THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
2010, 63 (5), 1014–1039

Psychology Press
Taylor & Francis Group

Applying an exemplar model to the artificial-grammar task: String completion and performance on individual items

Randall K. Jamieson
University of Manitoba, Winnipeg, Manitoba, Canada

D. J. K. Mewhort
Queen's University at Kingston, Kingston, Ontario, Canada

Jamieson and Mewhort (2009a) demonstrated that performance in the artificial-grammar task could be understood using an exemplar model of memory. We reinforce the position by testing the model against data for individual test items both in a standard artificial-grammar experiment and in a string-completion variant of the standard procedure. We argue that retrieval is sensitive to structure in memory. The work ties performance in the artificial-grammar task to principles of explicit memory.

Keywords: Artificial grammar; Exemplar theory; Implicit learning; Minerva 2; String completion.

Learning is difficult. Yet, people become sensitive to subtle regularities in the world without deliberate effort. Several theorists have explained the discrepancy by separate contributions from separate learning systems: one that handles deliberate learning and a second that learns contingencies automatically.

To address the possibility of automatic learning of contingencies, researchers often use a judgement of grammaticality (JOG) task. In a typical JOG task, participants study strings of letters ordered according to rules of a finite-state grammar, like the ones defined in Figure 1. After studying the strings, participants classify novel test strings as either grammatical or ungrammatical. Typically, people distinguish the two classes of stimuli (often achieving a score of about 60% correct), but cannot describe the grammar. Following Reber (1967), many researchers take the ability to sort grammatical from ungrammatical test strings as evidence that the participants have learned the grammar. Because participants cannot characterize the grammar, it is assumed that their knowledge of the grammar must be implicit (e.g., Cleeremans & Dienes, 2008; Dienes, Broadbent, & Berry, 1991; Knowlton,

Correspondence should be addressed to Randall K. Jamieson, Department of Psychology, University of Manitoba, Winnipeg, Manitoba, Canada. E-mail: randy_jamieson@umanitoba.ca

The research was supported by grants from the Natural Sciences and Engineering Research Council of Canada to both authors. We thank Phil Audette and Brian Hauri for their help in collecting the data. We thank Beth Johns and Matthew Kelly for comments on a draft of the paper. We thank Zoltan Dienes for providing us with his ranks and discussion of his data. Codes for the model and ample output are available from the first author.

http://www.psypress.com/qjep DOI:10.1080/17470210903267417

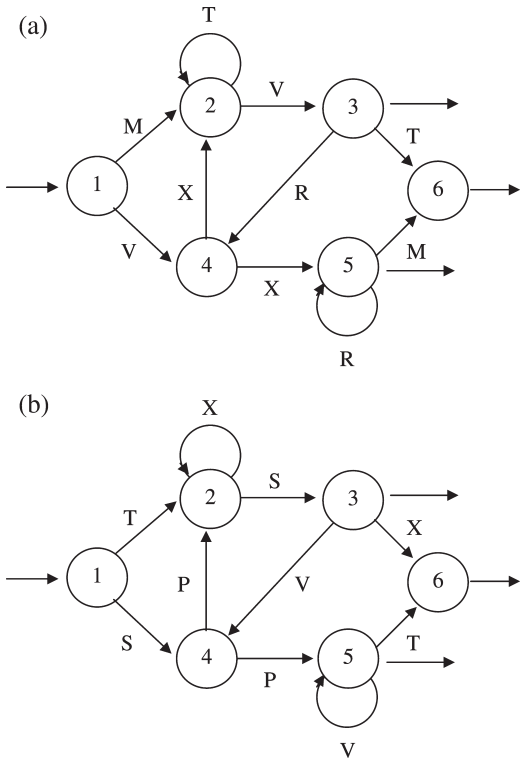

Figure 1. *Grammar a is redrawn from "Implicit and Explicit Knowledge Bases in Artificial Grammar Learning", by Z. Dienes, D. Broadbent, and D. Berry, 1991, Journal of Experimental Psychology: Learning, Memory, and Cognition, 17, p. 876. Copyright 1991 by the American Psychological Association. Adapted with permission. A grammatical stimulus is generated by starting at the leftmost node marked 1 and following the paths (indicated by arrows) until reaching one of the "exit paths" from Nodes 3, 5, or 6. When a path is taken, the associated letter is added to the end of the string. For example, moving from Nodes 1 to 2, 2 to 3, and 3 to 6 produces the string MVT. Any string of letters that cannot be produced with the grammar is "ungrammatical". Grammar b has the same structure as Grammar a but a different surface structure: The letters assigned to paths were remapped so that T replaced M, X replaced T, S replaced V, V replaced R, and P replaced X. Grammar a was used to construct items in Experiment 2. Grammar b was used to construct items in Experiments 1 and 3.*

Ramus, & Squire, 1992; Knowlton & Squire, 1994, 1996; Kuhn & Dienes, 2005; Manza & Reber, 1997; Reber, 1967, 1989).

Jamieson and Mewhort (2009a) have shown, however, that performance in a typical JOG task can be explained without assuming that participants know the grammar (either implicitly or otherwise). In their account, judgements of grammaticality are based on the probe's global similarity to memory of the studied items (see also Brooks, 1978; Pothos & Bailey, 2000; Vokey & Brooks, 1992). To illustrate the power of the global-similarity approach, Jamieson and Mewhort adapted a classic model of memory—Hintzman's (1986, 1988) Minerva 2—to the JOG task. The model matched human performance in three novel experiments and in three classic experiments (Dienes et al., 1991; Reber, 1967; Vokey & Brooks, 1992). In another paper, Jamieson and Mewhort (2009b) extended their account to predict performance in the serial reaction time (SRT) task.

An account's credibility includes not only its ability to explain phenomena of interest, but to handle new phenomena. Hence, to reinforce our global-similarity account, we challenged the model by applying it to two new phenomena. The first is a string-completion task derived from word completion in the implicit-memory literature. The second is a standard JOG task in which we ask whether the global-similarity account predicts not only overall performance but also individual differences among the test exemplars. Because both phenomena require the participant to apply knowledge derived from studying grammatical exemplars, they remain in the domain of the implicit-learning literature.

USING GRAMMATICAL KNOWLEDGE: STRING COMPLETION

Word completion is often used to examine implicit memory. In a typical word-completion task, word stems (e.g., BE_R) are presented, and participants are invited to complete the stem with the first word that comes to mind (e.g., BEER or BEAR). Priming is demonstrated when participants complete the stems to match words studied earlier in the experiment. Priming is said to be implicit when the participant is unable to recognize whether or not the word had been presented earlier in the experiment (Tulving, Schacter, & Stark, 1982).

Here, we adapt the word-completion procedure to the JOG task. Instead of measuring priming in recall, we measure participants' ability to fill in strings with an alternative to make it consistent with a grammar used to construct training items. For instance, after studying grammatical strings, participants might be tested with the fragment, *T_S*, paired with alternatives, *P*, *S*, *T*, *V*, and *X*. Their task is to select the alternative that makes the string grammatical. Presumably, if participants have learned the grammar, they will perform the task successfully.

How could participants supply a grammatical alternative? Proponents of the rule-learning view have no trouble with this possibility: Participants learn the grammar in training and apply that knowledge at test. The challenge for the account is to bring the principle to life in a competent model (see Cleeremans, Servan-Schreiber, & McClelland, 1989). Of course, selecting a grammatical alternative poses a similar challenge to Jamieson and Mewhort's (2005, 2009a, 2009b) account, because it forces us to generalize the existing model to a new paradigm.

The first experiment presents a string-completion experiment designed, first, to demonstrate that participants can complete test strings legally after studying a few grammatical exemplars and, second, to provide target data for the model. Subsequently, we test whether or not the exemplar model can match performance in the task.

EXPERIMENT 1: STRING COMPLETION

Experiment 1 examined participants' ability to complete test strings grammatically after studying a small number of grammatical training items. In the training phase, participants were asked to remember 20 exemplars: The strings were presented successively, each for 6 s. After the 20 exemplars had been presented, participants were told that the items had been constructed according to the rules of an artificial grammar. The participants were then given a string-completion test. On each trial of the string-completion test, an incomplete string was shown (i.e., one of the letters in a test string was replaced by an underscore). The participant selected one letter from a set of five to replace the underscore. The alternatives were the five consonants used in the training items. Following the string-completion test, participants were asked to articulate the rules of the grammar.

Method

Participants

A total of 126 students from the University of Manitoba undergraduate participant pool took part in the study. All participants reported normal or corrected-to-normal vision.

Apparatus

The experiment was administered on personal computers (PCs). Each PC was equipped with a 19-in. wide-screen monitor, a standard keyboard, and a standard mouse. Participants responded using the mouse to click on words displayed on the monitor and with the keyboard to report the rules of the grammar.

Materials

A total of 40 grammatical strings were constructed using the artificial grammar in the bottom panel of Figure 1 (i.e., Grammar b). The full set of materials is presented in Table 1.

A total of 20 of the items served as training items; the remaining 20 items served as test items. We deleted a letter from each test item and replaced it with an underscore. The letters deleted from the strings were selected purposefully so that the string could be completed legally by choosing one of the five choice alternatives and otherwise illegally (by choosing any one of the remaining four alternatives). This constraint facilitated a clear analysis of performance: For each string, the probability of selecting the legal alternative by chance was .2. To obtain data on the choice frequencies for each item, we presented the same training and test items to all 126 participants; however, the order in which items were presented to each participant was randomized.

Table 1. *The training and test strings used in Experiment 1*

Training strings	*Test strings*
SPSVP	S_S
SPSVPS	SPS_PV
SPSX	S_V
SPT	SP_T
SPVV	SPV_V
SPVVT	S_VVVT
SPVVVV	SP_X
SPXSVP	SPXX_
SPXSX	_PXXXS
SPXXSX	TSV_T
TSVP	_SVPV
TSVPS	TSVP_T
TSVPSX	TS_
TSVPVV	T_SVPS
TSVPXS	TXS_PV
TXS	TX_X
TXSVP	TXX_
TXSVPT	T_XSVP
TXXSX	TXX_S
TXXXXS	TX_XXSX

Note: All of the training strings are grammatical according to Grammar b in Figure 1. All of the test strings can be completed to be grammatical.

Procedure

Participants were tested in groups of 4 to 7. Each participant was seated in front of a different computer. The participants were told that they would be shown strings of letters and that they should try to remember each string.

The participant initiated the training phase by clicking on the message "Start" (displayed at the centre of the computer screen). After the participant clicked on the message, the screen was cleared for 750 ms. Immediately thereafter, the first training string was displayed at the centre of the screen for 6 s. Next, the screen was cleared, and after 750 ms the next string was displayed. The cycle repeated until all of the training strings had been presented.

Following the training phase, participants were told that the strings had been constructed according to the rules of an artificial grammar, and they were invited to perform a string-completion test. Participants were instructed to complete the test strings so that they were consistent with the artificial grammar that had been used to construct the training strings.

The participant initiated the test phase by clicking on a button marked OK. When the participant clicked on the button, the screen was cleared; 1 s later the first test string was displayed. The string was centred horizontally and was displayed about 2 cm above the centre of the screen. The instruction, "Select a letter" was presented approximately 3 cm below the string. The five response alternatives (i.e., *P*, *S*, *T*, *V*, and *X*) were presented approximately 1 cm below the instruction. When the participant clicked on a response alternative, the screen was cleared; 1 s later, the next string was displayed (the instruction and five alternatives were presented as well). The cycle continued until all of the 20 test strings had been presented.

Following the string-completion test, a text editor appeared on the computer screen. A message above the text editor instructed the participant to describe the rules that he or she thought had been used to construct the training strings.

Results

Table 2 presents the data from the string-completion task. Each row of the table shows one test string along with the choice frequencies for the five response alternatives (i.e., the number of participants that chose each alternative).

Participants completed strings legally on 41% of the trials (chance = 20%), significant by a binomial test, $p < .05$. One can look at the data in Table 2 to see the alternative that participants chose most frequently for each test string as well as the corresponding legal completions. As shown, the most frequently chosen alternative completed 14 of the 20 test strings legally (Items 1, 3, 4, 5, 6, 9, 10, 11, 13, 14, 16, 17, 19, and 20), for a success rate of 70% (chance = 20%).

Despite their ability to complete strings legally, none of the participants described the grammar. Indeed, most declined even to attempt a description. The result replicates the standard dissociation

Table 2. *Frequency with which choice alternatives were selected to complete each of the items in Experiment 1*

			Response alternatives					
Item number	*String*	*Legal completion*	*P*	*S*	*T*	*V*	*X*	$r_{obs,pred}$
1	S_S	P	**46**	4	15	17	44	.75
2	SPS_PV	V	1	8	43	**28**	46	.14
3	S_V	P	**68**	1	23	10	24	.87
4	SP_T	V	1	6	5	**66**	48	.53
5	SPV_V	V	18	13	31	**32**	32	.25
6	S_VVVT	P	**60**	0	3	33	30	.75
7	SP_X	S	2	**29**	34	39	22	.49
8	SPXX_	S	7	**30**	32	21	36	.43
9	_PXXXS	S	0	**56**	52	14	4	.98
10	TSV_T	P	**65**	16	2	9	34	.86
11	_SVPV	T	23	4	**85**	5	9	.97
12	TSVP_T	V	2	**30**	2	36	56	.65
13	TS_	X	36	4	5	**15**	66	−.26
14	T_SVPS	X	21	6	0	20	**79**	.85
15	TXS_PV	V	7	8	39	**33**	39	.22
16	TX_X	S	13	**61**	2	24	26	.82
17	TXX_	S	17	**60**	8	19	22	.92
18	T_XSVP	X	**19**	53	2	26	26	−.01
19	TXX_S	X	38	**3**	6	19	60	.31
20	TX_XXSX	X	22	35	2	17	**50**	.82

Note: The item selected by Minerva 2 is shown in bold in the choice frequency data. $r_{obs,pred}$ is the correlation between participants' choice frequencies (obs) and the model's mean echo intensities (pred) across the five response alternatives.

of performance and awareness in studies of judgement of grammaticality.

Participants' preference for legal alternatives combined with their inability to articulate the grammar suggests that they knew the grammar implicitly. We have argued earlier that participants do not learn the grammar; instead, performance reflects the global similarity of each probe to the participants' memory of studied exemplars (Jamieson & Mewhort, 2009a). Hence, the current results challenge our account to show that it can be extended to handle the new string-completion task.

Minerva 2

Our model is based on Minerva 2, a multitrace theory of retrieval from memory. When a participant studies an item, the event is encoded to memory as a unique trace. In the model, each trace is represented using a unique random vector of n elements. Each vector element takes one of two values: +1 or −1 with equal probability—that is, $p(+1) = p(-1) = .5$. An association between items is represented by concatenating the constituent item vectors to form a new vector of larger dimensionality.

Memory is treated as a matrix with one row (vector) for each studied event, or string of associated events. Encoding an event involves copying the corresponding vector to a new row in the memory matrix. Encoding is sometimes imperfect. Imperfect encoding is implemented by setting some vector elements to zero (indicating that the element is indeterminate or unknown). A parameter, L, controls the probability with which an element is stored. As L increases, encoding quality improves.

All retrieval is cued. When a retrieval cue is presented, it activates each trace in memory in proportion to its similarity to the cue. The activated traces are aggregated into a composite called the *echo*; the contribution of each trace to the echo is based on its activation.

The similarity of trace, i, to the probe, P, is computed as,

$$S_i = \frac{\sum_{j=1}^{n} P_j \times M_{ij}}{n_R}$$

where P_j is the value of the jth feature in the probe, M_{ij} is the value of jth feature of the ith row in memory, and n_R is the number of features in the vectors that are relevant to the comparison (i.e., the number for which either P_j or M_{ij} are nonzero). Like the Pearson r, the similarity of the ith item to the probe, S_i, is scaled to the interval $[-1, +1]$. Similarity equals $+1$ when the trace is identical to the probe, 0 when the trace is orthogonal to the probe, and -1 when the trace is opposite to the probe.

The ith trace's activation, A_i, is the cube of its similarity to the probe—that is,

$$A_i = S_i^3$$

The activation function exaggerates differences in similarity between a probe and items in memory by attenuating activation of exemplars that are not highly similar to the probe. This allows traces most similar to the probe to dominate the information retrieved. Note that the exponent in the activation function preserves the sign of the argument, S_i.

The information retrieved by a probe is a vector, C, called the echo. The echo is computed by weighting each of the $i = 1 \ldots m$ traces in memory by their activations and, then, summing all m traces into a single vector,

$$C_j = \sum_{i=1}^{m} A_i \times M_{ij}$$

The information in the echo is converted to a decision variable called echo intensity, I, by computing the cosine of the echo and probe.

A similar calculation can be used to compare subfields in the echo to potential responses. For example, one can retrieve an echo and, then, compute the cosine for a subset of the elements to the vector for each potential response alternative. This method is used to simulate cued recall. Jamieson and Mewhort (2009b) used it to select a response mapped to particular stimuli. We use it, here, to simulate performance in the string-completion task.

Simulation of Experiment 1

The basic implementation is the same as that in Jamieson and Mewhort (2009a). First, five vectors were constructed, one for each consonant that was used to construct the strings in Table 1 (i.e., one vector for each of *P*, *S*, *T*, *X*, and *V*). Each vector had dimensionality 20 and had values of $+1$ or -1 selected with equal probability to each element, $p(+1) = p(-1) = .5$. For each of the training and test strings, the appropriate vectors were concatenated to form a single vector. That is, each string was represented as a vector composed of six successive 20-element subfields, one subfield for each consonant in the string. For strings with fewer than six consonants, the empty subfields were filled with zeros. In test strings, the elements in the string that corresponded to the position of the missing letter were also set to zero.

We used vectors of dimensionality 20 for two reasons: first, to be consistent with our application of the model in previous work (Jamieson & Mewhort, 2009a, 2009b). Second, sufficient dimensionality is important to allow a gradation in encoding. To illustrate, a vector of 20 elements that has 1 element copied incorrectly to memory has 5% of its information deleted. A vector of 2 elements that has 1 element copied incorrectly to memory has 50% of its information missing.

To represent the situation after training, we stocked memory with a vector for each training string. Hence, following simulated study, memory was a 20×120 matrix. A test string was used as a retrieval probe, and an echo was retrieved.

To interpret the information in the echo, we computed the similarity of each response alternative—*P*, *S*, *T*, *X*, and *V*—to the relevant elements in the echo (i.e., the field corresponding to the position of the underscore). We selected for report the letter that had the highest similarity.

Finally, to obtain stable estimates for each of the choices, we conducted 100 replications of the procedure.

The model's choices are indicated by the entries shown in bold in Table 2. The values in Table 2 were computed for simulations with encoding quality, L, equal to 0.7. As shown, the model preferred the legal alternative for 16 of the 20 test strings, a success rate of 80% (chance = 20%). Like our participants, the model completed a majority of the test strings legally.

Next, we compared the model's decisions to those of our participants. Minerva 2's first choice matched the participants' first choice amongst the five alternatives for 12 of the 20 test strings (Items 1, 3, 4, 5, 6, 9, 10, 11, 14, 16, 17, and 20) for a success rate of 60% (chance = 20%). This success rate indicates that Minerva 2 and participants generally agree on what constitutes a legal completion.

The rightmost column in Table 2 shows the correlations between participants' choice frequencies and the model's mean echo intensities. These scores index the model's ability to anticipate the choice distribution for each item.[1]

As shown, the model did a good job of predicting the distribution of choices for 10 of the 20 test items (i.e., $r \geq .75$), a modest job of predicting the distribution of choices for 4 of the test items (i.e., $.5 \leq r \leq .74$), and a poor job of predicting the distribution of choices for the remaining 6 items (i.e., $r \leq .49$). The largest correlations were for strings that both participants and the model completed legally.

Correlations for Items 2, 5, 13, 15, 18, and 19 represent notable failures. However, the model's failure to explain participants' decisions for Items 2, 5, and 15 could not be more ironic: Whereas the model wanted strongly to complete the strings legally, participants preferred illegal completions.

Minerva 2's failure to match the distributions of participants' choices for Items 13, 18, and 19 underscores that the model is not a complete explanation of participants' behaviour. To illustrate, consider the item *TS_* (Item 13). Whereas the model wanted to complete the item with the letter *V*, participants preferred *X*. As decisive as participants were in choosing *X*, Minerva 2 insists on *V* because all of the training strings that started *TS* were followed by *V* in the third position.

The experiment demonstrates that participants can complete strings legally following brief exposure to grammatical training items. At first blush, the result suggested that participants had abstracted the grammar in training and then applied that knowledge at test. We have demonstrated, however, that string completion can be understood in terms of global similarity using Minerva 2's standard associative mechanism for cued recall. When a test item is presented it retrieves an aggregate of the items in memory. Information in the aggregate points to a legal completion.

As we noted earlier, a model's credibility depends on its ability to encompass new phenomena. The current example—the ability to complete strings legally—is a case in point. It illustrates yet another scenario—recall of grammatical sequences (Jamieson & Mewhort, 2005), JOG (Jamieson & Mewhort, 2009a), and choice reaction-time (Jamieson & Mewhort, 2009b)—in which an exemplar model handles a result that might otherwise have demanded a specialized learning mechanism. In the next section, we confront yet another challenge to the global-similarity position.

USING GRAMMATICAL KNOWLEDGE: INDIVIDUAL EXEMPLARS

Jamieson and Mewhort (2009a) showed that global similarity predicts overall performance in the JOG task averaged across both materials and participants. Dienes (1992) has argued, however, that a competent model should predict more than average performance. It should also predict performance

[1] We report correlations here as descriptive statistics. Due to the small number of observations across which they are computed, any inferential analysis is questionable.

on individual test items. In the context of a JOG task, it should predict which particular items people judge to be most and least grammatical.

Dienes's (1992) argument is, in practice, a corollary of the idea that a model's credibility reflects its ability to handle new phenomena. It is an important corollary because it forces an increasingly complete account of performance and, for that reason, helps to prevent a model from claiming spurious success—that is, from working for the wrong reason. Hence, we agree that predicting performance for individual items provides a stronger test than one based on an average across items. Accordingly, to take up Dienes's challenge, we assess how well global similarity can predict performance on individual items.

Dienes (1992)

Dienes (1992) used stimuli from Reber and Allen (1978) to compare several computational accounts of the JOG task. Reber and Allen's grammar is labelled (a) in Figure 1, and the complete set of stimuli is presented in Table 3.

Reber and Allen's (1978) grammar produced 41 strings, 20 of which served as grammatical training strings; 25 were selected as the grammatical test items (5 of the 25 grammatical test items served as both a training and test item).

Table 3. *Dienes's (1992) stimuli*

	Test strings	
Training strings	*Grammatical*	*Ungrammatical*
MTTTTV	VXV (1)	XVRVM/XVRXRR (16)[a]
MTTVT	**VXVRX (13)**	XRVXV (20)
MTV	**VXVRXV (9)**	XTTTTV (6)
MTVRX	VXVRXR (2)	VXX (23.5)
MTVRXM	**VXVT (17)**	VXMRXV (11)
MVRX	VXR (12)	VXRVM (10)
MVRXRR	VXRM (4)	VXRRT (2)
MVRXTV	VXRRR (14)	VXRT (3.5)
MVRXV	VXRRRM (7)	VVXRM (22)
MVRXVT	VXTV (11)	VRRRM (7)
VXM	VXTTV (16)	MXVRXM (5)
VXRR	VXTTTV (15)	MXVT (9)
VXRRM	**MVRXVT (18)**	MVRTR (15)
VXRRRR	MVRXM (8)	MMVRX (17)
VXTTVT	MVRXR (10)	MTVV (12)
VXTVRX	MVRXRM (3)	MTVRTR (13)
VXTVT	MVT (5)	MTM (23.5)
VXVRX	**MTV (24)**	MTRV (14)
VXVRXV	MTVRXV (20)	MTRVRX (8)
VXVT	MTVRXR (6)	MTTVTR (1)
	MTVT (20)	RVT (19)
	MTTV (22)	RRRXV (21)
	MTTVRX (20)	TXRRM (25)
	MTTTV (25)	TVTTXV (18)
	MTTTVT (23)	TTVT (3.5)

Note: These materials were used in Experiment 2. Grammaticality of the items is defined relative to Grammar a in Figure 1. Dienes's item ranks for the test items are presented in parentheses. Items shown in bold served as both training and test strings.

[a]The string *XVRXRR* appears in Dulany, Carlson, and Dewey (1984) and in Dienes's (1992) list of materials but not in Dienes et al.'s (1991) list of materials. The string *XVRVM* replaces *XVRXRR* in Dienes et al.'s list of materials. In the experiments reported in this paper, we used *XVRVM*.

A total of 25 ungrammatical test strings were created by introducing violations into grammatical strings.

Dienes (1992) aggregated data from several experiments that used Reber and Allen's (1978) materials. Based on the aggregate, he ranked the test strings from the most often endorsed as grammatical (rank = 1) to the least often endorsed as grammatical (rank = 25). Dienes's ranks are presented in parentheses in Table 3. Note that the ranks are computed separately for the grammatical and ungrammatical test items.[2]

Dienes (1992) studied Minerva 2's ability to predict the relative performance of individual items in the JOG task and found that its predictions were poor for both grammatical and ungrammatical items. He concluded that Minerva 2 (and related models, see Estes, 1986; Medin & Schaffer, 1978) could not explain participants' judgements in the JOG task.

Minerva 2's failure to handle performance on individual items stands in sharp contrast to its success in handling data averaged over items. In particular, it stands in marked contrast to our demonstration that global similarity provides a clear account of overall performance in the JOG task (Jamieson & Mewhort, 2009a). How can the same model work well on averaged data but poorly on individual items?

A potential answer concerns the way the model was implemented. Dienes (1992) used very different representation assumptions than Jamieson and Mewhort (2009a). Perhaps the paradoxical difference in results can be traced to the differences in representation. To consider that possibility, we provided his stimuli to our implementation of the model to see if our model could fit his rank-order data.

Simulation of Dienes (1992)

The procedure for the simulation was similar to that for the simulation of Experiment 1. For each training and test string, the appropriate vectors were concatenated to form a single vector. That is, each string was treated as a vector composed of six successive 20-element subfields, one subfield for each consonant in the string. For strings with fewer than six consonants, the empty subfields were filled with zeros.

To represent memory after training, we stocked the model's memory with a vector for each training string. Hence, memory after training was a 20×120 matrix. To simulate a judgement of grammaticality, a test string was used as a retrieval probe, the echo was retrieved, and the echo intensity for the entire string was computed by comparing the echo to the probe. We computed the mean echo intensity for each test string.

To obtain stable point estimates, we averaged the echo intensities for each probe across 100 independent simulations of the experiment and used the averaged scores to determine the rank order of the test items.

We used a Spearman rank correlation to quantify the match between the predicted ranks and the ranks reported by Dienes (1992). The correlations are presented in Table 4 as a function of encoding quality (controlled by the parameter L). We used the rank correlation firstly because of Dienes's precedent and secondly because rank data are ordinal.

As shown in Table 4, our results were very different from Dienes's results (1992). In his analysis, the correlations between the predicted and the empirical ranks were −.08 and .27 for grammatical and ungrammatical strings, respectively (see Dienes's Table 10). Dienes's implementation of Minerva assumed perfect encoding; hence, our condition most comparable to his comes from the simulations with $L = 1.0$. With $L = 1.0$, our correlations were −.15 and .61, for grammatical and ungrammatical strings, respectively. In short, our simulation produced a substantially better fit to the ungrammatical items (.61 vs. .27) and a marginally poorer fit to the grammatical items (−.15 versus −.08).

The poor fit for the grammatical items is not hard to understand. With perfect encoding (i.e.,

[2] We thank Professor Dienes for providing these ranks to us.

Table 4. *Rank order correlations between Minerva 2's predictions and the results from Dienes (1992)*

	L			
	.10	*.40*	*.70*	*1.0*
Grammatical	.32	.19	− .03	− .15
Ungrammatical	.39	.54	.63	.61

Note: Within-category correlations equal to or greater than .40 are reliable at the .05 level (two-tailed). Between-category correlations equal to or greater than .28 are reliable at the .05 level (two-tailed).

$L = 1$), our implementation of Minerva recognized and endorsed as grammatical the five test items presented both at study and at test: *MTV*, *VXVRX*, *VXVRXV*, *MVRXVT*, and *VXVT*. By contrast, Dienes's (1992) participants did not: They ranked *MTV*, a studied grammatical test item, 24 out of 25. Importantly, if his participants had recognized the studied items, they surely should have judged those items to be grammatical. After all, they were told that the training items were grammatical by definition.

To mimic the failure to recognize the studied items, we relaxed the learning parameter. As *L* was reduced, the model did a better job at fitting the empirical ranks for grammatical items.

What conclusions follow? On the one hand, one might conclude that the model works for ungrammatical items but does not work for grammatical ones. The paradox remains that the model does well at handling performance averaged over exemplars (see Jamieson & Mewhort, 2009a, especially the simulation of Dienes et al., 1991). On the other hand, it is clear that encoding quality is an important issue. Inasmuch as Dienes's (1992) participants did not recognize the studied items, a simulation that assumes perfect encoding ensures a mismatch between observed and predicted rankings.

The encoding-quality issue underscores a critical difference between the two implementations of Minerva 2. In Dienes's (1992) implementation, each consonant was represented by set of five input units, one unit for each letter, as is standard for a neural-network model. Each letter was coded by setting one unit within the set to a value of $+1$ and by setting the other units to values of -1. Using this coding system, the letters *M* and *V* might be represented as {+1 −1 −1 −1 −1} and {1 +1 −1 −1 −1}, respectively. A string of letters was coded by setting the identity of units that were hard-wired to a serial position within a string. That is, the identity of each letter was coded within a set of five units, and the letter's serial position within the string was coded in six successive clusters, one for each position in a string. In terms of the *M* versus *V* example, if the first set of five units indicates the letter in Position 1 of a string, and the second set of five units indicates the letter in Position 2, *MV* would be: Position 1 = {+1 −1 −1 −1 −1} and Position 2 = {−1 +1 −1 −1 −1}. Under other assumptions (called the co-occurrence representation), consonants were represented by a set of five units in which one unit was given a value $+1$ and the other units a value of zero. Using this system, the letters *M* and *V* might be coded { +1 0 0 0 0} and {0 +1 0 0 0}, respectively, with the string *MV* coded as: Position 1 = {+1 0 0 0 0} and Position 2 = {0 +1 0 0 0}.

Although Dienes's (1992) representation assumptions are used routinely in neural-network simulations, they make it difficult to represent noisy encoding of individual symbols within a string, especially within a multitrace model such as Minerva 2. Suppose, for example, that a participant could not remember what letter was in a particular position in a string. To represent the situation, one must be able to degrade the identity of the letter without introducing complications about position. In our implementation, the distortion was easy to arrange; each $+1/-1$ element in a trace was replaced by a value of 0 with probability $(1 - L)$. Replacing elements with 0s obscures information about each letter independently.

Using Dienes's (1992) representation assumptions, however, the same technique does not work. Consider the { +1 −1 −1 −1 −1} vector coding for *M*. Replacing the $+1$ value by a 0—that is, {0 −1 −1 −1 −1}—may appear to degrade the identity of the letter, but, because

the −1 values indicate what the letter is not, the vector still signals *M*. Replacing one of the other −1 values with the +1 value introduces competition between letters, but does not degrade the representation of either one. For example, {+1 +1 −1 −1 −1} indicates that the letter is either *M* or *V* (because it is not one of the other possibilities). With the co-occurrence representation, setting all elements to zero is equivalent to encoding nothing, $L = 0.0$.

Because participants did not recognize the studied test strings and because encoding quality will almost surely influence judgements of grammaticality, a simulation of the JOG task must be able to accommodate degraded memory of materials.

In addition to the representation issue, the use of aggregate data raises a second potential problem. Although Dienes (1992) was careful to confirm that the item ranks were consistent across the experiments, the training procedures in the experiments differed. In some cases, participants were told to search for rules during the training phase. In others, the participants performed a concurrent task during the training phase. Finally, participants were sometimes shown strings one at a time and were sometimes shown the strings in batch form. Even though the manipulations had little influence on performance as measured by ranks averaged over items and participants, aggregating data across training procedures makes it difficult to simulate the situation. If participants study strings under different instructions, how should the difference be represented in memory? What difference in representation follows from instructions to search for rules versus instructions to memorize exemplars? To build a convincing computational account, one needs to decide not only what extra factors to include but also how each factor should be weighted in the simulation.

To remedy the potential procedural questions, we collected fresh data with Dienes's (1992; see Reber & Allen, 1978) materials. The idea was to assess whether our model's failure reflected a problem with the data rather than a problem with the model itself.

EXPERIMENT 2: DIENES'S MATERIALS

The experiment was a standard JOG task. In the training phase, participants were asked to remember 20 training strings: The strings were presented successively for 6 s each. Following training, the participants were told the training strings were constructed according to the rules of an artificial grammar and that their task was to judge the grammaticality of test strings. On each test trial, a string was presented, and the participant rated its grammaticality on a scale from −100 to +100 using a slider. After the test strings had been rated, the participants were asked to articulate the rules of the grammar.

Method

Participants

A total of 39 students from the University of Manitoba undergraduate participant pool took part in the study. All participants reported normal or corrected-to-normal vision.

Apparatus

The apparatus was the same as that in Experiment 1.

Stimuli

We used the same stimuli as Dienes (1992; see his Table 3).

Procedure

Participants were tested in groups of 4 to 7. Each participant was seated in front of a different computer. The participants were told that they would be shown strings of letters and that they should try to remember each string.

The participant initiated the training by clicking on the message "Start" (displayed at the centre of the computer screen). After the participant clicked on the message, the screen was cleared for 750 ms. Immediately thereafter, the first training string was displayed at the centre of the screen for 6 s. Next, the screen was cleared, and after 750 ms the next string was displayed.

The cycle repeated until all of the training strings had been presented.

After the training phase, the participants were told that the strings had been constructed using rules and that their task was to rate the grammaticality of test strings. On each test trial, one of the test strings from Table 3 was presented at the centre of the screen. A line approximately 5 cm in length was displayed approximately 3 cm below the string: The line contained a slider positioned at its centre. The phrases "Rule Violating", "Unsure", and "Rule Conforming" were displayed at the left, centre, and right of the line, respectively. A button labelled "OK" was centred approximately 2 cm below the line.

To start the test, the slider was positioned at the centre (neutral) point on the line. To rate the grammaticality of the test string, the participant used the computer mouse to position the slider on the line and then clicked on the word "OK". Immediately thereafter, the screen was cleared and, 1 s later, the next string was displayed. If the participant did not move the slider from the neutral position before clicking on the word "OK", a message instructed the participant that he or she must move the slider from the neutral position in order to complete the trial (i.e., the computer would not let participants provide a rating of zero). The cycle continued until all of the test strings had been presented.

After the test strings had been presented, a text editor appeared on the computer's screen. A message above the text editor invited the participant to describe the rules that were used to construct the training strings.

Results

Participants' grammaticality ratings were converted to discrete classifications: Ratings greater than zero were classified as *grammatical*; ratings less than zero were classified as *ungrammatical*. Using the classified data, participants judged 60% ($SE = 2\%$) of the grammatical test strings and 45% ($SE = 2\%$) of the ungrammatical test strings to be grammatical. By a signal-detection analysis, discrimination was modest, $d' = 0.41$ ($SE = 0.07$), but reliably better than chance, $t(38) = 5.67$, $p < .05$. Although discrimination was weak, the results are consistent with data in other JOG tasks.

None of the participants could articulate the rules of the grammar; indeed, most declined to guess at the rules. In summary, the experiment reproduced the classic results of a JOG task: Participants discriminated test strings at about 60% correct but could not describe the grammar. We now turn to an analysis of performance for individual test strings.

Table 5 presents the mean grammaticality rating for each of the 50 test strings. Strings 1 through 25 are grammatical; Strings 26 through 50 are ungrammatical. We used the mean grammaticality ratings to rank the strings from "judged most grammatical" (rank = 1) to "judged least grammatical". Rank acceptability of the strings is shown for both within-category (i.e., the 25 grammatical test strings and the 25 ungrammatical test strings ranked separately) and between-category (i.e., all 50 strings ranked independent from grammatical status) comparisons.

Our results agreed, at least in part, with Dienes's (1992) results. The rank correlation for ungrammatical strings in our experiment and in Dienes's aggregated results was reasonable, $r(23) = .69$, $p < .05$. The correlation for grammatical strings, however, was negative and marginally significant at the $\alpha = .05$ level, $r(23) = -.36$, $p \approx .05$.

Our participants' ranks agreed with Dienes's (1992) ranks for the ungrammatical but not for the grammatical items. Inspection of the ratings in Table 5 makes clear the reason for the discrepancy for the grammatical strings. Almost all of our participants (95%) endorsed the string *MTV* as grammatical. Whereas our participants' judgements were lured by a familiar trigram (i.e., *MTV*), Dienes's participants were not. In our results, *MTV* had a larger mean rating than any of the other 49 test strings: The lower bound of the 95% confidence interval for the rating of *MTV* was 67.99, and the next highest mean ratings were 42.90 and 41.33 for strings *MTVRXV* and *MTVRXR* (note that both *MTVRXV* and *MTVRXR* begin with *MTV*). Test strings that included *MTV* (i.e., Strings 18, 19, 20, 21, 40,

Table 5. *Mean grammaticality ratings as well as within- and between-category ranks for individual items presented in the JOG task in Experiment 2*

		Grammaticality ratings		*Rank order acceptability*	
Item number	*Item*	*M*	*SE*	*Within-category*	*Between-category*
1	VXV	− 37.33	10.40	25	48
2	**VXVRX**	3.36	9.90	19	29
3	**VXVRXV**	17.79	10.59	12	16
4	VXVRXR	8.92	10.42	17	26
5	**VXVT**	1.56	11.23	20	30
6	VXR	20.85	11.35	9	11
7	VXRM	22.05	10.33	8	10
8	VXRRR	31.31	9.66	4.5	4
9	VXRRRM	10.95	11.27	15	21
10	VXTV	0.41	9.66	21	31
11	VXTTV	− 5.21	10.86	23	36
12	VXTTTV	9.23	10.68	16	25
13	**MVRXVT**	11.54	10.99	14	19
14	MVRXM	31.31	9.02	4.5	4
15	MVRXR	18.51	8.97	11	14
16	MVRXRM	− 5.03	9.96	22	34
17	MVT	− 19.62	10.85	24	41
18	**MTV**	81.26	6.77	1	1
19	MTVRXV	42.90	9.32	2	2
20	MTVRXR	41.33	9.79	3	3
21	MTVT	26.67	10.14	7	8
22	MTTV	3.46	12.40	18	28
23	MTTVRX	20.38	10.60	10	12
24	MTTTV	29.85	12.27	6	6
25	MTTTVT	17.18	11.22	13	17
26	XVRVM	− 26.79	9.57	22	46
27	XRVXV	− 9.38	10.28	15	38
28	XTTTTV	4.49	11.41	10	27
29	VXX	− 39.05	10.46	24	49
30	VXMRXV	25.64	10.65	2	9
31	VXRVM	27.03	8.66	1	7
32	VXRRT	9.69	10.10	8	23
33	VXRT	13.28	10.13	5	18
34	VVXRM	− 7.87	9.67	14	37
35	VRRRM	− 5.03	11.32	13	34
36	MXVRXM	18.10	9.76	4	15
37	MXVT	− 4.03	10.85	12	33
38	MVRTR	10.03	9.61	7	22
39	MMVRX	− 25.90	8.75	20	44
40	MTVV	− 23.51	9.23	19	43
41	MTVRTR	11.44	10.60	6	20
42	MTM	− 62.77	8.84	25	50
43	MTRV	− 1.97	10.15	11	32
44	MTRVRX	9.41	9.56	9	24
45	MTTVTR	18.64	9.70	3	13
46	RVT	− 13.82	11.01	17	40
47	RRRXV	− 35.77	9.23	23	47
48	TXRRM	− 23.46	9.30	18	42
49	TVTTXV	− 26.23	9.65	21	45
50	TTVT	− 9.92	11.35	16	39

Note: Items 1 through 25 are grammatical, and Items 26 through 50 are ungrammatical. Items shown in bold served as both training and test strings. JOG = judgement of grammaticality.

and 41) were rated as more grammatical ($M = 30.02$) than strings that included a recursive version of *MTV* (Strings 22, 23, 24, 25, and 45; $M = 17.90$) or that did not contain *MTV* ($M = -1.38$). There was one notable exception to the dominance of the trigram *MTV* on decision: Participants rated the ungrammatical string *MTVV* to be ungrammatical. Although we can only speculate, we suspect that participants treated *MTVV* as a "close miss" to *MTV* and rejected it on those grounds. This interpretation would be consistent with the fact that participants rated the string *MVT* to be the second most ungrammatical grammatical item. Finally, when giving their rule statements, 32 of our participants told us that *MTV* was in the training strings. A total of 12 of the 32 asserted that *MTV* was a rule of the grammar.

The correlation between the ranks predicted by Minerva 2 and our empirical ranks are presented in Table 6 as a function of encoding quality (L). For $L = 1.0$, the correlations for grammatical and ungrammatical strings were .07 and .57, respectively. Relaxing encoding quality produced an unimpressive improvement for the fit to the grammatical items, achieving $r = .09$ at best.

Minerva 2's failure for the grammatical strings is consistent with Dienes's (1992) conclusions. Now, however, we can track why it failed: The failure reflected our participants' focus on a single familiar trigram, *MTV*—a bias not shared by Minerva 2. The same bias may also explain why discrimination was weaker in our experiment than in other JOG tasks.

Table 6. *Rank order correlations between Minerva 2's predictions and the results of Experiment 2*

	L			
	.10	*.40*	*.70*	*1.0*
Grammatical	− .07	.04	.09	.07
Ungrammatical	.51	.57	.62	.57
All strings	.43	.52	.55	.52

Note: Within-category correlations equal to or greater than .40 are reliable at the .05 level (two-tailed). Between-category correlations equal to or greater than .28 are reliable at the .05 level (two-tailed).

Whereas Experiment 2 solved a potential problem associated with aggregating data collected with different training procedures, the experiment uncovered a new problem: Our participants were hyperfocused on a single familiar trigram, *MTV*. Minerva 2 has no mechanism with which to appreciate that *MTV* is the acronym of a popular television station. Virtually all other current computational theories would fail for the same reason.

To solve for the *MTV* problem, we conducted another experiment using the same training procedure as that of Experiment 2 and the same grammar. However, the new experiment changed the surface structure of the grammar so that *MTV* could not occur. Specifically, we remapped the consonants from Experiment 2 so that the letters *M*, *T*, *V*, *R*, and *X* in grammar a were replaced by letters *T*, *X*, *S*, *V*, and *P* in Grammar b. The remapped grammar is shown in Figure 1. Note, Grammar b is identical to Grammar a except for the letters assigned to the paths. After the remapping, the trigram *MTV* in Experiment 2 became *TXS* in Experiment 3.

EXPERIMENT 3: REMAPPED STRINGS

Experiment 3 was designed to remove the MTV effect that arose in Experiment 2. Participants studied 20 grammatical training strings and then rated the grammaticality of test strings. The materials were generated using the grammar from Experiment 2; however, the surface structure of the materials differed. Finally, participants were asked to describe the rules of the grammar.

Method

Participants

A total of 47 students from the University of Manitoba undergraduate participant pool took part in the study. All participants reported normal or corrected-to-normal vision.

Materials

The stimuli are shown in Table 7. The materials were constructed by remapping the strings from Experiment 2 using the following rules: $M \rightarrow T$, $T \rightarrow X$, $V \rightarrow S$, $R \rightarrow V$, and $X \rightarrow P$. Thus, *MTV* was rewritten as *TXS*. Remapping letters changed the surface structure of the materials, but left the grammatical structure intact (i.e., Grammars a and b in Figure 1 have the same structure; only the assignment of letters to the transitions is different).

Procedure

The procedures for the training phase and for the JOG task were the same as those in Experiment 2: Only the materials differed.

Table 7. *Training and test strings used in Experiment 3 with items constructed with Grammar b in Figure 1*

	Test strings	
Training strings	*Grammatical*	*Ungrammatical*
SPSVP	SPS	PSVST
SPSVPS	**SPSVP**	PVSPS
SPSX	**SPSVPS**	PXXXXS
SPT	SPSVPV	SPP
SPVV	**SPSX**	SPTVPS
SPVVT	SPV	SPVST
SPVVVV	SPVT	SPVVX
SPXSVP	SPVVV	SPVX
SPXSX	SPVVVT	SSPVT
SPXXSX	SPXS	SVVVT
TSVP	SPXXS	TPSVPT
TSVPS	SPXXXS	TPSX
TSVPSX	**TSVPSX**	TSVXV
TSVPVV	TSVPT	TTSVP
TSVPXS	TSVPV	TXSS
TXS	TSVPVT	TXSVXV
TXSVP	TSX	TXT
TXSVPT	**TXS**	TXVS
TXXSX	TXSVPS	TXVSVP
TXXXXS	TXSVPV	TXXSXV
	TXSX	VSX
	TXXS	VVVPS
	TXXSVP	XPVVT
	TXXXS	XSXXPS
	TXXXSX	XXSX

Note: Items shown in bold served as both training and test strings.

Results

As in Experiment 2, we converted participants' grammaticality ratings to discrete classifications: Ratings greater than zero were classified as *grammatical*, and ratings less than zero were classified as *ungrammatical*. According to the classified data, participants judged 63% ($SE = 2\%$) of the grammatical test strings and 45% ($SE = 2\%$) of the ungrammatical test strings to be grammatical. By a signal-detection analysis, discrimination was slightly better than that in Experiment 2, $d' = 0.47$ ($SE = 0.06$), and was reliably better than chance, $t(46) = 8.53$, $p < .05$. None of our participants articulated the rules of the grammar.

Again, the experiment reproduced the standard result: Participants discriminated test strings at a rate of about 60% correct but could not articulate the rules of the grammar. We now turn to performance on each of the 50 test strings.

The data for each of the 50 test strings are presented in Table 8. The rank acceptability of the test strings is shown for both within-category (i.e., for the 25 grammatical test strings and 25 ungrammatical test strings separately) and between-category (i.e., for all 50 strings independent of grammatical status) comparisons.

Although participants ranked *TXS* (the string remapped from *MTV*) highly, they were not hyperfocused on it. The six test strings that included the trigram *TXS* (Strings 18, 19, 20, 21, 41, and 42) were rated to be only slightly more grammatical ($M = 15.98$), than the test strings that included a recursive version of *TXS* (Strings 22, 23, 24, 25, and 45; $M = 5.13$) or the strings that did not include the trigram ($M = 5.06$). None of the participants mentioned *TXS* in their rule statements, nor did they systematically report any other grouping of letters. The mean of the ratings for the five studied test strings ($M = 36.82$, $SE = 4.44$) was much higher than the mean rating for the other 20 grammatical test strings ($M = 13.81$, $SE = 2.71$).

Reber and Allen (1978) reported that participants are especially sensitive to grammatical violations that occur at the beginning and ends of strings and are less sensitive to violations that

Table 8. *Mean grammaticality ratings as well as within- and between-category ranks for individual items presented in the JOG task in Experiment 3*

		Grammaticality ratings		*Rank order acceptability*	
Item number	*Item*	*M*	*SE*	*Within-category*	*Between-category*
1	SPS	− 3.09	9.62	24	35
2	**SPSVP**	35.30	8.22	3	3
3	**SPSVPS**	44.28	8.41	2	2
4	SPSVPV	16.49	9.67	14	17
5	**SPSX**	27.36	8.89	10	11
6	SPV	28.43	9.06	6	7
7	SPVT	21.64	9.19	12	15
8	SPVVV	0.55	10.28	23	32
9	SPVVVT	3.57	10.58	19	27
10	SPXS	28.06	8.34	8	9
11	SPXXS	6.87	9.52	17	23
12	SPXXXS	15.51	10.43	15	19
13	**TSVPSX**	27.62	9.22	9	10
14	TSVPT	29.11	8.47	5	6
15	TSVPV	28.32	8.96	7	8
16	TSVPVT	20.47	9.08	13	16
17	TSX	− 4.74	10.60	25	37
18	**TXS**	49.55	7.76	1	1
19	TXSVPS	30.02	9.24	4	5
20	TXSVPV	15.47	9.38	16	20
21	TXSX	6.81	10.04	18	24
22	TXXS	2.94	10.12	20	28
23	TXXSVP	25.89	9.82	11	13
24	TXXXS	2.66	11.01	21	29
25	TXXXSX	1.28	10.62	22	31
26	PSVST	− 6.17	9.22	13	38
27	PVSPS	− 1.43	9.25	10	33
28	PXXXXS	− 13.38	11.80	17	42
29	SPP	− 18.09	11.13	18	43
30	SPTVPS	8.70	9.87	6	22
31	SPVST	24.94	8.47	3	14
32	SPVVX	3.79	9.11	8	26
33	SPVX	27.11	8.21	2	12
34	SSPVT	− 22.89	8.17	20	45
35	SVVVT	− 4.11	10.06	12	36
36	TPSVPT	5.04	9.68	7	25
37	TPSX	35.02	7.86	1	4
38	TSVXV	− 2.98	10.24	11	34
39	TTSVP	− 19.60	10.18	19	44
40	TXSS	1.64	9.72	9	30
41	TXSVXV	− 7.64	9.60	15	40
42	TXT	− 24.32	10.42	21	46
43	TXVS	11.15	9.06	5	21
44	TXVSVP	16.28	9.22	4	18
45	TXXSXV	− 7.11	9.80	14	39
46	VSX	− 29.13	9.57	22	47
47	VVVPS	− 43.21	8.49	25	50
48	XPVVT	− 31.40	8.04	23	48
49	XSXXPS	− 10.87	10.09	16	41
50	XXSX	− 32.85	9.06	24	49

Note: Items 1 through 25 are grammatical, and Items 26 through 50 are ungrammatical (see Grammar b in Figure 1). Items shown in bold served as both training and test strings. JOG = judgement of grammaticality.

occur in the middle of strings. It is clear from Table 8 that participants were especially sensitive to rule violations at the first serial position in ungrammatical test strings. All of the grammatical training strings started with one of two letters, *T* or *S*, and participants rated all eight ungrammatical strings that began with a letter other than *T* or *S* negatively (Items 26, 27, 28, 46, 47, 48, 49, and 50). Unfortunately, we cannot evaluate other positional dependencies with our materials. To illustrate the problem, it is not clear what rules are violated by Item 45 in Table 8: *TXXSXV*. Taking the path from Node 1 to Node 2 to Node 2 to Node 3 in Grammar b in Figure 1, one might consider the string to be grammatical up to *TXXS* with the pursuing *X* making the string ungrammatical in order to arrive at Node 5 to complete the string with the letter *V*. By an alternative conception, one might take the path from Node 1 to Node 2 to Node 2 to Node 3 to Node 6 and treat the string as grammatical up to *TXXSX* with the final *V* making the string ungrammatical. In the former case, the first rule violation occurs internal to the string; in the latter, the first rule violation occurs at the end of the string. This ambiguity underscores the difficulties associated with any analysis that assesses the point at which a string becomes ungrammatical, at least when using a complicated grammar such as the one in Figure 1. Because of the ambiguity, analysis of performance depending on where in a string a rule is violated is speculative.

In previous work we used simpler grammars that imposed sequential but not positional constraints on strings. With these grammars, we could count the number of first-order sequential rule violations in a string. We demonstrated in an experiment that participants' likelihood of rejecting a string increases the greater the number of rule violations in it. We demonstrated, by simulation, that our adaptation of the Minerva 2 model accommodates the trend (see Jamieson & Mewhort, 2009a).

The string *MTV* was ranked highly in Experiment 1 because it was a familiar acronym. The string *TXS* was ranked highly in the current experiment because it was a studied string. By all appearances, remapping the letters solved the *MTV* problem observed in Experiment 1.

Simulation of Experiment 3

We computed rank correlations to quantify the fit between item ranks as predicted by Minerva 2 and item ranks from Experiment 3. The correlations are presented in Table 9 as a function of *L*. Figure 2 summarizes the relation between the model's prediction (based on mean echo intensity) and our participants' rating of grammaticality ($L = 0.7$).

In contrast to Experiment 2, our implementation of Minerva provides a good fit to the ranks for both the grammatical and ungrammatical strings. Indeed, as shown in Table 9, for larger values of *L*, the fit to the grammatical strings was better than the fit to the ungrammatical strings. In short, the model works well provided (a) that target data are collected with a single training procedure, (b) that the procedure described by the simulation matches the procedure in the experiment, and (c) that the strings do not include a salient idiosyncrasy that lures participants' decisions (e.g., MTV). The earlier failures to predict grammaticality ratings for individual strings in the JOG task did not reflect a problem with the model; they reflected a problem with the empirical estimates of grammaticality.

The important differences between Experiments 2 and 3 underscore Dienes's (1992) insistence that a competent model must predict more than average performance. In both experiments, overall performance—that is, performance averaged over exemplars

Table 9. *Rank order correlations between Minerva 2's predictions and the results of Experiment 3—remapped items*

	L			
	.10	*.40*	*.70*	*1.0*
Grammatical	.46	.53	.59	.56
Ungrammatical	.58	.54	.53	.46
All strings	.68	.70	.71	.67

Note: Within-category correlations equal to or greater than .40 are reliable at the .05 level (two-tailed). Between-category correlations equal to or greater than .28 are reliable at the .05 level (two-tailed).

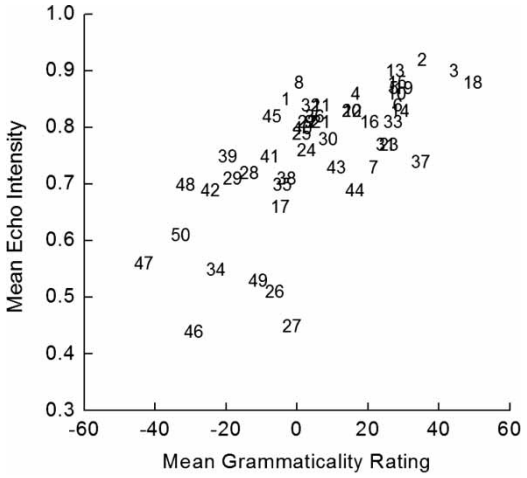

Figure 2. *The mean grammaticality rating (observed) plotted against mean echo intensity (simulated) for each of the 50 test strings in Experiment 3, L = .7. Numbers in the figure correspond to the numbered items in Table 8.*

and participants—was similar to performance in standard tasks from the literature. Indeed, discrimination of grammatical from ungrammatical probes was about the same, and there was no evidence of an overall difference in performance between Experiments 2 ($d' = 0.41$) and 3 ($d' = 0.47$), $t(84) = 0.72$, $p > .05$. But, as illustrated by differences in ranking individual items, processing underlying discrimination in the two tasks was very different. It follows that overall discrimination is a very insensitive criterion by which to judge a model's success. As Dienes argued, a competent model must do more. By the same token, the fact that people can discriminate grammatical from ungrammatical probes is very weak evidence concerning how they do so. Certainly it ignores much of the information available in the data.

In summary, the global-similarity principle successfully anticipated both the rank acceptability of test strings in the JOG task and performance in the string-completion task from Experiment 1. We suggest that participants infer a test string's grammaticality by its global similarity to the training items and that they use memory of the training items to select letters in the string-completion task. Of course, the model's success does not rule out the possibility that participants learned the grammar, but it does show that global similarity is an important contributor.

Experiment 1 substantiated our claim that participants' decisions in a string-completion task can be understood using standard principles of exemplar storage and retrieval. Experiments 2 and 3, however, taught us that an item analysis is susceptible to idiosyncrasies in materials (i.e., the MTV effect). To ensure that our data from Experiment 1 do not reflect an idiosyncrasy in our materials, and to ensure that our fit to the results of Experiment 1 reflects a general prediction from our model, rather than a fortunate fit for a specific set of materials, we replicated the string-completion task from Experiment 1 using different materials that were constructed according to the grammar in Figure 3.

EXPERIMENT 4: STRING COMPLETION

Experiment 4 was a replication of the string-completion task from Experiment 1; only the materials differed.

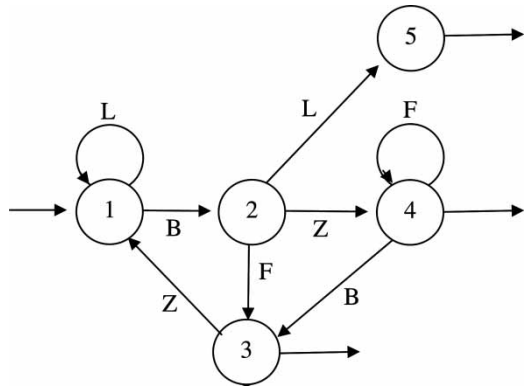

Figure 3. *The grammar is redrawn from "Intact Artificial Grammar Learning in Amnesia: Dissociation of Classification Learning and Explicit Memory for Specific Instances", by B. J. Knowlton, S. J. Ramus, and L. R. Squire, 1992, Psychological Science, 3, p. 175. Copyright 1992 by Wiley-Blackwell. Adapted with permission. A grammatical stimulus is generated by starting at the leftmost node, marked 1, and following the paths (indicated by arrows) until reaching one of the "exit paths" from Nodes 3, 4, or 5.*

Method

Participants

A total of 107 students from the University of Manitoba undergraduate participant pool took part in the study. All participants reported normal or corrected-to-normal vision.

Materials

Materials were generated using the grammar in Figure 3 borrowed from Knowlton et al. (1992). We generated 43 grammatical strings. Of the 43 grammatical strings, 19 were assigned at random to serve as training items; the remaining 24 items served as test items. The materials are presented in Table 10. For each of the 24 test strings, one letter was replaced by an underscore. The letters replaced were selected so that each of the four response alternatives, *B*, *F*, *L*, and *Z*, was replaced by an underscore an equal number of times across the set of test strings (i.e., 6 times each) and so that each string could be completed legally in one way only. By making each of the four letters the correct choice an equal number of times across the test set, we controlled for a potential bias to select a particular letter—a factor we had not controlled for in Experiment 1. No other efforts were made to balance properties in the items.

Table 10. *Materials from Experiment 4*

Training strings	*Test strings*
LBF	_BL
LBL	B_B
BZF	_LBL
BZFB	L_BZ
LBZB	BZ_F
LLBF	BFZ_L
LBZF	_FZBZ
LLLBZ	LLL_L
LLBZF	LB_FB
BFZBF	BZF_B
LLLBF	LB_FF
BZFFF	BFZ_BZ
LLBZB	LLLB_F
BZBZBZ	BZ_FFF
BZBZBL	LBZ_FF
LLLLBZ	LLBZ_F
BFZBZF	_LLLBL
LBFZBF	L_FZBZ
LBZFFB	LL_LBF
	LBF_BL
	BFZL_L
	BF_LBF
	BFZ_ZB
	BZFF_B

Note: Grammaticality is defined according to the grammar in Figure 2.

Procedure

The procedure was identical to that of Experiment 1; only the materials differed.

Results

Table 11 shows the choice frequencies (i.e., the number of participants that chose each alternative) for each of the test items. Participants completed test strings legally on 39% of the trials (chance = 25%), significantly better than chance by a binomial test, $p < .05$. The response alternatives chosen by the majority of participants completed 15 of the 24 test strings legally (Items 1, 2, 3, 4, 6, 7, 8, 9, 11, 13, 17, 18, 19, 20, and 22) for a success rate of 63% (chance = 25%). As in Experiment 1, our participants preferred to complete the strings by selecting the grammatical completion. Despite their ability to complete strings legally, none of the participants could describe the grammar. Indeed, most declined even to attempt a description. The result replicates the standard dissociation of performance and awareness in studies with artificial grammars.

Simulation of Experiment 4

We applied the model to the materials from Experiment 4 to the model. The procedure was identical to the simulation of Experiment 1; only the materials differed. The alternatives predicted by the model ($L = 0.7$) are shown in bold in Table 11. As shown, the model selected the legal alternative for 20 of the 24 test strings, a success rate of 83% (chance = 25%). Its first choice matched participants' first choice for 16

Table 11. *Frequency with which choice alternatives were selected to complete each of the items presented in the string-completion task in Experiment 4*

			Response alternatives				
Item number	*String*	*Grammatical completion*	*B*	*F*	*L*	*Z*	$r_{obs,pred}$
1	_BL	L	5	32	**53**	17	.84
2	B_B	Z	4	22	37	**44**	.83
3	_LBL	L	4	19	**38**	16	.76
4	L_BZ	L	6	34	**49**	18	.64
5	BZ_F	F	40	**15**	49	3	.37
6	BFZ_L	B	**46**	23	33	5	.71
7	_FZBZ	B	**53**	9	29	16	.97
8	LLL_L	B	**75**	10	9	13	.99
9	LB_FB	Z	8	7	27	**65**	.86
10	BZF_B	F	8	6	54	**39**	− .42
11	LB_FF	Z	6	20	40	**41**	.58
12	BFZ_BZ	L	**9**	55	37	6	− .91
13	LLLB_F	Z	5	20	12	**70**	.58
14	BZ_FFF	F	40	**15**	44	8	− .63
15	LBZ_FF	F	48	**25**	28	6	− .15
16	LLBZ_F	F	**66**	18	18	5	.84
17	_LLLBL	L	27	11	**57**	12	.97
18	L_FZBZ	B	**60**	3	26	18	.85
19	LL_LBF	L	14	10	**71**	12	.85
20	LBF_BL	Z	9	23	24	**51**	.95
21	BFZL_L	B	33	**14**	50	10	− .99
22	BF_LBF	Z	29	12	17	**49**	.96
23	BFZ_ZB	B	**34**	43	26	4	.34
24	BZFF_B	F	10	**8**	43	46	− .77

Note: The item selected by Minerva 2 is shown in bold in the frequency data. $r_{obs,pred}$ is the correlation between participants' choice frequencies (obs) and the model's mean echo intensities (pred) across the four response alternatives.

of the 24 test strings, a success rate of 67% (chance = 25%). The model's success rates indicate that it and participants agree in most cases on what constitutes a legal completion, even in cases when the choice common to both the model and the participants was illegal according to the grammar.

The rightmost column in Table 11 shows correlations between participants' choice frequencies and the model's mean echo intensities for each of the test strings. As shown, the model did a good job of predicting the distribution of choices for 12 of the 24 test items (i.e., $r \geq .75$), a modest job of predicting the distribution of choices for 4 of the 24 test items (i.e., $.5 \leq r \leq .74$) and a poor job of predicting the distribution of choices for the remaining 8 items (i.e., $r \leq .49$).

Minerva 2 failed to match participants' decisions for Items 5, 14, 15, and 24 because it preferred legal completions whereas participants preferred illegal completions. In some sense, then, one might argue that the model outperformed participants on these items.

Items 1, 3, 7, and 17 illustrate a more troubling mismatch between the model's and the participants' decisions. Each item required the participants to complete the first letter in a string. By a rational analysis, participants should complete all of these strings with either *B* or *L* (i.e., one of the two letters that began the 19 training strings). Although the majority of participants behaved rationally, some chose alternatives *F* and *Z* to complete these strings. Minerva 2, of course, predicted string completion from memory

of the training strings. Because all of the training strings begin with *B* or *L*, the model could not match the people who chose *F* or *Z*.

The simulation confirms that the global-similarity principle anticipates participants' ability to complete test strings legally and that it does a good job of predicting which alternative the majority of participants will select. The ability to predict string completion based on materials generated with diverse grammars (i.e., Dienes's, 1992 grammar borrowed from Reber & Allen, 1978; and Knowlton et al.'s, 1992, grammar) demonstrates that the principle is general rather than a fortunate fit to a particular set of items. Taken together, the simulations confirm our earlier and main conclusions. Rule-like behaviour emerges during retrieval of structured exemplars in memory. More generally, rule-like behaviour does not require a direct representation of the rules.

GENERAL DISCUSSION

In a JOG task, participants study grammatical training strings and then judge the grammaticality of test strings. According to Jamieson and Mewhort (2009a), people infer a test probe's grammaticality by calculating its global similarity to their memory of the training items: Test strings similar to the training strings are classified as grammatical. To implement their ideas, Jamieson and Mewhort adapted Minerva 2—a classic model of human memory—to the task (Hintzman, 1986, 1988). In the present work, we extended the analysis to show that the same principles anticipate which items participants will endorse most strongly as grammatical and which of *n* alternatives they will choose to complete a stem. Taken together it is clear that performance in the JOG task can be understood using the same basic principles that have long been used to understand retrieval from episodic memory.

Although we are encouraged that the principle of global similarity can anticipate results from implicit-learning paradigms, our implementation of the principles is far from a complete model. We were successful only after we ensured that our stimuli did not contain highly familiar acronyms. The point underscores a long-standing problem in the study of human memory: Ebbinghaus (1885/1964) developed nonsense syllables to examine learning, independently of prior knowledge. There is no principled way, however, to isolate participants from their preexperimental knowledge. At best, we can try to develop stimuli that offer limited opportunities to apply such knowledge (e.g., consonant strings), and we can hope that participants' use of preexperimental knowledge is idiosyncratic enough that we can safely aggregate across their data.

The same point underscores a serious limitation of all current models of the JOG task. Our model can capture the structure of events in our experiments, but, like all current models, it cannot capture the "MTV effect". To do so, it would have to know and apply real-world knowledge in addition to memory for the symbols used in the experiments. That is, we can understand the use of sequential structure among symbols in an experiment with artificial stimuli, but, lacking preexperimental knowledge, we do not have a satisfactory account of human memory. Using a distinction popularized by Tulving (1985), models need to bridge the gap between semantic and episodic memory.

The MTV effect also raises a challenge for empiricists. If the experiment were to be conducted elsewhere, some sequence of letters within our remapped stimuli might be familiar to participants (e.g., a local airport code, local postal code, acronym of a local business).[3] If so, the rank acceptability of the exemplars might differ from ours. The challenge for empiricists, then, is to ensure that their stimuli do not reintroduce the MTV effect. No current model—including

[3] Since these experiments were conducted, the Toronto Stock Exchange (TSX) has appeared regularly in the daily news. The regular discussion around the TSX has caused participants to treat the trigram *TSX* differently than they had in the past.

our implementation of Minerva 2—can appreciate such idiosyncrasies in materials.[4]

To the extent that the materials and results of Experiment 1, 3, and 4 avoid contamination by preexperimental knowledge, we offer them as a standard against which new models of the JOG task can be applied: As far as we can tell, our ranks escape the difficulties associated both with collapsing across different training regimes and with inclusion of idiosyncratic elements.

Throughout this and related papers (e.g., Jamieson & Mewhort, 2009a, 2009b), we used Minerva 2's processing mechanisms to calculate global similarity. Because *global similarity* may be confused with other uses of the term *similarity*, a comment is in order.

The term global similarity was first used in the context of recognition memory. In item recognition, participants study a list of items (usually words) and then classify probes as studied (targets) or unstudied exemplars (foils). Here, global similarity indexes the similarity of a probe to an aggregate of the studied material (the echo). Hintzman (1986, 1988) used vectors composed of random $+1$ or -1 elements to represent words used in experiments (Minerva vectors). Such vectors can hardly be expected to map onto any words in the real world, but that fact does not invalidate their use. An experiment involves sets of words selected from a normed pool, such as the Toronto Word Pool (Friendly, Franklin, Hoffman, & Rubin, 1982). Assuming that the words are chosen at random from the pool, they are nominally independent, but exhibit incidental similarities. Like words selected from a word pool, a set of Minerva vectors are orthonormal in expectation (nominally orthogonal), but exhibit incidental similarities. The critical point is that the similarity of a set of vectors maps nicely onto the pattern obtained by sampling words from a standard word pool. As a result, a probe's global similarity to the studied set is an indirect marker of its status. Critical to the perspective, of course, is that random vectors represent properties of a set of items rather than the properties of individual items.

To apply Minerva to the JOG situation, we used Minerva vectors to represent individual symbols and concatenated the symbol vectors to create vectors that represented both the studied grammatical exemplars and the probes. Here, global similarity indexes the structure of the studied exemplars. As before, the random vectors need not map onto the properties of individual symbols, because the concatenated vectors map onto the spatial (i.e., sequential) structure of the exemplars. If it were important to map both the similarity structure of the symbols and the structure of the exemplars, we could use Murdock's (1995) algorithm to adjust the similarity of the symbol vectors. For example, if an exemplar were composed of the symbols *CESO*, and participants were to encode it phonically, we could use Murdock's algorithm to construct vectors for *C* and *E* that are more similar to one another than they are to the vectors for *S* and *O*. Of course, if the symbols were represented in a visual code, one might prefer to make *C* and O similar rather than *C* and *E*.

Applied to the JOG task, global similarity indexes structure in the studied exemplars extracted in response to a particular probe. The structure is what the participant can apply from what has been learned during the study phase in response to the probe. Such structure is a kind of abstraction that arises during retrieval. Dual-process accounts (e.g., Knowlton et al., 1992; Reber, 1967) claim that abstracting the grammar is unlike episodic retrieval—indeed, the difference underlies their claim for two systems. Hence, to avoid confusion with the two-process claim for two systems, we are reluctant to argue that the echo is an abstraction of the grammar. Rather participants store studied exemplars. Abstraction, if that is the correct term, is a property of retrieval.

Others have used different concepts of similarity in the JOG task. To illustrate, Pothos and

[4] Redington and Chater (1996) discussed the problem in the JOG procedure as the *fixed-effect-fallacy*. In a proper JOG experiment, each item would be equally likely to serve as a training or test item. We did not rotate items in our procedure because we were attempting to discover rather than ameliorate item effects.

Bailey's (2000) similarity analysis involved four separate steps. First, participants rated the similarity of all possible pairs of exemplars. Second, Pothos and Bailey converted the ratings into distance measures in a multidimensional similarity space using Shepard's (1987) scaling algorithm. Next, they converted the distance measures to psychological similarities using Shepard's exponential transformation. Finally, they applied Nosofsky's (1988a, 1988b) exemplar model for classification to demonstrate that similarity is a reliable predictor of classification in the JOG task.

Other measures of similarity include associative chunk strength and edit distance. Associative chunk strength quantifies a test item's similarity in terms of the number of times its bigrams and trigrams appeared in the training set (see Johnstone & Shanks, 2001; Knowlton & Squire, 1996). Edit distance measures similarity in terms of the number of characters that must be changed in one string to make it identical to the other. Both have been used to predict judgements of grammaticality (see Pothos & Bailey, 2000).

Pothos and Bailey (2000) compared grammatical status, chunk strength, edit distance, and scaled similarity as predictors of classification in the JOG task. Because grammaticality remained a reliable source of variance, even after the other sources of variance had been taken into account, they concluded that similarity and grammaticality are independent predictors. Although they provided a measure of the reliability of the predictors, they did not provide a correlation matrix to indicate the overlap among the several predictors nor a standard measure (such as a root mean square error, RMSE) to indicate how well each model could account for the data.

There are several reasons why it is difficult to compare our approach against Pothos and Bailey's (2000) approach. Firstly, subjective similarity ratings—the primary data in Pothos and Bailey's (2000) classification—include many characteristics of the exemplars and symbols used to construct the exemplars. As we noted earlier, similarity might be based on phonological characteristics or on visual ones. Because the participants are free to base a rating of similarity on any dimension, subjective similarity ratings may also reflect some of same characteristics of the stimuli indexed in the edit-distance and chunk-strength measures. Shepard's (1987) scaling technique provides a statistical rationale for deciding the number of independent dimensions on which the ratings are based, but the interpretation of the dimensions remains subjective. In short, an approach based on similarity ratings leaves the basis of the rating an open question. Secondly, it is not clear how any of the measures overlap with global similarity as it is derived in Minerva 2. One difficulty in considering overlap is that global similarity is extracted as a response to a particular probe, whereas the similarity ratings used by Pothos and Bailey are based on properties of stimuli measured outside of the context of the JOG task. Global similarity, on the other hand, is taken dynamically in the context of a particular probe compared to incompletely encoded exemplars; it changes from probe to probe over the course of a simulation. Pothos and Bailey's approach provides a fixed, retrospective fit to data based on a stable pattern of similarity ratings. Such fits do not lend themselves to the kinds of analyses possible in a process account (e.g., similarity judgements under poor encoding). Finally, because Pothos and Bailey's final predictions are based on two prior steps—scaling to obtain distance measures and transformation to psychological space—it is hard to adjudicate the separate contributions of each step to the final answer.

Throughout this paper, and elsewhere (see Jamieson & Mewhort, 2005, 2009a, 2009b), we have argued for a *retrospective account* of performance in the artificial-grammar task. Participants answer questions about test items using memory of training exemplars. The position contrasts with the implicit learning view that proposes that the cognitive system abstracts the grammar *prospectively* in anticipation of future need.

Redington and Chater (2002) highlighted the same distinction earlier in terms of lazy versus eager learning. Lazy learning is passive. It involves storage of experiences and generalization of those experiences to new situations (e.g., judgements of

grammaticality). Eager learning is active. It involves active extraction of statistical regularities in service of future goals and aims.

We cannot help but see performance in the JOG and string completion tasks to be a lazy and retrospective process. At training, participants store the training items. At test, participants use their memory of the training items to answer the odd questions put to them. We find it difficult to imagine how an eager learning system would know to extract a grammar in anticipation of something so unusual as a judgement-of-grammaticality task. Of course, our use of Minerva 2 pushes the issue to the fore. The system eschews the need for a priori abstraction of structure and proposes instead that rule-like behaviour emerges from parallel retrieval of stored exemplars.

We have demonstrated that global similarity is relevant to characterizing performance in artificial-grammar tasks. Of course, we cannot discount that participants knew the grammar: For the same reason that it is not possible to prove a null hypothesis, it is not possible to disprove the implicit learning view. Rather, we are left to cling to parsimony inasmuch as the global similarity account ties performance in explicit and implicit learning tasks to a set of common principles. Although the problem of distinguishing the implicit learning and exemplar views is a difficult one, it is one that needs to be solved in a direct and convincing experiment.

We have argued that participants' performance in the JOG and string completion tasks reflects the participants' use of memory for the training exemplars. Our account has clear limitations. It has difficulty explaining participants' biases in decision (i.e., the *MTV* effect). It assesses similarity between a probe and trace as if the strings were left justified so that letters are compared by corresponding serial position within strings; certainly, participants' similarity assessments are not limited to that scheme. Whereas the model predicted performance for a majority of individual items in the JOG and string-completion tasks, it failed to predict peoples' judgements for others; the misfits suggest that participants rely on information or strategies in addition to global similarity. Finally, we have yet to apply the model to the entire range of artificial-grammar procedures: Thus far, we have shown that global similarity handles judgements of grammaticality (Jamieson & Mewhort, 2009a), serial reaction time performance (Jamieson & Mewhort, 2009b), and serial recall (Jamieson & Mewhort, 2005). To build a more complete account, we need next to tackle the classification–recognition dissociation for artificial-grammar materials in amnesia (e.g., Knowlton et al., 1992) and performance in the transfer version of the JOG task in which training and test strings are instantiated with different symbols (e.g., Brooks & Vokey, 1991; Manza & Reber, 1997; Redington & Chater, 1996). It remains to be seen whether or not the exemplar account can survive these other critical tests.

The work presented here, and the work on which it builds (Jamieson & Mewhort, 2005, 2009a, 2009b), provides a common account of putatively distinct "implicit" and "explicit" tasks. We are encouraged that parallel retrieval of instances explains performance across tasks often distinguished as "implicit" and "explicit".

Original manuscript received 11 November 2008
Accepted revision received 24 July 2009
First published online 21 October 2009

REFERENCES

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Lawrence Erlbaum Associates.

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: General*, *120*, 316–323.

Cleeremans, A., & Dienes, Z. (2008). Computational models of implicit learning. In R. Sun (Ed.), *The Cambridge handbook of computational modeling* (pp. 396–421). New York: Cambridge University Press.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. L. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, *1*, 372–381.

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

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.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger, Trans.). New York: Dover. (Original work published 1885)

Estes, W. K. (1986). Memory storage and retrieval processes in category learning. *Journal of Experimental Psychology: General*, *115*, 155–174.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods & Instrumentation*, *14*, 375–399.

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

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551.

Jamieson, R. K., & Mewhort, D. J. K. (2005). The influence of grammatical, local, and organizational redundancy on implicit learning: An analysis using information theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 9–23.

Jamieson, R. K., & Mewhort, D. J. K. (2009a). Applying an exemplar model to the artificial-grammar task: Inferring grammaticality from similarity. *Quarterly Journal of Experimental Psychology*, *62*, 550–575.

Jamieson, R. K., & Mewhort, D. J. K. (2009b). Applying an exemplar model to the serial reaction-time task: Anticipating from experience. *Quarterly Journal of Experimental Psychology*, *62*, 1757–1783.

Johnstone, T., & Shanks, D. R. (2001). Abstractionist and processing accounts of implicit learning. *Cognitive Psychology*, *42*, 61–112.

Knowlton, B. J., Ramus, S. J., & Squire, L. R. (1992). Intact artificial grammar learning in amnesia: Dissociation of classification learning and explicit memory for specific instances. *Psychological Science*, *3*, 172–179.

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.

Knowlton, B. J., & Squire, L. R. (1996). Artificial grammar learning depends on implicit acquisition of both abstract and exemplar-specific information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 169–181.

Kuhn, G., & Dienes, Z. (2005). Implicit learning of non-local musical rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1417–1432.

Manza, L., & Reber, A. S. (1997). Representing artificial grammars: Transfer across stimulus forms and modalities. In D. C. Berry (Ed.), *How implicit is implicit learning?* (pp. 73–106). Oxford, UK: Oxford University Press.

Medin, D. L., & Schaffer. M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Murdock, B. (1995). Similarity in a distributed memory model. *Journal of Mathematical Psychology*, *39*, 251–264.

Nosofsky, R. M. (1988a). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 700–708.

Nosofsky, R. M. (1988b). Similarity, frequency, and category representation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 54–65.

Pothos, E. M., & Bailey, T. M. (2000). The importance of similarity in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 847–862.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning & Verbal Behavior*, *6*, 855–863.

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

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

Redington, M., & Chater, N. (1996). Transfer in artificial grammar learning: A reevaluation.

Journal of Experimental Psychology: General, *125*, 123–138.

Redington, M., & Chater, N. (2002). Knowledge representation and transfer in artificial grammar learning (AGL). In R. M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness: An empirical, philosophical and computational consensus in the making* (pp. 121–143). Hove, UK: Psychology Press.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

Tulving, E. (1985). *Elements of episodic memory*. Oxford, UK: Oxford University Press.

Tulving, E., Schacter, D. L., & Stark, H. A. (1982). Priming effects in word-fragment completion are independent of recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *8*, 336–342.

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