition</th>
<th>Grammatical</th>
<th>Nongrammatical</th>
<th>Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Close</td>
<td>Far</td>
<td>Close</td>
</tr>
<tr>
<td>Experiment 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammar transfer</td>
<td>.58</td>
<td>.49</td>
<td>.45</td>
</tr>
<tr>
<td>Similarity transfer</td>
<td>.60</td>
<td>.46</td>
<td>.46</td>
</tr>
<tr>
<td>Combined</td>
<td>.59</td>
<td>.47</td>
<td>.46</td>
</tr>
<tr>
<td>Experiment 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation training</td>
<td>.44</td>
<td>.26</td>
<td>.38</td>
</tr>
<tr>
<td>Free-recall training</td>
<td>.37</td>
<td>.21</td>
<td>.33</td>
</tr>
<tr>
<td>Mnemonic training</td>
<td>.34</td>
<td>.25</td>
<td>.29</td>
</tr>
<tr>
<td>Combined</td>
<td>.38</td>
<td>.24</td>
<td>.33</td>
</tr>
<tr>
<td>Experiment 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Label</td>
<td>.47</td>
<td>.29</td>
<td>.46</td>
</tr>
<tr>
<td>Unique mnemonic</td>
<td>.36</td>
<td>.27</td>
<td>.40</td>
</tr>
<tr>
<td>Combined</td>
<td>.41</td>
<td>.28</td>
<td>.43</td>
</tr>
<tr>
<td>Experiment 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mnemonic test</td>
<td>.68</td>
<td>.46</td>
<td>.62</td>
</tr>
</tbody>
</table>

**Note.** Effect size estimates are point-biserial correlation coefficients.
literature, this large effect of specific similarity cannot be attributed strictly to unusual training conditions such as few training items or special emphasis on memorization.

A dual knowledge view has two resources to explain why both specific similarity and grammaticality affected both recognition and grammaticality judgments. The first is the possibility of incomplete learning of the grammar. Incomplete learning of the grammar was invoked by Reber (1989; Reber & Allen, 1978; Reber, Kassin, Lewis, & Cantor, 1980) to explain the observation that grammatical violations were more easily detected at the beginning and end of a string than in the middle. Reber's suggestion was that the subjects had learned less about the middle portions of the grammar than the ends. The same principle could be applied to explaining some degree of control by specific similarity; namely, given the limited number of training items, the grammar was not completely learned, with the result that the subjects fell back on specific item memory when grammatical knowledge was insufficient. The control by specific similarity observed in Experiment 1 might mean only that people have not yet acquired sufficient grammatical knowledge to determine all of the explainable variance.

In addition to incomplete learning, a dual knowledge position could suggest that the two types of information are inherently confusable. A strong feeling of familiarity for a particular item might result in a false judgment of grammaticality even though the grammar had been sufficiently learned to determine correct responses if the competition from familiarity were not present. One might expect such competition to be particularly strong if there were only a few training items and if they had been recently learned. Conversely, greater control by the knowledge of structure might be exerted with more training items or with greater time since training as found, for example, by Robbins et al. (1978) for dot patterns and R. Allen and Reber (1980) for artificial grammars. One could imagine further that the two sources of information might normally combine in mixtures determined by the salience of item retrieval rather than item memory affecting judgment only in the complete absence of relevant structural knowledge. The apparently greater relative effect of specific similarity on judgments of grammaticality than on recognition memory could be explained by observing that the recognition memory results are being compared only for the false-positive responses. These responses, averaging 29% of the new items, are selectively the trials for which specific item information is weak. If abstracted grammatical knowledge intrudes most strongly on trials for which item information is weak, then it would be expected that the grammaticality variable would account for more of the variance on these trials.

A distributive position also has several ways of dealing with the effects of both specific similarity and grammaticality on recognition memory and grammaticality judgments. As suggested earlier, the retrieval time averaging in the instance models would produce a correlation with the empirical variable of grammaticality. The empirical variable of specific similarity measures only whether a probe item is the same as a memory item on all but one letter in position. For this variable to capture all of the variance resulting from item memory, all such deviations would have to be subjectively equal. However, in a nongrammatical item, the deviation itself is dissimilar to the comparable location in any other item and is thus less likely than a grammatical item to recruit strength from other items. Furthermore, because all of the instance models cited have nonlinear generalization gradients, this "chorus of instances" is most likely to be observable precisely when the probe as coded is not very similar to one of the memory items as coded. That is, if the testing conditions are conducive to strong retrieval of a particular individual item, then the relative impact of the chorus of instances would not be expected to be especially strong.

In sum, the effectiveness of both specific similarity and grammaticality can be handled by both broad approaches. The grammaticality variable is a measure of abstracted knowledge of structure for the abstractive position and a measure of the pooling of multiple items at retrieval for the distributive position. The specific similarity variable is a measure of item knowledge for the abstractive position and a measure of the predominant evocation of a single memory item for the distributive position. Furthermore, both positions suggest that the relative effect of the grammaticality variable will be greatest when retrieval conditions do not encourage the clear retrieval of individual items. Finally, both positions suggest an effect of these variables on recognition memory.

However, this experiment does provide difficulties for one variety of a dual knowledge position, namely, that the two kinds of knowledge are under logical task control. By such a position, transfer judgments should have been more under control of the grammaticality variable when the subjects were instructed to sort on the basis of grammaticality than when they were asked to sort on the basis of specific similarity. The observed lack of a shift in task control with respect to grammaticality cannot be dismissed too lightly by an abstractive position. There has to be some degree of task control over the use of grammaticality if the two sources of knowledge are to be reasonably managed, and it is not immediately clear why these differences in instructions would not have elicited the suggested difference in behavior if the subjects had the two sources of information to use. All ratings for the specific similarity condition were made directly on a clearly labeled scale of similarity, so there was little chance that the subjects misunderstood or forgot the instructions. In addition, had the subjects simply forgotten the similarity instructions, it is unlikely that they would have demonstrated an enhanced effect of specific similarity in their transfer judgments. Mathews et al. (1989) reported a related lack of difference in transfer performance between similarity and grammar-instructed subjects that persisted through several cycles and hundreds of trials. This result suggests that the lack of task control demonstrated in our experiment is robust over considerable practice.

Experiments 2 and 3: Effect of Training Procedures

Another form of control of the availability and specificity of item information at retrieval is provided by variation in the type of coding operations required of the subjects. Because Experiments 2 and 3 differed only in the particular encoding
ITEM KNOWLEDGE IN ARTIFICIAL GRAMMARS

operations used, they are presented and discussed together. The effect of these experiments is to investigate Reber’s suggestion (Reber, 1989; Reber & Allen, 1978) that different types of training influence the efficacy of the implicit abstraction process. We suggest instead that the variations in performance do not reflect so much variations in the degree to which abstract learning occurs as variations in the effect of specific item retrieval. In Reber and Allen’s (1978) experiment, the transfer performance of subjects asked to learn the items as paired associates was contrasted with that of subjects asked simply to observe the items. Reber and Allen found that, relative to observation training, paired-associate training increased the accuracy of item retrieval on a test of recognition memory and decreased the accuracy of sorting new items for grammaticality. They suggested that the paired-associate training directed the subjects’ attention toward individuating the items: emphasizing minor differences and suppressing commonalities to provide distinctive retrieval cues for the labels. The task demands of a paired-associate task, then, would have the effect of suppressing the abstraction of the grammar, promoting a switch to Brooks’s (1978) suggested process of judging similarity to individual training items as a basis for grammaticality decisions. Extending this suggestion to the current experimental paradigm, the predictions are clear. Training that emphasizes the individuation of particular items should result in a greater effect of specific similarity and a decreased effect of grammaticality.

The prediction from the distributive position regarding item individuation training is less obvious. The apparent default assumption in the literature is that increased item learning leads to increased effects of specific items on tests of categorical transfer. This seems unlikely to be true in general. The process of emphasizing minor differences and suppressing commonalities among particular items, suggested by Reber and Allen (1978) as being the result of paired-associate learning, may also have resulted in the training items seeming less similar or being less available in response to transfer items. Thus, the poorer categorical transfer performance that Reber and Allen found from subjects given paired-associate training and on which Reber and Allen rested much of their argument may have occurred ironically because the subjects had fewer similar seeming items available in response to the transfer items than did subjects given the less individuated observation training. Under these conditions, then, the distributive position suggests a decreasing effect of specific similarity as a function of increasing item individuation. If so, there also should be fewer false alarms in recognition memory.

Experiments 2 and 3 investigated these ideas by comparing the effects of grammaticality and specific similarity across various training conditions thought to produce differing degrees of item individuation. In Experiment 2, three encoding conditions were used. In order of increasing item individuation, these were observation training, analogous to the observation condition used by Reber and Allen (1978); free-recall training, similar to that used in Experiment 1; and mnemonic training, a variant of the paired-associate training task. For this last condition, each training item was associated with a mnemonic phrase that resembled a newspaper headline such that the item would be an acronym of the key words of the phrase. For example, the List 1 training item MTTTWT was associated with the phrase “Montreal’s Thousands Take the TV Times,” and the List 2 training item MVX was associated with the phrase “Mandy Viewed X-rays.” The phrases were constructed to overlap partially with at least one other phrase in both thematic content and the words used, although no attempt was made to relate the consistency of the phrases to the grammatical consistency of the items. By making the mnemonic phrases for this condition less distinct, it was hoped that they would not necessarily destroy any chance that might otherwise have existed to see common patterns across the items.

In contrast, the two training conditions used in Experiment 3—label and unique mnemonic training—were designed to provide a unique encoding of each training item. For label training, each of the training items of a given list was associated with a unique animal name (e.g., beaver, moose) such that the animal referent was also unique (e.g., because moose was used, elk was not). Similarly, for the unique mnemonic condition, in which the mnemonics were constructed as in Experiment 2, the thematic content and the words used for each mnemonic phrase were unique to that phrase and item. Because of the greater specificity of the mnemonic phrases to their acronym items over the arbitrary pairing of items and animal names, unique mnemonic training was expected to result in greater item individuation than label training.

Both experiments used the basic paradigm described for Experiment 1 but with the following differences. First, 8 items rather than 16 were used in each training list. It appears that both the abstractive and distributive positions would hold that using a smaller number of training items would increase the subjects’ reliance on item-specific knowledge and, hence, specific similarity on a test of categorical transfer possibly at the expense of grammaticality. This result is well documented in the concept-learning and classification literature as the “category size” effect and is commonly interpreted as evidence for an increasing shift from instance to abstract knowledge as a function of increasing category size (e.g., Homa, Sterling, & Treppel, 1981). However, this effect also is consistent with the distributive position as we outline it here.

Second, the artificial grammar used to generate the training and grammatical transfer items is less complicated than that used in Experiment 1, bringing it more into line with the relatively simple grammars used in most other grammar studies but making it more difficult to control specific similarity orthogonal to grammaticality. To circumvent the resulting limited number and variation of items that may be generated, two approaches to producing close items were used: Two items were considered to be similar to one another if they differed from each other by no more than a single-letter substitution (e.g., VX and VXR), as in Experiment 1, or a single-letter addition (e.g., VX and VXRM). Finally, the recognition task followed rather than preceded the categorical transfer task in Experiments 2 and 3.

Method

Materials. Sixteen training items were selected from the pool of all possible items, three to eight letters in length, generated from the
Figure 2. Artificial grammar used in Experiments 2 and 3.

The artificial grammar shown in Figure 2. The training items were selected from the items that were three, five, and seven letters in length. Consequently, substitution transfer items were three, five, and seven letters in length, and addition transfer items were four, six, and eight letters in length. As in the previous experiment, each set of training and matched grammatical transfer items was chosen to be at least two letters in position different from the training items of any other matched set. The corresponding nongrammatical items were produced from the grammatical items by substituting the letters M, X, R, and T for the letters X, M, T, and R, respectively, in the grammatical items in the same letter position that the grammatical items differed from their matched training items; these substitutions are illegal according to the grammar. The training items and the corresponding transfer items are shown in Table 4.

As in the previous experiment, the training items were divided into two lists constructed such that the item lengths, the initial and terminal letters, and the grammatical subrules were reasonably matched across the two lists. Each training list was divided into two sublists of 4 items each, and five orders of the items in each sublist were generated. For the mnemonic training condition of Experiment 2 and the label and unique mnemonic training conditions of Experiment 3, the associated phrase or label of each training item was printed next to the item and separated with a dash. For the transfer test, the 32 substitution transfer items were divided twice into two 16-item lists, as were the 32 addition transfer items, for a total of eight transfer lists, with each item occurring twice across lists. Two recognition lists were constructed: One contained the List 1 training items, and the other contained the List 2 training items. For each list, a sample of 16 transfer items was chosen, balanced for grammaticality, alteration type (i.e., substitution or addition), and specific similarity.

Subjects: In Experiment 2, 96 university undergraduates, 32 in each of three training conditions participated in exchange for course credit in introductory psychology. They were tested in groups of 4 to 20, and subjects in the free-recall and mnemonic training conditions were deliberately tested together. Observation subjects were tested separately to control the time they had to study the items. In Experiment 3, 64 university undergraduates, 32 in each of the two training conditions, participated in exchange for either course credit in introductory psychology or $2.00 each. Subjects were tested in groups of 4 to 20, and training conditions were deliberately mixed within a testing session.

Procedure: The general procedure for both experiments followed that of Experiment 1: All instructions and materials were presented to the subjects in booklets, and one half of the subjects in each training condition received List 1, and the remainder received List 2. The sublist of each training list that was presented first was counterbalanced across subjects within a given training list, and the order of presentation of the five orders of a given sublist was systematically rotated across subjects within sublist and training list. Subjects received all five orders of a sublist before proceeding to the next sublist. Between sublists, the training instructions were repeated, and the subjects were informed that they would be receiving a new list of four items. Subjects were not informed about the number of training sublists they would receive or that their task consisted of anything more than the training task.

Subjects given observation training were asked to study the four items on each page for 90 s and to turn to the next page when signaled by the experimenter. Subjects in the free-recall condition of Experiment 2 were given the same training task as that given subjects in Experiment 1. Subjects in the mnemonics conditions of both Experiments 2 and 3 and in the label training condition of Experiment 3 were similarly treated except that they were asked to recall both the item and its associated phrase or name to ensure that they would attempt to use the associates during training.

Subjects received the categorical transfer test as outlined in Experiment 1. As in the previous experiment, each item was presented twice for a total of 128 separate judgments, although subjects were not informed of this fact. One half of the subjects in each training condition received the substitution items before the addition items, and the remaining subjects received them in the opposite order. Unlike the previous experiment in which subjects used a response scale, subjects in the present experiment were instructed to place a check mark next to each item that they believed obeyed the rules and to leave all other items blank. Following the categorical transfer test, subjects were given the recognition test appropriate to the training list received. They were instructed to indicate which of the 24 items they had seen during the training phase. With the exception of the training phase for observation subjects, all phases of the experiments were subject paced.

Results

Recognition. The major results of the recognition phase, analyzed as in the previous experiment, are shown in the middle two panels of Table 2. Considered as a recognition task, mnemonic training led to better recognition than the other two conditions in Experiment 2 with a mean proportion correct of $M = .79$. Actively attempting to learn the items, as subjects were asked to do in the free-recall condition, led to more accurate recognition ($M = .77$) than did merely observing them ($M = .74$). These differences were significant for false positives, $F(2, 93) = 7.10, MS_e = 0.037$, but not for hits ($F < 1$) as may be seen in Table 2. In Experiment 3, subjects trained with unique mnemonics were more accurate at identifying the items ($M = .85$) than were the label-trained subjects ($M = .76$). Again, the difference in accuracy was a function
of a significant difference in the mean rate of false positives, $F(1, 62) = 11.79, MS_e = 0.024$, although in this case, as is shown in Table 2, there was a marginally significant effect on hits as well, $F(1, 62) = 3.60, MS_e = 0.049, p = .06$.

The effects of both grammaticality and specific similarity on recognition false positives were significant for both Experiments 2 and 3. For Experiment 2, a significantly greater mean proportion of false positives was produced to grammatical items ($M = .13$) than to nongrammatical items ($M = .06$), $F(1, 93) = 26.62, MS_e = 0.019$, and to close items ($M = .12$) than to far items ($M = .06$), $F(1, 93) = 13.10, MS_e = 0.023$. Grammaticality was the larger of the two effects, although not significantly ($z = .99$). For Experiment 3, subjects produced a significantly greater mean proportion of false positives to grammatical items ($M = .09$) than to nongrammatical items ($M = .04$), $F(1, 62) = 6.26, MS_e = 0.017$, and to close items ($M = .09$) than to far items ($M = .04$), $F(1, 62) = 10.63, MS_e = 0.019$. Specific similarity was the larger of the two effects, although not significantly ($z = .50$). It evidenced a marginally significant interaction with training condition in the direction of a greater effect of specific similarity for subjects given label rather than unique mnemonics training, $F(1, 62) = 3.32, MS_e = 0.019, p = .07$. None of the remaining effects for Experiment 3 were significant.

For Experiment 2, there also was a significant interaction of grammaticality and specific similarity, $F(1, 93) = 4.40, MS_e = 0.016$, with the effect of grammaticality larger for close items than for far items, although just how much larger varied significantly as a function of training condition, $F(2, 93) = 3.81, MS_e = 0.016$. As is shown in Table 2, neither of these interaction effects disordinally affected the lower-order main effects of grammaticality and specific similarity. The remaining effects were not significant.

### Categorical transfer

The major results of the categorical transfer phase are shown in the second and third panels of Table 3. For both Experiments 2 and 3, the mean proportion of items labeled grammatical was subjected to a three-way ANOVA, with subjects nested within training condition as the random variate. The combination of the independent variables of grammaticality and specific similarity resulted in four cells per subject with 32 responses per cell. For Experiment 2, as in the previous experiment, the main effects of both grammaticality and specific similarity were reliable; a significantly greater proportion of grammatical items ($M = .31$) than nongrammatical items ($M = .27$) were called grammatical, $F(1, 93) = 9.11, MS_e = 0.020$, as were a significantly greater proportion of close items ($M = .36$) than far items ($M = .22$), $F(1, 93) = 171.64, MS_e = 0.011$. However, in contrast to Experiment 1, specific similarity was by far the significantly larger of the two effects ($z = 5.57$) as is shown in Table 3. For Experiment 3, the effect of specific similarity again was significant; subjects produced a significantly greater mean proportion of grammatical responses to far items ($M = .42$) than to close items ($M = .27$), $F(1, 62) = 105.99, MS_e = 0.013$. Unlike the previous experiments, however, there was no significant effect of grammaticality ($F < 1$); the mean proportion of grammatical items labeled grammatical was identical to that of nongrammatical items ($M = .35$).

For both Experiments 2 and 3, there was no significant main effect of training conditions ($Fs < 1$). Moreover, in neither experiment were subjects differentially affected by the grammaticality of the items as a function of training condition, respectively, $F < 1$ and $F(1, 62) = 1.35, MS_e = 0.017$. However, training condition did interact significantly with the specific similarity of the items in both experiments. As shown by the effect sizes in Table 3, subjects in Experiment
2 who had received mnemonic training were less influenced by the specific similarity of the items than were subjects in the observation or free-recall conditions, \( F(1, 93) = 3.79, MSe = 0.011 \). For Experiment 3, the effect of specific similarity, as in recognition, was larger for subjects given label rather than unique mnemonic training, \( F(1, 62) = 7.32, MSe = 0.013 \). The remaining effects were not significant.

**Discussion**

To test the effect of increased item individuation during training on later transfer, it must first be established that the training conditions produced differences in item differentiation. Taking recognition memory performance as an index of differentiation, these differences were demonstrated by the significant decline in false-alarm rates in both Experiments 2 and 3 across learning conditions. If increased emphasis on item learning during training produces less learning of an abstract grammar and by default a greater emphasis on memory for individual items, then we would expect an increasing effect of specific similarity in both the recognition memory and grammatical judgment tasks. Instead, the effect of specific similarity on categorical transfer in both experiments significantly decreased in response to increasing item individuation. Although not demonstrated for Experiment 2, the recognition results of Experiment 3 also showed a reliable decrease in the effectiveness of the specific similarity variable to determine false alarms. Thus, the first result of these experiments is that learning conditions that result in enhanced retrieval of specific item information can decrease the effectiveness of such information to support generalization. Apparently, items can be too well differentiated to support effectively transfer performance on new items.

However, what was not observed within either experiment was a reliable decrease in the effect of the grammaticality variable with increasing item differentiation. The effect of grammaticality was not even in the right direction in Experiment 2 for either recognition memory or grammaticality judgments or for recognition memory for Experiment 3; the results for grammaticality judgments in Experiment 3 were suggestive of a decrease but were not significant. The combination of a reliable decrease in the effect of the specific similarity variable and no consistent effect of the grammaticality variable does not encourage the idea that an increasing emphasis on learning individual items suppresses the tacit learning of an abstract grammar, transferring control to reliance on specific items, as suggested by Reber and Allen (1978). Instead, this pattern of results supports the notion that the poorer transfer performance in Reber and Allen's experiment of conditions with good item learning may have occurred not because the subjects switched from an abstract knowledge base to a less efficient, episodic one but rather because one episodic source of their judgments—generalization from remembered, specific exemplars—was rendered less effective.

Although there has been no differential effect of the grammaticality variable on categorical transfer within experiments, inspection of the effect sizes in Table 3 reveals a marked reduction in the size of the effect of grammaticality across experiments, from a large effect \( r_{pb} = .80 \) in Experiment 1 to a moderate effect \( r_{pb} = .30 \) in Experiment 2 to no effect at all \( r_{pb} = -.02 \) in Experiment 3. The overall effect size of the specific similarity variable, in contrast, remained relatively constant at about \( r_{pb} = .78 \). The same pattern is evident for the recognition false-alarm data in Table 2. The most obvious way to account for this substantial difference between Experiment 1 and Experiments 2 and 3 is by the reduction in the number of training items from 16 to 8. Fewer training items provide less evidence for a process that relies on commonality across items. As mentioned previously, this process could be one of abstraction of a grammar during training or postcomputation at the moment of retrieval. This large effect of category size was not directly tested in this study because it occurred across experiments. By contrast, the manipulations in Experiments 2 and 3 that were designed to influence grammaticality were relatively ineffective. The training tasks within Experiment 2 varied over what should have been an effective range according to both Reber and Allen (1978) and Vokey (1983) produced no significant effect at all. The reduction in effect size between Experiments 2 and 3 may reflect the influence of training task variations but again varied between experiments and so cannot be directly tested. This possibility is discussed further in Experiment 4.

Again, however, this reduction in the effect of grammaticality across experiments was not accompanied by a corresponding increase in the effectiveness of specific similarity. At the extreme, in the unique mnemonic condition of Experiment 3, the effect of grammaticality was completely eliminated, recognition memory was more accurate than in any other condition in these three experiments, and yet the effect size of specific similarity was numerically the smallest rather than the largest among these conditions. The difference in the response of the specific similarity and the grammaticality variables within and across experiments suggests that the use of one source of information is not contingent on the current output of the other for judgment about a particular item. For example, grammaticality is not used only to the extent that the transfer item fails to invoke retrieval of a specific, similar training item, and the use of specific similarity is not used only to the extent that grammaticality fails to be sufficient.

The results of these experiments also provide unique evidence that the pooling of information across items is important be it an abstracted grammar or retrieval time averaging. We previously (Vokey, 1983; and in a previous incarnation of the paper incorporating Experiments 2 and 3) entertained the possibility that the effects on grammaticality judgments described by Reber and Allen (1978) could be produced simply by variations in response bias. Because item similarity and grammaticality were confounded in their experiment, a loose criterion of similarity to known old items would be likely to include more grammatical than nongrammatical items. As a result, learning conditions that produced greater item differentiation and consequently more stringent criteria for responding “grammatical” on the basis of item similarity might result in lower accuracy. Conversely, a looser criterion for item similarity, as might be produced by the observation condition, would produce a response of “grammatical” to more neighbors, neighbors that would be more likely to be
grammatical than nongrammatical. However, such a mechanism would predict a strong, positive relation between the frequency of saying “grammatical” and the effect of our specific similarity variable. An inspection of the difference between Experiment 1 and Experiments 2 and 3 indicates that this was not the case. Consequently, although differences in bias resulting from differences in item individuation can explain some part of the effect of different learning conditions, they cannot by themselves eliminate a need to rely on some independent effect of grammaticality. McAndrews and Moscovitch (1985) reported an independent line of evidence that also eliminates the sufficiency of response bias as an explanation of the grammaticality effect. Their subjects were given forced choice between pairs of items contrasting all four combinations of the similarity and grammaticality variables. They found an effect of grammaticality after specific similarity was accounted for, which because of the forced-choice nature of their task could not be explained by response bias.

Experiment 4: Interactive Effect of Retrieval Conditions

Experiment 4 was designed to gain greater control over the availability and specificity of item information at test by directly manipulating the conditions of retrieval. Any differences observed in categorical transfer then would not be due to differential acquisition of the sources of knowledge but only to whether the test conditions differentially promoted their influence on the subjects’ judgments. The procedure we used is similar to one used in a series of experiments by Light and Carter-Sobell (1970; see also Hunt & Ellis, 1974): the heart of the paradigm for demonstrating encoding specificity effects in recognition memory.

In Light and Carter-Sobell’s (1970) paradigm, adjectives are associated with homograph nouns during the study phase to bias one of the possible meanings of the noun. For example, the subjects’ interpretation of the noun jam is biased toward one of its possible meanings by the adjective in strawberry jam. During a subsequent test of recognition, the homograph noun is presented either in a similar context that biases a similar interpretation (e.g., raspberry jam) or in a new context that biases a different interpretation of the noun (e.g., traffic jam). The subjects are required to indicate which of the nouns (not phrases) were encountered during training. Homograph nouns presented in similar contexts across training and test are better recognized than those for which the contexts have been changed from training to test. Recognition is further enhanced if the adjective–noun pairing is literally identical across training and test (Light & Carter-Sobell, 1970). Similar results have been obtained when the familiarity of the test contexts is unconfounded with the context manipulation, for example, with preexposure of the adjective traffic, although this increase in the overall familiarity of the test context appears to moderate the effect (Donaldson, 1981). This context effect, then, appears to be a result of inducing matches or mismatches between study and test in the way items have been processed rather than in the formal properties of the test situation.

The intent of this next experiment was to extend this encoding specificity design to a test of categorical transfer. Each training item was associated with a different context, and each transfer item, independently of its status as a grammatical or nongrammatical item, was associated either with an appropriate context that was similar to that of its matched training item or an inappropriate context that was similar to that of some training item other than the one to which the item was matched. To the extent that the retrieval of specific item information plays an important part in subjects’ transfer judgments, then by analogy with Light and Carter-Sobell’s (1970) results, subjects should respond preferentially to close items in appropriate contexts while generally rejecting all other items, including those that are formally close but associated with inappropriate contexts.

Such an effect of retrieval context is not a critical challenge to an abstractive position. There is no reason why an abstractive position could not add principles to cover retrieval time phenomena. However, that such proposals are not made routinely emphasizes a difference in the focus of the two approaches. Abstractive approaches focus on the stable properties of the abstracted knowledge base as an explanation of test performance and tend to treat retrieval effects as secondary. In contrast, part of the rationale of a distributive approach is that a distributed knowledge base can in principle respond easily to important local conditions, an ability that is particularly important when there are interactions that change the values of generally predictive cues (e.g., Brooks, 1987; Kahanman & Miller, 1986). The purpose of the current experiment, then, is to demonstrate the power of variations in the retrieval context and to emphasize the importance of considering the way previously acquired knowledge interacts with retrieval context.

Method

Materials. The materials were the same as those used in Experiment 1 with the exception of the modifications required to present the items with associated contexts during training, recognition, and categorical transfer. Sixteen unique themes were used as the biasing contexts of the items. Each theme was assigned to a training item in each of the two training lists and, as in the previous experiments, was expressed in a phrase resembling a newspaper headline such that the item would be an acronym of the key words of the phrase. For example, one of the themes involved the scenario of angry Muppets exiting a restaurant and was expressed as “Muppets Vacate Restaurants Visibly Miffed” for the List 1 training item MVVRM and as

\[ MVVRM \]

In a similar experiment, sets of animal names (e.g., moose, elk, deer, and lion, tiger, leopard) were used as the contextual associates. Although the effects of context appropriateness were in the expected direction, none proved statistically reliable for either the categorical transfer or recognition phases. The failure to find the expected effects even during recognition where Light and Carter-Sobell (1970) previously had demonstrated analogous effects may have resulted from the arbitrariness of the pairing of animal names and items. The use of phrases as the associated contexts in the present experiment in which each item is an acronym of the phrase was an attempt to render the associations less arbitrary, thereby boosting the magnitude of effect.
"Muppets eXit Restaurants Vehement, Verbose, (and) Visibly Miffed" for the List 2 training item MXRVVM.

To support the manipulation of context appropriateness across categorical status and specific similarity, each theme (for a given training list) was associated with a specifically similar transfer item and a dissimilar transfer item. For one half of these assignments, the specifically similar item was grammatical and the dissimilar item nongrammatical, and for the remaining assignments the reverse was true. For the resulting 6 items (across the two lists) of a thematic set, the theme was expressed for each item using key words unique to that set. Thus, to continue the earlier example, the specifically similar grammatical transfer item of MVRVM, MXRVVM, was associated with the appropriate contextual phrase "Muppets eXit Restaurants Visibly Miffed," whereas the specifically similar nongrammatical item, MTRVM, was associated with the inappropriate context from a different theme expressed as "Monkeys Trail Relatives (and) Vagrants Merrily." Across the 64 transfer items, each theme occurred four times, once with a close grammatical item and again with a close nongrammatical item for a subject trained with a given list and once with a far grammatical item and again with a far nongrammatical item. Thus, by themselves, the themes provided no basis for discriminating among the items with respect to either grammaticality or specific similarity. Furthermore, although one half of the transfer items were presented in appropriate contexts, this distinction is meaningful only for the close transfer items of subjects trained with a given list (i.e., because subjects trained with a given list never see the theme-to-training-item pairings of the alternate training list, all far items were, in a sense, associated with inappropriate contexts). However, the distinction for far items is retained for analysis because it provides direct control for potential a priori item, theme, and item–theme pairing differences.

Subjects. Thirty-two university undergraduates participated in exchange for either a course credit in introductory psychology or $3.00 each. They generally were tested in groups of 2 to 14, although a few were tested individually.

Procedure. The general procedure was the same as that for the grammar transfer condition in Experiment 1. For each phase, the items were presented separately with a dash from their associated contextual phrase. For training, the subjects were instructed to study carefully the four items and associated phrases on each page and then to recall the items and their associated phrases on a following blank page. After training, the subjects received the recognition and categorical transfer tests as outlined in Experiment 1. For both tasks, they were informed that, although they may find the phrases associated with each item to be helpful, their task was to judge the letter strings.

Results

Recognition. The major results of the recognition phase are shown in the bottom panel of Table 2. The subjects were remarkably accurate in recognizing the training items, scoring a mean proportion of $M = .96$ hits and only $M = .08$ false positives. It is clear from the extreme disparity between the hit and false-positive rates that the subjects rarely considered anything but the training stimuli (items and associated phrases) to be old and rarely missed recognizing them. Apparently, the effect of having the mnemonic phrases at both training and test is, as intended, to enhance retrieval of the specific training experiences when confronted with the training stimuli while, to a large degree, preventing this retrieval, or at least preventing misrecognition on the basis of such retrieval, for new items considered as a whole.

The false-positive data were subjected to a three-way ANOVA within subjects. The combination of the two levels per each of the three factors of grammaticality, specific similarity, and context appropriateness resulted in eight cells per subject with four responses per cell. The primary results for the effects of grammaticality and similarity are summarized in the bottom panel of Table 2, and those for the effect of context appropriateness may be found in the upper panel of Table 5. As in Experiment 1, the main effects of both grammaticality and specific similarity were significant; the mean false-positive rate for grammatical items ($M = .11$) significantly exceeded that for nongrammatical items ($M = .06$), $F(1, 31) = 6.49$, $MS_{s.e.} = .022$, and the mean false-positive rate for close items ($M = .11$) significantly exceeded that for far items ($M = .05$), $F(1, 31) = 9.55$, $MS_{s.e.} = .020$, with the effects again being strictly additive ($F < 1$). Specific similarity was the larger of the two effects, although not significantly ($z = .34$).

Context appropriateness was significant as a main effect; the mean false-positive rate to items presented with appropriate contextual phrases ($M = .10$) was almost twice that to items presented with inappropriate contextual phrases ($M = .06$), $F(1, 31) = 5.93$, $MS_{s.e.} = .024$. This effect was not independent of the specific similarity of the items, $F(1, 32) = 5.74$, $MS_{s.e.} = .017$; in fact, as is shown in Table 5 and assessed by Fisher's least significant difference (LSD) test, the main effect of context appropriateness (or, equivalently, the main effect of specific similarity) was the result of close items in appropriate context recruiting significantly more false positives than did either close items in inappropriate contexts or either type of far items, which did not differ significantly among themselves (LSD = .05). None of the remaining effects was significant.

Categorical transfer. The proportion of items receiving positive responses were subjected to a three-way ANOVA within subjects. The combination of two levels for each of the independent variables of specific similarity, grammaticality, and context appropriateness resulted in eight cells per subject with 16 responses per cell. The major results of the categorical transfer phase for grammaticality and specific similarity are shown in the bottom panel of Table 3. Both grammaticality and specific similarity were significant as main effects; the mean response rate to grammatical items ($M = .57$) was slightly but significantly greater than that to nongrammatical items ($M = .52$), $F(1, 31) = 6.80$, $MS_{s.e.} = .020$, whereas the mean response rate to close items ($M = .65$) was much greater than that to far items ($M = .44$), $F(1, 31) = 67.10$, $MS_{s.e.} = .020$. These effects were significant as main effects; the mean response rate to grammatical items ($M = .57$) was slightly but significantly greater than that to nongrammatical items ($M = .52$), $F(1, 31) = 6.80$, $MS_{s.e.} = .020$, whereas the mean response rate to close items ($M = .65$) was much greater than that to far items ($M = .44$), $F(1, 31) = 67.10$, $MS_{s.e.} = .020$.

Table 5
Mean Proportion of Items Given Positive Responses as a Function of Specific Similarity and Context
Appropriateness for the Recognition and Categorical Transfer Phases of Experiment 4

<table>
<thead>
<tr>
<th>Phase</th>
<th>Appropriate</th>
<th>Inappropriate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Close</td>
<td>Far</td>
</tr>
<tr>
<td>Recognition</td>
<td>.15</td>
<td>.06</td>
</tr>
<tr>
<td>Categorical transfer</td>
<td>.78</td>
<td>.44</td>
</tr>
</tbody>
</table>
0.039. Specific similarity was the significantly larger of the two effects ($z = 2.76$), and the two effects were additive, $F(1, 31) = 2.9$, $MS_e = 0.005$.

The results of the categorical transfer phase for context appropriateness are shown in the lower panel of Table 5. Items presented in appropriate contexts received a significantly greater mean proportion of positive responses ($M = .61$) than did items in inappropriate contexts ($M = .48$), $F(1, 31) = 48.46$, $MS_e = 0.022$. As with the results of the recognition phase, however, this effect interacted significantly with specific similarity, $F(1, 31) = 32.45$, $MS_e = 0.039$. As may be seen in Table 5, the main effects of both factors were primarily a function of elevated responding to close items in appropriate contexts; Fisher's LSD test indicates that subjects responded positively significantly more often to these items than to items in any of the remaining three cells (LSD = .07). None of the remaining effects was significant.

Discussion

The results of this experiment provide a clear demonstration of powerful effects of specific item retrieval on transfer. There was no effect of specific similarity unless the context was appropriate for retrieving the closest training item. When the context of a test item supported the retrieval of the specific, highly similar training item, the subjects much more often than not were willing to call the item grammatical and on recognition were twice as likely mistakenly to call the item old. If, alternatively, the context was more appropriate for the retrieval of some dissimilar item (i.e., as would be true for inappropriate close items and both types of far items), the specific similarity of the item was irrelevant.

Grammaticality continued to exert a small but independent effect. In this case, it was independent not only of specific similarity but of the appropriateness of the retrieval context. This independence of effect is important because it supports the contention of a source of knowledge informing the subjects' judgments separate from the retrieval of individual exemplars and under the control of different factors. None of the variables we manipulated within each of the experiments had any effect on the influence of this knowledge base. As noted in the discussion of Experiments 2 and 3, it is only across experiments that we find appreciable variation in the effect of grammaticality. The same was true here. Comparing the effect sizes in Tables 2 and 3 of the current experiment with those of Experiment 1 (in which the same training and transfer items were used) shows that, although the overall effect of specific similarity demonstrated a slight increase, the effect of grammaticality plummeted. We suspect that this reduction in the magnitude of the effect of grammaticality reflects a change in the salience of the retrieval of individual specific items across the experiments.

The unique aspect of this experiment is the provision of specific retrieval cues at test. These retrieval cues could easily have restricted retrieval to a single item for some (close) items and no retrieval at all for others. If strong item retrieval actively suppresses retrieval of the grammaticality information, an interaction between the variable of grammaticality and the variables affecting retrieval is expected. Far items in inappropriate retrieval contexts, for example, for which specific item retrieval would be at a minimum, should have evidenced a much larger effect of grammaticality than did close items in appropriate contexts. No such effect was detected.

General Discussion

The experiments used material that at least partially unconfounded the similarity of test items to particular training items and the similarity of the test items to some pooled representation of multiple items. The first result of these experiments is that both empirical variables—grammaticality and specific similarity—had an effect on the recognition and grammaticality judgments of subjects within the same task. The effect of specific similarity was not restricted to situations in which there were an unusually small number of training items or that placed an unusual focus on memorization by the subjects. Conversely, as shown in Experiment 4, providing retrieval conditions that are extreme in their ability to retrieve particular items did not eliminate an effect of grammaticality.

As a result of the robustness of these two variables, the key issue to be addressed is the way in which these types of knowledge are coordinated. There are several results among the present experiments that bear on this issue. First, the importance of individual training items in supporting grammaticality judgments on new items cannot be assumed to be positively related to the “goodness of learning” of the items; learning conditions that increased the differentiation of items as indexed by improved recognition memory resulted in smaller effects of specific similarity on the transfer tasks. Apparently, items can be too well differentiated to support effectively transfer performance on new items. Second, this decrease in the effect of specific similarity was not accompanied by an increase in the effect of grammaticality; it is perfectly possible to have combinations of learning and retrieval conditions in which the effectiveness of both sources of information for supporting generalization simultaneously increased or decreased. These results suggest that no scheme that tightly links the effects of the two variables in a compensatory relationship, be it degree of grammar learning or tightness of similarity criterion, will work by itself. Indeed, over most of the experiments here, the effects of specific similarity and grammaticality were additive. At an extreme, represented by the unique mnemonics condition of Experiment 3, the effect of the pooled source of knowledge was completely eliminated, but the effect of specific similarity was also reduced. However, this result, although it does not support Reber and Allen's (1978) suggestion of compensatory relation, certainly does support their notion that an extreme of item-specific encoding reduces the effect of grammaticality. The same conclusion follows from the reduction in the effect of grammaticality across Experiments 1 and 4. Finally, there is not the degree of task control over these two variables that would be expected if they normally served different purposes. In Experiment 1, there were no differences in the effect of grammaticality between subjects instructed to sort on the basis of grammaticality and those asked to sort on the basis of specific similarity. As mentioned, this lack of shift in task
control cannot be taken too lightly by an abstractive position; there has to be some degree of task control over grammaticality if the normal purpose of this form of information is to respond to generalizing tasks.

The salience of individual item information and retrieval time manipulations that we have discussed are unfortunate complexities from the abstractive point of view. In some sense, they are noise that the investigator must look through to examine the abstracted knowledge base that is presumed to be the mainstay of most classification decisions. However, we believe that this background assumption has deflected attention away from retrieval conditions and consequently away from cues that help to tailor knowledge to complex interactions. We do not think of the encoding-specific conditions of Experiment 4, for example, as a special case; they just bring under analytic control that which is inherent, but less obvious, in virtually any classification experiment. Note that this is not to question, necessarily, the existence of abstracted knowledge for any domain, only its completeness and usefulness as a description of the determinants of performance. As S. W. Allen and Brooks (1991) demonstrated with artificial materials and Brooks, Norman, and Allen (1991) showed with real-world stimuli, the retrieval of a single, prior interpreted instance can contribute markedly to performance even in domains with relatively simple, easily articulated, abstract characterizations that are known to the subjects and claimed by them to be the source of their behavior. From our perspective, the blending of item and abstractive information within a transfer session is not a minor modification of the abstractive position. To some extent, it forces the attention to the details of retrieval that we have been advocating. In the absence of attention to item knowledge and the conditions of retrieval, these effects likely are attributed to lapses of attention, random responding, or, when more successful than the rules themselves, intuition and deep insight.

There are probably many situations in which the learner is in the position of coordinating "on the fly" information from different prior knowledge sources into single categorical judgments. For such a process, the nature of the retrieval conditions and how they interact with the details of the original coding are of crucial importance. In fact, this is the major reason for our continued interest in working with artificial grammars. Any such investigation, however, requires the use of materials that unconfound the effects of knowledge of specific items and knowledge of structure. We believe that several of the manipulations that have been used in the literature would be interesting to examine with such a design. For example, several studies (e.g., Mathews et al., 1989; Reber, 1976) demonstrated successful transfer when the letters that express a given grammar are changed from training to test. Such results have been taken as evidence that very abstract relationships have been learned. However, when specific similarity is unconfounded with grammaticality, there is still an effect of specific item similarity on categorical transfer over changed letter sets (Brooks & Vokey, 1991).

The grammar materials provide a shifting pool of partially learned items that can interact with the specificity of the cuing conditions. These are exactly the conditions that would be important for investigation of the chorus of instances that is a prominent feature of many contemporary models of classification and recognition memory. Despite the difficulty of working with the grammar materials, some such preparation is important for developing the implications of multiple instance models and indeed any hybrid-knowledge approach to knowledge of structure.

References


Received April 26, 1990
Revision received July 22, 1991
Accepted August 15, 1991

---

MEMBERS OF UNDERREPRESENTED GROUPS: REVIEWERS FOR JOURNAL MANUSCRIPTS WANTED

If you are interested in reviewing manuscripts for APA journals, the APA Publications and Communications Board would like to invite your participation. Manuscript reviewers are vital to the publication process. As a reviewer, you would gain valuable experience in publishing. The P&C Board is particularly interested in encouraging members of underrepresented groups to participate more in this process.

If you are interested in reviewing manuscripts, please write to Leslie Cameron at the address below. Please note the following important points:

- To be selected as a reviewer, you must have published articles in peer-reviewed journals. The experience of publishing provides a reviewer with the basis for preparing a thorough, objective review.
- To select the appropriate reviewers for each manuscript, the editor needs detailed information. Please include with your letter your vita. In your letter, please identify which APA journal you are interested in and describe your area of expertise. Be as specific as possible. For example, "social psychology" is not sufficient—you would need to specify "social cognition" or "attitude change" as well.
- Reviewing a manuscript takes time. If you are selected to review a manuscript, be prepared to invest the necessary time to evaluate the manuscript thoroughly.

Write to Leslie Cameron, Journals Office, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.